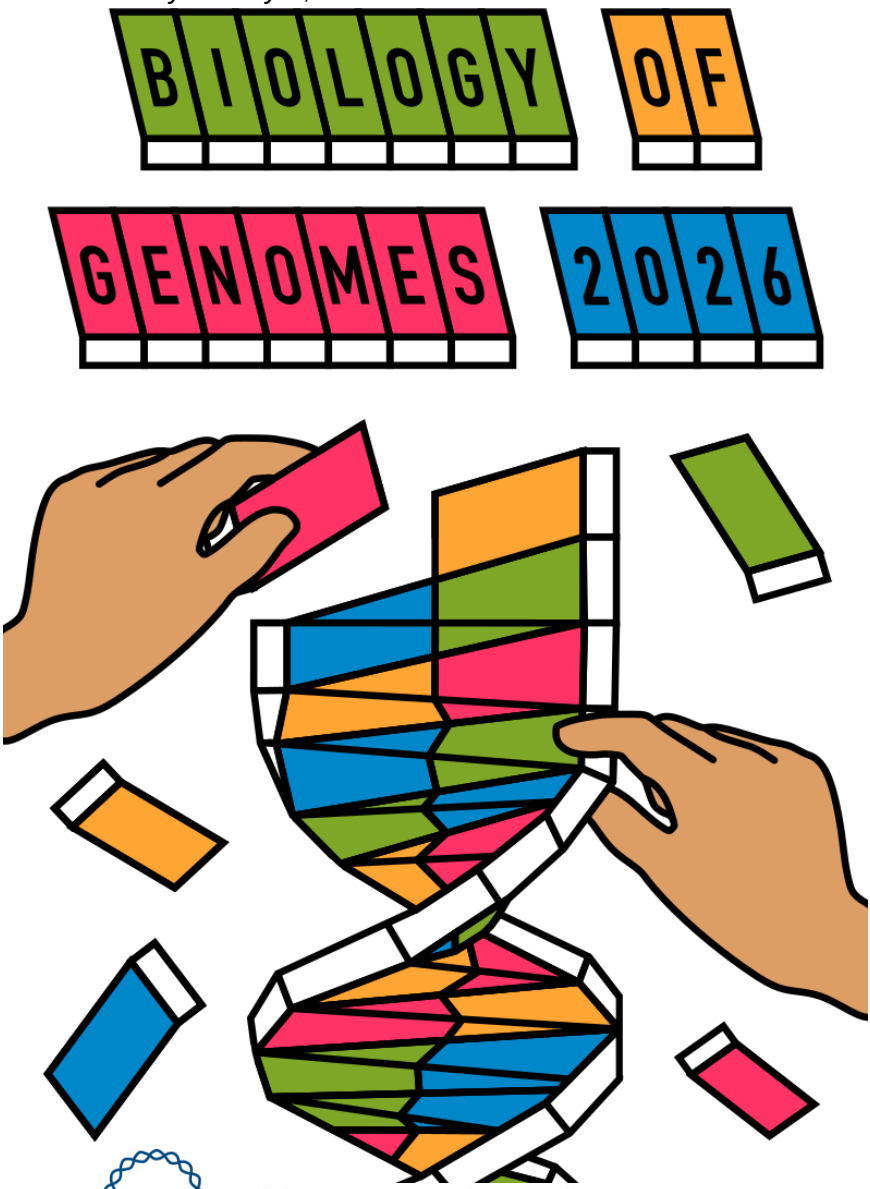
Abstracts of papers presented
at the 2026 meeting on

BIOLOGY OF GENOMES

May 5–May 9, 2026



CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2026 meeting on

BIOLOGY OF GENOMES

May 5–May 9, 2026

Arranged by

Alexis Battle, *Johns Hopkins University*
Matthew Meyerson, *Dana-Farber Cancer Institute*
Aaron Quinlan, *University of Utah*
Jenny Tung, *Max Planck Institute for Evolutionary Anthropology and Duke University*

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

This meeting was funded in part by the **National Human Genome Research Institute (NHGRI),** a branch of the **National Institutes of Health.**

Technical Contributors: **Oxford Nanopore; PacBio**

Contributing Sponsor: **Merck**

Scholarship contributor: **JXTX Foundation**

Contributions from the following companies provide core support for the Cold Spring Harbor Meetings Program:

Corporate Sponsors

Amgen
Calico Labs
New England Biolabs
Novartis

Cover: Cover designed by Iker Rivas-González, Max Planck Institute for Evolutionary Anthropology. Based on "Fold Your Own DNA," by Alex Bateman (2003). ****See back of book for your own foldable DNA****

BIOLOGY OF GENOMES
Tuesday, May 5– Saturday, May 9, 2026

Tuesday	7:30 pm – 10:30 pm	**1** Functional Genomics
Wednesday	9:00 am – 12:00 pm	**2** Complex Traits and Genomic Medicine
Wednesday	1:00 pm – 1:45 pm	NHGRI Discussion Panel
Wednesday	2:00 pm – 5:00 pm	**3** Evolutionary & Non-Human Genomics
Wednesday	5:00 pm	*Wine & Cheese Party*
Wednesday	7:30 pm – 10:30 pm	**Poster Session I**
Thursday	9:00 am – 12:00 pm	**4** Computational & Statistical Genomics
Thursday	1:30 pm – 4:30 pm	**5** Cancer Genomics
Thursday	5:00 pm – 6:00 pm	ELSI Panel and Discussion
Thursday	7:30 pm – 10:30 pm	**Poster Session II**
Friday	9:00 am – 12:00 pm	**6** Emerging Methods and Technologies
Friday	2:00 pm – 4:30 pm	**Poster Session III**
Friday	4:30 pm – 6:00 pm	GUEST SPEAKERS
Friday	6:30 pm	*Cocktails and Banquet*
Saturday	9:00 am – 12:00 pm	**7** Population Genomics

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 5—7:30 PM

SESSION 1 FUNCTIONAL GENOMICS

Chairpersons: **Yoav Gilad,** University of Chicago, Illinois
Yogesh Goyal, Northwestern University and Chan Zuckerberg Biohub, Chicago, Illinois

WEDNESDAY, May 6—9:00 AM

SESSION 2 COMPLEX TRAITS AND GENOMIC MEDICINE

Chairpersons: **Alexander Gusev,** Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts
Arbel Harpak, University of Texas at Austin

WEDNESDAY, May 6—1:00 PM

NHGRI PANEL DISCUSSION

An Overview of Changes and Opportunities at NIH and NHGRI

Moderator: **Ismail Safi,** NHGRI

Panelists from the NHGRI Division of Genome Science:

Alexander Arguello
Jyoti Dayal
Erin Ramos
Sarah Wheelan

NHGRI will share a brief overview of recent changes in policies and research priorities across NIH, with a focus on what these updates mean for the genomics research community. Topics will include emerging scientific focus areas, changes to peer review, and where to find funding opportunities, including for trainees and early career investigators. Following a short presentation, an interactive Q&A with NHGRI Leadership and Program Directors from the Division of Genome Sciences will give attendees the opportunity to ask questions and dive deeper into the topics presented.

WEDNESDAY, May 6—2:00 PM

SESSION 3 EVOLUTIONARY AND NON-HUMAN GENOMICS

Chairpersons: **Andres Bendesky,** Columbia University, New York, New York
Jeffrey Ross-Ibarra, University of California, Davis

Red Queen evolutionary dynamics in the endocrine system 15
Andres Bendesky.
Presenter affiliation: Columbia University, New York, New York.

Unheralded high MHC Class II polymorphism in the abundant Atlantic herring resolved by long-read sequencing
Minal Jamsandekar, Fahime M. Sangdehi, Florian Berg, Michael F. Criscitiello, Brian W. Davis, Marten Larsson, JingYi Li, Mats Petersson, Leif Andersson.
Presenter affiliation: Texas A&M University, College Station, Texas. 22

WEDNESDAY, May 6—5:00 PM

Wine and Cheese Party

WEDNESDAY, May 6—7:30 PM

POSTER SESSION I

See p. xviii for List of Posters

THURSDAY, May 7—9:00 AM

SESSION 4 COMPUTATIONAL AND STATISTICAL GENOMICS

Chairpersons: **Glennis Logsdon,** University of Pennsylvania Perelman School of Medicine, Philadelphia
Michael Schatz, Johns Hopkins University, Baltimore, Maryland

A global view of human centromere variation and evolution
Glennis A. Logsdon, Shenghan Gao, Keisuke K. Oshima, Shu-Cheng Chuang, Mark Loftus, Annalaura Montinaro, David S. Gordon, PingHsun Hsieh, Miriam K. Konkel, Mario Ventura.
Presenter affiliation: University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania. 23

Advancing rare disease diagnosis with long-read sequencing and pangenomics
Shloka Negi, Jean Monlong, Sarah L. Stenton, Seth I. Berger, Brandy McNulty, Ivo Violich, Jouni Sirén, Francesco Andreace, Sagorika Nag, Konstantinos Kyriakidis, Anne O'Donnell-Luria, Emmanuèle Délot, Karen H. Miga, Benedict Paten.
Presenter affiliation: UCSC Genomics Institute, Santa Cruz, California. 24

SESSION 5 CANCER GENOMICS

Chairpersons: **Rameen Beroukhim,** Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts
Elizabeth Murchison, University of Cambridge, United Kingdom

Detecting centromeric fusion events in cancer genomes
Jakob M. Heinz, Matthew Meyerson, Heng Li.
Presenter affiliation: Harvard Medical School, Boston, Massachusetts; Dana-Farber Cancer Institute, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 36

Methods to study modified T cell—cancer cell behaviors and interactions in live-cell killing assays
Barbara E. Engelhardt, Adam Weiner, Justin Adjasu, Cole Citrenbaum, Julie Tran, Stefanie Bachl, Scott Linderman, Julia Carnevale, Alexander Marson.
Presenter affiliation: Gladstone Institutes, San Francisco, California; Stanford University, Stanford, California. 37

The human pangenome reference reduces ancestry-related biases in somatic mutation detection
Chau V. Pham, Farida S. Abdelmalek, Tracy Hua, Kathleen E. Houlahan.
Presenter affiliation: McMaster University, Hamilton, Canada. 38

THURSDAY, May 7—5:00 PM

ELSI PANEL and DISCUSSION

The barn door is open---Defining boundaries of what to look for in pre-implantation testing of IVF embryos

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

Panelists: **Bogdan Pasaniuc,** Perelman School of Medicine, University of Pennsylvania
Jeremy Grushcow, Juniper Genomics
Stacey Pereira, Baylor College of Medicine

Over the past two decades, pre-implantation genetic testing (PGT) has become a routine component of in vitro fertilization (IVF), where it is used to screen fertilized embryos before they are transferred into an intended mother or surrogate's uterus**.** PGT generally focuses on identifying aneuploidy or highly penetrant Mendelian conditions. Some laboratories are also offering PGT that includes screening embryos for polygenic traits, or PGT-P, to collect and return polygenic risk scores (PRS) to estimate an embryo's future risk of complex diseases. As scientific and commercial capability in PGT and PGT-P grows, a central question emerges: What criteria should guide decisions as to what traits should be tested as part of IVF? At this time the answers to this question are

unsettled. Several considerations come into play, including the predictive accuracy and generalizability of different risk calculations; clinical considerations regarding the validity, utility, and demand for PGT-P results in the IVF setting, and parental views about what knowledge would meaningfully inform reproductive choices during IVF.

This session aims to foster a nuanced discussion about how to decide what diseases and traits are good candidates for PGT testing for IVF today, and whether or not some genomic information that seems inappropriate to return today will become more or less acceptable as knowledge and norms change.

The panel brings together three perspectives to explore whether criteria are needed, and if so, how they might be defined. First, **Bogdan Pasaniuc,** Director of the Center for Computational Biomedicine at the Perelman School of Medicine, University of Pennsylvania and a leader in population studies, computational biology and multi-omics will consider the meaning and meaningfulness of risk modeling for complex diseases in the IVF context. **Jeremy Grushcow,** CEO of the fertility genetics company Juniper Genomics will provide insight into how one company approaches decisions about what disease information is offered to prospective parents. **Stacey Pereira,** social scientist and Associate Professor at the Baylor College of Medicine Center for Medical Ethics and Health Policy, will draw on empirical data from her research on PGT-P, focusing on how IVF patients and infertility specialists evaluate which health conditions and traits are appropriate targets for PGT-P.

THURSDAY, May 7—7:30 PM

POSTER SESSION II

See p. xxix for List of Posters

SESSION 6 EMERGING METHODS AND TECHNOLOGIES

Chairpersons: **Nilah Monnier Ioannidis,** University of California, Berkeley and University of California, Santa Cruz
Winston Timp, Johns Hopkins University, Baltimore, Maryland

Leveraging long-read chromatin data to predict full-length RNA isoforms with deep learning
Gali Bai, Nigel Brigstocke, Colette Felton, Angela N. Brooks.
Presenter affiliation: University of California, Santa Cruz, Santa Cruz, California. 43

Chromosome-scale vole assemblies via CiFi-HiFi sequencing resolve a prairie vole-specific AVPR1A duplication
Mohamed Abuelanin, Gulhan Kaya, Juniper A.. Lake, Christine Lambert, Ksenia Krasheninnikova, Jo Wood, Kerstin Howe, Jonas Korlach, Devanand Manoli, Jessica Tollkuhn, Megan Y.. Dennis.
Presenter affiliation: University of California, Davis, Davis, California. 44

FRIDAY, May 8—2:00 PM

POSTER SESSION III

See p. xli for List of Posters

FRIDAY, May 8—4:30 PM

GUEST SPEAKERS

Janet Kelso
Max Planck Institute for Evolutionary Anthropology

Jonathan Pritchard
Stanford University

FRIDAY, May 8—6:30 PM

COCKTAILS and BANQUET

SESSION 7 POPULATION GENOMICS

Chairpersons: **Amy Goldberg,** University of California, Los Angeles
John Novembre, University of Chicago, Illinois

POSTER SESSION II

POSTER SESSION III

Integrating long-read RNA sequencing with genomics and phenomics to discover novel disease-relevant splice-altering genetic variants
David Wu, Feng Wang, Quan Sun, Xinjun Ji, Robert Wang, Joseph Park, Ryan Park, Stacy Woyciechowski, Lan Lin, William Gaynor, Yi Xing.
Presenter affiliation: CHOP, Philadelphia, Pennsylvania; University of Pennsylvania, Philadelphia, Pennsylvania. 249

Explainable sequence-based model reveals divergent transcription initiation rules in *Drosophila* and human
Ruoxuan Wu, Kseniia Dudnyk, Jian Zhou.
Presenter affiliation: University of Chicago, Chicago, Illinois. 250

Sex-stratified single-cell transcriptomic analysis reveals molecular and cellular signatures across multiple psychiatric disorders
Yan Xia, Ro Malik, Zhongzheng Mao, Nancy Fang, Declan Clark, Mark Gerstein.
Presenter affiliation: Yale University, New Haven, Connecticut. 251

VOUS—Variational Ornstein-Uhlenbeck Stochastics linking single-cell lineage tracing with dynamic gene expression
Jiawei Xing, Stephen Staklinski, Adam Siepel.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 252

Escape from X inactivation drives sex differences in gene expression
Carrie Zhu, Liaoyi Xu, Arbel Harpak.
Presenter affiliation: University of Texas at Austin, Austin, Texas. 253

Genetic regulation of cell type–specific chromatin accessibility shapes immune function and disease risk
Angli Xue, Jianan Fan, Oscar Dong, Hao Huang, Peter Allen, Eleanor Spenceley, Anna Cuomo, Albert Henry, Ling Chen, Elizabeth Dorans, Kyle K. Farh, Wei Zhou, Alkes L. Price, Gemma A. Figtree, Alex W. Hewitt, Daniel G. MacArthur, Joseph E. Powell.
Presenter affiliation: Garvan Institute of Medical Research, Sydney, Australia; University of New South Wales, Sydney, Australia. 254

Multi-layer omics studies to understand human immune system
Kazuhiko Yamamoto, Rintaro Fujimoto, Hiroki Kitaoka, Saya Hisano, Akari Suzuki, Yasuhiko Murakawa, Koshi Imami, Makoto Arita, Yosuke Isobe, Hiroshi Ohno, Shohei Asami, Shin-ichiro Fujii, Takeya Kasukawa, Jun Seita, Yukinori Okada.
Presenter affiliation: RIKEN, Yokohama, Japan. 255

Genomic mosaicism reveals developmental organization of sensory and sympathetic ganglia
Xiaoxu Yang.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 256

Profiling of internal variation of SINE-VNTR-Alu elements in the All of Us long-read cohort
Alex Yenkin, Yulia Mostovoy, Karan Jaisingh, Xuefang Zhao, Yongqing Huang, Fabio Cunial, Samuel Lee, Kiran Garimella, Michael Talkowski.
Presenter affiliation: Harvard University, Boston, Massachusetts; Massachusetts General Hospital, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts. 257

Functional dissection of circulating fatty acids-associated loci using CRISPR-based genetic perturbations
Ke Yi, Huifang Xu, Haifeng Zhang, Pengpeng Bi, Kaixiong Ye.
Presenter affiliation: University of Georgia, Athens, Georgia. 258

Satellite DNA fragility, epigenetic disruption, and extrachromosomal amplification converge to drive structural genome instability in canine osteosarcomas
Feyza Yilmaz, Wonyoung Kang, Sabriya A. Syed, Francis H. O'Neill, Jody T. Lombardi, Patrick Kwok Shing Ng, Ching C. Lau, Charles Lee.
Presenter affiliation: The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut. 259

A 200 million cell genome-wide perturb-seq atlas with CRISPRi, CRISPRa, and siRNA
Kwontae You, Alejandro Mendez Mancilla, Dulguun Amgalan, Eyal Ben David, Hong Gao, Jiang Zhu, Doyeon Kim, Emily Laubscher, Jonatan Perez, Ling Chen, Lenka Dohnalova, Marcos Nascimento, Sebastian Pineda, Zala Sekne, Wenhe Lin, Martijn Vochteloo, Lauren Varanese, Kyle Kai-How Farh.
Presenter affiliation: Illumina, San Diego, California. 260

< Back

< Back

BEYOND THE MEAN: GENETIC CONTROL OF GENE EXPRESSION FIDELITY AND DISPERSION

Yoav Gilad

University of Chicago, Medicine, Chicago, IL

For decades, studies of gene regulation have been built around a single quantity: the mean expression level of a gene, averaged across large populations of cells. Even with the advent of single-cell transcriptomics, most analyses have retained this framework by aggregating measurements into pseudobulk profiles, effectively reducing single-cell data back to population averages. As a result, a fundamental feature of single-cell measurements, the variability in expression among individual cells of the same type and state, has remained largely unexplored. Whether this cell-to-cell variation reflects technical noise or a biologically meaningful property of regulatory systems remains an open question with important implications for understanding gene regulatory fidelity, robustness, and threshold-dependent phenotypes.

We used single-cell RNA-seq to test directly whether dispersion of gene expression represents noise or regulated biological signal. Using a novel experimental platform based on heterogeneous differentiated cultures (HDCs), which generate a broad spectrum of human cell types in vitro, we first show that patterns of regulatory dispersion are highly structured across cell types. Genes with low dispersion are often shared among related cell types and are enriched for core functional and identity-related roles. Within each cell type, genes with higher expression levels tend to exhibit lower dispersion, indicating that precise regulatory control is preferentially imposed on genes that are most critical for cellular function.

To determine whether dispersion is under genetic control, we leveraged a comparative framework including human and chimpanzee iPSC lines, an allotetraploid human–chimpanzee line, and mixed-species HDCs. We show that interspecies differences in regulatory dispersion are primarily driven by cis-regulatory divergence, demonstrating that expression variability itself is subject to evolutionary tuning through local sequence changes.

Together, these results establish gene expression dispersion as a regulated and genetically encoded dimension of gene regulation that captures variation in regulatory fidelity. By moving beyond population averages and treating cell-to-cell variability as biological signal rather than technical artifact, our findings reveal an additional layer of regulatory information with broad implications for evolution, development, and disease.

PRIMATE-SPECIFIC ALU ELEMENTS SLOW HUMAN TRANSDIFFERENTIATION BY TITRATING CEBPA

Ramil Nurtdinov[1], Carme Arnan[1], Maria Sanz[1], Amaya Abad[1], Alexandre Esteban[1], Sebastian Ullrich[1], Rory Johnson[1], Sílvia Pérez-Lluch[1], Roderic Guigó[1,2]

[1]Center for Genomic Regulation, Computational Biology & Health Genomics, Barcelona, Spain, [2]Universitat Pompeu Fabra, Medicine and Life Sciences, Barcelona, Spain

Biological processes such as differentiation, development, gestation, and ageing are broadly conserved among closely related species, yet their tempo can diverge strikingly. Humans, for example, often execute equivalent programmes far more slowly than mice—a phenomenon previously attributed to differences in metabolism and/or protein stability. These explanations, however, may not fully account for how species tune the speed of otherwise conserved regulatory circuits. To investigate the genetic control of biological tempo, we exploited a cross-species model of transdifferentiation in which pre-B cells are converted into macrophages by the induction of the transcription factor CEBPA. Although the underlying programme is equivalent in human and mouse, its kinetics differ markedly: completion requires ~7 days in human cells but only ~3 days in mouse cells. Using this system, we uncovered a previously unrecognised mechanism that modulates process speed through the expansion of the primate-specific Alu repeats. Specifically, we found that in human, CEBPA is extensively recruited to Alu repeats, effectively titrating the factor away from canonical regulatory targets. This reduces functional CEBPA availability, diminishing regulatory efficiency and slowing the overall pace of transdifferentiation. To directly test causality, we engineered a CRISPR/dCas9-based strategy to selectively impair CEBPA binding at Alu elements in human cells. Disrupting this decoy recruitment increases CEBPA occupancy at canonical sites and accelerates the transdifferentiation programme, demonstrating that repetitive regions can actively tune the kinetics of a conserved biological process. To our knowledge, this constitutes the first example of a genetic mechanism capable of modulating the speed at which a complex biological programme unfolds, and the first demonstration that this speed can be experimentally manipulated via targeted intervention in repetitive DNA. Notably, Alu repeats harbouring strong CEBPA motifs have accumulated preferentially in ape genomes relative to other primates, suggesting a potential evolutionary link between transposable element expansion and ape-specific phenotypes, including longevity. Together, our findings reveal an unexpected layer of genomic encoding in which lineage-specific repeat landscapes reshape transcription factor availability and, in turn, drive species-specific shifts in developmental tempo.

MASSIVELY PARALLEL REPORTER ASSAY-INFORMED MODELING IMPROVES PREDICTION OF CONTEXT-SPECIFIC ENHANCER-GENE REGULATORY INTERACTIONS

Anat Kreimer[1,2], William Degroat[1]

[1]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, NJ, [2]Department of Biochemistry and Molecular Biology, Rutgers, The State University of New Jersey, Piscataway, NJ

Enhancers are cis-regulatory elements that drive context-specific gene expression, yet their target genes and modes of action remain incompletely defined. Because the majority of disease-associated variants lie in non-coding regulatory DNA, accurate, cell type-specific enhancer-gene (E-G) mapping is essential for interpreting genetic risk. However, existing E-G prediction frameworks lack the resolution needed to reliably capture context-dependent regulatory interactions. Massively parallel reporter assays (MPRAs) provide a direct measure of cis-regulatory activity, but their integration into genome-scale E-G models has remained limited.

Here, we introduce MPRabc, an MPRA-informed machine-learning framework that substantially improves the accuracy and context specificity of E-G interaction prediction. MPRabc integrates predicted MPRA activity, sequence-derived regulatory features, epigenomic signals, three-dimensional chromatin contact maps, and CRISPR-based perturbation training data. Benchmarking against CRISPR interference-validated regulatory interactions demonstrates that MPRabc consistently outperforms state-of-the-art models.

We generated high-resolution E-G networks for K562, HepG2, and hiPSC lines. We applied a modular, graph-based framework to identify regulatory substructures, map trait-associated variants and expression quantitative trait loci, and resolve transcription factor drivers of enhancer activity. Across cellular contexts, MPRabc accurately recovers lineage-defining regulatory programs, including GATA1 in K562, HNF1A/B in HepG2, and POU factor circuits in hiPSCs.

Together, these results establish MPRA-informed modeling as a scalable and generalizable strategy for decoding enhancer function, improving cell type-specific regulatory inference, and linking non-coding genetic variation to gene regulatory mechanisms across diverse cellular contexts.

FIBER-TEnCATS – A TARGETED APPROACH TO SIMULTANEOUSLY STUDY TRANSPOSABLE ELEMENT SEQUENCE, DNA METHYLATION, AND CHROMATIN ACCESSIBILITY

Katarina Pavlovic, Torrin McDonald, Alan P Boyle

University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI

Transposable elements (TEs) are repetitive genomic sequences that are highly polymorphic in the human population, can contain various transcription factor binding sites, influence gene expression, and contribute to disease. Their impact depends both on their genomic location and chromatin accessibility. Because of high sequence similarity among copies of evolutionarily young and most active TEs, short-read methods often struggle to fully recover their sequence and chromatin accessibility along the entire length of the element. Recent advances in techniques that combine long-read sequencing with different methyltransferases enable the recovery of chromatin accessibility, DNA methylation, and sequence information within the same cells and along the same fibers, reducing batch effects and improving read mappability compared to traditional short-read methods. However, most of these approaches rely on whole-genome sequencing, which remains costly and requires multiple flow cells to achieve adequate coverage, making them less accessible to many laboratories.

To study the sequence, DNA methylation, and chromatin accessibility in a TE-targeted manner, reducing costs and time while increasing coverage compared to whole-genome long-read sequencing methods, we developed Fiber-TEnCATS, a technique that combines our previous enrichment capture approach with a recent chromatin accessibility-related long-read method, Fiber-seq. Using this new method, we capture over 93% of genomic copies of the TE family of interest, with up to 3 mismatches to our guide RNA, in the T2T-HG002 genome. On average, our method provides more than twice the coverage over regions of interest compared to the standard Fiber-seq method. It also accurately reflects known chromatin accessibility patterns observed in ATAC-seq data and can recover signals over repetitive regions that ATAC-seq misses due to short-read multimapping issues. The long reads not only resolve the multimapping problem but also span enough of the flanking genomic sequence to enable haplotagging, allowing us to examine both the chromatin accessibility linked to TE insertion polymorphisms and the presence of heterozygous epialleles at TE loci. Finally, while in most cases DNA methylation and reduced chromatin accessibility at a locus go hand in hand, we identify hypomethylated CpG loci that maintain low chromatin accessibility, possibly safeguarding against the escape from TE silencing, and highlight the importance of studying these silencing mechanisms simultaneously.

THE UNREASONABLE INFORMATIVENESS OF GENE CO-FLUCTUATIONS

Yogesh Goyal

Northwestern University and Chan Zuckerberg Biohub, Chicago, IL

Gene expression fluctuations across single cells are often treated as noise, yet they encode rich information about the underlying regulatory architecture. Here, I will present two complementary frameworks from our lab that harness these fluctuations. First, we leverage information from recently divided sister cells ("twins"), identifiable via DNA barcoding, to infer gene regulatory networks. This framework, which we call TwINFER, discriminates regulatory from non-regulatory correlations, resolves causal directionality, and eliminates false positives in notoriously difficult fan-out and feed-forward loop motifs. Applied to lineage-barcoded hematopoiesis data, TwINFER refined network inference and flagged multi-state genes. Second, we apply linear response theory from statistical physics to predict transcriptome-wide perturbation outcomes using gene co-fluctuations in unperturbed cells. This framework, CIPHER, was validated across 11 large-scale Perturb-seq datasets covering >4,000 perturbations, recapitulating genome-wide responses to single and double perturbations. Importantly, the effect vanishes when gene-gene covariances are eliminated. Ongoing work in my lab further extends these ideas to learn transferable phenotype-to-genotype mappings across perturbation studies at scale. Together, our results demonstrate that baseline cellular variability is a quantifiable resource for dissecting regulatory logic.

EPIGENETIC CHARACTERIZATION OF PSEUDOGENES ACROSS HUMAN TISSUES

Yunzhe Jiang[1,2], Beatrice Borsari[1,2,3], Mark B Gerstein[1,2,4,5]

[1]Yale University, Program in Computational Biology and Biomedical Informatics, New Haven, CT, [2]Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT, [3]Universitat de Barcelona, Department of Genetics, Microbiology & Statistics, Barcelona, Spain, [4]Yale University, Department of Computer Science, New Haven, CT, [5]Yale University, Department of Statistics & Data Science, New Haven, CT

Pseudogenes have historically been regarded as non-functional remnants of genome evolution. However, relative to other noncoding genomic elements, their promoter architecture and epigenetic regulation remain incompletely understood. Here, we systematically characterize pseudogene promoters and compare them with those of protein-coding genes and long non-coding RNAs. To do this, we integrate matched transcriptomic and epigenomic data across 26 human tissues from the ENCODE EN-TEx project. We uniformly annotate promoters with chromatin features (histone modifications, chromatin accessibility, and DNA methylation), sequence motifs, and evolutionary conservation, generating an online catalog. Leveraging the catalog, we show that, across multiple tissues, transcribed unprocessed pseudogenes exhibit chromatin patterns similar to those of active protein-coding genes. In contrast, transcribed processed pseudogenes show a strikingly different pattern: most lack the canonical hallmarks of transcription (e.g., active histone marks) at their promoters. Instead, their promoters show increased overlap with LINE elements, enrichment for YY1-like binding motifs, and higher Hi-C contact frequency, particularly with distal enhancer-like regulatory regions. Together with their greater conservation (relative to unprocessed pseudogenes), these features suggest that the transcription of processed pseudogenes may require regulatory mechanisms distinct from canonical promoter-associated epigenetic activation.

BEYOND COPY NUMBER: THE REGULATORY ARCHITECTURE OF MITOCHONDRIAL DNA GENE EXPRESSION

Parisa Riahi[1], Sharwary Raghupathy[3], Bryan Le[1], Sol Taylor-Brill[1,2], Dylan Taylor[4], Rajiv McCoy[4], Shweta Ramdas[2], Arslan A Zaidi[1,2,5]

[1]University of Minnesota, Bioinformatics and Computational Biology Program, Minneapolis, MN, [2]University of Minnesota, Genetics, Cell Biology, and Development Department, Minneapolis, MN, [3]Indian Institute of Sciences, Centre for Brain Research,, Bangalore, India, [4]Johns Hopkins University, Department of Biology, Baltimore, MD, [5]University of Minnesota, Institute of Health Informatics, Minneapolis, MN

Mitochondrial DNA copy number (mtCN) is widely used as a biomarker for mitochondrial dysfunction and disease risk, yet its relationship to mtDNA gene expression (mtGE)—the primary functional output—remains poorly understood. Prior studies in heterogeneous tissues are susceptible to confounding by cell composition. To address this, we developed a rigorous directed acyclic graph (DAG) framework to test the mtCN-mtGE relationship using RNA- and DNA-seq data from 731 lymphoblastoid cell lines (LCLs) from the 1000 genomes Project.

Standard RNA-seq normalization induce spurious transcriptome-wide correlations with mtDNA genes, which are highly expressed in LCLs. Excluding all known mitochondrial genes from normalization eliminates these artifacts. Furthermore, expression PCs computed from this scheme capture axes of covariance driven by confounders but not mitochondrial function, preventing overcorrection. We find no correlation between mtCN and expression (beta=0.016,p=0.473), and no association with genetically predicted copy number. Our test is well calibrated: pseudogenes and lncRNAs – also excluded from normalization – showed no inflation (GC ~ 1) in correlation with mtCN, in contrast with previous studies.

We analyzed the mtCN-mtGE association in 17 GTEx tissues and found significant association only in whole blood (beta=0.17,p=1.4e-5). Using linear mixed models, we show tissues with lower mtDNA content exhibit more positive mtCN-mtGE associations, suggesting mtCN predicts mtGE only below a rate-limiting threshold. LCLs, with high mtDNA copy number, likely operate above this threshold, where accessible copies rather than absolute content become relevant.

To assess accessibility, we developed a population genetic method leveraging heteroplasmy drift between DNA and RNA to estimate the effective number of transcribed mtDNA copies—the transcriptional bottleneck. Our Bayesian model accounts for mutation, sequencing error, and cell passage drift, yielding unbiased estimates in simulations. In LCLs, we estimate ~700 effective copies (95%CI 671-980) from 813 total copies, indicating high accessibility that contrasts with HeLa cells where mtDNA is largely inaccessible, suggesting accessibility varies dramatically across cell types.

Our findings demonstrate that mtCN and expression are decoupled in most tissues, challenging copy number as a proxy for transcriptional output and highlighting the need to better understand mechanisms underlying mtCN-disease associations.

A UNIQUE LONGITUDINAL APPROACH TO OMICS DATA REVEALS DISTINCT FACETS OF SEX-SPECIFIC AGING

Cameron R Kelsey[1], Baptiste Sadoughi[1], Rachel M Petersen[2], Marina M Watowich[2], Angelina Ruiz Lambides[3], Cayo Biobank Research Unit[4], Michael J Montague[4], Lauren J Brent[5], Michael L Platt[4], James P Higham[6], Amanda J Lea[2], Noah Snyder-Mackler[1]

[1]Arizona State University, Centre for Evolution and Medicine, Tempe, AZ, [2]Vanderbilt University, Department of Biological Sciences, Nashville, TN, [3]University of Puerto Rico, Caribbean Primate Research Centre, San Juan, PR, [4]University of Pennsylvania, Department of Psychology, Philadelphia, PA, [5]University of Exeter, Centre for Research in Animal Behaviour, Exeter, United Kingdom, [6]New York University, Department of Anthropology, New York, NY

Aging is the most significant predictor of mortality and many diseases. The risk of these age-associated outcomes often differs between males and females yet the molecular mechanisms underlying these sex differences remains unclear. In part this is because aging is highly heterogeneous and longitudinal studies that properly characterize within-individual aging trajectories are lacking. To address this gap, we combined a unique, human-relevant study population–the rhesus macaques of Cayo Santiago–with a statistical approach that has not been applied to 'omics data to date, that allows us to differentiate within- ("aging") from between-individual ("age-associated") effects. We used DNAm data generated from whole blood for 163 individuals (1-5 samples per individual), capturing up to 80% of their median lifespan, and tested for age and sex effects at 109,820 DNAm regions (534,624 CpGs). We found that aging resulted in a general loss of DNAm across the genome, but this effect was context specific–there was significant enrichment (FDR < .05) of both hypomethylated regions at transcriptionally active sites and hypermethylated regions at transposable elements (e.g. SINEs). Aging effects for males and females were highly correlated (r = .76) and also resulted in an overall loss of DNAm. However, males showed accelerated hypermethylation (FDR < .05) with age whereas females showed accelerated hypomethylation (FDR < .05). In particular, DNAm increased in promoters of several key genes in the TWEAK and p38-MAPK pathways in males while females showed loss of DNAm at promoters of genes involved in inflammation and tumor suppression. This suggests suppression of genes regulating inflammation, oxidative stress, and tumor progression in males, but expression of genes that are potentially protective against cancer and inflammation-related disease in females. Overall, these findings highlight that within-individual aging may occur largely at regulatory regions and repetitive elements, and they reveal potential molecular mechanisms that may contribute to sex differences in aging outcomes and disease risk.

RIBOSOMAL DNA COPY NUMBER AND SEQUENCE POLYMORPHISMS SHAPE HUMAN PHYSIOLOGY AND DISEASE RISK

Anil Raj[1], Jordan S Brown[1], Nathaniel H Thayer[1], Manuel Hotz[1], Irene Lam[1], Nicole Fong[1], Elena P Sorokin[1], Marjola Thanaj[2], Daphna Rothschild[3,4], Jonathan K Pritchard[3,4], Maria Barna[3,4], David G Hendrickson[1]

[1]Calico Life Sciences LLC, South San Francisco, CA, [2]University of Westminster, Research Center for Optimal Health, School of Life Sciences, London, United Kingdom, [3]Stanford University, Dept. of Genetics, Stanford, CA, [4]Stanford University, Dept. of Biology, Stanford, CA

Variation in ribosomal DNA (rDNA) copy number and inter-copy sequence polymorphisms influence diverse physiological traits in model organisms, yet their consequences for human health remain poorly characterized. Here, we provide the largest analysis of 45S rDNA copy number to date, and the first population-scale characterization of 5S rDNA copy number, using whole-genome sequencing from 490,383 UK Biobank participants. Despite encoding components of the same molecular machine, these arrays vary independently, are each highly heritable, and associate with divergent phenotypes. Higher 45S copy number associates with common metabolic diseases, increased adiposity, and hematological signatures reminiscent of ribosomopathies. Molecular characterization reveals a coherent through line: we observe elevated secretory cell-derived proteins in plasma, altered proteostasis and translation gene programs across tissues, and increased glucose-stimulated insulin secretion in primary human pancreatic islets. In contrast, 5S copy number shows no disease associations but instead correlates with proportional organismal growth: increased lean mass and organ volumes. Here too, molecular signatures align with population-level findings, as tissue transcriptomics reveals changes to myogenic gene expression programs and altered fat metabolism. Beyond copy number variation, we characterized intra-individual sequence variants between the 45S rDNA copies, identifying heritable ribosome subtypes and low-heritability ribosomal RNA (rRNA) mutations. Common heritable variants in rRNA Expansion-Segments — regions protruding from the ribosome core — associate with distinct traits: es15l with adiposity, es39l with body dimensions, and es27l with blood-related traits and diseases. Rare-variant mutational burden in both the 18S and 28S expansion segments linked rRNA mutations to diverse diseases including cancer and acute myocardial infarction. This multi-scale convergence establishes that the two human rDNA arrays function as independent genetic factors with divergent consequences for cellular physiology and human health.

COMPARATIVE ANALYSIS OF HUMAN AND CHIMPANZEE LIVER CELL RESPONSES TO INNATE IMMUNE STIMULATION

Anna M Cormack[1,2], Kenneth Barr[2], Yoav Gilad [1,2]

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

Humans and chimpanzees differ in their responses to a number of infectious diseases, including hepatitis C virus (HCV). Compared to humans, chimpanzees spontaneously clear HCV at higher rates and rarely progress to severe liver disease, suggesting inter-species differences in hepatic innate immune responses.

To investigate the molecular basis of these differences, we used a panel of human and chimpanzee iPSCs to develop a liver-enriched heterogeneous differentiating culture (LHDC) system. LHDCs allow us to access multiple liver-relevant cell types from both species, including hepatocytes, hepatoblasts, cholangiocytes, macrophages, fibroblasts, and endothelial cells. Compared to existing liver organoid and liver-on-chip models, LHDCs exhibit greater cellular diversity and can be readily dissociated for single-cell RNA-seq, enabling scalable, cost-effective comparative analyses across species, cell types, and exposures.

To identify cell-type-specific immune responses that may underlie inter-species differences in infection outcomes, we exposed 5 human and 5 chimpanzee LHDCs to pathogen-associated molecular patterns that model bacterial and viral infection: lipopolysaccharide, a TLR4 agonist and component of Gram-negative bacterial cell walls, and polyinosinic:polycytidylic acid, a TLR3 agonist that mimics viral double-stranded RNA. We used single-cell RNA-seq to profile at least 10,000 single cells per individual per condition and performed differential gene expression analysis to identify conserved and divergent responses to immune stimulation within each cell type, including pathogen-specific transcriptional signatures.

By enabling direct single-cell-resolved comparisons of immune response programs across diverse liver cell types and immune stimuli, our approach reveals how evolutionary differences in regulatory architecture can shape divergent infection outcomes, including enhanced antiviral responses in chimpanzees.

SINGLE-CELL MULTIOMICS OF NEURONAL ACTIVATION REVEALS CONTEXT-DEPENDENT GENETIC CONTROL OF BRAIN DISORDERS

Lifan Liang[1], Siwei Zhang[2], Zicheng Wang[1], Hanwen Zhang[2], Chuxuan Li[2,3], Christina Thapa[2], Emily Oh[2], David Sirkin[2], Xiaotong Sun[1], Alexandra Barishman[2], Ada McCarroll[2], Alexandra C Duhe[2], Sheng Qian[1], Xiaoyuan Zhong[1], Brendan Jamison[1,2], Whitney Wood[2], Xin He[1], Jubao Duan[2]

[1]The University of Chicago, Department of Human Genetics, Chicago, IL, [2]Endeavor Health Research Institute, Center for Psychiatric Genetics, Evanston, IL, [3]University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, [4]The University of Chicago, 5Department of Psychiatry and Behavioral Neuroscience, Chicago, IL

Genome-wide association studies (GWAS) have identified hundreds of loci for neuropsychiatric disorders (NPD), yet the context-specific mechanisms mediating these effects remain largely elusive. We hypothesized that many risk variants act specifically during neuronal stimulation - a critical state for plasticity that is uncaptured in standard post-mortem profiles. To test this, we generated a comprehensive single-cell multi-omics dataset of neuronal activation using human iPSC-derived excitatory and inhibitory neuron co-cultures from 100 cell lines. We profiled gene expression and chromatin accessibility for over one million neurons across three time points (0, 1, and 6 hours), capturing the resting state, early response, and late response, respectively.

We identified thousands of "dynamic QTLs"—variants driving inter-individual variation in gene expression and chromatin accessibility specifically upon neuron stimulation. Notably, genetic effects for hundreds of genes and thousands of regulatory elements were undetectable at the resting state. In contrast to our baseline QTLs, these dynamic QTLs showed lower concordance with existing brain QTL resources (e.g., GTEx, PsychENCODE, ROSMAP), highlighting the unique potential of dynamic QTLs to uncover novel biological insights missed by steady-state atlases.

Crucially, QTLs detected at the stimulating states were highly enriched (10-60 fold) for NPD-associated variants, with caQTLs explaining a larger proportion of disease heritability than eQTLs. Integrative fine-mapping revealed abundant risk genes whose functional relevance is unmasked only by stimulation. Specifically, we identified *CPT1C* and *CROT* as putative drivers linking the dynamic regulation of lipid metabolism to NPD risk. These findings demonstrate that stimulating model systems can uncover latent genetic effects, providing a new mechanistic roadmap for interpreting non-coding variation in brain disorders.

THE IMPACT OF RARE DELETERIOUS MUTATIONS ON HUMAN LIFESPAN

Hong Gao[1], Joshua G Schraiber[1], Jacob C Ulirsch[1], Shu Tadaka[2], Daniel M Sanchez[3], Shan Dong[4], Heidi L Rehm[5], Shamil Sunyaev[6], Anne O'Donnell-Luria[5], Stephan J Sanders[4], Kyle K Farh[1]

[1]Illumina, Inc., BioInsight, Foster City, CA, [2]Tohoku University, 2Tohoku Medical Megabank Organization, Sendai, Japan, [3]M42, Inc., Abu Dhabi, United Arab Emirates, [4]University of California San Francisco, Psychiatry, San Francisco, CA, [5]Broad Institute of MIT and Harvard, Medical and Population Genetics, Boston, MA, [6]Harvard Medical School, Biomedical Informatics, Boston, MA

Natural selection has shaped the genetic history of our species, but its ongoing effects in present-day human populations remain unclear, particularly in view of recent technological and environmental changes which have markedly reduced premature mortality and doubled human life expectancy in less than 10 generations since the industrial revolution. Previous studies have identified genetic variants associated with phenotypic aspects of selection in present-day populations, including effects on reproductive success, and estimated selective constraint for protein-truncating variants. However, generalizing this work to cover all protein-coding variants has proven challenging due to the difficulty of predicting the effects of deleterious alleles, especially for missense variants, which constitute the vast majority of protein-altering variation in human cohorts.

Here, we integrate deep learning with demographic modeling to accurately estimate the heterozygous selection coefficient, PrimateAI-3D s_{het}, for nearly all ~70 million possible protein-coding single nucleotide variants in the human genome, observing the best performance when missense variants are stratified by their PrimateAI-3D pathogenicity predictions. In 454,712 individuals from the UK Biobank, we characterize the s_{het} burden of deleterious variants per person and examine their effects on phenotypic aspects of selection. On average, a person carries 3.3 rare genetic variants with $s_{het} > 2\%$, corresponding to a 5-month reduction in lifespan per variant. These variants act through hundreds of common diseases and often exist as intermediate-effect alleles in genes where full-penetrance variants would cause Mendelian disease.

After observing significant effects on lifespan, we explore mechanistic hypotheses for how the selective pressures on these variants arose in pre-industrial generations. We observe deleterious mutations accumulate faster than they are removed and have validated this finding in four independent large-scale cohorts: the All of Us cohort, the Vanderbilt University Medical Center cohort, the Tohoku Medical Megabank, and the Emirati Genome Program. The observed patterns were consistent across all cohorts, implying that historically, the effect of s_{het} burden on mortality was likely the primary driver facilitating the removal of deleterious variants, whereas in contemporary populations, their effects on disease and mortality risk are insufficient to oppose the accrual of new mutations.

MORE IS MORE: SHARED PHENOTYPES AMONGST SEX CHROMOSOME TRISOMIES HINTS DOSAGE-SENSITIVE EFFECT OF PSEUDOAUTOSOMAL REGIONS IN GENETIC MALES AND FEMALES

Aoxing Liu[1,2,3], Yining Wang[1,3], Wenhan Lu[1,2], Zhili Zheng[1,2], Konrad Karczewski[1,2], Mark J Daly[1,2,3]

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, [3]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Genetic males and females differ biologically in many ways. Nearly all the time, comparisons searching for sex differences are made directly between XX females and XY males, with observed differences, either in behaviors, biological processes, or diseases, attributed to the apparent differences in sex gonadal hormones and genetically, the dosage effect of X chromosome or the testis-determining factor gene residing on the Y chromosome. Unlike autosomal aneuploidy, which usually causes spontaneous abortion, sex chromosome trisomies (SCTs) are relatively tolerated and have a considerable prevalence among general populations such as biobank participants. Introducing these trisomies into the analysis can disrupt the perfect correlation between genetic sex and sex chromosome karyotypes, thereby providing a unique perspective on the phenotypic consequences of sex chromosomes.

We recently published a large biobank-based survey of the prevalence and phenotypic consequences of SCTs (PMID: 40840450). Among the notable findings were a high prevalence of all three SCTs (47,XXY; 47,XYY; 47,XXX), an overwhelming majority of carriers (>85%) lacking a genetic or karyotypic diagnosis, and a surprising similarity of non-reproductive related phenotypes shared across all SCTs, including increased height and elevated risk of asthma and multiple vascular diseases. We further ask whether the origin of the extra sex chromosome would have any impact on the phenotypic consequence; we test it in 47,XXY and see no differences regarding both the parent of origin and the identical or homologous status of the two X chromosomes.

Collectively, these shared phenotypes suggest a common biological mechanism driven by gene dosage of the ~3 Mb pseudoautosomal regions (PAR) shared between the X and Y chromosomes. To test this hypothesis, we perform the first meta-pheWAS for PAR variants in FinnGen, UKB, and All of Us; among the >2000 phenotypes examined, height and asthma emerge with the strongest PAR associations (P<5.0e-8). The strongest PAR association to height is adjacent to SHOX - independent analysis of UKB shows loss-of-function variants in SHOX strongly reduce height (P<1.5e-52), consistent with increased height of SCT carriers. To expand these insights, we perform proteomic profiling of >100 SCT carriers using Olink in FinnGen and UKB-PPP identifies >100 proteins shared across SCTs (P<9.2e-06), with the strongest associations for PAR-encoded proteins IL3RA, CSF2RA, CD99, and XG. Many SCT-shared proteins are correlated and share pQTLs. Notably, a locus at 9q34.2 that increases circulating levels of the PAR-encoded protein IL3RA as well as the vascular integrity-related protein TIE1, is associated with increased risk of multiple SCT-shared vascular diseases. Our findings suggest a broad dosage-sensitive effect of PAR on traits such as height, asthma, and vascular diseases.

EARLY AND CURRENT ENVIRONMENTS EXERT DISTINCT EFFECTS ON IMMUNE FUNCTION IN THE ORANG ASLI

Layla Brassington[1], Audrey M Arner[1], Grace Rodenberg[1], Nicholas Ryan[1], Diane Song[1], Tan Bee Ting A/P Tan Boon Huat[2], Izandis bin Mohd Sayed[3], Yvonne A Lim[2], Vivek V Venkataraman[4], Ian J Wallace[5], Thomas S Kraft[6], Amanda J Lea[1]

[1]Vanderbilt University, Biological Sciences, NSH, TN, [2]Universiti Malaya, Parasitology, KL, Malaysia, [3]Hospital Orang Asli, KL, Malaysia, [4]University of Calgary, Anthropology, CGY, Canada, [5]University of New Mexico, Anthropology, ABQ, NM, [6]University of Utah, Anthropology, SLC, UT

Humans evolved in environments characterized by subsistence foraging and hunting, high physical activity, and frequent exposure to diverse parasites. Today, many people experience radically different lifestyles due to industrialization, market integration, and urbanization. These lifestyles are linked to elevated metabolic and inflammatory disease risk relative to non-industrial environments, yet how industrialization shapes immune function within populations remains poorly understood. To address this gap, we worked with the Orang Asli of Peninsular Malaysia–an Indigenous population spanning a lifestyle gradient from subsistence horticulture and foraging to market-integrated, urban environments. We collected survey data on current and early-life environments, bulk PBMC RNA-seq data (n=922), single-cell RNA-seq data for deconvolution (n=4), and 13 circulating biomarker measurements (n=370). As economic transitions are occurring rapidly within Orang Asli communities, early and current industrialization exposure were only modestly correlated (R^2=0.26), allowing us to disentangle their independent effects on adult immune variation.

Current lifestyle exerted stronger effects on gene expression than early-life conditions, associated with 1,428 and 223 genes, respectively (10% FDR). Early-life genes were enriched for GO terms related to adaptive immunity, particularly T cell development and differentiation, consistent with predicted T cell abundance (p=0.002) and circulating IL-8 (p=0.01). In contrast, current lifestyle was more strongly associated with innate immune function, including dendritic cell abundance ($p=8.76\times10^{-7}$), metabolic pathway gene expression, and circulating IL-10, TNF-α, IFN-γ, and CRP ($p<0.05$). Using elastic net regression, we accurately classified individuals who spent their entire lives in urban or rural environments based on gene expression alone (AUC=0.94). When applied to individuals who transitioned between environments, the model generally classified individuals by their current rather than early-life environment, emphasizing the importance of adult conditions. Overall, our results show that early-life microbial and environmental exposures set long-term adaptive immune trajectories, even as current lifestyles strongly influence immune function.

RED QUEEN EVOLUTIONARY DYNAMICS IN THE ENDOCRINE SYSTEM

Andres Bendesky[1,2]

[1]Columbia University, Zuckerman Mind Brain Behavior Institute, New York, NY, [2]Columbia University, Department of Ecology, Evolution and Environmental Biology, New York, NY

The Red Queen paradigm describes how systems of interacting elements with conflicting interests often evolve: changes in one element lead to compensatory changes in another element, establishing a (fragile) equilibrium. Red Queen interactions have been described in several contexts — between pathogens and their hosts, between transposable elements and their host genome, and between paternally-silenced and maternally-silenced genes in the case of genomic imprinting. Here, we describe how the evolution of the hypothalamic-pituitary-adrenal (HPA) axis of *Peromyscus* mice has many features of Red Queen interactions. We discovered multiple evolutionary changes in the HPA axes of closely-related *Peromyscus* species that have large individual effects, but that in aggregate neutralize each other to achieve similar HPA axis states. This highlights a potential role for Red Queen dynamics in the evolution of endocrine systems and suggests the presence of underlying conflicts.

LONG-READ SEQUENCING REVEALS THE GENOMIC ARCHITECTURE OF ALTERNATIVE REPRODUCTIVE TACTICS IN SWORDTAIL FISHES

Gabriel A Preising[1,2], Tristram O Dodge[1,2], Daniel L Powell[3,2], John J Baczenas[1,4], Theresa R Gunn[1,2], Alexandra E Donny[5], Rhea Sood[1], Paola Fascinetto-Zago[1,2], Ryan Cross[6], Samantha M Mason[6], Emmarie P Alexander[7], Andrew J Harris[7], Kang Du[8], Carla Gutiérrez-Rodríguez[9], Oscar Rios-Cardenas[9], Molly R Morris[6], Molly Schumer[1,2,4]

[1]Stanford University, Biology, Stanford, CA, [2]Centro de Investigaciones Científicas de las Huastecas "Aguazarca" A.C, Calnali, Mexico, [3]Louisiana State University, Biological Sciences, Baton Rouge, LA, [4]Howard Hughes Medical Institute, Chevy Chase, MD, [5]University of Washington, Genome Sciences, Seattle, WA, [6]Ohio University, Biology, Athens, OH, [7]Texas A&M University, Biology, College Station, TX, [8]Texas State University, Xiphophorus Genetic Stock Center, San Marcos, TX, [9]Instituto de Ecología A.C., Red de Biología EvolutivaXalapa, Mexico

Alternative reproductive tactics exist across the tree of life. In such systems, individuals within the sex under stronger sexual selection can evolve diverse reproductive strategies, including coercive mating or female mimicry. These polymorphic strategies are driven by environmental and genetic variables to varying degrees. When reproductive tactics are strongly genetically linked, loci underlying this variation are predicted to accumulate in regions with reduced recombination rates, creating discrete phenotypic distributions for alternative morphs. Here, we characterize the genomic architecture of alternative reproductive tactics within swordtail fishes (*Xiphophorus*). Males of different swordtail species can be categorized as having large courting males, small coercive males, or being polymorphic for both male reproductive morphs. Using long-read sequencing, we generated genome assemblies for species exhibiting each of these scenarios. Within one polymorphic species, *X. multilineatus*, we performed a genome-wide association study for reproductive morph and identified a large, structurally variable region on the Y-chromosome that correlates with reproductive tactic. This region contains expansions of multiple genes, including the melanocortin-4 receptor (*mc4r*) which regulates feeding behavior in vertebrates. We explored the evolutionary history of this region and showed that across species, small coercive males exhibit marked reductions in *mc4r* copy number, while large courting males show high variance in *mc4r* copy number. The non-recombining region of the Y-chromosome containing these structural variants likely acted as a refuge for tactic-specific loci to accumulate, explaining their persistence over time. These results highlight structural variation of the Y-chromosome as a mechanism for driving the evolution of reproductive strategies.

ANCIENT HUMAN AND FAUNAL DNA FROM HOLOCENE ARCHAEOLOGICAL SEDIMENTS

Niall Cooke*[2], Gözde Atağ*[1,2], Roman Scholz[2,4], Kevin Nota[2], Matthias Meyer[2], Jozef Bátora[3], Knut Rassmann[4], Benjamin Vernot[1,2]

[1]University of Vienna, Evolutionary Anthropology, Vienna, Austria, [2]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany, [3]Slovak Academy of Sciences, Institute of Archaeology, Nitra, Slovakia, [4]German Archaeological Institute, Romano-Germanic Commission, Frankfurt, Germany

Most ancient DNA is retrieved from bones or teeth - but many sites lack such skeletal elements. Furthermore, bones and teeth leave a record primarily at the moment of death, but an individual sheds DNA throughout their entire life, in principle leaving a trace of their presence where they lived and worked. This DNA can be preserved in archaeological sediments, but this approach has thus far only been demonstrated in Pleistocene cave sediments. Here we present results from a wide-spanning project exploring the role of sedimentary ancient DNA taken directly from living spaces at twelve open-air Holocene settlements from throughout Europe. We captured and sequenced both human and faunal DNA in 183 sediment samples: remarkably, samples at all twelve sites contained ancient mammalian DNA, and ten sites yielded ancient human DNA, demonstrating the broad applicability of this method. We find that the faunal DNA is dominated by domestic taxa, in stark contrast to the wild animal remains recovered at these sites, suggesting that the sediment DNA originates from close association of humans with their domesticates. For 25 samples we were able to resolve human mitochondrial haplogroups; these haplogroups are consistent with the previously published skeletal literature for these regions, where such data is available. We then successfully enriched for human nuclear DNA in eight samples, revealing new insights into the ancestry of people who lived at three Neolithic and Bronze Age archaeological sites. At one of these sites we are able to connect an individual sediment sample with people buried hundreds of kilometers from the settlement, using only nuclear DNA isolated from the sediment sample. A key challenge when attempting to isolate and reliably analyze sedimentary nuclear DNA is distinguishing between genetic material originating from human or non-human sources. We outline and demonstrate novel methods to successfully overcome and mitigate the impact of faunal DNA on analysis.

MELANOMA IN A BENTHIC CATFISH SPECIES REPRESENTS A NEW TRANSMISSIBLE CANCER WITH MULTIPLE LINEAGES.

Julie A Dragon[1], Mark Henderson[3], Kevin Gori[2], Zoe Clarke[2], Elizabeth P Murchison[2]

[1]University of Vermont, Microbiology and Molecular Genetics, Burlington, VT, [2]University of Cambridge, Cambridge, United Kingdom, [3]University of Vermont, Rubenstein School of Environmental Sciences, Burlington, VT

Since 2012, brown bullheads (Ameiurus nebulosus, a type of catfish) in a lake spanning Vermont, USA and Quebec, Canada have shown a high rate of melanomas, suggesting a causal contaminant or contagion. We used whole genome sequencing to test the hypotheses that the melanomas were virally induced and/or that the bullhead in this lake had some genomic background predisposing them to this affliction. Instead, we found is a novel transmissible cancer. Tumor and matched healthy host tissues revealed that tumor mitochondrial and nuclear genomes are more closely related to each other than to their hosts or unaffected fish, supported by hundreds of thousands of genetic variants. These findings indicate that melanoma in these brown bullheads represents the fourth documented type of naturally occurring transmissible cancer in animals, after Tasmanian devils, dogs, and bivalve species. We have since identified a different clonal cancer lineage in another population of brown bullhead in Maine, and historical records, as well as recent social media posts, describing similarly afflicted bullhead throughout the northeastern US and Canada. This raises important questions about the cancer's origin and evolution, mode of transmission, and long-term impact on fish populations, as well the ecological implications for waterways in this part of North America.

ARG-BASED DEMOGRAPHIC INFERENCE REVEALS IMPACTS OF EUROPEAN COLONIZATION ON AMERICAN CROP DIVERSITY

Jeffrey Ross-Ibarra[1,2,3]

[1]University of California Davis, Evolution and Ecology, Davis, CA,
[2]University of California Davis, Center for Population Biology, Davis, CA,
[3]University of California Davis, Genome Center, Davis, CA

Ethnographic and archaeological data have long documented the human toll of European colonization, resulting in the deaths of perhaps 60-90% of indigenous populations. While archaeological data also highlight the concomitant loss of farmland, the genetic impacts of European colonization on crop diversity have been largely ignored. Here, using ancestral recombination graphs built from whole genome assemblies of multiple American crops, I show that colonization resulted in strong bottlenecks likely resulting in the loss of considerable allelic diversity.

LARGE-SCALE RE-WRITING OF AVIAN GENOMES

Anna C Lagani[1], Paolo Mita[1], Willian Silva[2], Matthew Biegler[3], Erich Jarvis[3], Dominic Wright[4], Teresa Davoli[1], Jef D Boeke[1]

[1]NYU Grossman School of Medicine, Institute for Systems Genetics, New York, NY, [2] Linkoping University, Biology, Linkoping, Sweden, [3]Rockefeller University, Genetics and Genomics, New York, NY, [4]Uppsala University, Molecular Genetics and Bioinformatics, Uppsala, Sweden

Genome engineering in birds has the potential to improve the efficiency and humaneness of poultry agriculture, promote manufacturing of biologics using eggs, and provide avenues for preventing or reversing the extinction of wild bird species. To date, most avian genome editing has been limited to small-scale changes, primarily in chickens. Here, we introduce a method for re-writing of large (>20,000 base pairs) genomic loci in chicken primordial germ cells (PGCs), a cell type which can be used to generate genome-edited birds using well-established methods. In this proof-of-concept project, we are genetically "re-wilding" the domestic White Leghorn chicken by re-writing four domestication-associated loci, replacing them with sequences from the wild Red Junglefowl genome. These include the BCO2 gene locus (associated with skin color differences), the PMEL gene locus (responsible for feather pigmentation), a locus near the SEMA3A gene (potentially contributing to behavioral patterns), and the TSHR gene locus (contributing to differences in egg incubation time). We will identify additional loci potentially linked to domestication-related traits and re-write these loci in the White Leghorn genome to dissect their function. Once our genome re-writing approach is optimized in chickens, we will apply it in other avian species to attempt genetic rescue of endangered and extinct birds. For example, we aim to de-extinct the passenger pigeon by re-writing key genome loci in the band tailed pigeon, a close living relative. Overall, developing approaches for large-scale avian genome engineering will enable conservation efforts and be of interest to industries that rely on birds.

REGULATION OF A STATE OF ‘SUSPENDED ANIMATION’ IN KILLIFISH

Christopher He[1], Rui Xiong[1], Stephanie Gagnon[1], Rogelio Barajas[1], Param Priya Singh[1,2]

[1]University of California, San Francisco, Anatomy, San Francisco, CA, [2]Bakar Aging Research Institute, Anatomy, San Francisco, CA

Extremophiles—species that live in extreme environments—have evolved unique adaptations for survival. Understanding how extreme adaptations evolve can reveal new resilience pathways with important ramifications for survival, stress resistance and aging in all organisms. The African killifish is an extremophile for embryo survival. Killifish lives in ephemeral ponds that completely dry up for ~8 months each year. To survive this annual drought, they have evolved a form of long suspended animation, with embryos entering diapause and subsisting in the mud during the dry season. Diapause embryos survive for months, even years—longer than adult life—without any detectable tradeoff for future life. Remarkably, diapause embryos already have complex organs and tissues, including a developing brain and heart. Hence, diapause provides long-term protection to a complex organism. However, the mechanisms underlying the evolution and regulation of cell-type specific protective mechanisms in diapause are unknown. We performed single cell and bulk multi-omics in the embryos of killifish during diapause and development states. We find that diapause evolved by a recent remodeling of regulatory elements at very ancient gene duplicates (paralogs) present in all vertebrates. By integrating chromatin accessibility and gene expression dynamics at the single cell level, we identified cell type specific transcription factors underlying diapause entry and maintenance. CRISPR-Cas9-based perturbations identify key transcription factors including REST/NRSF and FOXOs that are critical for the global regulation of diapause gene expression program. Many of these factors (e.g. REST) are also implicated in aging, Alzheimer's disease and stress resistance in human neurons, suggesting that the mechanisms discovered by studying killifish diapause can uncover new mechanisms to counter aging. Overall, our work identifies cell-type specific mechanisms that have the potential to promote long-term survival by activating suspended animation programs in other species.

UNHERALDED HIGH MHC CLASS II POLYMORPHISM IN THE ABUNDANT ATLANTIC HERRING RESOLVED BY LONG-READ SEQUENCING

Minal Jamsandekar [1], Fahime M Sangdehi[2], Florian Berg[3], Michael F Criscitiello [4], Brian W Davis[1], Marten Larsson[2], JingYi Li[1], Mats Petersson[2], Leif Andersson[1,2]

[1]Texas A&M University , Veterinary Integrative Biosciences, College Station, TX, [2]Uppsala University , Department of Medical Biochemistry and Microbiology , Uppsala, Sweden, [3]Institute of Marine Research in Bergen , Marine Institute, Norway, Norway, [4]Texas A&M University , Department of Veterinary Pathobiology, College Station , TX

Major Histocompatibility Complex (MHC) genes are the most polymorphic in vertebrate genomes due to their important function of presenting a diversity of peptides from pathogens to initiate an adaptive immune response. Here, we have characterized MHC class II genes and explored their polymorphism and evolution in one of the most abundant vertebrates in the world, Atlantic herring (Clupea harengus). The vast population size and schooling behavior make Atlantic herring an attractive target for pathogens. Hence, we hypothesized that it would maintain exceptionally high MHC polymorphism. We used PacBio HiFi long-read whole genome sequencing data of 14 individuals from three different geographic regions in the Atlantic Ocean and Baltic Sea. The analysis identified nine MHC class II loci distributed across four chromosomes. We found two distinct lineages of class II genes, a highly polymorphic one denoted DA and a non-polymorphic DB lineage, arranged as alpha-beta gene pairs (DAA-DAB or DBA-DBB). The DA genes showed extremely high nucleotide diversity in exon 2, strong signatures of positive selection (dN/dS >> 1) as well as copy number variation. Structure prediction revealed that all highly polymorphic amino acid residues occurred in the predicted peptide binding cleft. Two of the most polymorphic loci showed distinct allelic groupings (supertypes), with high sequence similarity within supertypes and high sequence divergence between supertypes. Most of the haplotypes had genes from different supertypes, thus maintaining diverse class II repertoire in a single individual. There was also a highly significant nonrandom association of DAA and DAB alleles within supertypes, strongly suggesting coevolution of DAA and DAB alleles forming the peptide binding domain. This study reveals that the herring MHC class II genes are among the most, if not the most, polymorphic so far described in vertebrates. Their exceptional polymorphism surpasses that of human due to a larger gene repertoire, copy number variation, and pronounced sequence divergence among alleles. This unheralded polymorphism is most likely explained by the combined effects of the vast population size, minimizing genetic drift, and strong pathogen-driven balancing selection.
BioRxiv: https://www.biorxiv.org/content/10.1101/2025.06.08.658498v1

A GLOBAL VIEW OF HUMAN CENTROMERE VARIATION AND EVOLUTION

Glennis A Logsdon[1], Shenghan Gao[1], Keisuke K Oshima[1], Shu-Cheng Chuang[1], Mark Loftus[2], Annalaura Montinaro[3], David S Gordon[4], PingHsun Hsieh[4], Miriam K Konkel[2], Mario Ventura[3]

[1]University of Pennsylvania Perelman School of Medicine, Department of Genetics, Philadelphia, PA, [2]Clemson University, Department of Genetics & Biochemistry, Institute for Human Genetics, Clemson, SC, [3]University of Bari, Aldo Moro, Department of Biosciences, Biotechnology and Environment, Bari, Italy, [4]University of Minnesota, Department of Genetics, Cell Biology, and Development, Institute for Health Informatics, Twin Cities, MN

Centromeres are essential for accurate chromosome segregation during cell division, yet their highly repetitive sequence has historically hindered their complete assembly and characterization. Consequently, the full spectrum of centromere diversity across individuals, populations, and evolutionary contexts remains largely unexplored. Here, we address this gap in knowledge by assembling and characterizing 2,110 complete centromeres from a diverse cohort of individuals representing 5 continental and 28 population groups. By developing a novel suite of bioinformatic tools tailored for centromeric regions, we uncover previously unknown variation, including 226 novel centromere haplotypes and 1,870 new α-satellite higher-order repeat (HOR) variants. We find that mobile element insertions are present in 30% of centromeres, with chromosome 16 harboring *Alu* insertions within the kinetochore site at a 17x higher frequency than expected. While most centromeres have a single kinetochore site, 6% of them have di-kinetochores and <1% have tri-kinetochores, which we confirm with long-read CENP-A CUT&RUN, DiMeLo-seq, and multi-generational inheritance. We further show that the position of the kinetochore is not random and is, instead, closely associated with the underlying sequence and structure. To understand the nature of evolutionary change, we compared 2,110 complete human centromeres to 5,747 centromeres recently assembled by the Human Pangenome Reference Consortium. We show that centromeres have a >20x variation in mutation rate, and a subset of centromeres show evidence of introgression from archaic hominins. We validate these mutation rates in a 4-generation family, spanning 28 members and 483 accurately assembled centromeres and show that the kinetochore site is the most rapidly mutating region in the centromere, with 2.6x more single-nucleotide variants than the rest of the centromeric α-satellite HOR array on average. We propose a model that reveals an ‘arms race’ between centromeric sequence and proteins, with frequent mutations within the site of the kinetochore that lead to changes in genetic and epigenetic landscapes and, ultimately, rapid evolution of these critically important regions.

ADVANCING RARE DISEASE DIAGNOSIS WITH LONG-READ SEQUENCING AND PANGENOMICS

Shloka Negi[1], Jean Monlong[1,2], Sarah L Stenton[3,4], Seth I Berger[5], Brandy McNulty[1], Ivo Violich[1], Jouni Sirén[1], Francesco Andreace[1,6], Sagorika Nag[1], Konstantinos Kyriakidis[1], Anne O'Donnell-Luria[2,3,7], Emmanuèle Délot[8], Karen H Miga[1], Benedict Paten[1]
[1]UCSC Genomics Institute, Biomolecular Engineering and Bioinformatics, Santa Cruz, CA, [2]Institut de Recherche en Santé Digestive, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France, [3]Center for Mendelian Genomics, Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, [4]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, [5]Children's National, Research Institute, Washington, DC, [6]Institut Pasteur, Université Paris Cité, Sequence Bioinformatics Unit, Paris, France, [7]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, [8]Institute for Clinical and Translational Science, University of California, Irvine, CA

More than 50% of families with suspected rare monogenic diseases remain unsolved after whole-genome sequencing. Long-read sequencing (LRS) could help bridge this diagnostic gap by capturing variants inaccessible to short-read sequencing (SRS). However, cost and the lack of streamlined genomic workflows limit widespread clinical adoption. We present a cost-efficient nanopore-based LRS diagnostic framework that combines an optimized sequencing protocol with a scalable bioinformatics workflow called Napu (Nanopore Analysis Pipeline for U).
To evaluate the additional diagnostic yield of LRS, we sequenced 98 samples from 41 rare-disease families using nanopore sequencing, achieving ~36x coverage and a 32-kb read N50 per sample from a single flow cell. Napu generated phased genome assemblies, small- and structural variant (SV) calls, and methylation profiles for all samples. On average, LRS covered coding exons in ~280 genes and ~5 known Mendelian disease genes not covered by SRS. Compared with SRS, LRS identified additional rare, functionally annotated variants, including SVs and tandem repeats, and completely phased 87% of protein-coding genes. It also detected additional de novo variants and distinguished postzygotic mosaic from prezygotic de novo events. LRS established diagnostic variants in 11 probands, spanning de novo and compound heterozygous variants, large-scale SVs, and epigenetic modifications. We also propose a new test for congenital adrenal hyperplasia (CAH) that leverages the human pangenome and LRS to accurately resolve phased rearrangements in the RCCX locus, providing a more accurate approach than existing diagnostics.
To further improve scalability and reduce computational cost, we introduce a pangenome-guided assembly framework that leverages haplotype imputation from the human pangenome. A new sample is modeled as a mosaic of existing pangenome haplotypes with sample-specific variation, followed by hybrid de novo assembly to reconstruct the true diploid genome. As a proof of concept, we demonstrate complete reconstruction of complex pathogenic RCCX haplotypes underlying CAH. As we scale to whole-genome sequencing, this approach will provide a faster and less expensive alternative to conventional de novo assembly and assembly-based SV calling.

MITIGATING CATASTROPHIC FORGETTING IN GENOMIC FOUNDATION MODELS WITH CONTINUAL LEARNING

Alan Murphy, Masayuki Nagai, Peter Koo

Simons Center, Quantitative Biology, Cold Spring Harbor, NY

Genomic deep learning models trained on DNA sequence have demonstrated substantial potential for predicting regulatory activity, interpreting noncoding variants, and designing synthetic regulatory elements. However, most existing genomic foundation models are trained almost exclusively on the human reference genome, limiting their exposure to regulatory diversity and impairing generalisation to unseen loci, novel cellular contexts, synthetic constructs, and disease-relevant variants. A common response has been to fine-tune these models on new datasets, such as personalised genomes paired with gene expression measurements. Yet this strategy introduces a fundamental limitation: catastrophic forgetting, in which previously learned regulatory knowledge is overwritten during adaptation to new data. Here, we demonstrate for the first time that fine-tuning Enformer, a widely used genomic foundation model, on personalised genomes leads to pervasive catastrophic forgetting across genome-wide regulatory predictions. More broadly, repeated fine-tuning yields a growing collection of specialised models, rather than updating a single model that accumulates regulatory knowledge over time.

To address these challenges, we propose continual learning as a principled training framework for genomic models. Continual learning encompasses a class of methods, including experience replay, regularisation-based strategies, and dynamic architectures, that allow models to incorporate new regulatory data while preserving previously acquired knowledge. Rather than fragmenting into task-specific models through repeated fine-tuning, continual learning enables both forward transfer, in which prior knowledge facilitates learning on new loci or perturbations, and backward transfer, in which newly learned regulatory logic refines earlier predictions. We evaluate continual learning in two complementary settings: first, by adapting a large generalist model, Enformer, to CRISPRi-style perturbation screens of combinatorial cis-regulatory elements; and second, by adapting a specialist transcription initiation model, ProCapNet, to MPRA-style perturbation screens that mutagenise cis-regulatory elements. Across both settings, continual learning substantially mitigates catastrophic forgetting and improves generalisation to diverse forms of genetic variation and out-of-distribution prediction tasks. Together, these results indicate that continual learning offers a practical approach for incorporating newly generated functional genomics data into existing models without overwriting prior regulatory knowledge. By introducing continual learning into genomic modeling, we outline a training paradigm that aligns with the iterative nature of biological data generation and supports the development of shared, incrementally updatable models of gene regulation.

IMPERFECT SEQUENCE MATCHING IS ASSOCIATED WITH HOMOLOGY-DIRECTED DOUBLE STRAND BREAK REPAIR

Simona Dalin[1,2], Sophie Webster[1,2], Rose Gold[1], James Haber[3], Marcin Imieliński[4], Gaddy Getz[1,5], Rameen Beroukhim[1,2]

[1]Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, [2]Dana Farber Cancer Institute, Medical Oncology Department, Boston, MA, [3]Brandeis University, Department of Biology , Waltham, MA, [4]New York University, Pathology Department, New York, NY, [5]Massachusetts General Hospital, Pathology Department, Boston, MA

Rearrangements are genomic alterations that arise when accurate double-strand break (DSB) repair mechanisms fail, causing deletions, insertions, inversions, and translocations. They affect larger regions of genomes than any other variant, covering 300,000x more area than SNVs. However, SNVs have been far better studied, as they are easier to detect. Several mechanisms of DSB repair involve perfectly matching bases immediately adjacent to rearrangement breakpoints, termed microhomology (MH). MH's role in DSB repair and rearrangement formation is well-characterized. However, sequences beyond the perfectly matching bases have not been systematically studied and current models of DSB repair ignore these flanking regions. We hypothesized that imperfect sequence matching (ISM) beyond the MH adjacent to breakpoints may also contribute to DSB repair. We detected and classified ISM near breakpoint junctions via a novel algorithm using Smith-Waterman alignment and a statistical model of background ISM levels. We applied our algorithm to germline and somatic rearrangements in the Cancer Genome Atlas (TCGA) dataset and found more than half of germline rearrangements have significantly more ISM than expected by chance. In contrast, <1% of somatic rearrangements have significant ISM. However, in both contexts, ISM is abundant in rearrangements putatively formed by single-strand annealing (SSA) and homologous recombination (HR), using >20bp of MH. Furthermore, ISM sequence features suggest they may assist with heteroduplex formation during repair. Strikingly, ISM is enriched adjacent to germline SVs with breakpoints in programmed meiotic DSB hotspots. We conclude that ISM plays a major role in homology-directed DSB repair and is a distinguishing feature of genome evolution between germline vs. cancer contexts. Better understanding of its role will illuminate mechanisms of disease formation and therapeutic opportunities that result from improper DSB repair.

HAPLOTYPE ARCHITECTURE SHAPES PHENOTYPIC DIVERSITY ACROSS PLANT AND ANIMAL PANGENOMES

Katharine M Jenike[1], Nicole Brown[2], Sam Kovaka[2], Mattias Benoit[3], Robin Burns[1], Frances Chen[4], Tyler Collins[2], Blaine Fitzgerald[5], Iacopo Gentile[5], Anat Hendelman[5], Delphine Larivière[6], Srividya Ramakrishnan[2], Hagai Shohat[5], Anton Nekrutenko[6], Elinor K Karlsson[4], Zachary B Lippman[5,7], Ian R Henderson[1], Michael C Schatz[2]

[1]University of Cambridge, Plant Sciences, Cambridge, United Kingdom, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]Université de Toulouse, Castanet-Tolosan, France, [4]Broad Institute, Cambridge, MA, [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [6]The Pennsylvania State University, State College, PA, [7]Howard Hughes Medical Institute, Cold Spring Harbor, NY

Pangenomes provide a powerful lens into adaptation and phenotypic variation across agriculture, human health, and fundamental biology. Structurally diverse sequences revealed by pangenomes are increasingly recognized as potential contributors to phenotypic diversity, yet understanding how these sequences are organized and relate to traits remains poorly resolved. Despite rapid progress in genome sequencing, methods for analyzing pangenomes are computationally demanding and have shown limited ability to resolve how complex genetic variation is organized into functional units that shape phenotypes.
Addressing this need, we introduce Panagram (https://github.com/kjenike/panagram), an ultrafast reference-free platform for assembling and annotating the haplotype architecture within a pangenome consisting of evolutionarily coherent blocks of shared sequence variation. The core metric of Panagram is the pan-kmer bitmap that quantifies local sequence similarity across samples and enables rapid, on-the-fly clustering and analyses of haplotypic blocks. These capabilities provide a unified framework for association testing, introgression detection, and the flexible discovery of biologically meaningful variation across pangenomes.
We apply Panagram to multiple plant and animal systems using both short and long read sequencing, including all 605 vertebrate genomes from the Zoonomia Project spanning mammalian and avian clades. In Zoonomia, we use Panagram to identify 1163 genes associated with longevity and metabolic function ($p<4.8e6$). Then in A. thaliana, we analyze 143 globally diverse genomes across 44 environmental conditions, revealing thousands of haplotypic associations ($p<10e10$) within previously known and newly characterized genes. In Solanum, Panagram accurately identifies introgressions in the indigenous crop S. aethiopicum underlying widespread exchange of disease-resistance genes and other agronomically important loci. Finally, using six newly assembled domesticated and mixed-breed genomes in Canis, Panagram places them within the broader evolutionary context of Zoonomia and uncovers breed-specific haplotype blocks, extensive tracts of nonreference sequence, and lineage-restricted variants associated with morphology and disease. These analyses demonstrate how Panagram unifies haplotype-centered pangenome assembly, visualization, and trait association to illuminate architectures that shape phenotypic variation across plants and animals.

FUNDAMENTAL ERRORS IN SINGLE CELL VELOCITY ANALYSIS ARISING FROM THE OMISSION OF CELL GROWTH

Vishal Shah[1], Hia Ming[2], Brian Cleary[1,2,3,4]

[1]Boston University, Bioinformatics, Boston, MA, [2]Boston University, Biomedical Engineering, Boston, MA, [3]Boston University, Computing and Data Sciences, Boston, MA, [4]Boston University, Biology, Boston, MA

A multitude of new methods that measure and analyze single cell “RNA velocity” offer an incredible promise: by estimating kinetic parameters of expression for each gene in many single cells in a population, one can, in principle, piece together observed short-term changes in each cell and map long-term expression trajectories that were never directly observed, charting the paths that single cells take in dynamic processes and providing a foundation for understanding the phenotypic space that cells can occupy. Despite the robust and ongoing development of methods, prevailing RNA velocity frameworks have failed to account for a fundamental aspect of cellular dynamics: cell growth. We show that this significantly constrains the applicability of RNA velocity analysis in proliferating contexts, including many developmental systems of interest. In a growing population, biomass (including RNA and other macromolecules of the cell) is constantly accumulating. This is true too at the single cell level: biomass accumulates from the beginning of cell cycle to the end before division brings daughter cells roughly back to the same size and state. This implies that to keep up with cell growth we expect a homeostatic velocity (defined in the terms of production and degradation) that is positive, which is at odds with the conventional estimation, interpretation, and uses of velocity. We demonstrate systematic errors in interpretation and estimation that arise from ignoring cell growth. We show how inefficient detection and sampling similarly give rise to systematic artifacts. Analysis of existing datasets confirms near universal presence of such artifacts. Finally, we point the way forward for correcting some of these issues and highlight that explicitly accounting for cell growth in the RNA velocity framework can lead to new biological insights. In particular, this view shows that cell growth rate can be a global regulator of gene inducibility, in the sense that inducing large changes in abundance is “easy” in slow growing and “hard” in fast growing cells.

SPATIALLY RESOLVED HOST–MICROBIOME INTERPLAY IN EED USING THE *HOMIC* FRAMEWORK

Mateusz Garbulowski[3], Sara Fernández[1], Sanja Vicković[1,2,3]

[1]New York Genome Center, New York, NY, [2]Columbia University, Department of Biomedical Engineering and Herbert Irving Institute for Cancer Dynamics, New York, NY, [3]Science for Life Laboratory, Uppsala University, Department of Immunology, Genetics and Pathology, Beijer Laboratory for Gene and Neuro Research, Uppsala, Sweden

Environmental enteric dysfunction (EED) is an intestinal condition in children primarily driven by chronic pathogen exposure, which leads to inflammation, impaired nutrient absorption, and poor growth outcomes. Its development arises from limited access to nutritious food and healthcare, along with poor sanitation and hygiene, spanning from pregnancy to early childhood. However, the mechanisms linking maternal gut health, fetal growth, and the subsequent development of EED remain largely unknown.

In this study, we apply high-resolution spatial host–microbiome sequencing (SHM-seq) to characterize bacterial communities associated with EED-related host transcriptional programs during pregnancy. SHM-seq simultaneously maps host gene expression and microbial composition at high spatial resolution, revealing local host–microbiome microenvironments in intact tissues. The complexity of analyzing such host–microbiome data motivates the development of user-friendly, integrated pipelines for data preprocessing and downstream analysis, particularly in spatial transcriptomics. To address this challenge, we developed homic, a host–microbiome analysis framework implemented as a Python package with complementary R functions. Its core feature is a deep learning model that enhances Kraken2 taxonomic classification accuracy by leveraging simulated SHM-seq data. In addition, homic implements a suite of spatial microbiome analysis methods, including data cleaning (host read decontamination), data quality assessment (saturation analysis and species richness metrics), taxonomic profiling (identification of highly abundant bacterial taxa), and network inference (co-abundance networks via generalized boosted regression trees).

Enabling spatially resolved analysis of host–microbiome microenvironments, homic provides a comprehensive framework to explore bacterial communities and their links to host transcriptional programs. Applied to SHM-seq data from EED, it uncovers localized host–microbe relationships that may underlie disease pathophysiology. The Python and R implementations, along with full documentation, are publicly available.

SINGLE-CELL SPLICING ANALYSIS WITH ISSAC UNCOVERS CELL TYPE-SPECIFIC AND CELL STATE-DEPENDENT sQTLs

Yuntian Zhang[1], Wenjing Wang[2], Zhi Yang Tan[2], Yihan Tong[2], Chang Xu[2], Chi Tian[2], Gao Wang[3], Boxiang Liu[1,2]

[1]National University of Singapore, Biomedical Informatics, Singapore, [2]National University of Singapore, Pharmacy and Pharmaceutical Sciences, Singapore, [3]Columbia University , Neurological Sciences, New York, NY

Single-cell RNA sequencing enables comprehensive profiling of gene expression and splicing at cellular resolution, revealing cell type-specific and cell state-dependent regulation (variation within cell types based on their functional states). While genetic studies of expression (eQTLs) in single cells are well established, the genetic regulation of alternative splicing in single cells remains challenging. Existing single-cell splicing QTL (sQTL) studies perform pseudobulk aggregation using bulk analysis methods, which reduces power to detect cell type-specific sQTLs and cannot capture cell state-dependent splicing regulation. Here, we introduce ISSAC to directly quantify metacell-level splice site usage and map cell type- and cell state-specific sQTLs through generalized linear mixed models. In real-world benchmarking on peripheral blood single-cell data, ISSAC identified 1.5- to 2.4-fold more cell type-specific sQTLs than pseudobulk sQTL analysis, and uniquely enabled cell state-dependent sQTL discovery. We applied ISSAC to a harmonized aging brain resource consisting of approximately 3 million dorsolateral prefrontal cortex (DLPFC) single nuclei from 722 donors. ISSAC identified 42,998 independent cis-sQTLs across seven major cell types and 16,861 independent cis-sQTLs across 67 subcell types, with ~70% of sGenes showing no overlap with eGenes. We identified 369 independent sQTLs whose genetic effects were mediated by various cell states such as dendrite development and synaptic signaling. Additionally, we uncovered 194 Alzheimer's-biased and 207 sex-biased sGenes, as well as 142 risk genes that colocalized with neurological disorders including Alzheimer's disease, Neuroticism, Amyotrophic lateral sclerosis, Parkinson's disease, Lewy body dementia and Schizophrenia. Specifically, we functionally validated a causal variant rs11549690 modulating TRPT1 exon 7 skipping to influence neuroticism risk.

THE IMPACT OF NEGATIVE SELECTION ON SVs IN CANCER GENOMES

Shahab Sarmashghi[1,2,3], Ellie R Kim[3,4], Wolu Chukwu[1,2,3], Andrew Cherniack[1,3], Alison Taylor[5], Rameen Beroukhim[1,2,3,4]

[1]Dana-Farber Cancer Institute, Medical Oncology, Boston, MA, [2]Dana-Farber Cancer Institute, Cancer Biology, Boston, MA, [3]Broad Institute, Cancer Program, Cambridge, MA, [4]Harvard Medical School, Medicine, Boston, MA, [5]Columbia University Vagelos College of Physicians and Surgeons, Pathology and Cell Biology, New York, NY

A major goal of cancer genomics has been to identify driver genetic alterations that contribute to cancer development and progression. However, genetic alterations that are under negative selection can also be highly informative because they indicate potential vulnerabilities of cancer cells. In the case of structural variants (SVs), individual genetic alterations can alter many genes--often by generating somatic copy-number alterations (SCNAs) that extend over large genomic loci--entangling positive effects on driver genes with negative "constraining" effects on essential or toxic genes. This creates both a problem and opportunity. The problem is that proper detection of the driver effects of genetic alterations requires disentangling the negative effects they may also engender. The opportunity is that such disentanglement provides insight into negative selection and cancer vulnerabilities. We have developed a method, called TRISCUT, that evaluates both the positive and negative selective effects of SCNAs at every copy-number level to form a unified accounting of the consequences of genetic alterations on cancer progression. Applied to 11,000 cancer genomes, we detect over 250 genomic loci under selection, including novel driver events and cancer vulnerabilities and also an unprecedented resolution into driver and constraining genes at known loci.

GENOME-WIDE CHARACTERIZATION OF CLONAL HEMATOPOIESIS REVEALS EXTENSIVE NON-CODING PUTATIVE DRIVER MUTATIONS

Joshua S Weinstock[1], Karen Conneely[1], Janghee Woo[2], Marios Arvanitis[3], Mitchell J Machiela[4], Cameron Russell[1]

[1]Emory University, Department of Human Genetics, Atlanta, GA, [2]Emory University, Department of Hematology, Atlanta, GA, [3]The Ohio State University, Division of Cardiology, Columbus, OH, [4]National Cancer Institute, Division of Cancer Epidemiology and Genetics, Rockville, MD

As humans age, we acquire somatic mutations in our blood, leading to clonal hematopoiesis (CH). Despite the prevalence of CH in aged individuals, recent searches for selective sweeps in single-cell derived colonies have revealed that most clones have expanded without a known driver mutation. This extensive, unexplained CH motivated our search for novel driver mutations across ~490K blood whole genome sequences from the UK Biobank. We searched across variants with a minor allele count of at least 10 (~147M variants) to discover alleles that are enriched in aged individuals. We identified 45 variants including known CH driver genes (e.g., DNMT3A, ASXL1, SF3B1) and 35 novel variants.

Among the novel age-associated variants, we identified 72 carriers of a somatic mutation in the TERT promoter. Although previously reported in pan-cancer analyses, TERT promoter mutations have typically been excluded from population searches for CH. We also observed a somatic intronic insertion in UGT2B7 in 1,165 carriers, a cluster of IGH point mutations, and centromeric variation. We conducted a phenome-wide association study (PheWAS) among 30 common disease phenotypes to characterize the phenotypic correlates of these mutations, finding 965 links between somatic mutations and common diseases, including 37 protective associations. After estimating the total liability scale variance explained of common diseases by CH mutations, we found that non-canonical CH contributed 28% of the variance explained. We then performed a genome-wide association study of the IGH mutations, finding that common germline variation at GRAMD1B is strongly associated with IGH mutations, and finally a proteome-wide association study to characterize the plasma proteomic correlates of CH. Overall, we characterize CH at both increased breadth and resolution and characterize the entire cascade from upstream germline risk haplotypes to downstream clinical correlates. We release our summary statistics in a publicly accessible portal, somatic.emory.edu.

BREAK-INDUCED REPLICATION DRIVES TELOMERES TO RECOMBINE WITH DNA SATELLITES DURING TELOMERE CRISIS

T. Rhyker Ranallo-Benavidez[1], Yi-An Chen[1], Noelle H Fukushima[1], Ogechukwu Mbegbu[1], Szehoi Chan[2], Tianpeng Zhang[2], Floris P Barthel[1]

[1]The Translational Genomics Research Institute (TGen), Bioinnovation and Genome Sciences, Phoenix, AZ, [2]University of Virginia School of Medicine, Department of Radiation Oncology, Charlottesville, VA

Telomere dysfunction drives genome instability through breakage-fusion-bridge (BFB) cycles, yet BFB alone does not fully characterize the genomic consequences of telomere crisis. A definitive molecular signature linking genome alterations directly to telomere dysfunction remains elusive. To address this fundamental question, we developed a model of crisis by introducing HPV E6/E7 into normal human astrocytes (NHA). Short-read sequencing revealed accumulating copy number alterations over successive population doublings (PD), while multiplexed FISH (M-FISH) showed dynamic cytogenetic abnormalities strikingly enriched at acrocentric chromosomes. Southern blotting and immunofluoresence showed PD-dependent telomere shortening and anaphase bridges, consistent with telomere dysfunction. However, conventional alignment and variant calling failed to capture specific translocation events from sequencing data, despite their abundance on M-FISH, reflecting the inability of standard bioinformatic approaches to detect highly subclonal events and/or events that cannot be confidently mapped to a reference genome. To address this, we developed KaryoScope, a single-molecule reference-free long-read analysis approach inspired by DNA FISH. Applying KaryoScope to NHA cells in crisis, we observed rare but unmistakable long-range recombination events between telomeres and human satellite DNA, including classic satellites (hsat1-3) and alpha satellites. In cells immortalized by alternative lengthening of telomeres (ALT), satellite recombination was highly abundant but events completely disappeared in cells immortalized by telomerase reactivation. Remarkably, inducing telomeric damage in ALT-positive U2OS cells using TRF1-FokI significantly enhanced telomere-satellite recombination and prompted telomeric recruitment of RAD52, implicating break-induced replication (BIR) as the driving mechanism. In a clinical cohort of 20 ALT-positive astrocytoma patients, we observed thousands of telomere-satellite recombination events, indicating an important but underappreciated role in ALT tumor development. Our results demonstrate that telomere crisis invariably scars genomes with a signature reflecting an underlying BIR process.

A NEAR-COMPLETE PANCREATIC TUMOR AND NORMAL GENOME ASSEMBLY-BASED BENCHMARK FOR PERSONALIZED GENOMICS

Justin M Zook[1], Jennifer McDaniel[1], Justin Wagner[1], Chunlin Xiao[2], Keith Oshima[3], Glennis Logsdon[3], Genome in a Bottle Consortium[1]

[1]National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD, [2]NIH, NCBI, Bethesda, MD, [3]University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA

In cancer samples, reference gaps and germline variants have obscured somatic variants and methylation changes in repetitive regions like homopolymers, tandem repeats, centromeric satellites, telomeres, and segmental duplications. To resolve these somatic variants, including complex rearrangements, we construct and curate near-complete haplotype-resolved assemblies of a broadly-consented hypodiploid pancreatic cancer cell line and matched normal tissues. The tumor assembly fully and accurately recapitulates all 35 chromosomes that are consistently observed across tumor cells with spectral karyotyping. We polish the assemblies with accurate short reads to correct errors in homopolymers and short tandem repeats. By directly comparing near-complete tumor and normal assembled haplotypes, we discover many variants missed by typical methods, developing curated clonal somatic small variant, structural variant, and copy number variant benchmarks. We find that most somatic LINE insertions originate from a rare hypomethylated germline LINE insertion present in the normal assembly but absent in GRCh38, highlighting the necessity of personalized references. We precisely resolve sequences of translocations and inversions, most of which are complex, including telomeric and acrocentric translocations. The most complex is a translocation between two different haplotypes of chromosome 19 and 22 involving nested foldback inversions. We confirm 1,340 small variants, 32 insertions and 62 deletions >50 bp, and 10 translocation and inversion breakpoints that do not have clear GRCh38 coordinates, mostly in centromeric satellite regions. Additionally, somatic translocations result in two dicentric chromosomes that each have two putative kinetochores, based on hypomethylation patterns from native long reads. While the focus of this work is characterizing somatic variants present in most cells across passages, these matched tumor and normal assemblies provide personalized references for future work characterizing subclonal variation across different passages and “clonal” lines derived from single cells. Overall, this paired tumor and normal assembly uncovers >4,000 variants altering >1 Mbp of sequence in repetitive regions that have been hidden by reference gaps and germline variants. These challenging variants constitute over half of small indels, large insertions and deletions, and translocations, providing the first comprehensive, curated Genome in a Bottle Consortium benchmark resource for cancer genomics.

TRANSMISSIBLE CANCER: WHEN CANCER CELLS BECOME INFECTIOUS AGENTS

Elizabeth Murchison

University of Cambridge, Cambridge, United Kingdom

Cancer arises when mutations drive cells of the body to proliferate abnormally and to invade the body's tissues. Most cancers exist only within the bodies of the individual hosts that spawn them; rarely, however, cancers in certain animal species acquire the potential to spread between individuals. In such transmissible cancers the cancer cells themselves become infectious agents of disease. Elizabeth Murchison will discuss recent research on the origins and evolution of naturally occurring transmissible cancers in animals.

DETECTING CENTROMERIC FUSION EVENTS IN CANCER GENOMES

Jakob M Heinz[1,2,3], Matthew Meyerson[3,4,5], Heng Li[1,2]

[1]Harvard Medical School, Department of Biomedical Informatics, Boston, MA, [2]Dana-Farber Cancer Institute, Department of Data Science, Boston, MA, [3]Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, [4]Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, [5]Harvard Medical School, Department of Genetics, Boston, MA

Centromeres are essential for proper chromosome segregation, yet their long satellite repeats made them inaccessible to genomic studies until the development of long-read whole-genome sequencing (WGS) technologies, which enabled the Telomere-to-Telomere consortium to assemble human centromere sequences in 2022. However, standard aligners still often fail to unambiguously map reads to repeat regions, causing many reads that support a centromere-spanning structural variant (SV) to be filtered out by downstream analyses. Consequently, centromeres remain systemically underrepresented in cancer genome analyses.

The Human Pangenome Research Consortium's 466 complete genomes have led to significant advances in understanding centromere evolution and germline variation, but they have seen limited use for the study of somatic centromeric mutations. Therefore, we developed a method to rescue ambiguously mapped reads in repetitive DNA by compiling a k-mer database of common centromere-specific sequences and scoring read segments using a maximum log-likelihood framework to detect chromosome-specific centromeric signatures. This approach allows us to identify reads containing centromeric sequences from multiple chromosomes and merge supporting reads into breakpoint calls.

We validated this approach on PacBio HiFi data from the COLO829 melanoma line, identifying a previously karyotyped t(1;3)(q12;p21) centromeric fusion that standard SV callers miss. We are now scaling to an additional 44 paired tumor/normal long-read datasets to quantify the frequency and spectrum of centromere-associated fusions and developing a short-read adaptation to enable cohort-scale screening of the ~8,000 WGS samples in The Cancer Genome Atlas. By making somatic centromeric rearrangements detectable, this work aims to characterize the recurrence of centromere-involving SVs across diverse cancer types.

METHODS TO STUDY MODIFIED T CELL—CANCER CELL BEHAVIORS AND INTERACTIONS IN LIVE-CELL KILLING ASSAYS

Barbara E Engelhardt[1,2], Adam Weiner[1], Justin Adjasu[2], Cole Citrenbaum[3], Julie Tran[1], Stefanie Bachl[4], Scott Linderman[3], Julia Carnevale[4], Alexander Marson[1,4]

[1]Gladstone Institutes, Gladstone Institute, San Francisco, CA, [2]Stanford University, Biomedical Data Science, Stanford, CA, [3]Stanford University, Statistics, Stanford, CA, [4]UCSF, Medicine, San Francisco, CA

Cancer immunologists use killing assays – images across time of modified immune cells co-cultured with cancer cells – to quantify the relative efficacy of modifications to immune cells on the number of cancer cells over time. However, the resulting videos show complex behaviors that are discarded after extracting counts across frames. In prior work, we performed cell segmentation and tracking on modified T cells co-cultured with A375 cancer cells across bright field and RFP channels, where the RFP marks cancer cell nuclei. We built models to quantify length of interactions between T cells and cancer cells, to capture changes in cellular morphology before and after interactions, to quantify proliferation rates, to estimate the rate that T cells inhibit cancer cell proliferation, and to distinguish T cell proliferation from T cell recruitment. These phenotypes capture morphology, motility, and simple behaviors and allow us to consider additional therapeutically-relevant factors when designing new T cell therapies beyond transcriptomic profiles.

Our new work extends these phenotypes in two biologically important ways. First, we design a graph neural network to combine live-cell imaging videos with transcriptomics to estimate – based on paired imaging and expression data such as from Cellanome – cell states, cytokine levels, and morphology-related genes that correlate with cell morphology and dynamics. With this method, we estimate relative time in active states, cytokine production and release, and cellular health of every cell; furthermore, we identify differences in phenotypes across different T cell modifications or cell types. Second, we extend the MoSeq2 state space model, which identifies the “behavioral syllables” ab initio from videos of one or two individuals (e.g., flies, worms, mice) that can be grouped together to characterize the behavioral semantics of the organisms. In particular, we extend the MoSeq2 model to both allow as many as a thousand interacting individuals in one video (here, two types of cells). We identify a number of behavioral syllables in the data that are differential across T cell modification and experimental variables, and that can be combined to identify complex cellular behaviors including evading detection, pack hunting, and fleeing. Our work opens the door to engineering specific behaviors into immune cell therapies through complex behavioral phenotyping of the cells.

THE HUMAN PANGENOME REFERENCE REDUCES ANCESTRY-RELATED BIASES IN SOMATIC MUTATION DETECTION

Chau V Pham[1,2], Farida S Abdelmalek[1,2], Tracy Hua[1,2], Kathleen E Houlahan[1,2]

[1]McMaster University, Department of Biochemistry and Biomedical Sciences, Hamilton, Canada, [2]McMaster University, Centre for Discovery in Cancer Research, Hamilton, Canada

Accurate somatic mutation detection underpins effective biomarker discovery and clinical implementation. Miscalling somatic mutations can exacerbate health disparities, particularly when the likelihood of miscalling is linked to genetic ancestry. Current genomics workflows designed to identify somatic mutations run the risk of potentiating bias towards individuals of European descent. This is in part because genomics workflows require alignment of sequencing reads to a reference genome. Commonly used human reference genomes collapse extensive genetic variability into a single linear genome of which 70% is derived from one donor. Unsurprisingly, these linear genomes fail to capture the full spectrum of genetic variation, which can lead to misalignment of sequencing reads particularly for individuals underrepresented by the linear reference genomes. To address this shortcoming, the Human Pangenome Reference Consortium released the first draft of the human pangenome reference containing 47 phased, diploid assemblies from a diverse catalogue of individuals. The human pangenome reference has shown increased accuracy in detecting germline variants, but it remains to be seen if this benefit will translate to somatic mutation detection. To address if the incorporation of the human pangenome reference in somatic mutation detection workflows can mitigate ancestry-related biases, we systematically benchmarked somatic single nucleotide variant (SNV) detection leveraging the human pangenome in 30 whole exome sequenced bladder tumours with matched blood tissue from The Cancer Genome Atlas (TCGA). We selected bladder cancer because it is a highly point mutated tumour and has well documented differences in incidence, prognosis and somatic mutation landscape linked to genetic ancestry. We found somatic SNV detection leveraging the human pangenome reference outperformed the linear reference, most notably in individuals of East Asian ancestry where we observed, on average, a 20% improvement in detection accuracy as measured by F1-score. Improvements to detection accuracy in individuals of European ancestry were marginal. The increase in accuracy was attributed to reduced germline contamination and reduced reference bias. Finally, we demonstrate somatic SNV detection leveraging the pangenome increases precision comparable to ensemble approaches that take the consensus across multiple tools. Thus, leveraging the pangenome mitigates the need for computationally expensive ensemble approaches. Taken together, these data motivate the use of the human pangenome reference into genomics workflows to ensure equitable and accurate somatic mutation detection.

PERPLEXITY: AN ENTROPY-BASED METRIC FOR QUANTIFYING DIVERSITY IN MULTIOMIC DATA

Megan D Schertzer[1,2], Stella H Park[2], Jiayu Su[3], Gloria Sheynkman[1], David A Knowles[2,3,4]

[1]UVA, Dept. of Molecular Physiology and Biological Physics, Charlottesville, VA, [2]New York Genome Center, Faculty Labs, New York, NY, [3]Columbia University, Dept. of Systems Biology, New York, NY, [4]Columbia University, Dept. of Computer Science, New York, NY

High-throughput omics technologies now routinely detect hundreds of thousands of distinct molecular species. The complexity of RNA and protein output, including extensive isoform diversity and post-transcriptional/post-translation modifications, varies dramatically across genes, cell types, and disease states. Researchers face overwhelming filtering choices during data analysis such as expression thresholds, usage ratio cutoffs, and sample representation requirements. Variance in even one of these dimensions can significantly impact downstream results. Many analysis pipelines lack a consensus on filtering parameters, and in the absence of ground truth to distinguish biologically relevant low abundance molecules from noise, no single parameter set can be objectively justified. This complicates biological interpretation and limits cross-study comparisons.

Omics output is continuous, not binary. Rather than refining arbitrary thresholds and discarding data, we propose a fundamentally different approach to quantify and compare molecular complexity: perplexity. Originally introduced in ecology to measure species diversity, perplexity estimates the effective number of species in a community weighted by their relative abundances. When applied to omics data, perplexity provides an interpretable and systematic metric that condenses a complete abundance distribution into a single value that represents the effective number of distinct molecular variants–whether transcript isoforms or post-translational states. Here, we apply this framework to 124 ENCODE4 PacBio long-read RNA-sequencing datasets across 55 human cell types. We quantify isoform diversity at multiple molecular levels: (1) gene potential–the number of detected isoforms per gene, (2) gene perplexity, (3) protein-coding transcript perplexity, (4) predicted protein (ORF) perplexity, and (5) tissue-specific usage. Among genes with multiple protein-coding transcripts (n = 12,658), we observe an average of 2.1 effective protein isoforms (range: 1.0-32.6), with 4,593 exhibiting tissue-specific isoform expression patterns. We further demonstrate perplexity's utility in proteomics, where it quantifies the effective number of isoforms a given peptide could have originated from based on relative transcript abundances.

Perplexity is easy to compute, straightforward to interpret, broadly applicable to diverse omics modalities, and immediately deployable in existing workflows.

SPACE-TAG ENABLES SPATIAL CHROMATIN PROFILING WITH CUT&TAG

Chao Yan[1], Ruiyang He[2], Sara Fernadez[1], Sanja Vickovic[1,2]

[1]New York Genome Center, Vickovic Lab, New York, NY, [2]Columbia University, Department of Biomedical Engineering and Herbert Irving Institute for Cancer Dynamics, New York, NY

Spatial epigenomics maps chromatin states within tissue context, yet current methods remain limited by accessibility and scale. We developed SPACE-Tag, a spatial CUT&Tag framework that converts epigenetic marks into poly-adenylated RNA through T7-based linear amplification, enabling seamless integration with standard spatial transcriptomic workflows. Applied to the mouse brain, SPACE-Tag resolved activating and repressive chromatin landscapes, faithfully recovering gene-regulatory and anatomical features, and showed strong correlation with existing bulk and single-cell reference data. Extending the analysis to explore spatially-resolved gene regulatory architecture, SPACE-Tag identified three spatial super-enhancer modules enriched for transcription-factor motifs that separate cortex, hypothalamus and fiber tract programs. In addition, we report the first broad-scale tissue-wide cis-regulatory element enrichment analysis resolving a high-resolution functional map of the mouse cortex, including visual, auditory, and somatosensory regions. Finally, leveraging the scalability of SPACE-Tag, we generated an integrated multimodal map of the aging brain, jointly profiling five distinct histone epitopes and gene expression.

ctrl-PASTE DELIVERS TRANSGENES AT HIGH COPY NUMBER BY TARGETING REPETITIVE SEQUENCES ACROSS THE HUMAN GENOME

Kousuke Mouri[1], Ryan Tewhey[1,2,3]

[1]The Jackson Laboratory, Mammalian Genetics, Bar Harbor, ME, [2]University of Maine, Graduate School of Biomedical Sciences and Engineering, Orono, ME, [3]Tufts University School of Medicine, Graduate School of Biomedical Sciences, Boston, MA

The use of high copy number screens has been critical for functional characterization assays aimed at decoding the genome. However, existing approaches have fundamental tradeoffs limiting their utility. Episomal assays achieve high copy number but lack native chromatin context. In contrast genomically integrated approaches such as lentiviral vectors or transposon systems capture local chromatin effects but impose restrictions on the type of cargo that can be delivered, introduce noise due to the randomness of integration sites, or fail to achieve efficient high copy integration, all of which reduce statistical power to detect subtle regulatory effects. No existing tool simultaneously provides flexible transgene structure, directed genomic integrations, and high copy number.
Here we present ctrl-PASTE, a method that resolves this tradeoff by targeting repetitive sequences in the human genome for multi-copy transgene delivery. Our approach builds on the previously developed PASTE framework, which combines prime editing to install an attB landing site at a defined locus with the BxbI serine integrase to deliver a cargo vector of arbitrary size into that site. By directing PASTE to repetitive elements present at high copy number per haplotype, ctrl-PASTE achieves multi-copy integration.
To identify targets with high integration rates, we screened 4,241 prime editing guide RNAs (pegRNAs) spanning multiple repeat classes, including SINEs, LINEs, retrotransposons, and tandem repeats, with genomic copy numbers ranging from 10 to 54,000 per haplotype. Following transfection into HEK293T cells, we captured integrated attB sites together with their flanking genomic sequences and used targeted sequencing to identify which pegRNAs contributed productive integrations. This screen yielded multiple effective pegRNAs, including LINE-1-targeting guides that achieve 8-11 integrated transgene copies per cell.
To evaluate how reporters for *cis*-regulatory elements (CREs) perform at integrated at LINE elements, we delivered a massively parallel reporter assay construct into LINE-1 loci via ctrl-PASTE and observed CRE activity comparable to matched episomal controls (r=0.82), demonstrating that these genomic sites support robust, quantitative reporter expression. Together, these results establish ctrl-PASTE as an alternative high -copy -number transgene delivery strategy that is not limited by cargo restrictions or the stochastic effects of random integrations.

GENOTYPING THE DISTAL JUNCTION OF THE rDNA IDENTIFIES ROBERTSONIAN TRANSLOCATION CARRIERS AND UNVEILS HIDDEN STRUCTURAL POLYMORPHISM

Arang Rhie*[1], Juhyun Kim*[1], Francisco Rodriguez-Algarra[2], Steven Solar[1], Sergey Koren[1], Dmitry Antipov[1], Caralynn M Wilczewski[3], Justin Paschall[3], Tamara Potapova[4], Tyra G Wolfsberg[5], Sumeeta Singh[5], Sandra O Del Castillo Del Rio[2], Clesson Turner[3], Vardhman Rakyan[2], Adam M Phillippy[1], Human Pangenome Reference Consortium (HPRC).

[1]Genome Informatics Section, Center for Genomics and Data Science Research, NHGRI, NIH, Bethesda, MD, [2]The Blizard Institute, School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, [3]Reverse Phenotyping Core, Center for Precision Health Research, NHGRI, NIH, Bethesda, MD, [4]Stowers Institute for Medical Research, Stowers, Kensas City, MO, [5]Bioinformatics and Scientific Programming Core, Office of Scientific Core Facilities, NHGRI, NIH, Bethesda, MD

Balanced Robertsonian translocation (ROB) is the most common chromosomal rearrangement, with an estimated occurrence of 1 in 800 in newborn studies. Carriers are at increased risk of cancer and often diagnosed at fertility clinics after facing recurrent miscarriages, infertility, or aneuploid offspring. Genotyping carriers with sequencing reads has been challenging because of gaps and misrepresentation of the translocation fusion site in the current human genome reference. Only recently, telomere-to-telomere (T2T) human genomes successfully revealed sequences of the acrocentric short-arms, where the most common ROB fusion site is located. A ROB results in loss of two ribosomal DNA (rDNA) arrays and its adjacent distal sequences, including the highly conserved distal junction (DJ). Here, we present “DJCounter”, a novel method to type ROB carriers from short and long reads by copy number estimating the DJs. The method successfully genotypes ROB carriers in previously identified samples using alignments to T2T-CHM13v2, GRCh38 and a reference free approach. Applying the method to a cohort of healthy newborns and family members (n=4,172), we find candidates of unidentified ROB carriers, matching the reported 1 in 800 ratio. Similarly, the ratio was replicated in the UK BioBank (n=490,416). In addition to the ROB carriers, we report cases of one DJ loss or gains of DJs and their frequencies from the two cohorts and the 1KGP (n=3,202), along with the underlying structural composition as found in the Human Pangenome Reference Consortium (HPRC) release 2 assemblies. As this is the first time reporting the polymorphic copy number frequency of the DJs, we conclude a new phenotypic association study is an imminent future direction to understand its clinical implications.

*Contributed equally

LEVERAGING LONG-READ CHROMATIN DATA TO PREDICT FULL-LENGTH RNA ISOFORMS WITH DEEP LEARNING

Gali Bai[1], Nigel Brigstocke[2], Colette Felton[1], Angela N Brooks[1]

[1]University of California, Santa Cruz, Biomolecular Engineering, Santa Cruz, CA, [2]University of California, Santa Cruz, Computer Science and Engineering, Santa Cruz, CA

Predicting RNA isoforms from genomic sequences remains a fundamental challenge. Current models focus on individual components such as transcription start sites (TSSs) or splice sites, rather than complete isoform structures. Using paired long-read chromatin and RNA sequencing in yeast, we found that single-molecule chromatin architecture and sequence are correlated with RNA isoform structure. We then developed a framework for full-length isoform prediction using long-read chromatin data. To validate this chromatin-aware approach, we curated long-read multiomic datasets from human cell lines to train TSS and splice site prediction models. Incorporating epigenetic signals as an additional input channel improved accuracy in predicting both TSS positions and expression levels compared to sequence-only models. These results demonstrate that epigenetic signals provide critical regulatory information and establish a foundation for full-length isoform prediction with long-read chromatin data.

CHROMOSOME-SCALE VOLE ASSEMBLIES VIA CiFi-HiFi SEQUENCING RESOLVE A PRAIRIE VOLE-SPECIFIC AVPR1A DUPLICATION

Mohamed Abuelanin[1], Gulhan Kaya[1], Juniper A. Lake[2], Christine Lambert[2], Ksenia Krasheninnikova[3], Jo Wood[3], Kerstin Howe[3], Jonas Korlach[2], Devanand Manoli[4], Jessica Tollkuhn[5], Megan Y. Dennis[1]

[1]University of California, Davis, Genome Center, MIND Institute, and Department of Biochemistry & Molecular Medicine, Davis, CA, [2]Pacific Biosciences, Menlo Park, CA, [3]Wellcome Sanger Institute, Hinxton, United Kingdom, [4]University of California, San Francisco, Department of Psychiatry and Behavioral Sciences, San Francisco, CA, [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Repetitive and structurally complex genomic regions are a primary source of new gene functions, yet they remain underrepresented in most reference assemblies. CiFi couples chromosome conformation capture (3C) with PacBio HiFi sequencing to produce multi-contact concatemer reads that enhance mapping across repetitive regions (McGinty et al. 2025), previously applied to a fruit fly (~0.6 Gb). Here, we extend CiFi to mammalian genomes for the first time, generating haplotype-resolved, chromosome-scale assemblies for the monogamous prairie vole (Microtus ochrogaster) and its non-monogamous congener, the meadow vole (M. pennsylvanicus) (~2.5 Gb), from a single PacBio Revio SMRT Cell per species.

The prairie vole assembly spans 2.53 Gb across 97 scaffolds (scaffold N50: 89.2 Mb; contig N50: 39.2 Mb; 0.073% gaps), resolving the complete karyotype (2n = 54) across 28 chromosome-scale scaffolds and incorporating 520.8 Mb of sequence previously unplaced in the current reference (MicOch1.0; 6,335 scaffolds; 8% gap). The meadow vole assembly spans 2.36 Gb across 105 scaffolds (scaffold N50: 125.8 Mb). Both species achieved QV>70, 99.3% BUSCO completeness, and annotation of ~20,500 protein-coding genes. CiFi benchmarking against Hi-C shows fewer scaffolds (63 vs. 171), cleaner contact maps, and half the gap content with 4-6x less input.

Over 30 years of research on pair bonding in prairie voles has elucidated regulatory, pharmacological, and neural circuit mechanisms at the vasopressin 1a receptor gene (AVPR1A), with early BAC-based sequencing suggesting a duplicate paralog absent from the current reference. Segmental duplication annotations reveal two AVPR1A genes in the prairie vole assembly: a full-length and a truncated paralog ~900 kb apart. Comparison with the meadow vole assembly shows only a single full-length paralog, suggesting a prairie vole-specific duplication. Read-depth copy number estimation across eight vole species and 17 social behavior candidate genes supports this finding: prairie vole AVPR1A is at a higher copy than in all other species and uniquely so among all genes surveyed. These assemblies place AVPR1A homologs in chromosomal context for the first time, enabling assessment of their potential roles in driving divergent social behaviors between these species.

GENOMIC INSIGHT INTO CHROMOSOMAL FUSIONS AND ADAPTATION TO LEAF-EATING IN HOWLER MONKEYS

Amy Goldberg

University of California-Los Angeles, Los Angeles, CA

Howler monkeys are one of the most folivorous primate, routinely consuming mature leaves and relying on caeco-colic hindgut fermentation estimated to provide up to 31% of daily energy requirements. This difficult diet is poor in nutrients and high in toxins, posing adaptive pressures on their physiology. Here, we combine de novo long-read assemblies and population resequencing of four howler species to test for genetic signatures of adaptation to folivory across timescales. We identify convergence and divergence of pathways between howlers and colobines, another folivorous primate group. Further, even in the absence of scaffolding, we reconstruct chromosomal fusions by mapping long howler contigs to multiple outgroups. This approach recovers 10 of 12 known fusions from cytological data, and proposes over 20 new ones. Demographic reconstruction also suggests extreme species-tree to gene-tree discordance, with large scale ancient migration or hybrid speciation inferred to be the most likely cause based on approximate Bayesian computation. In sum, we provide new genomic resources for an understudied branch of the primate tree, and demonstrate how deeper sequencing of multiple related species can elucidate the genetic basis of traits shared across a genus.

CONCERTED EVOLUTION AND UNORTHODOX RECOMBINATION OF HUMAN SUBTELOMERES

Andrea Guarracino[1], Erik Garrison[2]

[1]Translational Genomics Research Institute, Bioinnovation and Genome Sciences Division, Phoenix, AZ, [2]University of Tennessee Health Science Center, Genetics, Genomics, and Informatics, Memphis, TN

Human subtelomeric regions are among the most dynamic and structurally complex parts of our genome, yet their interchromosomal relationships have remained difficult to characterize due to the limitations of both assembly completeness and alignment methodology. Here we present the most comprehensive survey of subtelomeric sequence relationships to date, leveraging 466 near-complete haplotype assemblies from the Human Pangenome Reference Consortium (HPRC) version 2. To analyze these regions, we introduce the implicit pangenome graph, a reference-free alignment approach that performs all-to-all pairwise comparisons across haplotypes—sampling approximately 12% of all possible combinations—without imposing chromosomal partitioning or positional bias. This yields a truly unbiased view of interchromosomal homology across the pangenome, where every haplotype serves as its own point of reference, allowing a systematic and universal view of human subtelomeric evolution.

A genome-wide survey of alignment identity reveals extended regions of interchromosomal homology spanning tens to hundreds of kilobases at nearly all subtelomeres—comparable in scale to canonical pseudohomologous systems such as PAR2 on the sex chromosomes. This dramatically expands the scope of known pseudohomologous regions in the human genome to include almost all subtelomeric regions. Cladistic analysis based on neighbor-joining trees of subtelomeric similarity uncovers both expected relationships—Xp/Yp and Xq/Yq via the pseudoautosomal regions, acrocentric short arms—and novel associations, including strong 10p–18p homology, a tightly linked clade involving 22q, 21q, 19q, 1q, 13q, and 17q, and extended DUX4-containing homology between 4q and 10q with wide copy number diversity. A large clade of many chromosome arms shares homology at moderate similarity, suggesting broad ongoing interchromosomal exchange. Principal component and community detection analyses of the similarity matrix further resolve subtelomeric clustering across human populations. We hypothesize that these patterns are maintained by recombination facilitated by the physical proximity of subtelomeres at the nuclear envelope, and evaluate this using Hi-C-derived three-dimensional genome maps. Our work exposes the extent to which ongoing recombination shapes these highly dynamic and poorly understood regions of the genome.

LEVERAGING ANCESTRAL RECOMBINATION GRAPHS TO DETECT ADAPTIVE DIFFERENCES AMONG GENE DUPLICATES

Charlotte M LeMay[1,2,3], Liaoyi Xu[2,3], Arbel Harpak[2,3]

[1]University of Texas at Austin, Computer Science, Austin, TX, [2]University of Texas at Austin, Integrative Biology, Austin, TX, [3]University of Texas at Austin, Dell Medical School, Population Health, Austin, TX

Segmental gene duplicates offer a golden opportunity for evolutionary innovation, with famous examples including gene families coding for hemoglobin, opsins, and antifreeze proteins. Until recently, the reliance on short read sequencing limited our ability to resolve differences between highly-similar duplicate gene sequences, study their evolution and detect such adaptations. With the advent of long-read sequencing, there is new potential for this study. However, appropriate genealogical inference methods for the study of gene families using these data are underdeveloped. In particular, segmental gene duplicates experience rapid interlocus gene conversion, the copying of sequence tracts from one gene and "pasting" onto the paralogous gene, which violate assumptions made by state of the art genealogical inference methods such as Ancestral Recombination Graphs (ARG). The local absence of observed gene conversions can also indicate beneficial differences between duplicate genes: if differences are beneficial, selection should act against gene conversion events that homogenize these differences. Here, we develop an ARG-based method for studying duplicated genes. Our method incorporates within-species polymorphism data which provides a resolution improvement over phylogenetic methods. We apply our method to simulated data, and show that it may be used to infer interlocus gene conversion rate, as well as detect selectively maintained "islands" with little gene conversion. Finally, we applied our method to detect IGC islands in human long-read sequencing data from the 1000 Genomes Project to catalogue gene conversion islands in the human genome and hypothesize about the adaptation they may point to.

DISTINCT MECHANISMS OF CNV FORMATION AT HUMAN CHROMOSOME 15Q13.3

Wolfram Höps*[1], David Porubsky*[2,3], DongAhn Yoo[2], Michelle de Groot[1], Amber den Ouden[1], Ronny Derks[1], Kendra Hoekzema[2], Maria del Pilar Caro Martin[4], Alessandro De Falco[5], Nicola Brunetti[5], Christian Schaaf[4], Evan Eichler*[2,6], Christian Gilissen*[1]

[1]Radboud University Medical Center, Department of Human Genetics, Nijmegen, Netherlands, [2]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [3]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, [4]University Hospital Heidelberg and Heidelberg University, Institute of Human Genetics, Heidelberg, Germany, [5]Telethon Institute of Genetics and Medicine (TIGEM), Department of Translational Medicine, Pozzuli, Italy, [6]University of Washington, Howard Hughes Medical Institute, Seattle, WA
*authors contributed equally

Human chromosome 15q13.3 is a hotspot for recurrent copy-number variations (CNVs) associated with neurodevelopmental disorders, epilepsy and schizophrenia. Segmental duplications (SDs) facilitate these rearrangements but have also impeded their sequence-level resolution, leaving mutational mechanisms unresolved.

We generated long-read, haplotype-resolved de-novo assemblies for 10 patient-parent trios carrying long ("BP4-BP5", n=6) or short ("CHRNA7", n=4) 15q13.3 CNVs, including five *de novo* events. We compared these structures to 581 diverse human haplotypes to assess structural diversity, and also to six primate species to investigate evolution of the 15q13.3 locus.

Our assemblies fully resolve all 15q13.3 CNVs and most parental alleles. Both types of 15q13.3 CNVs arise predominantly through non-allelic homologous recombination mediated by inversion polymorphisms that reconfigure SD architecture (inv-β for BP4-BP5, inv-ɣ for CHRNA7). While most CNVs are structurally distinct, three de-novo BP4-BP5 breakpoints cluster in a 2kbp recombination hotspot enriched in PRDM9 motifs. Inv-β is common in European samples (58% allele frequency) but rare in East Asian samples (2%), indicating strongly population-stratified susceptibility to BP4-BP5 CNVs. Inv-ɣ, mediating CHRNA7 CNVs, occurs in ~19% of individuals but lacks comparable population stratification. Comparison with six primate species shows rapid evolutionary turnover at 15q13.3, with CNV-promoting expansions being exclusive to humans.

Parental inversion polymorphisms determine 15q13.3 CNV risk in offspring, providing a mechanistic basis for recurrent rearrangement susceptibility and creating predictable breakpoints and distinct molecular forms of the "same" CNVs. Similar mechanisms are likely at play at many recurrent CNV loci in humans.

WHEN CLUSTERS MISLEAD: VISUALIZING OVERLAP AND UNCERTAINTY IN DIMENSIONALITY REDUCTION OF THE 1000 GENOMES PROJECT

Jasmine Liu[1], Alex Diaz-Papkovich[2], David Laidlaw[1], Sohini Ramachandran[2,3]

[1]Brown University, Department of Computer Science, Providence, RI, [2]Brown University, Data Science Institute, Providence, RI, [3]Brown University, Department of Ecology, Evolution, and Organismal Biology, Providence, RI

Dimensionality reduction (DR) methods, such as principal component analysis (PCA) and uniform manifold approximation and projection (UMAP), are widely used in population genetics to visualize patterns of population structure and ancestry. While these methods provide intuitive, low-dimensional maps of high-dimensional genetic data, they can also overemphasize discrete separation between populations and convey a sense of precise individual placement, making such visualizations susceptible to misinterpretation and misuse.

We present a visualization framework of genotype data from the 1000 Genomes Project that integrates DR plots with a spectrum of complementary views, ranging from variant-level overlap to individual-level uncertainty. Inspired by prior work on visualizing human genetic diversity[1], our dashboard combines Euler diagrams that summarize shared and population-specific genetic variants with PCA and UMAP scatterplots of these variants to observe how SNP selection influences the formation of population clusters. Additionally, we offer DR plots that reveal how individual placements depend on genetic context. Rather than treating an individual as a fixed point, we represent points as distributions across common-variant subsamples, allowing viewers to identify which individuals are more or less variable. Contour overlays illustrate the range of positions for selected individuals, emphasizing that a sample's location in an embedding is the result of decisions around data selection and filtering.

By linking these views, the dashboard enables a comprehensive, interpretive workflow. Euler diagrams provide a high-level overview of genetic similarity among groups of individuals. DR plots show how structure emerges from the shared or unshared variants. Lastly, stability overlays reveal where the structure is visually unstable, which tends to be concentrated at visual boundaries of population clusters and among admixed populations. Together, these coordinated views help explain why populations that strongly overlap in variant space can appear separated in low-dimensional projections, and why such separations should be interpreted with caution.

Our method enables population geneticists to reframe the visualization of SNP data, highlighting DR plots as context-dependent summaries whose interpretability varies across the genome. By making uncertainty and overlap explicit, the dashboard promotes a broader understanding of the apportionment of human genetic diversity, discourages essentialist interpretations of population clusters, and supports responsible use of visualization in genomics research and communication.

[1]Kitchens, J., Coop, G. "Visualizing Human Genetic Diversity." (2023) https://james-kitchens.com/blog/visualizing-human-genetic-diversity

CRYPTIC STRUCTURAL VARIATION IN THE MUCIN PAN GENOME AND DISEASE IMPLICATIONS

Elizabeth G Plender[1,2], Jiadong Lin[1], Timofey Prodanov[3], Isaac Wong[1], Katherine M Munson[1], Wanda K O'Neal[4], Tobias Marschall[3], Jesse D Bloom[2,5], Evan E Eichler[1,5]

[1]University of Washington, Genome Sciences, Seattle, WA, [2]Fred Hutch Cancer Center, Basic Sciences Division, Seattle, WA, [3]Heinrich Heine University, Institute for Medical Biometry and Bioinformatics, Dusseldorf, Germany, [4]University of North Carolina at Chapel Hill, Marsico Lung Institute, Chapel Hill, NC, [5]University of Washington/Fred Hutch, Howard Hughes Medical Institute, Seattle, WA

Mucins are large glycoproteins that provide hydration and barrier functions in epithelial tissues. Although genetically heterogenous, all mucins contain an exon of variable number tandem repeats enriched for glycosylated serines and threonines. Limitations in short-read sequencing have hindered exploration of how these VNTRs affect protein function and epithelial disease states. We leverage long read genomes (n = 296) from the Human Pangenome Reference Consortium and the Human Structural Variation Consortium to characterize 17 canonical mucin loci across populations. We accurately assemble at minimum 572 haplotypes (97%) of each gene and construct their phylogenetic history. The secreted mucins *MUC2/5AC/5B* harbor VNTRs that trend shorter in human alleles than the non-human primates ($p<2e-16$), while longer alleles are common in tethered mucins *MUC1/3A/7/12/22* ($p<0.05$). *MUC4* harbors the highest number of distinct length alleles (n=240), while *MUC12* has the largest size differential between alleles (55,233 base pairs) and the largest predicted protein (23,080 amino acids). Eleven of the mucins (*MUC1/2/3A/4/5AC/5B/6/12/17/19*) are significantly population stratified ($p<0.05$). In the mucin region of chr3q29, we find 19 structural configurations due to a recurrent inversion that mediates copy number variability in *MUC20*, with 40% of haplotypes harboring 2 or more complete paralogs. Additionally, we genotyped the canonical mucins using Locityper with an expanded reference set, yielding ~95% or more genotype concordance for allele haplogroups in 15 loci (excluding *MUC2* and *MUC12*). We genotyped short read sequencing data from 4,637 patients in the Cystic Fibrosis Foundation using this method and found a significant association between short alleles of *MUC1* and severe cystic fibrosis (pFDR = 0.0056). These findings deconvolute mucin variation via a phylogenetic pangenome approach and underscore the importance of future disease association.

HERITABILITY OF GERMLINE MUTAGENESIS IN 40 LARGE THREE- AND FOUR-GENERATION PEDIGREES

Michael E Goldberg, Alexis C Garretson, Camila Goclowski, Thomas A Sasani, Hannah C Happ, Julia Ostrander, Lynn B Jorde, Deborah W Neklason, Aaron R Quinlan

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

Germline mutations are the basis for genetic disease and underlie all heritable phenotypic variation on which evolution can act. Estimating mutation rate is therefore critical for modeling disease burden and selection. Mutation rate is a polygenic trait, affected by both genetic and environmental factors that modulate DNA damage, repair, and replication pathways. In the human germline, parental sex and age strongly predict mutation rate; few studies, however, have been able to identify loci that commonly affect it or measure its heritability. Nevertheless, higher germline mutation rate is associated with earlier mortality, hinting at shared architecture between mutagenesis and health.

Here, we present results from the first phase of the Gametes Through Generations (GTG) project, which comprises new whole genome sequencing of >1000 individuals from 22 4-generation and 18 3-generation CEPH/Utah pedigrees. Each family's 3rd and 4rd generations include a median of 8 and 4 children per couple, respectively. Prior studies of mutation rate heritability in humans have been limited to single-nucleotide mutations observed in trios. The large GTG pedigrees allow us to measure germline mutation in hundreds of individuals, eliminate false positives, and assign mutations to a parent-of-origin.

Because germline mutagenesis is a low count Poisson process, its inherently low signal-to-noise ratio clouds inference of heritability, especially in trios. Indeed, we find much higher power to detect nonzero heritability in GTG using simulated data and test traits. The large number of children per couple also allows us to measure repeatability, a statistic that marks an upper bound for its heritability, scales negatively with shot noise, and has never been inferred for mutation rates. We find that, even in GTG, the high variance in mutation rate contributes to a low repeatability of 0.15 and 0.51 for maternal and paternal mutation rates, respectively. Thus, detection of nonzero maternal mutation rate heritability may be impossible given ours and other modern datasets.

Accordingly, initial results using GTG DNMs show low to zero heritability for both maternal and paternal mutation rate. While preliminary, these findings indicate that the GTG dataset provides novel resolution critical for accurately estimating the heritability of germline mutagenesis and, as we will present, related genomic traits. The sequencing of these pedigrees has broad implications for both molecular evolution and genomic medicine, helping quantify an individual's unique risk for mutational burden.

THE AGOUTIC AGENT FOR LONG-READ GENOMIC PROCESSING AND ANALYSIS

Elnaz Abdollahzadeh[1,2], Ali Mortazavi[1,2]

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

Long-read sequencing technologies have transformed transcriptomics and epigenomics, yet the operational complexity of processing these data, spanning pipeline configuration, workflow management, and multi-step result interpretation, remains a significant barrier, particularly for non-computational researchers. Here we present the Agoutic, which is a Large Language Model (LLM) Agent built on the Model Context Protocol (MCP) that supports end-to-end processing and analysis of long-read genomic DNA (gDNA), direct RNA (dRNA), cDNA, and Fiber-seq samples. Agoutic coordinates planning, data retrieval, Nextflow pipeline execution via Dogme, and interactive result analysis through four MCP-connected servers exposed in a notebook-like interface. Domain-specific skill prompts guide the agent through bioinformatics workflows, while human-in-the-loop approval gates ensure that critical decisions such as pipeline execution remain under researcher control. The agent dynamically discovers tools via MCP, transparently routing requests across local clusters or consortium-level genomic databases to maintain reproducibility and auditability.
We applied Agoutic to uniformly reprocess ENCODE PacBio and Nanopore datasets with updated GENCODE annotations to identify differentially expressed novel isoforms across different conditions. For the subset of dRNA samples, we further characterized RNA modification patterns. By externalizing these workflows into a consistent, LLM-accessible interface, Agoutic lowers the barrier for researchers to run sophisticated long-read analyses while preserving the rigorous provenance required for publication. The agent does not replace the researcher but empowers them to focus on the underlying biology rather than the mechanics of pipeline execution and data wrangling.
Keywords: long-read sequencing, LLM agent, Model Context Protocol, Nextflow, bioinformatics orchestration, RNA modifications, genomic analysis

COMPREHENSIVE BENCHMARKING OF SOMATIC MUTATION DETECTION BY THE SMaHT NETWORK

The SMaHT Network, Alexej Abyzov[1]

[2]Mayo Clinic, Quantitative Health Sciences, Rochester, MN

Somatic mosaicism is increasingly recognized as a fundamental feature of human biology, yet the detection of somatic mutations remains challenging. The SMaHT Network conducted four large-scale benchmarking experiments to evaluate sequencing technologies, experimental approaches, and computational methods for detecting diverse somatic mutations. Cumulative sequencing coverage exceeded 1,000× with short reads and 100-400× with long reads for each of nine analyzed samples. We defined optimal strategies for integrating bulk short- and long-read sequencing for mutation detection and demonstrated that using donor-specific assemblies and human pangenome improved variant calling and extended mutation catalogs to challenging genomic regions. We benchmarked six duplex-seq technologies and showed that single-cell sequencing resolves cell type-specific mutational patterns and heterogeneity. Our results indicate that bulk, single-cell, and duplex analyses are complementary – and leveraging all three provides comprehensive characterization of mosaicism within a tissue. Together, these findings provide a roadmap for accurate, genome-wide somatic mutation discovery and analysis.

LEARNING CELL STATES FROM CO-EXPRESSION MODULES IN SINGLE-CELL DATA.

Sandesh Acharya[1,3,4,5], Dinghao Wang[2], Jiami Guo[1,3,4,5], Qingrun Zhang[2,4,5,6]

[1]University of Calgary, Department of Medical Science, Calgary, Canada, [2]University of Calgary, Department of Mathematics and Statistics, Calgary, Canada, [3]University of Calgary, Department of Cell Biology and Anatomy, Calgary, Canada, [4]University of Calgary, Alberta Children's Hospital Research Institute, Calgary, Canada, [5]University of Calgary, Hotchkiss Brain Institute, Calgary, Canada, [6]University of Calgary, Data Science Advisory Unit, Calgary, Canada

Genes do not function in isolation; they form intricate networks within cells, influencing various cellular processes like differentiation, development, and response to stimuli. This interplay involves the coordinated regulation of multiple genes, which contributes to the establishment or maintenance of distinct cell types. We propose that each cell type can be characterized by sets of genes that exhibit strong correlations with one another. Building upon this premise, we have developed a novel biclustering algorithm tailored for single-cell data analysis, leveraging gene-gene correlation patterns. Correlated gene networks are first identified from expression data using a simplified WGCNA-based framework, after which cells are partitioned into two groups using these gene signatures via K-means clustering. This process is applied recursively to each subcluster until clusters become sufficiently small or no additional correlated gene networks are detected, producing a dendrogram that captures hierarchical relationships between cell populations along with the gene modules that distinguish them.

To validate the effectiveness of this approach, we applied the algorithm to seven independent single-cell RNA-seq datasets spanning diverse biological contexts, as well as simulated single-cell datasets. Our method outperformed state-of-the-art clustering methods such as Seurat, SC3, Sincera, and Runibic based on the Adjusted Rand Index (ARI), while consistently identifying biologically meaningful clusters. Importantly, the gene modules identified at each iterative step included both differentially expressed and non-differentially expressed genes that collectively encode cell identity and function. Notably, many lowly expressed but biologically critical genes, including transcription factors and signalling molecules, were captured within these gene networks, highlighting the method's ability to detect subtle regulatory programs often missed by conventional approaches. Moreover, these gene modules provide functional insight into the regulatory mechanisms underlying each cell population and can be leveraged for downstream analyses such as trajectory inference, enabling the identification of lineage relationships, transitional cellular states, and dynamic biological processes in single-cell data.

THE DEVELOPMENTAL GTEx RESOURCE ENABLES THE DISCOVERY OF DISEASE-ASSOCIATED GENES

Sofia Salazar-Magaña[1], Temidayo Adeluwa[1,2], Sarah Sumner[1], Rebecca Keener[3], Winona Oliveros-Diez[4], Hae Kyung Im[1,2], and the dGTEx Consortium[5]

[1]The University of Chicago, Department of Medicine, Section of Genetic Medicine, Chicago, IL, [2]The University of Chicago, Genetics, Genomics, and Systems Biology, Chicago, IL, [3]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [4]KTH Royal Institute of Technology, Department of Gene Technology, SciLifeLab, Solna, Sweden, [5]The Broad Institute, MIT and Harvard, Cambridge, MA

The Developmental Genotype-Tissue Expression (dGTEx) project collected a resource of pediatric (ages 0–18) human tissues and corresponding gene expression data to enable the study of transcriptome regulation and variation in childhood. Unfortunately, small dGTEx sample sizes preclude traditional transcriptome-wide association studies (TWAS) to find disease genes. We circumvented this issue by leveraging DNA sequence-to-epigenome methods and developed tissue-specific SNP-based predictors of gene expression in 72 contexts, spanning 24 tissues and 4 developmental stages (ages 0–2, 2–8, 8–12, and 12–18).

Following the scPrediXcan approach (Zhou et al. 2025), we first trained deep learning-based predictors of gene expression, achieving high correlations with observed expression (median correlation = 0.85 across the genome). We translated these models into SNP-based predictors to address the computational burden of running TWAS at scale and to maintain compatibility with current summary-based TWAS methods. The deep learning– and SNP-based models were highly correlated (median correlation = 0.89).

To examine the value of dGTEx resources, we performed TWAS of 136 GWAS traits (99 adult traits and 37 pediatric traits) across all contexts and compared dGTEx TWAS performance against the traditional elastic net predictors based on adult GTEx. We also developed scPrediXcan models for adult GTEx to compare developmental and adult results using the same prediction framework for a fair comparison.

Using Open Targets (Koscielny et al. 2017) gene-disease associations as a proxy for ground truth, we computed the precision and recall of these predictors and found that our dGTEx TWAS yielded higher precision than the traditional elastic net models trained in adult GTEx. The precision of dGTEx and adult GTEx scPrediXcan-based models were similar.

dGTEx-based TWAS yielded at least 10% more discoveries than adult GTEx-based TWAS across all traits and at least 14% more across pediatric traits. This highlights the value of the dGTEx resource for investigating the biology of pediatric complex diseases.

In summary, by leveraging the dGTEx dataset and scPrediXcan, we have developed predictors of gene expression across multiple developmental stages that show better power than adult GTEx models.

CELL-RESTRICTED EXPRESSION OF RARE VARIANT-ASSOCIATED GENES UNDERLYING PROTECTION AGAINST ALZHEIMER'S DISEASE PATHOLOGY

Quadri Adewale[1,2], Eric Sun[1,2,3], Isabel Castanho[1,2], Pourya Naderi[1,2], Winston Hide[1,2]

[1]Beth Israel Deaconess Medical Center, Pathology, Boston, MA, [2]Harvard Medical School, Pathology, Boston, MA, [3]McGill University, Bioengineering, Montreal, Canada

Background: A substantial proportion of individuals retain normal cognitive function despite harboring advanced Alzheimer's disease (AD) neuropathology, a phenomenon referred to as cognitive resilience. Although large-scale genome-wide association studies have demonstrated that common AD risk variants predominantly converge on microglial and astroglial immune pathways, the cellular mechanisms mediating naturally occurring genetic resilience remain poorly understood.
Methods: To investigate how rare genetic variation may preserve cognitive function, we integrated whole-genome sequencing from multiplex AD families with single-nucleus transcriptomic profiles derived from the dorsolateral prefrontal cortex. Using extreme phenotype sampling, we stratified subjects into AD, resilient, and control. We prioritized putatively regulatory rare variants categorized as either risk-associated or protective. Variant-to-gene mapping was performed using PsychENCODE-validated regulatory annotations and compared against AlphaGenome directional effect predictions. We then mapped directionally annotated variants to cell-state-specific transcriptional programs observed in AD and resilient individuals.
Results: AlphaGenome-based gene assignment demonstrated stronger concordance with PsychENCODE regulatory annotations than nearest-gene mapping approaches, consistent with the long-range regulatory effects of enhancers. Risk-associated rare variants predominantly mapped to genes predicted to be upregulated in reactive glial states, promoting astrocyte activation and microglial inflammatory programs (e.g., EGFR, CLEC2D). In contrast, protective rare variants displayed highly restricted expression patterns, with significant enrichment in neuronal populations. Notably, protective variants were biased in expression preserving specific subsets of inhibitory interneurons, including SST+ subtypes, potentially maintaining excitatory–inhibitory balance. These findings suggest that genetic resilience is not merely characterized by reduced microglial activation, but instead reflects active homeostatic stabilization within neuronal circuits. Protective variants predicted to increase gene expression preferentially mapped to transcriptional programs relating to neuron–oligodendrocyte interaction and genes involved in synaptic structural support (including NRCAM and LRRC4C). Importantly, these cell-type–restricted effects appear to be independent of APOE4 carrier status.
Conclusion:
Rare-variant genetic architecture in AD reveals distinct, cell-state-restricted transcriptional effects. Systematic investigation of protective genetic mechanisms within defined neuronal populations provides a framework for prioritizing therapeutic strategies that emulate naturally occurring cognitive resilience.

FROM SNP TO SIGNALING: GENETIC MODIFIERS OF ODORANT RECEPTOR ACTIVATION BY ONION AND GARLIC COMPOUNDS

Jeremy L Aguilar[1,2], Mona A Marie[2], Hiroaki Matsunami[2]

[1]Duke University Trinity College, Biology, Durham, NC, [2]Duke University School of Medicine, Molecular Genetics and Microbiology, Durham, NC

Odorant receptors (ORs), found on olfactory sensory neurons in the nasal cavity, interact with complex volatile compounds, enabling animals to detect and differentiate both important olfactory and gustatory cues. Though, with humans possessing around 400 ORs and mice over 1,000, each detecting to various degrees, deciphering the intricacies of the olfactory code has proven challenging. Additionally, an added layer of genetic variability amongst individuals produces an array of phenotypic irregularities, at times causing dietary preferences as olfactory sensation works in tandem with gustation to create our perception of flavors. Previous genome-wide association studies (GWAS) have offered single nucleotide polymorphisms (SNPs) as possible explanations for dietary inclinations towards or against certain foods, possibly affecting dietary health. Yet, experimental validation of which OR, and whether the alteration affects receptor activity or expression, is lacking, as the localization of these SNPs in OR clusters has not been dissected. Additionally, poor cell surface expression of ORs *in vitro*, a common hurdle in studying their activation, further contributes to the lack of experimental validation. Thus, we pinpointed previously discovered SNPs to their respective gene (OR4K17, believed to perceive garlic and onion) and identified a ligand, Allyl Methyl Sulfide (found in both garlic and onion) that activates OR4K17 using luciferase assay odorant screens, guided by previous Mass Spectrometry (MS) food profiles. Now, by inducing all known human SNPs with site-directed mutagenesis, we have determined ligand-induced activation is altered, providing a probable explanation for the phenotypic observations in human food preference, being influenced by olfaction and not just gustation. Lastly, we have utilized Alphafold3 to demonstrate the ligand-receptor activation mechanism in both the wild type and mutant OR. Our data aids in further comprehending the role genetics play in determining an individual's olfactory perception and gustatory proclivities, altering an individual's sensitivity to certain foods, consequently influencing their diet and overall health.

MODEL-BASED INFERENCE OF REGIONAL AFRICAN CONTRIBUTIONS IN AFRICAN-AMERICAN GENEALOGIES USING TRANSATLANTIC SLAVE TRADE VOYAGE RECORDS

Kennedy Agwamba, Noah Rosenberg

Stanford University, Biology, Stanford, CA

The Transatlantic Slave Trade was transformative in shaping the genetic landscape of modern North America. Today, millions of people in the United States who identify as Black or African-American descend from forcibly transported captives arriving from multiple coastal regions of Africa. However, the systematic erasure of genealogical records under slavery has obscured links between modern descendants and their African ancestry. Here, we integrate data from the Transatlantic Slave Trade Database with a mechanistic model of African-American demographic history to estimate the regional composition of transported African genealogical ancestors. Using time-resolved records of embarkation from eight major African coastal regions, we model how ancestry is transmitted across generations to quantify the proportional contributions of each region to the pedigrees of mid-20th century African-American descendants. Our results indicate that ancestry reflects substantial contributions from multiple regions, with particularly large fractions tracing to the Bight of Biafra, Senegambia, and West Central Africa. Notably, the relative genealogical impact of the various regions does not strictly mirror the total numbers of captives transported. Differences in arrival timing and subsequent population growth substantially shape long-term ancestry patterns, amplifying some regions' genealogical contributions relative to others. These findings provide quantitative insight into the structured African ancestry embedded within African-American family trees and demonstrate how integrating historical records with mechanistic admixture models can recover genealogical patterns. More broadly, this framework offers a generalizable approach for reconstructing structured ancestral contributions in admixed populations.

LONG-READ ISOFORM SEQUENCING REVEALS EXTENSIVE TRANSCRIPTOMIC DIVERSITY IN BIPOLAR DISORDER AND SCHIZOPHRENIA

Nirmala Akula[1], Andy Qi[2], Qing Xu[3], Pavan Auluck[3], Stefano Marenco[3], Francis J McMahon[1]

[1]National Institutes of Health, Human Genetics Branch, National Institute of Mental Health, Bethesda, MD, [2]National Institutes of Health, National Institute of Aging, Bethesda, MD, [3]National Institutes of Health, Human GenetiHuman Brain Collection Core, National Institute of Mental Health, Bethesda, MD

The brain transcriptomic architecture of complex neuropsychiatric disorders such as bipolar disorder (BD) and schizophrenia (SCZ) remains incompletely resolved due to limitations of short-read RNA sequencing in reconstructing full-length isoforms. Although prior short-read analyses identified differential expression and alternative splicing events in BD and SCZ (Akula et al., 2021), the structural diversity of transcripts—including novel isoforms and complex splice variants—has not been comprehensively characterized.

Here, we applied long-read RNA sequencing to profile isoform-level variation across 135 postmortem brain samples (52 controls, 47 BD, 36 SCZ). Full-length cDNA libraries were sequenced using the PacBio Iso-Seq platform to generate high-fidelity, full-length cDNA reads. Data were processed through the Iso-Seq workflow, including primer trimming (lima), refinement and polyA/concatemer removal (isoseq refine), and transcript classification and filtering against reference annotations (pigeon).

Across samples, sequencing yielded consistent depth ($1.7–2.2 \times 10^7$ reads per sample), enabling robust isoform discovery. We identified millions of transcript features spanning diverse structural classes, including novel features within known reference loci (1,050,439), multi-exon transcripts with at least one annotated splice junction (809,234), intergenic transcripts (560,637), and retained introns (142,418). Following stringent filtering based on transcript length and read support (≥3–5 reads per sample), a high-confidence combined annotation was generated for downstream quantification and differential analysis using StringTie and Ballgown.

These results highlight the extensive and previously underappreciated isoform complexity present in the human brain. Ongoing analyses focus on differential isoform expression between diagnostic groups and integration with proteomic data from the some of the same samples to estimate the fraction of novel transcripts that are translated and refine protein isoform annotation. By resolving full-length transcript structures, this work advances isoform-resolved neuropsychiatric genomics and provides a foundation for identifying high-resolution molecular targets for therapeutic development.

MEXICAN BIOBANK ANALYSES OF ARCHAIC INTROGRESSION REVEAL GEOGRAPHIC STRUCTURE AND SIGNALS OF ADAPTIVE INTROGRESSION

Valeria Anorve-Garibay[1], Jazeps Medina-Tretmanis[1], Lourdes Garcia-Garcia[3], Maria Tusie-Luna[3,4], Maria Avila-Arcos[5], Andres Moreno-Estrada[7], Mashaal Sohail[6], Diego Ortega-Del Vecchyo[5], Emilia Huerta-Sánchez[1,2]

[1]Brown University, Center for Computational Molecular Biology, Providence, RI, [2]Brown University, Department of Ecology and Evolutionary Biology, Providence, RI, [3]Encuesta Nacional de Salud. Instituto Nacional de Salud Publica, INSP, Cuernavaca, Mexico, [4]Unidad de Biologia Molecular y Medicina Genómica, Universidad Nacional Autónoma de Mexico, CDMX, Mexico, [5]Universidad Nacional Autónoma de Mexico, Laboratorio Internacional de Investigación sobre el Genoma Humano, Queretaro, Mexico, [6]Universidad Nacional Autónoma de Mexico, Centro de Ciencias Genomicas, Cuernavaca, Mexico, [7]Centro de Investigación y Estudios Avanzados del IPN, Unidad de Genomica Avanzada, Irapuato, Mexico

Research has shown that out-of-Africa modern humans interbred with archaic hominins ~50,000 years ago. Detection of archaic introgressed segments has been carried worldwide mainly leveraging samples from the 1,000 Genomes Project populations, including American individuals such as Mexicans in Los Angeles. However, this cohort does not capture the extensive geographic and genetic diversity within Mexico, limiting our ability to explore archaic ancestry variation and assess its contribution to complex traits. Here, we characterize archaic introgression in 5,833 imputed genomes from the Mexican Biobank (MXB), one of the largest nationwide datasets from the Americas analyzed for archaic ancestry. We evaluated whether imputed genotype data can reliably detect introgressed segments by benchmarking against high-coverage whole-genome sequencing (WGS) data from 50 individuals. We observed that imputation recovered 99.3% of high-confidence archaic segments identified from WGS and increased sensitivity for detecting introgressed tracts. After applying stringent filtering, we retained 3.36 million segments covering 0.96 Gb (38%) of the callable genome. MXB individuals carry on average 74 Mb (1.45%) of archaic ancestry (1.31% Neanderthal and 0.06% Denisovan). We observed that archaic ancestry follows a north-south gradient within Mexico and is positively correlated with Indigenous American ancestry. We assessed the impact of post-colonial admixture on the distribution of introgressed segments by investigating how different combinations of continental ancestry harbor varying levels of introgression. We find that archaic segments are enriched on AMR,AMR and AMR,EUR ancestry tracts, demonstrating that recent admixture redistributed archaic variation across Mexico. To investigate adaptive introgression, we identified regions in the top 1% of introgressed haplotype frequency. We identify a high-frequency Neanderthal-derived haplotype at HDLBP (47%) that is significantly enriched on Indigenous American local ancestry tracts ($p=2.7\times10^{-12}$). At this locus, AMR ancestry is strongly associated with lower cholesterol levels (-19.57 mg/dL, $p=2.7\times10^{-4}$), linking archaic introgression to ancestry-specific lipid variation. Our results show that imputed genomes enable large-scale inference of archaic introgression, reveal how recent admixture reshaped archaic variation in the Americas, and identify candidate loci through which archaic DNA continues to influence human biology.

A NOVEL METHOD FOR INFERRING DEMOGRAPHIC HISTORY AND STRUCTURE FROM THE DISTRIBUTION OF HETEROZYGOUS SITES

Tommaso Stentella[1], Paul Etheimer[2], Florian Massip[2], Michael Sheinman[3], Peter F Arndt[1]

[1]Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Berlin, Germany, [2]Mines Paris, PSL University, Centre for Computational Biology, Paris, France, [3]Weizmann Institute of Science, Department of Physics of Complex Systems, Rehovot, Israel

In the last two decades, several methods to infer the demographic history of a population from whole genome sequence data have been developed. Despite these efforts, current methods remain computationally demanding, limiting researchers' ability to efficiently explore parameter space and to estimate confidence intervals of model parameters. As a consequence, it is difficult to quantitatively study natural populations, where the true history is unknown, and to decide whether inferred demographic events are accurate. This is especially true if data are limited, for instance when only one genome is available.

Here, we consider the distances between heterozygous sites in single diploid samples, deriving novel and simple analytical results for their distribution under varying population histories. These theoretical expectations allow us to efficiently fit and compare a large number of parametric models of varying complexity. The most probable model output by our method automatically partitions the demographic history into an optimal number of epochs having arbitrary duration, with parameters and confidence intervals estimated jointly. Moreover, we provide a distribution of histories that are compatible with the data.

We show that our method can quickly infer thousands of unique demographic histories, for instance from each of the single genomes in the 1000 Genomes Project. Such analyses reveal fine-scale variation at the individual level within and across populations, additionally allowing us to explore population structure through time. Our new method and analysis enhance our understanding of the extent to which demographic history can be reconstructed from whole genome sequence data in any species.

RESOLVING COMPLETE INVERSION STRUCTURES WITH PAV 3

Peter A Audano[1], Christine R Beck[1,2]

[1]The Jackson Laboratory, Genomic Medicine, Farmington, CT, [2]The University of Connecticut, Health Center, Farmington, CT

The largest genomic alterations, structural variants (SVs), affect 1.3% percent of the human genome per individual. Balanced inversions reverse a region of the genome, and although they account for less than 1% of SVs per individual, inversions are responsible for 41% of structurally-variable bases. Inversions can be mediated by large inverted repeats, mobile elements, or they can be part of larger complex SVs (CSVs) making them among the most difficult SVs to identify. We recently updated our assembly-based variant caller, PAV 3, to improve callsets for all variant types and identify CSVs. Additionally, PAV uses a combination of alignment and k-mer–density methods with sensitivity for the largest and smallest inversions. We find an average of 61 balanced inversions per sample ranging from 200 bp to 4 Mbp across the Human Genome Structural Variation Consortium (HGSVC) assemblies. Improved methods now correctly resolve CSVs replacing partial SVs with their full structure including rearrangement breakpoints. Per haplotype, 16 inversions on average identified by PAV 2 are now reclassified as part of CSVs. Including CSV-derived inversions, PAV 3 has 99% precision and 92% recall compared to PAV 2 with missing inversions falling in largely uncallable regions. PAV 3 continues to push the frontier of assembly-based variant calling by producing more complete callsets. With improvements in both speed and accuracy, PAV 3 is designed to support genomics as long-reads are applied with increasing frequency.

BIGBRAIN: DECODING THE *TRANS*-REGULATORY ARCHITECTURE OF EXPRESSION AND SPLICING USING 10,725 POSTMORTEM HUMAN BRAIN TRANSCRIPTOMES.

Kailash BP[1], Aline Réal[2], Winston H Dredge[1], Beomjin Jang[1], Derek Lamb[3], Benjamin Z Muller[1], Brielin C Brown[3], Jack Humphrey[1], David A Knowles[2,4], Towfique Raj[1]

[1]Genetics and Genomics, Icahn School of Medicine at Mount Sinai, NY, NY, [2]New York Genome Center, NYGC, NY, NY, [3]Genetics, University of Pennsylvania, Philadelphia, PA, [4]Computer Science, Columbia University, NY, NY

Background: The transcriptional complexity of the human brain is genetically controlled through a network of *cis* and *trans*-acting mechanisms. While *cis*-acting genetic variants are well-studied, *trans*-acting variants remain less explored. *Trans*-QTLs typically exert effects indirectly, often through *cis*-regulated mediators embedded in broader gene regulatory networks (GRNs). Studying these indirect *trans*-regulated effects is particularly challenging due to their small effect sizes and huge genome-wide multiple testing burden. To address this, we assembled **BigBrain**, a uniformly processed, large-scale resource comprising **10,725** bulk RNA-seq samples and genotype data from **4,656** individuals across **11** public cohorts. By harmonizing transcriptomic and genetic data across diverse brain regions, ancestries, and disease states, BigBrain enables robust QTL mapping and serves as a discovery platform for nominating causal genes and reconstructing regulatory networks underlying brain function and disease risk. Using BigBrain, we previously characterized *cis*-acting genetic effects on gene expression, splicing (Réal et al. 2025), and RNA editing (Dredge et al. 2025), establishing links to neurological disease. We now leverage BigBrain to systematically map *trans*-QTLs using both expression and splicing phenotypes.
Results: The meta-analysis substantially improved detection power compared to individual cohort analyses, highlighting the benefit of large-scale data harmonization. We identified **784** significant *trans*-eQTLs for **210** genes and **4,150** *trans*-sQTLs (junction-level associations) for **300** genes at a Bonferroni p-value threshold. Only **27** genes were shared between the *trans*-eQTL and *trans*-sQTL gene sets, suggesting distinct gene-regulatory mechanisms acting in *trans*. Notably, we found that **rs1044595**, the lead variant of the PSP GWAS locus 1q25.3 (STX6), which lies in an oligodendrocyte enhancer, exhibits *trans*-association with *OPALIN*, an oligodendrocyte marker gene, supporting an oligodendrocyte-specific *trans*-regulatory mechanism acting at this locus.
Conclusion: We present the most extensive catalogue of *trans*-acting genetic effects on the human brain transcriptome to date. Ongoing Mendelian Randomization analyses aim to link these associations to putatively causal GRNs, offering new insights into the indirect pathways that shape brain function and neurological disease risk.

FUNCTIONAL RIBOSOMAL DNA ARRAYS MARK THE ENDS OF A SUBSET OF HUMAN ALT CHROMOSOMES

Ogechukwu Mbegbu[1], Szehoi Chan[2], Yi-An Chen[1], T. Rhyker Ranallo-Benavidez[1], Noelle H Fukushima[1], Tianpeng Zhang[2], Floris P Barthel[1]

[1]The Translational Genomics Research Institute (TGen), Bioinnovation and Genome Sciences, Phoenix, AZ, [2]University of Virginia School of Medicine, Department of Radiation Oncology, Charlottesville, VA

Ribosomal DNA (rDNA) is organized in tandem repeat arrays on acrocentric chromosomes, forming the structural basis of the nucleolus. Despite its repetitive nature and susceptibility to recombination, its involvement in telomere maintenance in human cancer has not been explored. We developed KaryoScope, a single-molecule reference-free long-read analysis approach inspired by DNA FISH that detects complex sequence alterations in repetitive genomic regions otherwise inaccessible to conventional bioinformatic methods. Using KaryoScope, we found multi-unit rDNA repeat arrays juxtaposed to telomeric sequences in ALT cell lines (e.g., U2OS and VA13) but never in primary or telomerase-positive lines. These observations were replicated in 20 ALT-positive astrocytomas, demonstrating this is a pervasive and clinically relevant phenomenon. Consistently, in U2OS metaphase FISH, we found rDNA instead of telomere signal at the terminal end of a significant subset of chromosomes. Curiously, these metaphases were often also characterized by extrachromosomal heterochromatin deposits consisting of telomere and alpha satellite DNA, which was confirmed by KaryoScope. Analyzing single rDNA-containing molecules independent of telomere sequence, we furthermore found frequent multi-unit rDNA repeat arrays juxtaposed to non-acrocentric chromosome-specific sequences, representing complex structural variant involving rDNA repeats. Proteomic analysis of telomeric chromatin comparing ALT-positive to ALT-negative lines demonstrated telomeric recruitment of nucleolar effector RNA pol I and centromeric effectors CENP-B, CCAN and KMN in ALT cells, consistent with rDNA occupying chromosome termini and telomeres and alpha satellite DNA cohabitating extrachromosomal heterochromatin deposits. Telomere-C chromatin interaction analysis indicated that rDNA and telomeric sequences are frequently found in close spatial proximity, providing a substrate for recombination. These results suggest that replication-stressed ALT telomeres may exchange their terminal cap for an rDNA array, revealing a previously unknown complex chromosome rearrangement triggered by telomere instability.

A CELL TYPE-SPECIFIC POLYGENIC RISK METHOD REVEALS DISTINCT CELLULAR CONTRIBUTIONS AND GENETIC SUBTYPES FOR PRIMARY OPEN-ANGLE GLAUCOMA

Michelle A Bartolo, Inas F Aboobakar, Janey L Wiggs, Ayellet V Segrè

Mass Eye and Ear, Harvard Medical School, Department of Ophthalmology, Boston, MA

Primary open-angle glaucoma (POAG), characterized by optic nerve degeneration, is a leading cause of irreversible blindness with no cure. While elevated intraocular pressure (IOP) is a major risk factor, a third of patients have IOP in the normal range, highlighting the need to understand cellular mechanisms underlying disease. Hundreds of risk loci have been associated with POAG, most in noncoding regions, limiting biological interpretation. Polygenic risk scores (PRS) improve POAG risk prediction, but current models do not account for the functional context of variants in disease-relevant cell types. Here, we developed a biologically-informed PRS method that integrates cell type-specific gene expression and permutation analysis to investigate cellular contributions to POAG. Using single-nucleus RNA sequencing data from retina, macula, optic nerve head (ONH), and anterior segment tissues, we defined cell type-specific gene sets based on differential expression (fold-change>2, false discovery rate<0.1). We constructed cell type-informed PRS (ctPRS) by applying PROSPER to variants within ±100 kb of each gene to capture regulatory effects. We tested associations between each ctPRS and POAG risk in European UK Biobank participants (14,468 cases, 149,343 controls) using logistic regression adjusted for age, sex, and top genotype principal components. Since the association p-value was correlated with gene count (Pearson r=0.57, p=5E-10), suggesting gene set size bias, we implemented a competitive permutation framework matched on gene count and length (10k permutations per cell type) to derive empirical p-values. After permutation, correlation was reduced and nonsignificant (r=0.14, p=0.16). Of 101 cell types tested, 35 had significant POAG associations (q<0.1), including anterior segment fibroblasts and endothelia, retinal astrocytes, and ONH vascular cells. To explore genetic heterogeneity, we applied k-means clustering to POAG case ctPRS profiles from significant cell types and tested cluster associations with IOP. We identified 3 patient subgroups (IOP-associated, immune/macroglial, and vascular) with differential IOP associations consistent with physiologically distinct genetic subtypes. This method links polygenic risk to specific ocular cell types, reveals neuro-susceptibility POAG processes, and provides a generalizable strategy to integrate functional genomics into polygenic modeling. We will extend this framework to non-European UK Biobank participants to compare POAG-ctPRS across ancestries.

A SCALABLE PLATFORM FOR SINGLE-CELL CO-PROFILING OF THE TRANSCRIPTOME AND GENOTYPE

Jan Bergmann, Gabija Lauciute, Paulius Matulis, Jokubas Tamoliunas, Domas Rupkus, Emile Pranauskaite, Patrick Rolli, Vaida Zukauskiene, Andrius Sinkunas, Rapolas Zilionis

Atrandi Biosciences, R&D, 6421 Lakemont Ct, NY

Single-cell RNA sequencing (scRNAseq) has become a routine tool for profiling cellular composition and transcriptional responses across health and disease. Beyond transcriptomic analysis, scRNAseq has also been used to detect expressed genetic variants and infer perturbation identity in CRISPR–Cas-based high-throughput screens through guide RNA sequencing. However, these transcriptome-based measurements face several limitations. They are restricted to transcribed regions, depend on sufficient expression levels to overcome the inherent sparsity of scRNAseq data, and, in the context of genetic perturbations, the detection of a guide RNA transcript does not confirm successful editing at the target locus.
To address these limitations, we developed a high-throughput single-cell RNA & DNA co-assay based on our Semi-Permeable Capsule (SPC) technology, enabling highly parallel, multistep processing of single cells. The method couples whole-transcriptome profiling with targeted genotyping by multiplex PCR amplicon sequencing in the same cells, directly linking genotype to transcriptional state, confirming CRISPR edits at genomic targets, and functionally characterizing engineered or naturally occurring mutations.
We profiled >100,000 primary human cells across multiple peripheral blood mononuclear cell (PBMC) donors using co-sequencing of RNA and amplicons targeting 10 SNP-containing loci. Over 85% of captured cells yielded genotypes, enabling donor deconvolution from variant calls in the amplicons and robust mapping of genetic background to cell-state heterogeneity. Designed for scalability, the assay supports user-defined PCR panels targeting transcribed and non-transcribed loci and is readily extensible in both cell numbers and amplicon breadth. We will present results demonstrating the utility of the platform for dissecting genotype–phenotype relationships at single-cell resolution.

ORGANISATIONAL PRINCIPLES OF LONG NON-CODING RNAs REVEALED BY EXON DELETION

Sarang Bhutada[1,2,3,4], Hugo Guillen-Ramirez[2,3], Tina Uroda[2,3], Ines Bravo[2,3], Michela Coan[1,2,3,4], Rory Johnson[1,2,3,4]

[1]UCD, School of Medicine, Dublin, Ireland, [2]UCD, School of Biology and Environmental Science, Dublin, Ireland, [3]UCD, Conway Institute for Biomolecular and Biomedical Research, Dublin, Ireland, [4]UCD, Systems Biology Ireland, Dublin, Ireland

Long non-coding RNAs (lncRNAs) regulate cell phenotypes in health and disease, yet how function is encoded in their sequence remains poorly understood. Current models propose a modular architecture composed of discrete functional elements, but this is based on a limited set of paradigmatic examples and methods for mapping function to sequence are limited in scope and resolution. Here, we establish a high-throughput CRISPR-Cas9 strategy for dissecting lncRNA functional architecture at exon resolution. Using cell fitness as a phenotypic readout, we screened 358 exons from 107 lncRNAs across four human cell lines. We report that (1) a large proportion of exons have no detectable function, (2) a minority of exons are functional in any given cell line (19–111 exons), equivalent to one-fifth of total transcript nucleotides on average, and (3) functionality is enriched towards the 5' end of the transcript. We demonstrated through statistical and experimental analyses that lncRNA function depends on transposable elements, microRNA response elements and RNA binding protein sites. These sub-genic functional maps expand the catalogue of experimentally defined lncRNA functional elements by an order of magnitude, illuminate molecular mechanisms and broadly support a modular organisation for lncRNAs.

INFERRING EPISTASIS IN EVOLUTIONARY ACCUMULATION PROCESSES

Dmitry Biba, David McCandlish

Cold Spring Harbor Laboratory, Quantitative Bilogy, Cold Spring Harbor, NY

Cancer progression, viral evasion from immunity, evolution of antibiotic resistance, and antibody affinity maturation are all processes that can be adequately modeled as a sequential accumulation of "forward" mutations. Knowing how different mutations interact, i.e. identifying preferential orders of accumulation, can allow researchers to predict and/or influence the progression of the process under study. Despite their prevalence and importance, the general theory of unidirectional accumulation processes is currently lacking. In this work, we (1) develop such a theory, analogous but not equivalent to the theory of fitness landscapes and (2) devise a method for inference of context-dependent accumulation rates. We come up with a definition for epistatic interactions between mutations in an accumulation process and show different ways of parametrizing them. Additionally, we derive the constraints on a process whose accumulation rates are determined by a static fitness landscape. Our theoretical analysis informs a prior we use for non-parametric Bayesian inference of accumulation rates. The inferred rates match the data closely where data is abundant, allowing for a highly expressive fit, while data-sparse regions are dominated by the prior. In the future we plan to extend this method to provide more detailed information about the nature of the process, including the stability of the underlying fitness landscape and the mode of evolution (e.g. strength of selection).

WIDE-SCALE PHARMACOGENOMIC STUDY HIGHLIGHTS GLUCOCORTICOID-RELATED GENES ASSOCIATED WITH DRUG RESPONSE PHENOTYPES

Malgorzata Borczyk, Marcin Piechota, Paula Konowalska, Pawel Pienkowski, Sylwia Grubarek, Jacek Hajto, Rafal Kafel, Dzesika Hoinkis, Michal Korostynski

Maj Institute of Pharmacology Polish Academy of Sciences, Laboratory of Pharmacogenomics, Krakow, Poland

Ineffective or poorly tolerated pharmacotherapy imposes a substantial burden on healthcare systems, as limited efficacy and adverse effects affect the majority of patients. Genetic variation plays a significant role in this phenomenon, with current pharmacogenetic (PGx) guidelines reducing the incidence of adverse drug reactions by up to 30% (PREPARE study). However, these guidelines predominantly address common variants in genes involved in drug metabolism, while pharmacodynamic variants, particularly those in brain-expressed genes that mediate drug action, remain severely underrepresented in PGx GWAS. This underrepresentation largely results from the rarity of such variants and modest sample sizes of previous studies. Population-scale biobanks now enable large-scale PGx research, yet deriving reliable drug-response phenotypes from real-world data remains challenging. We hypothesized that carefully curated prescription dosage and treatment duration can serve as robust proxies for therapy tolerance and acceptance. We leveraged primary care records from the UK Biobank (~38 million prescriptions, response phenotypes for 307 drugs) and tested 1.3 million common SNPs (MAF > 0.5%) as well as aggregated rare variants (loss-of-function, missense) and established pharmacogenetic variants to achieve comprehensive genomic coverage. Applying REGENIE two-step regression with hold-out validation, we identified dozens of novel pharmacogenetic loci. In the interpretation stage we focused on genes within the glucocorticoid receptor (GR) signaling pathway and mechanisms underlying neuropharmacotherapies - an area we have extensively investigated in prior animal studies. We prioritized GR-related loci in the context of pain management and psychiatric drug responses. Our findings include variants in genes highly expressed in the central nervous system that are associated with pain-management and antidepressant therapies. Illustrative of this are rare loss-of-function variants in the GR-regulated gene *CAMK1G* linked to significantly higher dosing of amitriptyline (beta = +28 mg, $p = 5.06 \times 10^{-6}$; Bonferroni-corrected p = 0.04). These observations were further explored using a multi-trait graphical inference framework (CI-GWAS) to elucidate the interplay between pain pharmacotherapy and GR signalling. Overall, this study demonstrates that well-curated electronic health record data can drive the discovery of novel pharmacodynamic loci and provide mechanistic insights into drug action.

TRANSPLACENTAL TRANSFER EFFICIENCY REVEALS COMPARTMENTALIZED MATERNAL-FETAL TRANSCRIPTIONAL RESPONSES THAT MEDIATE PFAS EFFECTS ON PERINATAL OUTCOMES

Sean T Bresnahan[1], Hannah E Yong[2], Sierra Lopez[3], Jerry Kok Yen Chan[4], Shiao-Yng Chan[2], Elana R Elkin[5], Jonathan Y Huang[3], Arjun Bhattacharya[1]

[1]The University of Texas MD Anderson Cancer Center, Epidemiology, Houston, TX, [2]A*STAR Institute, Human Development & Potential, Singapore, [3]University of Hawai'i at Mānoa, Public Health Sciences, Honolulu, HI, [4]Duke-NUS Medical School, Obstetrics & Gynaecology, Singapore, [5]San Diego State University, Public Health, San Diego, CA

Per- and polyfluoroalkyl substances (PFAS) are ubiquitous environmental contaminants that cross the placental barrier and are associated with reduced fetal growth, though the molecular mechanisms remain incompletely understood. This is partly due to structural heterogeneity among PFAS compounds, which vary in chain length, functional groups, and transplacental transfer efficiencies (TPTE). Using our tissue-specific long-read transcriptome assembly, we analyzed isoform-resolved expression in n=124 term placentas with matched PFAS concentrations in maternal mid-gestation and cord blood. We examined parent-of-origin expression (POE) using allele-specific transcript quantification in trio genotyped samples, revealing that PFAS disrupts genomic imprinting in an isoform-specific manner. Gene-level aggregation masked these effects; isoform-level analysis detected shifts in POE biases with exposure. Using our novel SCENIC (Synthetic Confounding Evaluation for Negative Control Inference Comparison) framework, we evaluated several methods across confounding scenarios, finding instrumental variable approaches most robust to unmeasured confounding. Causal mediation analysis revealed that PFAS influences birth weight predominantly through perturbation of co-expression network hubs rather than differentially expressed transcripts. Critically, mediation strength, network topology, and compartmentalization scale systematically with TPTE. High-TPTE compounds recruit more mediators and segregate maternal from fetal regulatory programs within shared networks, with outcome-specific architectures that aligns with maternal-fetal conflict theory. These findings reveal TPTE as an organizing principle predicting both extent and mechanistic architecture of placental responses to toxicant exposure, including disruption of imprinting processes, with implications for risk assessment and therapeutic target identification.

TARGETTED GP-SCV MACHINE LEARNING LINKS GENOTYPE TO PHENOTYPE IN HEART DISEASE, ALLOWING FOR INDIVIDUALIZED INTERVENTION PLANS

Nava Ehsan, Ben C Calverley, William E Balch

Scripps Research, Department of Molecular and Cellular Biology, La Jolla, CA

Cardiovascular diseases (CVDs) are the world's biggest killer (an estimated 17.9 million deaths annually) and are caused by a myriad of factors from mutations in at least nine different genes to a wide range of phenotypes – both changeable (e.g. weight, diet) and not (e.g. age, ethnicity). Understanding how all of these factors relate to one another is vital to being able to address CVD in a precise and individualized manner, not simply through blanket one-size-fits-all interventions that work for the mythical "median" individual. The most widely used of these interventions is statins – used in patients worldwide. Its use is broad and not guided by individual factors in many cases.

In recent years, numerous studies have developed polygenic and clinical risk scores for CVD, but they largely lack the ability for causal interpretation. Although we understand that many variants are associated with disease, there is a huge gap in understanding the fundamental disease biology and finding the "why" behind genetic risk. We must understand not simply the levels of risk but also its causes, and on an individual basis, in order to best decide on the right interventions.

This is a task that requires approaches that are both precise and interpretable. For just this purpose, we developed Gaussian Process driven spatial covariance (GP-SCV). Through machine learning, GP-SCV generates genotype-phenotype landscapes based on probabilistic uncertainty and weighted proximity, using relatively sparse input information to then create genotype-phenotype crosstalk graphs that reveal causal genotype-phenotype relationships. We use the gold-standard UK BioBank datasets to examine mutations within three key genes (LDLR, APOB, and PCSK9) known to the main cause of familial hypercholesterolemia and their links to over one hundred phenotypes. From this we reveal environmental-clinical phenotype indicators specific to individuals with mutations in particular genetic loci, with highly different responses to statins in their quantified risk levels.

We are therefore able to take a holistic view of an individual's genetic and phenotypic profile and highlight their suitability for different pharmaceutical interventions, as well as showing which modifiable phenotypes should be the focus of non-pharmaceutical interventions. Unlike risk management, which simply estimates risk level, we provide a view of the dynamics of CVD from a precision approach, paving the way for more targeted prevention and treatments through a covariant view of central dogma.

LEVERAGING CLINICAL GENOME SEQUENCING DATA FOR METAGENOMIC ANALYSIS IN PATIENTS WITH INBORN ERRORS OF IMMUNITY

Wenjia Cao[1], Justin Lack[1], Jia Yan[2], Morgan Similuk[2], Steven Holland[2]

[1]National Institute of Allergy and Infectious Diseases , Integrated Data Sciences Section, Bethesda, MD, [2]National Institute of Allergy and Infectious Diseases , Division of Intramural Research, Bethesda, MD

The National Institute of Allergy and Infectious Diseases (NIAID) Centralized Sequencing Program (CSP) provides clinical research genomic services for patients with inborn errors of immunity (IEI) and complex disease etiologies. Since 2021, patients enrolled in the program have received whole genome sequencing at a median target depth of 30X, with the goal of establishing a genetic diagnosis as well as providing opportunities for independent research. IEI are marked by disruptions in immune function, leading to pervasive opportunistic infections. Comprehensive early detection of pathogens is necessary for improved treatment and diagnosis. To extend the utility of a clinical GS dataset, we sought to characterize metagenomic calls in patients with IEI and explore the dynamic interactions between the microbiota and immunodeficiencies. Furthermore, we hope this pipeline can timely address the diagnostic gaps in patients with IEI and other conditions predisposing to opportunistic infections.

We used CentrifugeR to perform metagenomic classification on 2,000 GS samples from the NIAID CSP, encompassing patients and unaffected relatives as controls. The dataset focused on blood samples from participants spanning a wide age range and diverse ancestral backgrounds. To validate pathogen detection from the pipeline, we compared our results against a subset of patients who had undergone commercial targeted metagenomic testing. Our pipeline successfully replicated all significant pathogen calls identified by the commercial tests. Clustering of all samples based on metagenomic calls on species level successfully segregates immunodeficient patients and healthy controls into distinct clusters, suggesting global metagenomic abundances differ between patients and controls. In addition, we tested for association of bacterial and viral taxa with clinical and phenotypic variables. Our study demonstrates that whole genome sequencing data can yield accurate pathogen calls, highlighting its potential as a valuable data modality that can be analyzed from large collections of clinical genome sequencing.

SINGLE-CELL GENOMICS DECONTAMINATION WITH CELLSWEEP

Maya Caskey[1], Joseph Rich[1], Ryan Weber[2], Ali Mortazavi[2], Lior Pachter[1], Ingileif Hallgrimsdottir[1]

[1]California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA, [2]University of California Irvine, Department of Developmental and Cell Biology, Irvine, CA

Single-cell genomics technologies enable high-throughput profiling of cells, but technical contamination remains a major obstacle to accurate downstream analysis. Free-floating ambient molecules released from lysed cells and global bulk contamination introduced during library preparation can both distort molecular profiles. These artifacts can obscure cellular identities and reduce the reliability of differential analysis or clustering results. We present an efficient and effective approach to removing ambient and bulk contamination that can be applied to data generated from a wide variety of technologies. We show that our tool, CellSweep, outperforms other methods for removing artifacts using numerous benchmarks.

CHARACTERIZING DIFFERENCES IN GENE EXPRESSION VARIABILITY BETWEEN HUMANS AND CHIMPANZEES

Alexander Chen[1], Brendan Jamison[1], Kenneth Barr[2], Xin He[1,3], Yoav Gilad[1,2,3]

[1]University of Chicago, Genetics, Genomics, and Systems Biology, Chicago, IL, [2]University of Chicago, Genetic Medicine, Chicago, IL, [3]University of Chicago, Human Genetics, Chicago, IL

Most studies of gene regulation focus on mean expression levels across populations of cells. However, gene expression levels often vary substantially from cell to cell – even among cells of the same type and from the same organism. While this cell-to-cell variability is typically dismissed as noise, emerging evidence suggests that some of it is genetically regulated. A key question is whether such regulation acts in *cis*, through local sequence variants, or in *trans*, through broader regulatory environments. To investigate this, we used a comparative genomics approach and single-cell RNA-sequencing data from 5 differentiated cell types derived from 3 humans, 3 chimpanzees, and 1 human–chimpanzee allotetraploid induced pluripotent stem cell line. Because both genomes are exposed to the same cellular environment in the allotetraploid line, this system allows us to isolate *cis* effects on gene expression variability.

We quantified mean-corrected dispersion for each gene in each cell type using *Memento*, a method designed to estimate variability in single-cell data. Using this framework, we identified thousands of differentially variable (DV) genes between species in the diploid lines, ranging from 790 to 907 per cell type (median 841), of which 15–25% were recapitulated in the allotetraploid line across cell types. The persistence of differential variability between the human and chimpanzee alleles in the shared cellular environment of the allotetraploid line indicates that a substantial fraction of gene expression variability differences are driven by *cis*-regulatory divergence. This result shows that expression variability is not solely an emergent property of cellular state or *trans*-acting factors but can be encoded by local genetic variation. Consistent with this interpretation, genes with low dispersion across cell types and species are enriched for housekeeping functions and are significantly depleted for features associated with regulatory complexity, including enhancer length, intron number, and transcription start site multiplicity, relative to genes with higher dispersion. These patterns suggest that tight control of expression variability is under selective constraint and represents an important dimension of gene regulation.

The observation that a substantial fraction of regulatory variability maps to *cis*-regulatory differences further implies that gene expression variability itself is a selectable trait. Selection may therefore act not only on mean expression levels, but also on variability as an independent regulatory mechanism contributing to phenotypic divergence between species.

MACHINE LEARNING REVEALS TISSUE-AGNOSTIC AND REGION-SPECIFIC ISOFORM AGING MARKERS IN THE HUMAN HIPPOCAMPUS

Xingyi Chen[1], Beril Erdogdu[2], Mihaela Pertea[2,4], Stephanie C Hicks[2,3,4,5]

[1]Johns Hopkins University, Department of Applied Math & Statistics, Baltimore, MD, [2]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [3]Johns Hopkins University, Department of Biostatistics, Baltimore, MD, [4]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [5]Johns Hopkins University, Center for Imaging Science, Baltimore, MD

Recent work has demonstrated that transcript isoform usage can accurately predict human brain age in the dorsolateral prefrontal cortex (DLPFC), highlighting isoform usage as a predictor of human aging. However, it remains unclear that such isoform-based aging markers are consistent across brain regions or are region-specific. Particularly, human hippocampus, a region key to memory and cognitive aging, has not been evaluated in this context. Here, we investigate whether isoform-based aging markers are tissue-agnostic or region-specific, with a focus on human hippocampus while also comparing to DLPFC. Using bulk RNA sequencing data spanning from prenatal development to late adulthood, we construct machine learning models based on isoform fractions to predict biological age. Models are trained and validated across both brain regions, enabling assessment of cross-regional generalization. Differential transcript usage (DTU) testing is implemented to infer predictive isoform features: both a discrete approach (hypothesis testing) and a continuous approach (trajectory analysis) are employed. Preliminary analyses indicate that isoform-based models achieve robust age prediction performance in the hippocampus. However, when using the DLPFC model on the hippocampus data, we observe limited transferability in model performance. This finding challenges the assumption that isoform aging markers are universally conserved. Feature overlap analyses reveal a small subset of isoforms with consistent age-associated usage patterns across regions, alongside hippocampus-specific markers that do not generalize. Trajectory analyses further reveal distinct age-dependent isoform dynamics, including gradual and switch-like patterns. At the conference, we will present a comprehensive evaluation of tissue-agnostic and tissue-specific isoform markers, including cross-region validation results and comparison across machine learning models. Together, this work clarifies and strengthens the robustness and regional specificity of isoform-level aging features and their use in biological age modeling.

UNRAVELING PUBERTY-DRIVEN IMMUNE CELL DYNAMICS AND ASTHMA PATHOPHYSIOLOGY AT SINGLE-CELL RESOLUTION

Yixuan Chen[1], Cynthia Kalita[1], Ali Ranjbaran[2], Julong Wei[2], Julian Bruinsma[3], Gabrielle Garlicki[1], Henriette Mair-Meijers[2], Samuele Zilioli[3], Roger Pique-Regi[2], Francesca Luca[1]

[1]University of Chicago, Human Genetic, Chicago, IL, [2]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [3]Wayne State University, Psychology, Detroit, MI

Puberty is a critical developmental period marked by profound hormonal, metabolic, immune, and physiological changes that coincide with shifts in asthma prevalence. Despite strong epidemiological evidence linking pubertal development to sex differences in asthma, little is known about how puberty shapes immune gene expression and regulatory programs at cell-type resolution. We studied a cohort of 300 children with asthma (130 females, 170 males; ages 10–18 years) from the Asthma in the Life of Families Today (ALOFT) study, profiling peripheral blood mononuclear cells using single-cell RNA sequencing. Across over 700,000 cells spanning 12 immune cell types, we systematically assessed associations between gene expression and age, pubertal stage, and menarche status. We identified puberty-associated gene expression changes that were highly sex- and cell-type–specific. When considering pubertal development stages (analogous to Tanner stages), females exhibited greater transcriptional remodeling than males, with the strongest effects observed in monocytes, including 130 and 209 differentially expressed genes associated with age and puberty, respectively (10% FDR). Genes upregulated with pubertal progression in female monocytes were enriched for interferon-responsive, cytokine signaling and bacterial response pathways, consistent with a shift in immune state during pubertal maturation.

To investigate regulatory mechanisms underlying these changes, we integrated single-cell ATAC-seq data from 16 participants. Among puberty- and age-associated genes in female monocytes, we observed significant enrichment of active transcription factor motifs from the AP-1/NF-κB, STAT/IRF, and GATA families, along with enrichment of estrogen receptor motifs detectable only at single-cell resolution. Regularized regression models demonstrated that transcription factor motif programs could predict puberty-associated differential expression, identifying 21 active motifs predictive of age-associated changes and 32 motifs predictive of puberty-associated changes, including estrogen receptor motifs, supporting a regulatory role for coordinated transcription factor networks. Notably, estrogen receptor expression levels remained largely stable across puberty, suggesting that estrogen-associated effects are mediated through changes in regulatory activity rather than expression. Overall, our findings suggest that puberty is associated with targeted, cell-type–specific immune transcriptional reprogramming in children with asthma, most prominently in female monocytes, providing mechanistic insight into how pubertal development may shape immune function and contribute to sex differences in asthma and immune-mediated disease risk later in life.

UNCOVERING DISEASE RESISTANCE GENE DIVERSITY IN SORGHUM THROUGH HIGH-QUALITY ASSEMBLIES AND FULL-LENGTH TRANSCRIPTOMICS

Kapeel Chougule[1], Sharon Wei[1], Zhenyuan Lu[1], Andrew Olson[1], Lydia Tressel[1], Nicholas Gladman[1], Michael Regukski[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA ARS NAA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY

Sorghum (Sorghum bicolor) is a vital cereal crop for food security, livestock feed, and bioenergy, grown worldwide under challenging climates. However, its productivity is continually threatened by pathogens and pests, including anthracnose, grain mold, and more recently the sugarcane aphid, which since 2013 has caused devastating losses across the U.S. sorghum belt. Breeding for durable resistance depends on identifying and deploying natural resistance (R) genes, yet these genes are among the most difficult to annotate due to their rapid sequence diversification, structural clustering, and the limitations of conventional genome annotation methods. High-quality reference assemblies are essential to resolve complex R-gene regions, which often contain large structural insertions, tandem duplications, and copy number variation. Resources such as SorghumBase now host assemblies and annotations for diverse sorghum accessions, including many lines with disease resistance traits. Building on this genomic foundation, we leveraged the growing diversity of sorghum accessions sequenced with long-read technologies to systematically characterize R-genes at pangenome scale. We first conducted genome-wide scans using the NLR-Annotator tool across 72 sorghum assemblies, identifying >350 candidate NLR loci. To complement this, we applied PlantNLR, a protein-based pipeline that detects canonical NLRs as well as NLRs with integrated domains (NLR-IDs), which may act as decoys to trap pathogen effectors. Integration of both methods revealed substantial overlap while also uncovering unique candidates from each. Comparative genomics using NB-ARC gene trees from SorghumBase then enabled the classification of NLR families, highlighting clear lineage-specific expansions and contractions that shape sorghum's immune diversity. To improve structural accuracy, we incorporated long-read full-length transcript sequencing (PacBio Kinnex and Revio platforms). Across resistant accessions, these data refined splice junctions and UTR boundaries, enhanced sensitivity of detection, and revealed extensive alternative splicing, intron retention, and antisense transcription at NLR loci. These transcriptomic insights not only improve annotation quality but also provide functional evidence for diversification among closely related paralogs. Together, these results demonstrate that the combination of high-quality assemblies, pan-genomic analyses, and full-length transcriptomics is essential for capturing the true diversity of sorghum disease resistance genes. This integrative framework lays the groundwork for identifying novel R-gene variants, understanding adaptive mechanisms, and accelerating the breeding of cultivars with durable resistance. This work is supported by USDA-ARS grant 8062-21000-051-000D

PANGENEINDEXER: A SCALABLE FRAMEWORK FOR CONSISTENT GENOME ANNOTATION ACROSS CROP PANGENOMES

Kapeel Chougule[1], Sharon Wei[1], Zhenyuan Lu[1], Andrew Olson[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA-ARS NAA, Robert W. Holley Center for Agriculture and Health, Ithaca, NY

As the number of high-quality crop assemblies continues to expand, the challenge of producing accurate, consistent, and scalable gene annotations across accessions has become increasingly critical. Single-reference annotations often fail to represent the full coding potential of a species, while traditional de novo pipelines, although precise, are computationally intensive and limited in sensitivity. To address this challenge, we developed PangeneIndexer, a pan-gene indexing workflow that leverages orthology-based gene family trees and representative models to efficiently propagate annotations across diverse assemblies. The workflow begins with the construction of gene family trees using the Ensembl Compara pipeline, from which representative models are selected based on curation priority and structural quality. During index construction, PangeneIndexer filters split or fused predictions and prioritizes curated models, ensuring both sensitivity and structural accuracy. These models form the foundation of a pan-gene index. Gene structures are then projected to new accessions with Liftoff and refined using PASA, which integrates transcriptomic evidence to resolve gene boundaries and splicing structures. We applied this framework to multiple crop pan-genomes, including 26 maize, 29 rice, and 18 sorghum accessions, generating pan-gene sets that partition into core, soft-core, and shell components. Compared to de novo annotations, the pan-gene index workflow demonstrated significant improvements in both coverage and uniformity, while reducing per-genome annotation time from 1–2 weeks to 2–3 days. In maize benchmarking, the pan-gene index encompassed over 116,000 pan-genes, partitioned into ~21,000 core, ~70,000 soft-core, and ~15,000 shell genes, reflecting the broad diversity across 26 accessions. Similarly, rice and sorghum pan-genomes showed tens of thousands of additional loci uncovered relative to reference-only annotations. Importantly, PangeneIndexer also recovered unique loci per accession with transcriptome or protein evidence support and also improved coverage for lineage-specific genes. To facilitate curation and evaluation, we employed the Gramene gene tree visualization platform, which enables rapid detection of inconsistent models within pan-gene families. This integration supports downstream evolutionary and functional studies while providing a scalable framework for incorporating additional accessions.
This work is supported by USDA-ARS grant 8062-21000-051-000D

ITERATIVE DESIGN OF TRAINING DATASETS FOR GENERALIZABLE SEQUENCE-TO-FUNCTION MODELS

Trevor Christensen*, Yash V Mundewadi*, Peter Koo

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

Predicting regulatory element activity from DNA sequence is a central challenge in biology. While modern sequence-to-function models achieve strong performance on held-out genomic sequences, they fail to generalize reliably to perturbations and synthetic sequences. This generalization gap reflects a fundamental limitation of current training data: genome-wide profiling assays sample only a narrow and biased slice of regulatory sequence space, leaving models without exposure to the genetic variation needed to learn transferable regulatory rules. Improving generalization therefore requires not just more data, but strategically selected data that better samples the distribution of sequences models will encounter at deployment. Here, we introduce S2F-LearningLoop, a modular platform for systematically exploring how training data composition and selection strategy affect sequence-to-function model performance. Using in-silico oracle models as experimental proxies, we compare a spectrum of approaches ranging from informed biological baselines that leverage prior knowledge of regulatory sequence features, to model-guided active learning strategies that use uncertainty and diversity criteria to identify maximally informative training sequences. Both the composition of candidate sequence pools and the criteria used to select from them are treated as separable design choices, enabling a structured exploration of how each contributes to model improvement. Performance is evaluated across test sets spanning genomic sequences, low-shift perturbations, high-shift synthetic variants, and purely random sequences, capturing the range of distribution shifts that arise in regulatory genomics applications. The goal of this platform is to identify which sequence generation and selection strategies most efficiently improve model generalization, ultimately informing the design of prospective experiments.

(* = authors contributed equally)

ESTABLISHING THE WOODCHUCK *(MARMOTA MONAX)* AS A SINGLE-CELL MODEL OF HEPATITIS B-DRIVEN HEPATOCELLULAR CARCINOMA

Zoe A Clarke[1,2], Jawairia Atif[3,4], Xinle Wang[3], Dustin J Sokolowski[1,5], Ciaran K Byles-Ho[6], Ruth Isserlin[2], Lewis Y Liu[3,4], Lawrence Wood[3,4], Damra Camat[3,4], Yijia Liu[7], Ariya Shiwram[3], Sharon J Hyduk[4], Sai Chung[3,4], Michael D Wilson[1,6], Jared T Simpson[1,5], Ian D McGilvray[4], Sonya A MacParland[3,4,7], Gary D Bader[1,2,8]

[1]Department of Molecular Genetics, University of Toronto, Toronto, Canada, [2]The Donnelly Centre, University of Toronto, Toronto, Canada, [3]Department of Immunology, University of Toronto, Toronto, Canada, [4]Ajmera Transplant Centre, General Hospital Research Institute, Toronto, Canada, [5]Ontario Institute for Cancer Research, (OICR), Toronto, Canada, [6]Genetics and Genome Biology, The Hostpital for Sick Children, Toronto, Canada, [7]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada, [8]Department of Computer Science, University of Toronto, Toronto, Canada

Hepatocellular carcinoma (HCC), a primary form of liver cancer, is one of the deadliest cancers currently affecting humans. The most common cause of HCC is a chronic hepatitis B virus (HBV) infection. Although vaccines and antivirals for HBV exist, many individuals remain infected or vulnerable to a chronic infection and the rate of cancer development from such an infection remains high. To improve the prognosis of this cancer, we need to improve our understanding of the tumour microenvironment to develop novel, effective therapeutics. Considering the availability and experimental challenges of working with human tissue, the progression of therapeutics is often aided by a comprehensive animal model.

In my work, I present the eastern woodchuck, *Marmota monax*, as a single-cell compatible animal model that reflects key qualities of human liver disease. The woodchuck is one of the few mammalian species naturally susceptible to HCC development when infected with a close relative of HBV, the woodchuck hepatitis virus (WHV). To establish the woodchuck as a single-cell compatible model, I annotated a newly sequenced woodchuck genome which was then used to study individual cells in the woodchuck liver and blood. Quality control challanges with liver single-cell RNA sequencing data led to an extensive investigation into how to better screen for empty droplets. Finally, cells were analyzed from both healthy and chronically infected tissue to understand the phenotype of diseased woodchuck cells when compared to a healthy reference. Diseased woodchuck liver cells exhibit many of the same characteristics of diseased human liver cells, including a significant increase in T cell exhaustion signatures.

COLLECTIVE MODES ORGANIZE EVOLUTIONARY DYNAMICS UNDER A NON-TRIVIAL GENOTYPE-TO-PHENOTYPE MAP

Aedan Brown[1], Sarah Datta[2], Pankaj Mehta[3], Brian Cleary[4,5,1]

[1]Boston University, Biomedical Engineering, Boston, MA, [2]Boston University, Math, Boston, MA, [3]Boston University, Physics, Boston, MA, [4]Boston University, Computing and Data Sciences, Boston, MA, [5]Boston University, Biology, Boston, MA

At every scale and in every domain of life, complex traits with a non-trivial genotype-to-phenotype map underlie fitness. Understanding how such traits evolve remains a central challenge in biology. Here, we investigate an alternative to a typical population genetic approach, which can abstract away the genotype-to-phenotype map and obscure the role of biological constraints. We establish a genotype-to-phenotype-to-fitness map underpinning a model complex trait (flux through a metabolic network), building on the successful and quantitatively tractable framework of Flux Balance Analysis. Through evolutionary simulation and theory, we find the surprising result that explicit representation of complexity gives rise to simple and reproducible dynamics. Even in highly degenerate systems where many phenotypes and genotypes can produce the same fitness, repeated evolutionary trials with the same network exhibit a preferred direction of evolution. We identify evolutionary collective modes (EvCMs) – linear combinations of genes with high, constant selective pressure – as organizers of the long-term evolutionary dynamics and the origin of this simplicity. EvCMs arise through the interaction of physical constraints, evolvability, and the requirements for reproduction. The interaction of these elements couples constituent parts, so that selection on individual genes becomes incoherent (or not easily interpreted), while also giving rise to distinctive evolutionary dynamics. The dynamics proceed through regimes, each with its own EvCM. Within a regime, selection on genes is conditional on the current genotype, with selective pressure hopping from gene to gene over time, while selection on the EvCM is independent of the genotype and therefore constant throughout the regime. Inspired by these results, we develop a computational framework for predicting the EvCMs of a metabolic network by combining ideas from Flux Balance Analysis, optimization, and evolutionary theory. This framework relates the structure of an EvCM to the expected mutational spectrum of genes, allowing us to make connection with long-term evolution experiments. We compare our theoretical predictions to experimentally observed mutation counts in the Lenski lines and identify signatures of EvCMs, suggesting that the genotype-to-phenotype maps stemming from metabolism likely place strong constraints on evolutionary dynamics.

INTEGRATING GWAS WITH A MULTIMODAL ATLAS OF THE FEMALE REPRODUCTIVE TRACT REVEALS CRITICAL CELL TYPES AND PATHWAYS IN GYNECOLOGICAL DISORDERS

Céleste E Cohen[1], Ana Paredes[1], Valentina Lorenzi[1], Christina Kim[1], Miriam Baumgarten[2], Cecilia Icoresi-Mazzeo[1], Cecilia Lindskog[3], Ariella Shikanov[4], Saher S Hammoud[5], Carl A Anderson[1], Luz Garcia-Alonso[1], Roser Vento-Tormo[1,6]

[1]Wellcome Sanger Institute, Hinxton, United Kingdom, [2]Cambridge University Hospital NHS FT, Cambridge, United Kingdom, [3]Uppsala University, Department of Immunology, Genetics and Pathology, Cancer Precision Medicine Research Program, Uppsala, Sweden, [4]University of Michigan, Obstetrics & Gynecology, Ann Arbor, MI, [5]University of Michigan, Department of Human Genetics, Ann Arbor, MI, [6]University of Cambridge, Cambridge Stem Cell Institute, Cambridge, United Kingdom

Reproductive disorders affect over 30% of women and have severe consequences for fertility and systemic health but remain largely understudied. These disorders include polygenic ovarian and uterine pathologies, such as polycystic ovary syndrome and uterine fibroids, which are shaped by dynamic hormone-driven remodelling of reproductive tissues over time. Here, we leverage reproductive tissue-specific gene expression and chromatin accessibility data to interpret genetic associations with common gynecological disorders and investigate their aetiologies. We present a unified multi-omic atlas of the human female reproductive tract, comprising over 1 million cells from more than 150 donors profiled by single-cell RNA-seq, single-nucleus ATAC-seq, and spatial transcriptomics across the uterus, fallopian tubes, and ovaries. Using a multimodal analytical framework, we integrate this atlas with GWAS summary statistics for multiple reproductive disorders to delineate the gene-regulatory mechanisms of risk variants and pinpoint disease-relevant cell types, including endometrial stromal cells in uterine disorders causing heavy menstrual bleeding. This approach identifies genetically supported disease pathways and highlights putative therapeutic targets, emphasizing the value of expanding GWAS in reproductive phenotypes. Together, this work provides a foundational resource for studying female reproductive biology and establishes an integrative framework for advancing insight into neglected gynecological conditions.

CONCERTED GENE–ENVIRONMENT INTERACTIONS ACROSS LOCI IN COMPLEX TRAITS

Sylvia Dai[1,2], Yanina Kuzminich[1], Gouri Rajaram[1], Hakhamanesh Mostafavi[1]

[1]NYU Grossman School of Medicine, Center for Human Genetics & Genomics, New York, NY, [2]New York University Abu Dhabi, Division of Science, Abu Dhabi, United Arab Emirates

Understanding how genetic effects are modified by environmental exposures is critical for explaining variability in complex traits. Gene–environment interaction studies typically examine one exposure at a time, which is often underpowered due to multiple-testing burden. However, complex traits are shaped by correlated environmental factors that may combine to influence genetic effects, and under an omnigenic architecture these exposures may modulate many loci simultaneously rather than individually. Motivated by this intuition, we constructed composite environmental predictors for a range of traits in the UK Biobank from diverse exposures, including dietary patterns, socioeconomic status, physical activity, and other behavioral and environmental variables that capture the aggregate contribution of modifiable factors.

We tested interactions between these composite lifestyle scores and genome-wide significant loci identified in previous genome-wide association studies (GWAS). We observe significant interactions for several traits, including LDL cholesterol, urate levels, and body mass index (BMI). Across loci, we observe concerted patterns of gene–environment interaction: genetic effects tend to be systematically larger among individuals with higher exposure, reminiscent of the amplification model previously reported for gene-sex interactions. Consistent with recent reports, we find that these interactions are sensitive to phenotype scale; however, there is no universal transformation (e.g., log or inverse-normal) that fully removes these interaction signals. In addition, a subset of SNPs exhibits interaction effects that deviate from the shared cross-locus pattern, suggesting locus-specific mechanisms beyond global amplification.

In summary, our results indicate that gene–environment interactions are pervasive and polygenic, that their interpretation cannot be attributed solely to scale artifacts, and more broadly demonstrate novel ways to conceptualize genetic interactions.

TISSUE AND CELLULAR SPATIOTEMPORAL DYNAMICS IN COLON AGING

Aidan C Daly[1,2], Francesco Cambuli[1], Tarmo Aijo[2], Britta Lotstedt[1,3,4], Nemanja D Marjanovic[3,5], Sara Fernandez[1], Olena Kuksenko[3,6], Matthew Smith-Erb[1], Daniel Domovic[1], Nicholas Van Wittenberghe[3], Eugene Drokhlyansky[3], Gabriel K Griffin[3,7], Hemali Phatnani[1,6], Richard Bonneau[2,8,9], Aviv Regev[3,5,9], Sanja Vickovic[1,3,10,11]

[1]New York Genome Center, Technology and Innovation, New York, NY, [2]Flatiron Institute, Center for Computational Biology, New York, NY, [3]Broad Institute, Klarman Cell Observatory, Cambridge, MA, [4]KTH Royal Institute of Technology, Science for Life Laboratory, Stockholm, Sweden, [5]Massachussetts Institute of Technology, Biology, Cambridge, MA, [6]Columbia University Irving Medical Center, Neurology, New York, NY, [7]Brigham and Women's Hospital, Pathology, Boston, MA, [8]New York University, Center for Data Science, New York, NY, [9]Genentech, Genentech, San Francisco, CA, [10]Columbia University, Biomedical Engineering, Herbert Irving Institute for Cancer Dynamics, New York, NY, [11]Uppsala University, Immunology, Genetics and Pathology, Uppsala, Sweden

Tissue structure and molecular circuitry in the colon can be profoundly impacted by systemic age-related effects, but many of the underlying molecular cues remain unclear. We built a cellular and spatial atlas of the colon across three anatomical regions and 11 age groups, encompassing ~1,500 mouse gut tissues profiled by spatial transcriptomics and ~400,000 single nucleus RNA-seq profiles. We developed a computational framework, cSplotch, which learns a hierarchical Bayesian model of spatially-resolved cellular expression associated with age, tissue region, and sex, by leveraging histological features to share information across tissue samples and data modalities. Using this model, we identify cellular and molecular gradients along the adult colonic tract and across the main crypt axis, and multicellular programs associated with aging in the large intestine. Our multi-modal framework for the investigation of cell and tissue organization can aid in the understanding of cellular roles in tissue-level pathology.

LEARNING TRANSFERABLE NEURONAL REGULATORY GRAMMAR FROM MASSIVELY PARALLEL REPORTER ASSAY DATA ACROSS CELL TYPES AND ACTIVITY STATES

William DeGroat[1], Anat Kreimer[1,2]

[1]Rutgers University, Center for Advanced Biotechnology and Medicine, Piscataway, NJ, [2]Rutgers University, Department of Biochemistry and Molecular Biology, Piscataway, NJ

Understanding how cis-regulatory elements (CREs) orchestrate cell type– and cellular state–specific gene regulation remains a central challenge in regulatory genomics. In neurons, regulatory logic varies across differentiation and stimulation states, yet most models of CRE activity are trained in a single context and generalize poorly across biological conditions, limiting their utility for interpreting noncoding variation.

We developed a sequence-based framework to learn transferable neuronal regulatory grammar from massively parallel reporter assay (MPRA) data collected across neuronal cell types and depolarization states. We trained a multi-task convolutional neural network (CNN) with a shared sequence encoder and context-specific output heads, enabling the model to disentangle regulatory features shared across contexts from those that modulate activity in a state-specific manner. In parallel, we implemented an interpretable gradient-boosted model based on positional *k*-mers and sequence-derived features to provide motif-level and spatial resolution of regulatory determinants.

To further dissect context-dependent regulatory grammar, we developed two complementary neural modeling frameworks. First, we implemented an allele-aware Siamese CNN in which paired reference and alternative sequences were processed through a shared encoder and combined via a difference-based regression head to directly predict allelic effects within each neuronal context, enabling causal modeling of variant impact. Second, we designed a factorized multi-task CNN that decomposed CRE activity into additive cell type–specific baseline and stimulation-dependent components. Using a held-out context framework, the model was trained on three neuronal conditions and evaluated on an unseen cell type–stimulation combination to test whether regulatory features could be recombined to predict activity in a novel context. Together, these architectures improved cross-context robustness relative to single-task baselines while preserving interpretable sequence-level attributions.

Finally, we integrated predicted CRE activity into our MPRabc model to construct neuronal cell type–specific enhancer–gene regulatory networks. By combining sequence-derived activity with chromatin accessibility, three-dimensional contacts, and regulatory annotations, this framework reconstructed transcription factor–driven programs across differentiation and activity states. Collectively, these results establish a scalable strategy for learning transferable regulatory representations in neurons and for systematically interpreting noncoding variation across biological contexts.

PHYLOGENETIC COMPARATIVE ANALYSIS OF APOBEC3 Z-DOMAIN GENE FAMILY EVOLUTION: IMPLICATIONS FOR BAT IMMUNITY

Brenda Delamonica[1,2], Piotr Mieczkowski[3], Simon Anthony[4], Tanya Lama[5], Mani Larijani[6], Liliana Davalos[1]

[1]Stony Brook University, Ecology and Evolution, Stony Brook, NY, [2]IRACDA, CIE, Stony Brook , NY, [3]UNC, Genetics, Chapel Hill, NC, [4]UC Davis, Pathology, Microbiology, and Immunology, Davis, CA, [5]Smith College, Biology, Northampton, MA, [6]Simon Fraser University, Molecular Biology and Biochemistry, Burnaby, Canada

APOBEC3 genes encode enzymes (A3Z) involved in innate immunity, with 1-3 copies in most non-primate mammals, 7 family members (A3A-H) in primates, and even more APOBEC3-like enzymes found in bats. Although this gene subfamily expansion has been implicated in unique features of bat immunity, only eight of ~1500 bat species have bene studied and hypothesized links between APOBEC3 evolution and bat-virus interactions remain to be tested. Here we deploy ExTRaCT (Exon Targeted Retrieval and Classification Toolbox) to search for sequences encoding the catalytic region, or Z-domain, of APOBEC3 genes in 102 species representing all 21 bat families, reconcile gene trees and species to infer the evolution of Z-domains in bats, and relate Z-domain evolution to viral diversity with known bat reservoirs. We found evolutionary convergence in Z3 domain loss, high birth-death turnover in the Z1 and Z2 domains that leads to increased APOBEC3 diversity, and the number of viruses detected by a species correlated negatively with Z-domain speciation and positively with its duplication. Our results demonstrate that bat-virus interactions are better explained by evolutionary processes than by the number of Z3 domains, highlighting the potential role of evolution in shaping such relationships.

MODELING NONLINEAR AND INTERACTION EFFECTS OF SPATIOTEMPORAL AND OTHER NON-GENETIC FACTORS IMPROVES PHENOTYPIC PREDICTION FOR COMPLEX TRAITS

Ross DeVito[1], Melissa Gymrek[1,2,3]

[1]University of California San Diego, Department of Computer Science and Engineering, San Diego, CA, [2]University of California San Diego, Department of Medicine, San Diego, CA, [3]University of California San Diego, Department of Pediatrics, San Diego, CA

Adjusting for non-genetic factors can improve genetic association testing and polygenic prediction, yet most studies rely on linear adjustments for a small set of covariates. Location and time covariates are typically not included in genetic studies, but they can serve as useful proxies for environmental exposures such as climate, pollution, and behavioral differences. These features especially highlight the weakness of linear covariate adjustments, which cannot capture their cyclical or nonlinear effects nor interactions between covariates. To evaluate the benefits of modeling nonlinear covariate effects, including spatiotemporal features, we adopted a null model approach in which an auxiliary nonlinear model predicts the phenotype from non-genetic covariates alone. This prediction is then included as an additional covariate in downstream analysis, allowing association studies, polygenic scores, and fine-mapping to account for nonlinear covariate effects and interactions without modifying existing tools. Using 16 continuous phenotypes from the UK Biobank, we first evaluated gradient boosted decision tree (GBDT) and neural network null models with combinations of additional time-of-day, time-of-year, and birth or home location covariates and determined how well these models could predict phenotypes from covariates alone compared to linear models. A GBDT null improved average R^2 by 4.3% over a linear null across all phenotype-covariate set pairs, and models including spatiotemporal covariates achieved the best performance for all phenotypes. Incorporating nonlinear null predictions with additional spatiotemporal features improved polygenic scores for all phenotypes, with median improvements of 7.3% for BASIL and 15.5% for PRS-CS over the standard approach. We additionally tested three case-control phenotypes and observed improvement beyond the 95% CI of the standard approach for depression. Finally, model interpretation identified both known and novel covariate interactions and nonlinear effects. This highlighted several important cases which linear adjustments would be unable to capture, including examples where phenotype means exhibit opposite age-related trends between sexes, cyclical temporal patterns for vitamin D reflecting seasonal variation, and complex spatial dependencies between birth and home locations. Together, these results demonstrate that a small addition to existing genetic analysis workflows can improve predictive performance and provide new insights on environmental effects for complex traits.

COLOCBOOST: INTEGRATIVE MULTI-OMICS QTL COLOCALIZATION MAPS REGULATORY ARCHITECTURE IN AGING HUMAN BRAIN

Xuewei Cao[1,2], Haochen Sun[4], Ru Feng[2], Rahul Mazumder[3], Gao Wang[2], Kushal K Dey[1], Carlos Buen Abad Najar[5], Yang Li[5], Philip L de Jager[6], David Bennett[7]

[1]Memorial Sloan Kettering Cancer Center, Computational and Systems Biology, New York, NY, [2]Columbia University, Columbia Neurology, New York, NY, [3]Massachusetts Institute of Technology, Sloan School of Management, Boston, MA, [4]Mt Sinai , Icahn School of Medicine, New York, NY, [5]University of Chicago, Biological Sciences Division, Chicago, IL, [6]Columbia University, Neurology, New York, NY, [7]Rush University, Rush Alzheimer's Disease Center, Chicago, IL

Multi-trait QTL (xQTL) colocalization has shown great promises in identifying causal variants with shared genetic etiology across multiple molecular modalities, contexts, and complex diseases. However, the lack of scalable and efficient methods to integrate large-scale multi-omics data limits deeper insights into xQTL regulation. Here, we propose ColocBoost, a multi-task learning colocalization method that can scale to hundreds of traits, while accounting for multiple causal variants within a genomic region of interest. ColocBoost employs a specialized gradient boosting framework that can adaptively couple colocalized traits while performing causal variant selection, thereby enhancing the detection of weaker shared signals compared to existing pairwise and multi-trait colocalization methods. Across a broad range of simulation designs involving multiple phenotypes, ColocBoost exhibited 2.9 to 5.2-fold higher power over existing methods, while maintaining significantly lower FDR. Even for colocalization involving two phenotypes, ColocBoost showed improved performance over coloc, particularly for weaker effect size signals. In a split sample analysis of ROSMAP eQTL samples into 2 groups, ColocBoost showed 43.5% higher recovery rate of the colocalization events compared to the coloc method. We applied ColocBoost genome-wide to 17 gene-level single-nucleus and bulk xQTL data from the aging brain cortex of ROSMAP individuals (average N = 595), encompassing 6 cell types, 3 brain regions and 3 molecular modalities (expression, splicing, and protein abundance). Across molecular xQTLs, ColocBoost identified 16,503 distinct colocalization events, exhibiting 10.7(±0.74)-fold enrichment for heritability across 57 complex diseases/traits and showing strong concordance with element-gene pairs validated by CRISPR screening assays. When colocalized against Alzheimer's disease (AD) GWAS, ColocBoost identified up to 2.5-fold more distinct colocalized loci, explaining twice the AD disease heritability compared to fine-mapping without xQTL integration. This improvement is largely attributable to ColocBoost's enhanced sensitivity in detecting gene-distal colocalizations, as supported by strong concordance with known enhancer-gene links, highlighting its ability to identify biologically plausible AD susceptibility loci with underlying regulatory mechanisms. Notably, several genes including BLNK and CTSH showed sub-threshold associations in GWAS but were identified through multi-omics colocalizations which provide new functional support for their involvement in AD pathogenesis.

MAPPING GENETIC ESSENTIALISM OF ETHNICITY AND NATIONALITY ACROSS 25 YEARS OF WIKIPEDIA DATA

Alex Diaz-Papkovich[1], Abigail Kuntzleman[2], Sohini Ramachandran[3]

[1]Brown University, Data Science Institute, Providence, RI, [2]Brown University, Center for Computational Molecular Biology, Providence, RI, [3]Brown University, Ecology, Evolution, and Organismal Biology, Providence, RI

Human genetics research has become easier than ever to access, largely because of the internet and open access publishing. However, there is limited research into how it lives in the information ecosystem. In a time where recent genetics research has been misused by companies to claim prediction of behaviour from DNA, by politicians to foment xenophobia, and by extremists to support race-based murder, it is critical to understand how the public accesses our work, and whether it is being presented to support genetic essentialism—the idea that group identity is a fixed, biological reality determined by DNA.

We present the first large-scale data-driven study on how human genetics research is consumed online. Wikipedia, a free online encyclopedia, is one of the internet's largest sources of information and a training set for large language models such as ChatGPT. Analyzing how genetics is invoked in Wikipedia pages on ethnicity offers insight into whether genetics research is being used to promote genetic essentialism. Since Wikipedia pages are editable and all revisions are online, we can also study temporal trends in text about genetics and ethnicity.

We identified 6,540 Wikipedia pages about ethnicities, analyzed over 2.8 million historical revisions from 2001 to December 31, 2025, and via text mining identified keywords related to genetics (e.g. "haplo", "DNA"). We found that 970 (14.8%) Wikipedia pages about ethnic groups had genetics keywords, that 414 (6.3%) pages had sub-sections dedicated specifically to the genetics of ethnic groups, and that both have become more common over time. Using pageview data, we found that genetics terms and sections are significantly ($p < 10^{-22}$) more likely to appear in more popular Wikipedia pages. We also analyzed the associated "talk" pages for each Wikipedia page, where readers and editors discuss page contents. Of 56,098 discussions about ethnic groups, 5,654 (10.1%) mentioned genetics.

We analyzed 137 pages about nationalities and found that 93 (67.8%) had genetics keywords, and 72 (52.5%) had sub-sections specifically about the genetics of the nationality. Analyzing the citations used in the genetics sub-sections of pages about nationalities, we found that there are no dominant citations: 613 of 667 citations appeared in only one page, and the most frequent citation appeared in 10 pages.

Taken together, our results suggest that genetics research is being collated to present a genetic essentialist view of nationality and ethnicity.

DEVELOPMENTAL AND CROSS-SPECIES REGULATION OF ALTERNATIVE SPLICING ACROSS HUMAN AND NON-HUMAN PRIMATE TISSUES

Laura Domenech[1], Philipp Rentzsch[2], Winona Oliveros[2], Fairlie Reese[3], Diego Garrido-Martín[4,5], Sanna Gudmundsson[1,2], Roderic Guigó[4], Marta Melé[3], Tuuli Lappalainen[2,6], François Aguet[1], Kristin Ardlie[1], dGTEx Consortium[1]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]Science for Life Laboratory & Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden, [3]Barcelona Supercomputing Center, Barcelona, Spain, [4]Centre for Genomic Regulation, Barcelona, Spain, [5]Department of Genetics, Microbiology and Statistics, Barcelona, Spain, [6]New York Genome Center, New York City, NY

Alternative splicing is a major driver of transcriptomic diversity and is dynamically regulated across tissues and developmental stages. While adult splicing landscapes have been extensively characterized, the developmental trajectory of splicing regulation in humans and its evolutionary conservation across primates remains incompletely understood. Moreover, adult-derived transcript metrics such as proportion expressed across transcripts (pext), widely used for clinical variant interpretation, are limited in their ability to capture early developmental isoform usage.

Here, we leveraged the developmental Genotype-Tissue Expression (dGTEx) resource together with age-matched macaque and marmoset datasets (NHP-dGTEx), including prenatal stages, to characterize developmental splicing programs across humans and non-human primates.

Using splice junction-based cluster modeling across human tissues, we identified hundreds to thousands of age-differentially spliced genes (DSGs) per tissue, with the most striking changes observed in testis and muscle. DSGs are enriched in tissue-relevant biological pathways, and are consistently overrepresented among genetically constrained genes, highlighting the functional and evolutionary importance of developmental splicing programs.

Orthogonal analyses of transcript isoform usage revealed coordinated isoform redistribution across developmental timepoints, accompanied by global shifts in isoform diversity. These changes are concordant with tissue-specific developmental shifts in exon-level expression (pext). In addition, analysis of millions of ClinVar variant sites revealed differences between dGTEx-derived and adult GTEx pext values, underscoring the importance of developmental context for variant interpretation. Cross-species comparisons further demonstrated that many developmental splicing patterns identified in humans are conserved in macaques and marmosets, while also revealing lineage-specific regulatory events.

Together, these results establish a cross-species framework for understanding developmental isoform regulation and its evolutionary dynamics across humans and non-human primates.

PAN-EPIGENOME REPRESENTS EPIGENOMIC DIVERSITY

Zheng Dong[1], Juan Macias-Velasco[1,2], Juan Jiang[1], Xiaoyu Zhou[1,3], Wenjin Zhang[1], Ting Wang[1,2,3]

[1]Washington University School of Medicine, Department of Genetics, St. Louis, MO, [2]Washington University School of Medicine, McDonnell Genome Institute, St. Louis, MO, [3]Washington University School of Medicine, The Edison Family Center for Genome Sciences and Systems Biology, St. Louis, MO

The Human Pangenome Second Release provides a foundation to investigate epigenetic dynamics across the breadth of human genetic diversity, which is what we define as the human pan-epigenome: a reference of human epigenetic variation.

The first draft human pan-epigenome is constructed from 440 long-read, haplotype-resolved DNA methylomes of lymphoblastoid cell lines from 26 populations. Each methylome recovers ~2.70M more CpGs beyond GRCh38, significant improving the comprehensiveness of human epigenome landscape. On average, 1.18M CpGs per methylome are gained or lost due to genetic variation—75.7% from single-nucleotide polymorphisms, 21.8% from structural variants, and 2.5% from indels—highlighting the direct impact of genetic variants on the epigenetic landscape. These genetically distinct CpGs also display dynamic methylation patterns and functional enrichments, suggesting an previously uncharted layer of genome regulation. The pan-epigenome map thus defines epigenetic variations associated with CpG gain/loss, methylation differences, and their combination.

By leveraging graph-based, single-nucleotide–resolution coordinates for non-reference CpGs, our pan-epigenome map uncovers a greatly expanded spectrum of epigenetic variations across populations. It functionally connects epigenetic diversity, genetic variation, and transcriptional regulation, highlighting the complex, multifaceted patterns of epigenetic regulation in relationship to phenotypic variation. Genetic variants associated with CpG methylation dynamics are enriched in specific biological, medically implicated, and evolutionary processes, and strongly support the epigenetic buffering hypothesis, underscoring the reciprocal impact of genome and epigenome on each other. These findings pave the way for broader pan-epigenomics studies in human health and diseases.

POPULATION-SCALE LONG-READ RNA SEQUENCING REVEALS ISOFORM DIVERSITY AND REGULATORY VARIATION

Hope E Eden*[1], Margaret R Starostik*[2], Jonas A Gustafson*[3], Katherine M Munson[4], Rebecca Martin[5], Kaitlyn Sun[4], Joy Goffena[3], Zev Kronenberg[6], Stacy L Musone[6], Jocelyne Bruand[6], Elizabeth Tseng[6], Devin K Schweppe[4], Rob Patro[7], Evan E Eichler[4,8], Winston Timp[1], Rajiv C McCoy[2], Danny E Miller[3,9,10]

[1]Johns Hopkins University, Dept of Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, Dept of Biology, Baltimore, MD, [3]University of Washington, Dept of Pediatrics, Seattle, WA, [4]University of Washington, Dept of Genome Sciences, Seattle, WA, [5]Seattle Children's Research Institute, Genomics and Spatial Biology CoLab, Seattle, WA, [6]Pacific Biosciences , Menlo Park, CA, [7]University of Maryland, Dept of Computer Science, College Park, MD, [8]University of Washington, HHMI, Seattle, WA, [9]University of Washington, Dept of Laboratory Medicine and Pathology, Seattle, WA, [10]University of Washington, Brotman Baty Institute for Precision Medicine, Seattle, WA

*Authors contributed equally

Background: Alternative splicing is a key mechanism underlying variation in human traits and disease. Previous RNA sequencing studies in humans have been limited by short read lengths that constrained isoform characterization, as well as narrow ancestry representation. Long-read RNA-seq, including PacBio Kinnex, enables direct and allele-specific observation of complete isoforms that are challenging to detect with short reads.
Materials and Methods: We generated PacBio Kinnex long-read RNA-seq data from lymphoblastoid cell lines from 903 individuals spanning all 26 populations of the 1000 Genomes Project (1KGP). These include 607 samples that overlap with published Illumina short-read RNA-seq data, enabling comparison between technologies.
Results: Sequencing yielded a median of 10.9 million full-length non-chimeric reads per sample with an average length of >2 kb, indicating a preponderance of full-length isoforms. Individuals expressed an average of 52,862 isoforms (11,209 genes) at ≥ 1 TPM, including 35,943 known and 15,691 novel isoforms (absent from GENCODE or CHESS annotations), several of which we validated using mass spectrometry data from a subset of 10 samples. By integrating existing 1KGP genotype data, we mapped quantitative trait loci (QTLs) associated with long-read-quantified gene expression and transcript usage, identifying numerous associations obscured in short-read data. Notable examples include transcript usage QTLs at the immune locus OAS1, where introgression and diversifying selection have shaped splice site polymorphism, driving associations with viral infection and severity.
Conclusion: Long-read RNA sequencing of diverse populations expands knowledge of isoform diversity and provides insight into the genetic basis of human gene regulation and evolution.

SYSTEMATIC IDENTIFICATION AND CHARACTERIZATION OF TRANSCRIPTIONAL SILENCERS Across Viral Genomes

Mohamed Y ElSadec[1], Benedetta D'Elia[2], Tommy Taslim[2], Susan Kales[3], Ryan Tewhey[3], Juan I Fuxman Bass[1,2]

[1]Boston University, Bioinformatics, Boston, MA, [2]Boston University, Molecular Biology, Cell Biology and Biochemistry, Boston, MA, [3]The Jackson Laboratory, Bar Harbor, ME

Precise and tunable control of gene expression depends on both activating and repressive mechanisms. While promoters and enhancers have been extensively characterized, silencers remain comparatively understudied despite their essential roles in transcriptional repression, dosage control, and safeguarding against inappropriate gene activation. Previous work from our group showed that viral genomes are enriched for activating cis-regulatory elements, suggesting that a complementary repressive layer must exist to maintain transcriptional balance within these compact genomes. Moreover, stage-specific control of gene activation and repression is critical for viral persistence and the ordered progression of the lytic expression cascade. To systematically identify viral silencers, we performed Massively Parallel Reporter Assays (MPRA) tiling the genomes of 34 double-stranded DNA viruses and retroviruses across multiple promoter contexts and cell lines. We observed a high density of silencers distributed throughout most viral genomes. These elements frequently cluster near activating regions, are often compact, and exhibit diverse transcription factor motif signatures. Some silencers act in a promoter- or cell-specific manner, while others are broadly repressive; a subset are bifunctional, switching between repression and activation depending on context. Motif enrichment analysis revealed that while many viral silencers share transcription factors with human silencers, others appear to be regulated by distinct sets of transcription factors. Together, this work delivers the first systematic atlas of transcriptional silencers across viral genomes, revealing a pervasive and previously unappreciated repressive regulatory layer that shapes viral gene-expression programs.

LIKELIHOOD-BASED GEOMETRIC MODELING OF DUPLEX NANOPORE READS ENABLES ACCURATE STR MOSAICISM QUANTIFICATION

Ingrid Flaspohler[1], Melissa Englund[2], Alan Boyle[1,2]

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

Somatic mosaicism in short tandem repeats (STRs) is a key driver of phenotypic variability and disease progression in repeat expansion disorders. Progressive expansion of pathogenic alleles within specific cell types is thought to modulate disease onset and severity. However, measuring STR mosaicism is confounded by sequencing errors, particularly in nanopore sequencing as quality declines in repetitive regions. Disentangling technical noise from genuine biological variation is therefore essential for interpreting STR dynamics. In this study, we develop a statistical framework that separates sequencing-induced variability from true somatic mosaicism by modeling paired forward and reverse reads from individual DNA molecules. Using a plasmid time-course dataset as a controlled system with no biological mosaicism expected at baseline, we model the joint distribution of Nanopore duplex reads and identify stable, locus-specific error geometries. These experiments reveal reproducible, motif-dependent patterns of strand-level error and enable calibration of error variance as a function of repeat length. We define mosaicism as excess variation beyond the expected clonal peak error envelope derived from this calibration. Applying this model to plasmid samples collected over successive generations allows longitudinal tracking of repeat variation. We observe a progressive increase in non-clonal signal over time, consistent with biological accumulation of repeat changes. Stable error profiles across time provides internal validation that increased mosaicism reflects true diversification rather than a drift in technical noise. We then apply this approach to patient-derived datasets from individuals with Huntington's disease (HD) and spinocerebellar ataxia type 27B (SCA27B). For each sample and allele, we estimate the modal repeat length and use plasmid-derived calibrations to predict the expected error variance at that length. Mosaicism is then quantified as the fraction of reads exceeding this predicted error distribution. In both HD and SCA27B, we observe clear length-dependent increases in mosaicism, consistent with established models of repeat instability. Expanded alleles exhibit greater excess variance than shorter alleles in the same individuals, supporting biological repeat expansion rather than sequencing-induced error. Together, these results demonstrate that duplex-based likelihood modeling can disentangle Nanopore sequencing error from true somatic mosaicism and enable sensitive and interpretable quantification of repeat instability.

INTERPRETABLE GENETIC RISK MODELS FOR NON-LINEAR RARE AND COMMON VARIANT INTERACTIONS

Willard W Ford[1], Zachary Rodriguez[2], Sandra Lapinska[1,3], Bogdan Pasaniuc[3,4], Theodore G Drivas[2,3]

[1]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, Philadelphia, PA, [2]Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine, Philadelphia, PA, [3]Department of Genetics, Perelman School of Medicine, Philadelphia, PA, [4]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, Philadelphia, PA

Polygenic Risk Scores (PRSs) have potential to dramatically improve clinical diagnostic capabilities: PRSs are measurable from birth, often provide information orthogonal to traditional risk factors, and can show effect sizes comparable to established predictors. Despite this, they remain mostly unadopted in the clinic due to their limited precision, difficulty transferring across populations, and ambiguous interpretability. Established PRS methods linearly aggregate summary statistics from Genome Wide Association Studies to predict genetic risk, maintaining interpretability but ignoring potential genetic interactions. In contrast, recent deep learning PRS models account for genetic interactions and outperform their linear counterparts, but in doing so they lose interpretability.

In contrast to these deep learning methods which predict genetic risk directly and waste compute relearning linear genetic associations, we propose a new method that predicts the residuals of any existing linear PRS. By taking advantage of large consortium efforts and technical advances in data curation, namely GTEx and AlphaMissense, our method first aggregates rare and common variant effects at the gene level, deriving missense, loss-of-function, and predicted gene expression scores for each gene for each individual. We then train an interpretable interaction model on our gene level embeddings that builds within- and between-gene interactions on top of PRS-derived linear genetic effects to improve overall precision.

In this work we show evidence that interpretable non-linear models, including random forest and shallow deep learning models, trained to predict linear PRS residuals, can efficiently and interpretably learn genetic interactions to improve risk predictions. To prevent data leakage we use linear PRS scores from eMerge, then perform cross-fold validation with a held-out test set inside AllOfUs for non-linear model training and evaluation. To evaluate population specific genetic effects we train our highest performing models across population structure stratified samples. We also complete preliminary validation inside Penn Medicine Biobank to demonstrate generalizability.

Crucially, our residual model structure allows for investigation of exclusively non-linear interactions, improving power to discover genetic interactions relative to other non-linear models. We additionally argue that our modular framework can, in principle, be used for investigating the residuals of any PRS including admixture or population specific tools. We hope that our unique method involving both rare and common variants to predict PRS residuals will enable further adoption of genetic risk scores to potentially improve clinical outcomes and research progress.

APPLYING PRECISION MEDICINE IN PRADER–WILLI SYNDROME THROUGH COMPREHENSIVE GENOME AND PHARMACOGENOMIC PROFILING

Manavalan Gajapathy[1], Brandon M Wilk[1], Donna M Brown[1], Caroline Vrana-Diaz[2], Gurpreet Kaur[1], Jessica Bohonowych[2], Deeptha Srirangam[1], Jaimie L Richards[3], Tarun Karthik Kumar Mamidi[1], Shaurita D Hutchins[1], Bitota Lukusa-Sawalena[1], Theresa V Strong[2], Elizabeth A Worthey[2]

[1]University of Alabama at Birmingham, Center for Computational Genomics and Data Science, Department of Genetics, UAB Marnix E. Heersink School of Medicine, Birmingham, AL, [2]Foundation for Prader-Willi Research, Covina, CA, [3]University of Alabama at Birmingham, Division of General Internal Medicine and Population Science, Department of Medicine, Birmingham, AL

Prader–Willi syndrome (PWS) is a rare neurodevelopmental disorder caused by disrupted gene expression within chromosome 15q11–q13, characterized by marked clinical heterogeneity. Diagnostic testing confirms the underlying molecular mechanism but does not capture additional molecular factors that may influence phenotype or treatment response. We studied the broader clinical utility of genome sequencing (GS) in 50 individuals with molecularly confirmed PWS recruited through the Global PWS Registry.

The study, including informed consent and sample collection, was conducted remotely, enabling participation from 27 states; an approach not feasible with traditional recruitment models. Dried blood spot and buccal samples were used for GS and pharmacogenomic (PGx) analyses with an n-of-1 precision medicine framework integrating genomic, PGx, and phenotypic data. Caregiver-reported features were mapped to the Human Phenotype Ontology to support structured genotype–phenotype analyses. Validated analytical pipelines detected and annotated small variants, structural variants (SV), copy number variants (CNV), regions of homozygosity, and PGx haplotypes. Small variants were prioritized using a transcript-aware, explainable neural network framework with expert manual review, and clinically relevant findings were returned through remote genetic counseling services.

Phenotypes varied across the cohort. GS confirmed previously clinically determined PWS mechanism in all participants and further refined molecular subtype assignments in several, including reclassification of SV within deletion categories. CNV analyses revealed additional previously unrecognized structural complexity at the PWS locus among participants with both Type I and Type II deletions. Candidate SVs were identified outwith the chromosome 15 region. GS also identified returnable ACMG-73 secondary findings and rare variants outside the PWS locus that may contribute to common PWS phenotypes and comorbidities. GS-based pharmacogenomic profiling yielded clinically actionable genotype–drug associations in a substantial subset of participants, identifying individuals at elevated risk of non-response or adverse drug reactions to medications commonly prescribed in PWS.

Overall, these results indicate that GS can broaden the clinical assessment of PWS beyond confirming the primary molecular defect. By distinguishing the underlying pathogenic mechanism from additional genomic and pharmacogenomic variation that may influence symptoms or treatment response, GS could provide a more nuanced framework for individualized interpretation and management.

ALTERNATIVE SPLICING OF HER2 SHAPES ANTIBODY-DRUG CONJUGATE RESISTANCE IN BREAST CANCER.

Gabriela D Guardia[1], Carlos H dos Anjos[1], Aline Rangel-Pozzo[2], Filipe F dos Santos[1], Alexander Birbrair[3], Paula F Asprino[1], Anamaria A Camargo[1], Pedro A Galante[1]

[1]Hospital Sirio-Libanes, Centro de Oncologia Molecular, Sao Paulo, Brazil, [2]University of Manitoba, Dept of Physiology and Pathology, Winnipeg, Canada, [3]University of Wisconsin-Madison, Department of Dermatology, Madison, WI

Breast cancer (BC) is a heterogeneous disease that can be molecularly classified based on the expression of ERBB2 (HER2) and hormone receptors. Although HER2-targeted therapies, including trastuzumab, antibody-drug conjugates (ADCs), and tyrosine kinase inhibitors, have significantly improved clinical outcomes, both primary and acquired resistance remain major challenges that limit long-term therapeutic benefit. Addressing these barriers is critical for enhancing treatment stratification and achieving durable patient responses. Alternative splicing is a post-transcriptional regulatory mechanism that increases transcript and protein isoform diversity, generating products with distinct functions, subcellular localizations, structural properties, and ligand or antibody-binding capacities. In this work, HER2 alternative splicing isoforms were comprehensively characterized, their expression evaluated in primary BC samples and cell lines, and their potential contribution to resistance to anti-HER2 therapies investigated. The catalog of HER2 protein-coding isoforms was expanded from 13 to 90, revealing extensive isoform diversity associated with distinct protein domain architectures, predicted subcellular localizations, structural configurations, and differential preservation of antibody-binding regions. Integration of transcriptomic profiling from 561 primary BC samples with mass spectrometry evidence revealed a complex HER2 isoform landscape, including previously unrecognized transcripts and protein-supported isoforms not detected by routine clinical assays. Analysis of HER2 isoform expression in BC cell models sensitive or resistant to trastuzumab and ADCs demonstrated that drug-resistant cells shift isoform usage toward transcripts predicted to encode proteins lacking key antibody-binding domains. These findings broaden the current understanding of HER2 biology and support the role of alternative splicing as a previously underappreciated mechanism contributing to resistance to anti-HER2 therapies, particularly ADCs. The expanded HER2 isoform landscape highlights the importance of isoform-resolved analyses for advancing precision oncology and improving the design and implementation of targeted cancer therapies.

WHEN AND WHY DO SEQUENCE-TO-FUNCTION MODELS FAIL FOR PERSONAL GENOME PREDICTION TASKS?

Jake T Galvin[1], Alexis J Battle[2]

[1]Johns Hopkins University, Biology, Baltimore, MD, [2]Johns Hopkins University, Biomedical Engineering, Baltimore, MD

Sequence-to-function (S2F) models trained on the reference genome have the ability to prioritize functional genetic loci, including expression quantitative trait loci (eQTLs). However, they struggle on the more challenging task of predicting gene expression levels from personal genomes, which harbor many variant combinations. S2F model predictions from personal genomes are sometimes anticorrelated with observed gene expression across individuals. We investigated the basis of this phenomenon and the common patterns of errors to motivate future improvements for S2F models. We evaluated Borzoi on the Multi-ancestry Analysis of Gene Expression (MAGE) dataset, consisting of paired personal genome and transcriptome data from 731 individuals. Restricting analysis to 6,639 genes with fine-mapped eQTLs, we found that predictions for 2,066 genes (31%) were anticorrelated with observed expression across individuals.

To understand what drives anticorrelated predictions, we examined the impact of lead eQTL variants. Performing in silico mutagenesis (ISM) for each lead eQTL, we demonstrated that when Borzoi predicts the wrong sign for a lead eQTL, it is also more likely to yield anticorrelated predictions for its corresponding gene. In fact, 50% of the anticorrelated genes had a lead eQTL whose sign was predicted incorrectly. Borzoi made fewer sign errors on high-PIP eQTLs; however, both highly correlated and anticorrelated genes were enriched for these eQTLs.

We next examined how the distribution of model signal across variants influences predictive accuracy, measured using the Gini coefficient for each gene, assessed on ISM scores across variants near that gene. Genes with high Gini scores, indicating regulatory signal was concentrated in fewer variants, exhibited more extreme cross-individual correlation and anticorrelation, whereas genes with broadly distributed signal across many variants showed attenuated correlation, closer to zero, indicating difficulty integrating dispersed variant effects into a gene-level prediction. Further, variants in proximity to high-scoring variants likewise received stronger attributions often contributing to poor performance.

The most extreme model failures on personal genomes arise from sign errors in individual variants prioritized by S2F models, but predictive performance for many genes is also limited by difficulty integrating signal across multiple variants. Together, these findings characterize the common sources of error for S2F predictions from personal genomes, informing their downstream applications and pointing to patterns that can guide future improvements to the models themselves.

QUANTIFYING EARLY POST-ZYGOTIC MUTATION VARIABILITY IN LARGE, MULTI-GENERATION PEDIGREES.

Camila L Goclowski, Michael E Goldberg, Alexis C Garretson, Tom A Sasani, Hannah C Happ, Julia Ostrander, Lynn Jorde, Deborah W Neklason, Aaron R Quinlan

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

Mutations that occur early in development can have broad consequences throughout the germline and soma. These post-zygotic mutations result in mosaicism and are associated with healthy genetic variation as well as the occurrence and transmission of several genetic diseases, including developmental abnormalities, neurological disorders, and cancer predisposition syndromes. Early post-zygotic mutations (ePZMs) that occur before germline specification can lead to mosaicism across multiple tissue lineages and be transmitted to multiple offspring. It is only in the last decade that ePZMs have been studied in detail, and initial estimates suggest that at least 7% of previously characterized de novo mutations (DNMs) are actually ePZMs. However, these estimates have largely relied on genome sequencing of small families. As a result, little is known about the contribution of ePZMs to overall mosaicism, the rates and spectra of these mutations, how they vary within and across families, or the factors that modulate ePZM burden.

The CEPH/Utah pedigrees, a collection of nearly 40 large, multi-generational families, provide an ideal framework for investigating ePZM dynamics and variability by enabling detection and validation of candidate ePZMs across multiple offspring. Prior work from our group examining 3-generation pedigrees revealed that 10% of variants typically characterized as germline DNMs are actually ePZMs with distinct mutational spectra; our recent long-read sequencing analysis of a single four-generation family suggests that up to 16% of putative germline DNMs are ePZMs. Together, these works highlight the existing range in ePZM measurements within and across sequencing technologies.

We recently sequenced the fourth-generation members from 22 CEPH/Utah pedigrees, raising the cohort size to ~1,000 individuals. This expansion empowers comprehensive characterization of ePZMs across multiple generations and the dynamics of ePZM burden with parental age. We will present our progress in measuring variability in ePZM burden across families and quantifying the relationship with maternal and paternal age. Ultimately, our work offers a powerful multigenerational framework for contrasting the rates and patterns of germline DNMs with those of ePZMs—an underappreciated yet significant and consequential source of genetic variation.

HORIZONTAL TRANSFER OF NUCLEAR DNA IN TRANSMISSIBLE CANCER

Kevin Gori, Elizabeth P Murchison

University of Cambridge, Department of Veterinary Medicine, Cambridge, United Kingdom

To date, numerous infectious cancers have been discovered in the wild. The oldest known, Canine Transmissible Venereal Tumour (CTVT), is a disease that infects primarily free roaming dogs.

Originating as a tumour in an Asian dog over six thousand years ago, for millennia CTVT has spread among dog populations as a contagious allograft. The CTVT cancer cells that infect modern dogs are direct descendants of the transformed cells of the progenitor animal and carry clonal copies of its genome. However, it is possible that a tumour cell can acquire DNA from normal cells non-clonally, through a process of horizontal gene transfer.

By exploiting the increased genetic diversity between cancer and host that is characteristic of transmissible cancers, we have recently identified the signature of horizontal gene transfer in a lineage of modern tumour samples. Using genomic sequence analysis, cytology and population genetics, we trace the source of this signal to a transfer of a highly rearranged fragment of DNA that was incorporated by CTVT from a host dog that lived in the Middle East over two thousand years ago.

GRAPH GENOME-BASED ATAC-SEQ ANALYSIS REVEALS HAPLOTYPE-SPECIFIC ACCESSIBLE CHROMATIN IN STRUCTURAL VARIANT REGIONS.

Andy Gu[1], Matthew Jensen[1], Heng-Le Chen[1], Yuhang Chen[1], Jiaqi Li[1], Timur Galeev[1], Yaxi Yang[1], Eric Yang[1], Anna Su[1], Alp Namalan[1], Isabella Wu[1], Tai Michaels[1], Michael Schatz[2], Joel Rozowsky[1], Mark Gerstein[1]

[1]Yale University, Molecular Biophysics & Biochemistry, New Haven, CT, [2]Johns Hopkins University, Department of Computer Science, Baltimore, MD

Comprehensive analyses of non-coding variants have provided valuable insights into the genomic mechanisms underlying complex traits. Among these, structural variants (SVs) are under-represented in functional genomic studies despite their outsized role towards disease and ability to alter multiple regulatory elements. Accurate characterization of genomic alterations in regulatory regions requires methods that account for SVs, which remain challenging to represent in linear reference genome contexts despite advances in long-read sequencing technologies.

To better survey the effects of SVs on gene regulation, we constructed personalized diploid graph genomes and re-aligned functional genomics datasets from the EN-TEx personal epigenome resource, which combines tissue-specific gene regulation data from GTEx with diverse functional data modalities from ENCODE. For the revised EN-TEx+ Resource, we developed reproducible graph genome workflows to enable haplotype-resolved peak calling for 511 tissue-specific datasets across seven data modalities (ATAC-seq, DNase-seq, methylation, and ChIP-Seq for three histone marks and CTCF) and 25 tissues from four healthy donors. We mapped a subset of datasets to a range of graph- and linear-based personal and reference genomes, finding that diploid personal graph genomes maximized the probability of mapping reads and identifying genomic elements. Using graph-based methods, we identified hundreds of accessible chromatin peaks within insertion sequences across donors that were systematically missed by linear reference approaches. Notably, a significant proportion of these peaks localized to SVs unique to specific haplotypes, representing "haplo-novel" regulatory elements invisible to the linear reference and only discoverable when surjecting graph-based peaks onto linear coordinates. A subset of haplo-novel peaks corresponded with chromatin conformation changes and outlier expression of nearby genes, highlighting the regulatory impact of SVs.

Ultimately, our approach emphasizes that SVs often carry novel or duplicated regulatory elements, underscoring their importance in gene regulation, and demonstrates the utility of personalized graph genomes for mapping functional genomics datasets. We are making the EN-TEx+ datasets available as an open-source and AI-ready resource, ensuring that genome assemblies and functional datasets are compatible with the Human Pangenome Reference.

DIVERGENT ONCOGENIC AND IMMUNE EVASIVE CANCER CELL REPROGRAMING IN MYXOID/ROUND CELL LIPOSARCOMA

Evan Seffar[1,2], Rodrigo Gularte-Mérida[1], George Li[1], Narasimhan P Agaram[3], Francisco Sánchez-Vega[2], Samuel Singer[1]

[1]Memorial Sloan Kettering Cancer Center, Department of Surgery, New York, NY, [2]Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics, New York, NY, [3]Memorial Sloan Kettering Cancer Center, Department of Pathology, New York, NY

How a single oncogenic event drives malignant transformation and shapes the tumor ecosystem remains a central question in cancer biology. Fusion-driven sarcomas offer an ideal model to address this, as they are defined by pathognomonic chromosomal translocations that generate singular oncogenic drivers. Myxoid/round cell liposarcoma (MRCLS) is among the most common of these tumors, predominantly affecting young adults and accounts for ~30% of liposarcomas. Over 90% of cases harbor a FUS:DDIT3 fusion, and is sufficient to initiate tumorigenesis. Despite its well-defined genetic origin, the downstream transcriptional programs orchestrated by this fusion remain poorly understood. To address this, we integrated single-cell RNA-seq with short- and long-read sequencing to identify fusion-positive cells, characterize their transcriptional states, and map tumor–microenvironment interactions across six MRCLS patients. Despite a shared oncogenic driver, cancer cells consistently diverged into two transcriptional states: (1) CTAg+ stem-like cells, marked by reactivation of cancer-testis antigens (*PRAME, CTAG2*) and expression of mesenchymal/adipose stem markers (*CD73*, *CD90, PDGFRA*); (2) RTK+ cells, marked by receptor tyrosine kinase activation (*IGF1R, FGFR2, PPARG*) and upregulation of NOTCH signaling. This bifurcation was conserved across patients despite variability in FUS breakpoints, suggesting that the DDIT3 active domain under FUS promoter controls the dual transcriptional programs. Both states deploy immune evasion strategies, though with differing strengths. CTAg+ cells had ~60% lower interferon-gamma (IFNγ) expression compared to normal stromal cells, while RTK+ cells had near- undetectable INFγ expression. Both populations downregulated MHC class I and II genes, with RTK+ cells displaying lower expression compared to CTAg+. All tumors exhibited depletion of $CD8^+$ T cells and enrichment of tumor-associated macrophages compared to normal tissue. Together, our findings reveal that a single oncogenic fusion can deterministically program complementary functional states—one preserving stemness and adaptive potential, the other partially differentiated with heightened oncogenic signaling—while orchestrating a multifaceted immune escape. This study provides the first single-cell atlas of fusion-driven MRCLS and advances understanding of how oncogenic fusions shape tumor heterogeneity and immune landscapes at cellular resolution.

LABEL-FREE LOCAL HAPLOTYPE EMBEDDINGS RECOVER CAUSAL GENETIC EFFECTS FROM LD-LINKED TAGS ACROSS POPULATIONS

Hersh V Gupta, Mariko Isshiki, Srilakshmi M Raj

Albert Einstein College of Medicine, Genetics, New York, NY

Introduction: GWAS and PRS transfer poorly across ancestries due to identifying tag rather than causal SNPs via linkage disequilibrium, not ancestry-specific biology. Existing solutions suffer two flaws: (1) discrete ancestry labels introduce reference panel bias and fail for admixed individuals, and (2) global genomic representations cannot capture local LD variation. We propose localized, label-free numerical embeddings of haplotypes to capture local sequence variation and enable transferability beyond discretized ancestry.

Methods: We simulated 20K European and 20K African samples (hapnest) with single-variant phenotypes to isolate LD-driven transfer failure before comparing in complex polygenic architectures. Causal variants (MAF 1-5%) were simulated across chromosomes 1-6; 5,900 variants with LD proxies (highest R^2-based tag with MAF>5%) were retained. We implemented local PCA on phased haplotypes, testing window sizes (0.25-5 cM) with ~1 SNP per 0.01 cM, and measured variance recovery (h^2_observed/h^2_true) in held-out samples. Effect adjustments were compared across unadjusted tag SNPs and adjustments for global continental labels, global PCA, and local PCA.

Results: Local PCA parallelizes efficiently (chromosome 1, 40K samples, 700 variants: 40 minutes at 0.25 cM windows). Applied to real data (1000 Genomes, HGDP, SGDP), local PCA achieved dosage reconstruction R^2>0.45 versus <0.05 compared to other methods in variants MAF 1-5%. It also captures localized selection signatures (e.g., LCT sweep), confirming capture of local LD structure.

In simulated data, local PCA achieved 10× improvement in variance recovery over unadjusted tag SNPs at low MAF (<2%) in held-out samples. Optimal performance occurred in 0.25 cM windows with 80 PCs. Larger windows required more PCs (175 at 1 cM, ~100 at 5 cM) but showed severe degradation (1 cM: 60% loss at low MAF, 25% at high MAF; 5 cM: 95% and 70% loss). In held-out African samples, variance recovery reached 20% (low MAF) and 45% (4-5% MAF) with 80 local PCs, versus 2%/10% (unadjusted/labels) and 2%/15% (global PCA). Performance held in 89% European-trained models applied to African test sets across all heritabilities. Global PCA degraded beyond 5 PCs and reversed effect estimates, suggesting global ancestry descriptions fail to capture local variation, though this remains to be tested in real-world data.

Conclusion: Label-free local embeddings recover effect sizes with optimal window sizes matching population LD extent, validating extension to complex phenotypes and real data. Future work explores non-linear embeddings for improved recovery.

PLACENTAL GENOMIC SIGNATURES FOR SOCIOECONOMIC INDICATORS IN US PREGNANT WOMEN

Tesfa D Habtewold[1], Richard J Biedrzycki[2], Prabhavi Wijesiriwardhana[1], Kunal Kathuria[1], Fasil Tekola-Ayele[1]

[1]National Institute of Child Health and Human Development, National Institutes of Health, Epidemiology Branch, Bethesda, MD, [2]National Institute of Child Health and Human Development, National Institutes of Health, Division of Population Health Research, Bethesda, MD, [3]National Institute of Child Health and Human Development, National Institutes of Health, Epidemiology Branch, Bethesda, MD, [4]National Institute of Child Health and Human Development, National Institutes of Health, Epidemiology Branch, Bethesda, MD, [5]National Institute of Child Health and Human Development, National Institutes of Health, Epidemiology Branch, Bethesda, MD

Background. Socioeconomic differences are associated with increased morbidity and mortality related to cardiometabolic diseases across the lifespan. However, their molecular mechanisms are poorly understood. In this study, we investigated placental genetic expression markers associated with socioeconomic indicators.

Methods. Bulk RNA sequencing was performed using the Illumina HiSeq2000 sequencing platform with 100 bp paired-end reads on 80 placental samples collected at delivery as part of the NICHD Fetal Growth Studies. Socioeconomic status was estimated using median (above median vs. below median) values of the education isolation index (EII), neighborhood deprivation index (NDI), and birthplace (non-US born vs. US). A negative binomial regression model was applied using the edgeR package in R to identify differentially expressed placental genes while adjusting for maternal age, fetal sex, health insurance type, cell types, and genotype and RNA principal components. Significant differentially expressed genes (DEGs) were defined for each comparison as having a Benjamini-Hochberg adjusted p-value of less than 0.05.

Results. Higher EII and NDI were associated with 26 and 16 DEGs, respectively (pFDR < 0.05). Birthplace was associated with 18 DEGs. The TRAV5 gene, which is essential for immune response, was a common DEG for EII and NDI. These DEGs were previously reported to be associated with type 2 diabetes, hypertension, and metabolic biomarkers. Significant tissue-specific genetic regulation was not observed in any of the DEGs.

Conclusion. Multiple genes were differentially expressed related to socioeconomic indicators. Social and environmental exposures could induce sustained transcriptional alterations in inflammatory, metabolic, and vascular regulatory pathways that ultimately contribute to cardiometabolic dysfunction.

NEW UCSC GENOME BROWSER FEATURES: FREE STORAGE SPACE FOR TRACK HUB ANNOTATION FILES, ON-THE-FLY LIFTOVER AND AN INTERACTIVE EDITOR FOR GENOME ANNOTATIONS

Maximilian Haeussler[1], Hiram Clawson[1], Brian Raney[1], Galt Barber[1], Jairo Navarro[1], Gerardo Perez[1], Anton Nekrutenko[2], Jonathan Casper[1], Luis R Nassar[1]

[1]UCSC, Genomics Institute, Santa Cruz, CA, [2]Pennsylvania State University, Dept. of Biochemistry and Molecular Biology, University Park, PA

As the number of high-quality genome assemblies grows rapidly - driven by the T2T Consortium, the Human Pangenome Reference, and large-scale comparative projects - and the number of human annotation tracks increases daily, so does the need for tools that make genome annotations accessible, portable, and easy to create. We present three new features of the UCSC Genome Browser that address these challenges.

First, we now provide free upload hosting for user-generated annotation files. Track hubs have long allowed researchers to display their own data (BAM, VCF, BigBed, BigWig, etc.) on the Genome Browser, but required maintaining a lab web server. Users can now upload up to 10 GB of annotation files directly to UCSC through an interactive web interface or command-line tool, with automatic generation of the hub configuration file. The resulting hubs are stored on fast infrastructure, can be saved into stable session links for sharing via manuscripts or with collaborators, and are compatible not only with the UCSC Genome Browser but also with IGV and other genome browsers. We can increase the quota generously upon request, and hubs of broad interest can be promoted to our Public Hubs listing or converted into native Browser tracks, as was done for TOGA gene models across hundreds of vertebrate genomes and for resources such as JASPAR and ENCODE4.

Second, we extended the Browser's long-standing liftOver functionality to support dynamic, on-the-fly coordinate conversion between any pair of assemblies for which a whole-genome alignment is available. Using precomputed alignments made with a new workflow that Galaxy provides, users can now browse one assembly while viewing annotations projected from another. For example, the rich annotation catalog of GRCh38 can be displayed directly on the T2T-HG002 diploid assembly, enabling immediate biological interpretation of new references without waiting for dedicated annotation efforts.

Third, we have added an interactive annotation editor designed for users who need to mark up a small number of features — such as candidate variants, primer binding sites, or probe locations — without the overhead of creating data files or working on the command line. The editor supports custom fields, sharing between users, and in-place editing, lowering the barrier for clinicians or students to communicate annotations.

Together, these features reflect a broader effort to make the Genome Browser not only a visualization platform but a collaborative workspace suited to the scale and diversity of modern genomics.

THE GENETIC BASIS OF SUSCEPTIBILITY, RESISTANCE, AND TOLERANCE TO *SALMONELLA* INFECTION

Christopher J Harbort[1], Bärbel Raupach[1], Denise Monack[2], Arturo Zychlinsky[1]

[1]Max Planck Institute for Infection Biology, Cellular Microbiology, Berlin, Germany, [2]Stanford Medicine, Microbiology and Immunology, Stanford, CA

Typhoid fever is a major disease burden worldwide, caused by infection with Salmonella enterica. Infection outcomes are varied: individuals can display high susceptibility (both morbidity and mortality), resistance (prevention or rapid clearance), or in some cases tolerance (asymptomatic infection). Despite the absence of disease symptoms, tolerant individuals may withstand a high bacterial burden, a combination that greatly increases transmission. Uncovering the genetic determinants of disease outcomes and tolerance will be important for prevention and treatment of not only Typhoid fever, but potentially myriad infectious diseases with diverse presentations.

Inbred mice offer a convenient model to uncover and characterize such genetic determinants of infection, but lack the diversity and relevance of the human population. Genome-wide association studies in humans with typhoid fever, however, have identified only a few large candidate regions, lacking the statistical power to hone in on genes. Furthermore, tools to pinpoint the mechanisms behind these associations and determine causality are lacking. To bridge the gap between the diversity of the human population and an experimentally tractable infection model, we are using a highly genetically diverse, outbred mouse stock combined with high resolution genetic mapping. The increased resolution of this system allows pinpointing QTL regions with sub-megabase precision, to identify genes as opposed to large regions.

We are combining extensive disease phenotyping, multi-omics, and genome reconstruction of a large cohort of genetically individual mice to map complex infection outcomes to genomic regions. A systems genetics approach integrating multi-omics data and genetic association mapping of complex phenotypes will provide important new insights to infection outcomes at the systems level. Importantly, we can then use the controlled diversity of this system to test mechanistic hypotheses uncovered from our analysis, a step beyond what is possible with association mapping in humans.

ISOFORM-LEVEL FINE-MAPPING IN TWAS USING LONG-READ-INFORMED PRIORS

Taylor Head, Sean Bresnahan, Arjun Bhattacharya

University of Texas MD Anderson Cancer Center, Epidemiology, Houston, TX

Transcriptome-wide association studies (TWAS) show limited replicability and elevated false-positive rates in part due to uncertainty in transcript annotation at the isoform level. Current reference annotations like GENCODE are comprehensive but include increasing numbers of transcripts that are weakly supported or unexpressed in a given tissue. This can inflate expression model complexity and obscure true underlying regulatory mechanisms. To evaluate this, we quantified expression across 48 GTEx tissues and observed substantial discordance (up to 50%) in TWAS-prioritized genes solely from changes in GENCODE annotation versions. In contrast, through long-read (LR) RNA-seq, we can directly observe full-length isoforms and provide an accurate and tissue-specific isoform space in principle. However, costs severely limit sample size in LR which can prevent full capture of the isoform repertoire and often preclude direct use as a standalone reference. This motivates a new isoform-level fine-mapping framework that leverages LR RNA-seq evidence without requiring requantification of short-read data. We propose a Bayesian model that jointly estimates posterior probabilities of causality for individual isoforms, total gene expression, and direct variant effects that incorporates isoform-specific priors derived from LR transcriptomic features. By prioritizing biologically-supported isoforms rather than expanding the dimension of multi-isoform expression models, this approach can improve causal inference in disease-specific tissues and contexts. We evaluate power and type I error using a robust simulation framework and benchmark our method against existing TWAS fine-mapping methods with planned application to publicly available short-read RNA-seq datasets.

A COMPLETE GENOME FOR THE COMMON MARMOSET.

Prajna Hebbar[1], Hailey Loucks[1], Joanna Malukiewicz[2], DongAhn Yoo[3], Murillo Rodrigues[4], Karina Ray[4], Tamara Potapova[5], Don Conrad[4], Benedict Paten[1]

[1]University of California Santa Cruz, Biomolecular Engineering, Santa Cruz, CA, [2]University of Hamburg, Department of Biology, Hamburg, Germany, [3]University of Washington, Department of Genome Sciences, Seattle, WA, [4]Oregon National Primate Research Center, Division of Genetics, Portland, OR, [5]Stowers Institute for Medical Research, Kansas City, MO

The common marmoset is a fascinating tiny New World Monkey (NWM) that is a pivotal species to investigate questions regarding primate evolution and human disease, such as Alzheimer's. Here we present the first telomere-to-telomere (T2T) reference genome for the common marmoset, adding over 88 Mb of sequence and resolving challenging genomic regions. An additional near T2T assembly from a second unrelated individual provides a total of four high quality haplotypes for analysis. The improved contiguity and accuracy of these assemblies enable unprecedented insights into complex and rapidly evolving genomic regions such as centromeres, sex chromosomes, ribosomal DNA structure, and major histocompatibility complex (MHC). For the first time, we fully resolved all marmoset centromeres, revealing dimeric alpha satellites with chromosomal specificity and stratified inactive layers documenting ancestral centromere turnover. One interesting feature of the sex chromosomes is that the Y chromosome, but not the X chromosome, carries ribosomal DNA, creating a sexually dimorphic copy number. We discovered multiple novel, marmoset-specific MHC genes that are predicted to protect against pathogens encountered in its environment. Leveraging this complete reference, we further identified over 900 previously unannotated, transcribed protein-coding genes specific to the marmoset lineage, including genes implicated in neurodevelopmental processes. Together with additional long-read marmoset assemblies, these genomes were used to construct a marmoset pangenome, providing a robust reference framework for short-read mapping across diverse individuals. In this talk, I will elucidate how this resource will improve the utility of the common marmoset as a biomedical model organism and fill key gaps in our understanding of primate evolution.

SEA ROBINS AS A MODEL FOR EVOLUTIONARY INNOVATIONS

Alex Zhang, Amy L Herbert

University of Chicago, Organismal Biology and Anatomy, Chicago, IL

Although advancements in genetics and genomics have transformed the field of evolutionary biology, much of this research has focused on traits that are reduced or lost. In contrast, the molecular mechanisms underlying novel or gained traits in wild species remain less understood. Moreover, while decreasing sequencing costs have enabled comparative genomic analyses across many fascinating species, functionally testing these findings in wild vertebrates remains challenging. To address these questions, we have been developing sea robins as a model for studying dramatic skeletal, sensory, and nervous system innovations in vertebrate evolution. Sea robins are saltwater fish that exhibit numerous evolutionary innovations, including expanded, wing-like pectoral fins and leg-like structures that allow the fish to walk and taste food on the ocean floor. We crossed two readily accessible sea robin species with notable trait differences, including significant pectoral fin size variation, to produce hybrids and explore the genetic basis of species-specific differences. We performed RNA-sequencing and allele-specific expression analysis on the pectoral fins of both species and F1 hybrids. We found that the gene *aldh1a2*, which is involved in retinoic acid synthesis, was significantly upregulated in the species with larger pectoral fins and appears to be regulated in *cis*-. Notably, *aldh1a2* has putatively been linked to increased forelimb size in other species, while reduced expression is associated with smaller wings in the flightless emu. When we disrupted *aldh1a2* in sea robins using CRISPR-Cas9 genome editing, we found that larvae developed without pectoral fins. Using ATAC-sequencing data from zebrafish pectoral fins, we have identified several putative *aldh1a2* enhancers which are conserved in both sea robins. Notably, several candidate enhancers show marked sequence divergence in the sea robin species with enlarged pectoral fins. To test for regulatory activity, we will clone species-specific enhancer sequences upstream of a minimal promoter driving GFP, inject into sea robin embryos, and quantitatively compare reporter expression levels. We will then use CRISPR-Cas9 genome editing to perturb these enhancer regions in each species to assess their *in vivo* contribution to *aldh1a2* expression and pectoral fin development. Together, these experiments will identify species-specific *aldh1a2* enhancer differences and, more broadly, test the functional relevance of genes and elements identified through large-scale sequencing studies.

DEVELOPMENT OF A HIGH-THROUGHPUT CUT&RUN PLATFORM FOR EPIGENOMIC MAPPING OF RARE PRIMARY IMMUNE CELLS

Allison R Hickman[1], Matthew R Marunde[1], Danielle Maryanski[1], Carolina Lin Windham[1], Courtney Barnes[1], Liz Albertorio-Saez[1], Dughan J Ahimovic[2], Michael J Bale[2], Juliana J Lee[3], Steven Josefowicz[2], Michael-Christopher Keogh[1]

[1]EpiCypher, Inc, EpiCypher, Durham, NC, [2]Weill Cornell Medicine, Department of Pathology and Laboratory Medicine, New York, NY, [3]Harvard Medical School, Department of Immunology, Boston, MA

Understanding immune cell differentiation is central to gene and cell therapy research. Regulation of chromatin structure drives many of these processes, leading to the widespread application of chromatin accessibility (ATAC-seq) and transcriptional (RNA-seq) profiling for immune cell characterization. However, these assays often fail to provide mechanistic insight and can lack the granular detail required to define rare and/or novel immune cells.

Epigenomic features – such as histone post-translational modifications (PTMs) and chromatin-associated proteins – mark distinct genomic features (e.g., transcriptional promoters or enhancers) and regulate chromatin structure, gene expression, and cell function. Mapping these features provides a rich context to study cell fate and has great potential for discovering new biomarkers and drug targets. Notably, the traditional chromatin mapping technology (ChIP-seq) is impractical for immune cell studies due to poor sensitivity, high background, and low throughput. Additionally, sample processing methods for ChIP vary and require substantial expertise, complicating its use in tightly controlled, multi-site studies.

Here, we present a breakthrough CUT&RUN assay for rapid, ultra-sensitive profiling of FACS-sorted primary immune cells. Our workflow generates reliable profiles from <10,000 cells per reaction, and is supported by a rigorous optimization strategy, high-quality antibodies, and robust spike-in controls. Further, by automating our CUT&RUN protocol, we were able to increase throughput and standardize sample handling. The resulting **autoCUT&RUN** workflow was applied at-scale in collaboration with ImmGen, a multi-site project building a comprehensive 'omic database of mouse immune cells. We have generated >1,500 epigenomic profiles from >100 sorted primary mouse immune cell types, demonstrating the reliability of our process across diverse samples and labs.

Our CUT&RUN approach allows researchers to fully leverage the epigenome to define cell identity. This technology has broad applications in cell and gene therapy research, including CAR T cells, Cas9/dCas9 off-target modifications, exhausted T cells, and induced pluripotent stem cells.

FUFIHLA: A TOOL FOR FULL-FIELD HLA TYPING FROM LONG READ DATA

Jingqing Hu[1], Qian Qin[1,2,3], Heng Li[1,2,4], Ying Zhou[1]

[1]Dana-Farber Cancer Institute, Department of Data Science, Boston, MA, [2]Harvard Medical School, Department of Biomedical Informatics, Boston, MA, [3]Brigham and Women's Hospital, Division of Rheumatology, Inflammation, and Immunity, Boston, MA, [4]Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA

Motivation: Allele typing for Human Leukocyte Antigen (HLA) genes has many important clinical applications. Popular short-read typing can only accurately distinguish alleles at the peptide level, which potentially limit our understanding of the effect of variants in non-coding region. Long read data has been proved to be useful in typing HLA alleles in full resolution, but only a few tools are publicly available and with significant limitations in practical application.
Results: We developed FuFiHLA, a lightweight open-source software, to type HLA alleles. Currently it supports typing alleles of six HLA genes (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, and HLA-DQB1) from long reads. Evaluation using 232 PacBio HiFi WGS samples from HPRC shows that FuFiHLA achieves 99.6% accuracy in the full field allele typing and QV as 51.7 for consensus allele sequence construction. Additional testing on four Nanopore R10 reads demonstrates slightly reduced accuracy in the fourth field.

DRUG REPURPOSING THROUGH DEEP LEARNING-BASED PREDICTION OF TRAIT-RELEVANT TRANSCRIPTION FACTORS

Xiaoqin Huang, Di Huang, Ivan Ovcharenko

National Library of Medicine, National Institutes of Health, Division of Intramural Research, Bethesda, MD

Background
Traditional drug development is costly, time-consuming, and marked by high attrition, whereas drug repurposing offers a more efficient strategy by identifying new uses for existing compounds. Although many computational approaches integrate gene expression and genome-wide association study (GWAS) data, most overlook how noncoding variants regulate gene expression. Because most GWAS variants reside in regulatory elements such as enhancers and act through transcription factor (TF)-mediated gene regulation, we developed an integrative framework to prioritize candidate drugs by linking trait-associated regulatory variants to TF-centered transcriptional programs and downstream drug-induced gene expression responses.
Results
Cell type-specific deep learning models accurately predicted enhancers (AUROC: 0.91-0.98, AUPRC: 0.58-0.86). Attribution-based motif interpretation recovered broadly acting and cell type-specific TF programs across contexts. In a breast cancer case study, we prioritized approved and investigational compounds that exhibited strong therapeutic concordance with FOXA1-associated transcriptional perturbations. Over 80% of these candidates consistently reversed key breast cancer-associated pathways, including cell cycle, hormone response, and growth signaling. Integration of curated drug-gene interaction data and literature evidence further refined this set to a final group of high-confidence candidates, including proteasome inhibitors (ixazomib, bortezomib, carfilzomib), metabolic modulators (pitavastatin, dorsomorphin), antimetabolites (floxuridine), a nucleoside metabolic inhibitor (clofarabine), and a topoisomerase inhibitor (camptothecin). Several of these compounds have prior clinical or preclinical evidence supporting their relevance in breast cancer.
Conclusion
We present a deep learning-based, regulatory variant-informed framework that connects genetic risk to therapeutic hypothesis generation through TF-centered regulatory modeling and transcriptional perturbation analysis. By integrating enhancer prediction, TF motif interpretation, pathway-level reversal assessment, and drug-gene interaction data, this approach enables context-specific drug prioritization grounded in regulatory mechanisms. Our framework provides a generalizable strategy for drug repurposing and establishes a foundation for extending regulatory variant-guided therapeutic discovery to a broad range of complex traits.

WHOLE EXOME SEQUENCING IN PERINATAL STROKE: PATHOGENIC/LIKELY PATHOGENIC YIELD ACROSS FIVE VASCULAR SUBTYPES

Jaan M Huik[1], Norman Ilves[1], Nigul Ilves[1], Sander Pajusalu[2], Rael Laugesaar[3], Triin Alter[1], Tiina Kahre[2], Ulvi Vaher[1], Pille Kool[1], Dagmar Loorits[4], Pilvi Ilves[1]

[1]University of Tartu, Institute of Clinical Medicine, Department of Radiology, Tartu, Estonia, [2]University of Tartu, Institute of Clinical Medicine, Department of Genetics and Personalized Medicine, Tartu, Estonia, [3]University of Tartu, Institute of Clinical Medicine, Department of Pediatrics, Tartu, Estonia, [4]Tartu University Hospital, Department of Radiology, Tartu, Estonia

Background: The genetic contribution to perinatal stroke is increasingly recognized, but subtype specific evidence remains limited by small cohorts, heterogeneous stroke classification and inconsistent genetic methods. We assessed diagnostic yield of all major vascular subtypes of perinatal stroke with whole exome sequencing (WES) using standardized classification and uniform variant interpretation.
Methods: Children (n=193) with five vascular subtypes—arterial ischemic stroke (AIS), hemorrhagic stroke (HS), cerebral venous sinus thrombosis (CSVT), periventricular venous infarction (PVI), and periventricular hemorrhagic infarction (PVHI)—were recruited from the Estonian Pediatric Stroke Database (1994–2025). Stroke subtype was confirmed via multidisciplinary consensus review of neuroimaging. We performed WES in all children and prioritized rare coding and splice-site variants within two gene sets: (1) vascular wall integrity/stroke mechanisms and (2) coagulopathy pathways. We classified variants following ACMG/AMP criteria. Differences in pathogenic or likely pathogenic diagnostic yield across stroke subtypes were assessed using Chi square test.
Results: Final cohort included 48 children with AIS, 14 with HS, 9 with CSVT, 60 with PVI and 62 with PVHI. We identified P/LP variants in all subtypes, with an overall diagnostic yield of 10.9% (21/193).
Pathogenic/likely pathogenic (P/LP) variants, variants of uncertain significance (VUS), and negative (Neg) findings were distributed as follows:
AIS (n=48): P/LP 6.3% (3/48); VUS 14.6% (7/48); Neg 79.2% (38/48)
HS (n=14): P/LP 21.4% (3/14); VUS 0% (0/14); Neg 78.6% (11/14)
CSVT (n=9): P/LP 22.2% (2/9); VUS 0% (0/9); Neg 77.8% (7/9)
PVI (n=60): P/LP 13.3% (8/60); VUS 5% (3/60); Neg 81.7% (49/60)
PVHI (n=62): P/LP 8.1% (5/62); VUS 14.5% (9/62); Neg 77.4% (48/62)
Overall (n=193): P/LP 10.9% (21/193); VUS 9.8% (19/193); Neg 79.3% 153/193
We observed no statistically significant differences in P/LP yield between stroke subtypes (Chi square test, p=0.31), suggesting a uniform genetic contribution regardless of vascular mechanism. VUS were identified in 9.8% (19/193) of the cohort, representing a significant substrate for future reclassification.
Conclusion: P/LP variants were detected across all perinatal stroke subtypes, supporting a clinically meaningful genetic contribution to all stroke subtypes. VUS are frequent representing candidates for future reclassification through segregation and functional studies. We propose that diagnostic WES should be integrated into the standard clinical workup for all perinatal stroke children to improve diagnostic precision and family counseling.
Funded by: Estonian Research Council PRG1912

EXPANDING AND IMPROVING THE GENCODE HUMAN REFERENCE ANNOTATION

Tobias Hunt, Jose M Gonzalez, Ryan Merritt, Jane Loveland, Jonathan M Mudge, Adam Frankish

EMBL-EBI, Cambridge, United Kingdom

The GENCODE consortium provides comprehensive reference annotations of all human and mouse protein-coding genes, pseudogenes, long non-coding RNAs, and small RNAs. Accurate gene annotation is essential for both genome biology and clinical genomics, as incomplete or erroneous annotation can propagate significant downstream analytical errors.

Although GENCODE is a well-established community resource, the rapid emergence of new sequencing technologies and diverse biological datasets presents ongoing opportunities to expand and enhance the GENCODE gene set for the benefit of its global user base.

Recent developments include:

1. Integration of long-read transcriptomic data: We have developed a suite of bioinformatic pipelines to automatically incorporate large volumes of PacBio and Oxford Nanopore (ONT) long-read RNA sequences into the GENCODE gene set. These efforts have expanded our catalog of novel splice variants and improved the consistency of annotated 5′ and 3′ UTRs.
2. Improved annotation of small RNAs: By standardizing our workflows and updating the annotation of small RNA families, we have enhanced the accuracy and usability of GENCODE as a reference resource for this important class of genes.
3. Identification of non-canonical ORFs: Using ribosome profiling and immunopeptidomics data, we have generated a high-confidence catalog of "non-canonical" open reading frames (ORFs) that may represent functional translation events within lncRNAs and UTRs of protein-coding genes.
4. Extension to new genome assemblies: We have provided GENCODE annotation for the T2T-CHM13 human genome by lifting over and refining models from GRCh38, while also adding novel gene models where appropriate. In parallel, we are developing strategies to project our annotation onto the emerging Human Pangenome reference, thus enabling the exploration of its genetic diversity and previously unresolved genomic regions.

Together, these initiatives extend the depth, accuracy, and scope of the GENCODE catalog, reinforcing its role as a gold-standard resource for both fundamental genome research and clinical genomics applications.

The GENCODE genesets are the default Human and Mouse annotation used in the Ensembl and UCSC genome browsers and can also be downloaded from www.gencodegenes.org.

on behalf of the GENCODE consortium

PARENTAL KINSHIP LANDSCAPES SHAPE THE EPIGENOME, DECELERATE EPIGENETIC AGING, AND ALTER THE BRAIN TRANSCRIPTOME IN *PEROMYSCUS*

Kim-Tuyen Huynh-Dam[1], Xiaoyu Feng[1], Celia Jaeger[2], Ioulia Chatzistamou[3], Hippokratis Kiaris[1,2]

[1]Department of Drug Discovery and Biomedical Sciences, College of Pharmacy, University of South Carolina, Columbia, SC, [2]Peromyscus Genetic Stock Center, University of South Carolina, Columbia, SC, [3]Department of Pathology, Microbiology and Immunology, School of Medicine, University of South Carolina, Columbia, SC

While the genetic consequences of parental relatedness are well-documented, the systematic impact of kinship on the epigenetic, transcriptomic, and proteomic architecture remains poorly characterized. Leveraging long-term captive colonies of deer mice (*Peromyscus*) and available detailed pedigrees, we applied a multi-omic strategy to explore how parental kinship modifies the molecular profile of offspring.

Utilizing DNA methylation (DNAm) profiling of ~37,000 CpG sites, we demonstrated that parental kinship—quantified via rigorous pedigree analysis—imprinted distinct epigenetic signatures capable of high-accuracy kinship prediction. Intriguingly, co-analysis with epigenetic age estimators revealed that increased kinship significantly delayed epigenetic aging, resulting in an estimated 13% reduction in epigenetic age. Through Genome-Wide Association Studies (GWAS), we identified specific genomic loci associated with this epigenetic age variance, suggesting a genetically encoded modulation of the epigenetic clock.

Our data highlighted a profound sex-dimorphism: males exhibited heightened sensitivity to parental kinship, characterized predominantly by CpG hypermethylation. These epigenetic shifts corresponded with significant perturbations in the brain transcriptome, particularly within pathways governing nervous system development and anatomic-specific gene regulation. We further validated these findings through plasma proteomic profiling of F1 and F2 hybrids—representing discrete levels of parental kinship—and experimental stress-response assays in cultured cells and mice. Collectively, these results identify parental kinship as a potent, yet underappreciated, modifier of the molecular phenome, with significant implications for understanding the biology of isolated populations and the evolution of aging.

GENETIC ARCHITECTURE OF miRNA EXPRESSION IN HUMAN BRAIN AND ITS CONTRIBUTION TO BRAIN DISORDERS

Arun Patil*[1], Anandita Rajpurohit*[1], Yong Kyu Lee[1], Carly Montoya[1], Carrie Wright[1], Geo Pertea[1], Thomas M Hyde[1,2,4], Joel E Kleinman[1,4], Joo Heon Shin**[1,2], Daniel R Weinberger**[1,2,3,4], Taeyoung Hwang**[1,2,3]

[1]Lieber Institute for Brain Development, Translational Neuroscience, Baltimore, MD, [2]Johns Hopkins University School of Medicine, Department of Neurology, Baltimore, MD, [3]Johns Hopkins University School of Medicine, Solomon H. Snyder Department of Neuroscience, Baltimore, MD, [4]Johns Hopkins University School of Medicine, Department of Psychiatry and Behavioral Sciences, Baltimore, MD

MicroRNAs (miRNAs) regulate a majority of protein-coding genes, yet the genetic determinants of miRNA expression in the human brain remain poorly defined. We profiled miRNA expression in 995 human brain tissues spanning four regions (DLPFC, mPFC, hippocampus, and caudate) and two ancestries (African and European), including neurotypical controls and individuals with schizophrenia, major depressive disorder, and bipolar disorder.

Brain regional variation was the dominant source of miRNA expression differences, exceeding the effects attributable to ancestry or psychiatric diagnosis. miRNA expression quantitative trait locus (miR-eQTL) analysis identified numerous cis-regulatory variants in a region- and ancestry-dependent manner, including 150 miR-eQTLs detected exclusively in African ancestry samples in DLPFC, largely driven by ancestry-specific minor allele frequencies. Across tissues and ancestries, miRNAs exhibited cis-heritability comparable to that of protein-coding genes. Multivariate adaptive shrinkage (MASH) analysis revealed substantial cross-region sharing of miR-eQTL effects, though sharing was weaker than that observed for mRNA eQTLs. Intragenic miRNAs showed positive co-regulation with host transcripts but limited evidence for feed-forward targeting of host genes.

miR-eQTL variants were significantly enriched in oligodendrocyte-active enhancers and OLIG2 binding regions, implicating lineage-specific regulatory architecture. Integration with 165 genome-wide association studies using TWAS and Mendelian randomization identified 15 miRNAs with putative causal effects on psychiatric and neurodegenerative traits, including bipolar disorder and Alzheimer's disease.

Together, these results define the genetic architecture of miRNA regulation in the human brain and implicate miRNAs as mediators of genetic risk for complex neuropsychiatric disorders.

* A.P. and A.R. contributed equally to this work. ** J.H.S., D.R.W., and T.H. jointly supervised the study.

LEVERAGING MULTI-ANCESTRY GENE EXPRESSION MODELS AND TWAS TO DISCOVER GENES AND PATHWAYS IN ASIAN CARDIOMETABOLIC DISEASES

Pritesh R Jain*[1], Konstanze Tan*[1], Marie Loh[1,2], John Chambers[1,2,3]

[1]LKC Medicine, Nanyang Technological University, Population and Global Health, Singapore, Singapore, [2]Imperial College London, Department of Epidemiology and Biostatistics, London, United Kingdom, [3]Precision Health Research, PRECISE, Singapore, Singapore

Translating genetic findings into biological insights is hindered by a lack of large-scale functional genomic data for non-European populations. While Transcriptome wide association studies (TWAS) bridge the gap between genetic variation and gene expression (eQTL), existing tools rely heavily on European datasets. To address this, we developed a trans-ancestry polygenic risk score (PRS) for gene expression using summary-level data to better predict transcriptomic variation across diverse ancestries.

We use the largest available eQTL studies from different continental populations (EUR [N ~32,000], AFR [N~684], AMR [N~320], Asian [N ~1225]) and generate trans-ancestry weights using the PRSCSx tool. We validate the performance of this multi-ancestry weights against PRSCS derived weights in the 1000 genomes MAGE dataset (N=731). We then use the weights to perform TWAS for six key cardio-metabolic traits in the HELIOS dataset, a multi-ethnic Asian cohort with genomic and clinical data.

The Trans-ancestry PRS outperformed ancestry-specific PRS in African (4.51% vs 4.07%, $P = 7.9x10^{-4}$), East Asian (6.97% vs 3.11%, $P = 5.9x10^{-158}$),South Asian (7.66% versus 3.24%, $P = 1.7x10^{-185}$), and Admixed Americans (8.43% vs 6.23%, $P = 7.9x10^{-4}$) populations, while marginally lower than ancestry-specific PRS in Europeans (7.37% versus 7.51%, P = 0.43). Contrary to expectation, European PRS outperformed ancestry-matched PRS in East and South Asian populations and equivalent to the trans-ancestry PRS. TWAS analyses of cardio-metabolic traits identified 51 gene-trait associations across four cardiometabolic traits ($P< 1.03x10^{-6}$). This includes associations of *GGPS1* with BMI, *SLC19A1* with triglyceride levels, and 22 gene associated with HDL Cholesterol near the *LDLR* loci.

At current sample sizes of publicly available eQTL summary statistics, The trans-ancestry gene expression weights are equivalent to the European specific weights and perform better than ancestry specific weights for non-European populations. This emphasizes the inadequate ancestral representation in eQTL studies, and limits PRS evaluation and TWAS application. Future efforts should focus on increasing the availability of eQTL data for non-European populations, improving PRS models by including more variants and population-specific LD. This will enable optimized TWAS studies to identify novel genes and pathways linked to complex traits in Asian and other non-European Populations.

EVIDENCE FOR REGULATORY GENE EXPRESSION VARIABILITY IN HUMAN CELL TYPES

Brendan Jamison, Alexander Chen, Kenneth Barr, Yoav Gilad

University of Chicago, Section of Genetic Medicine, Chicago, IL

By measuring gene expression across diverse biological contexts, gene regulatory studies have identified transcriptional programs that underlie development, cell identity, and cellular function. These studies have largely focused on mean expression levels measured across populations of cells. Transcriptional variability between individual cells, which is often substantial, is typically overlooked or treated as uninformative noise. If, however, variability itself encodes regulatory information, it could provide insight into penetrance, regulatory stability, and phenotypic robustness.
If gene expression variability is a regulated trait, dispersion should vary systematically across cell types. To test this prediction, we generated single-cell RNA-seq data from human iPSC-derived heterogeneous differentiating cultures of cardiovascular cell types (cardiac HDCs). We sequenced 75,000 cells from each of three cardiac HDC lines to a depth of 100,000 reads per cell. Gene expression variability was quantified using mean-corrected dispersion estimates obtained with Memento. Dispersion was strongly correlated with cell identity ($r = 0.9–0.99$ across six cell types), and we identified genes that show differences in dispersion between cell type. Genes associated with low regulatory dispersion in specific cell types were enriched for functions characteristic of those cell types. This observation motivated an analysis of genes with low dispersion across all cell types, under the expectation that tightly regulated genes might exhibit distinct functional and regulatory properties. Consistent with this hypothesis, such genes show high gene–gene connectivity and are enriched for housekeeping functions. Relative to highly dispersed genes, genes with low dispersion have shorter enhancers, fewer transcription start sites, and are depleted for TATA promoter motifs, consistent with simpler regulatory architectures. In addition, lowly dispersed genes are depleted for eQTLs, suggesting reduced tolerance for genetic perturbation of their regulation.
Together, the cell-type specificity of dispersion and the shared properties of lowly dispersed genes indicate that gene expression variability is a regulated feature of transcriptional control. In this framework, low variability reflects high regulatory fidelity, defined as the precision with which cells achieve and maintain an optimal expression level. Elucidating the genetic and mechanistic basis of regulatory variability will be important for understanding how expression robustness is maintained and may ultimately suggest new strategies for modulating the expression of disease-associated genes.

BENCHMARKING POLYGENIC RISK SCORE ALGORITHMS FOR CROSS-ETHNIC TRANSFERABILITY

Peilin Jia[1,2], Shuhua Li[1,2,3]

[1]China National Center for Bioinformation, Application and Development Department, Beijing, China, [2]Beijing Institute of Genomics, Chinese Academy of Sciences, Application and Development Department, Beijing, China, [3]University of Chinese Academy of Sciences, School of Future Technology, Beijing, China

Polygenic risk scores (PRS) are pivotal for elucidating the genetic architecture of complex traits and advancing personalized medicine. However, the clinical utility of PRS is hindered by limited cross-ancestry transferability, primarily due to the underrepresentation of non-European cohorts in discovery samples. Identifying the determinants of predictive accuracy across populations is essential to mitigate disparities and develop equitable genomic tools. To address this translational gap, we systematically evaluated 12 state-of-the-art PRS methods through extensive simulations and real-world datasets from UK Biobank. Beyond performance evaluation, we dissected several important factors that influence cross-ancestry transferability. We demonstrated that population differentiation (Fst) and genetic architecture scale factor (α) are key drivers of transferability decay. These results provide a quantitative framework for optimizing PRS algorithms, underscoring that incorporating ancestry-aware genetic architecture is essential for equitable performance across diverse populations. Our findings offer critical guidance for building robust, generalizable genomic tools and advancing the equitable implementation of precision medicine worldwide.

THE GENETIC, EPIGENETIC AND TRANSCRIPTIONAL LANDSCAPES OF TRANSPOSABLE ELEMENTS IN HUMAN PANGENOMES

Juan Jiang, Xiaoyu Zhuo, Ronghan Li, Juan Macias-Velasco, Ting Wang

Washington University School of Medicine, Department of genetics, Saint Louis, MO

Transposable elements (TEs) comprise nearly half of the human genome and remain an important source of genetic and regulatory variation. However, reference bias and short-read genotyping have limited population-scale resolution of TE diversity, particularly for polymorphic mobile element insertions (MEIs). Leveraging a multi-ancestry human pangenome of 232 individuals with haplotype-resolved assemblies, matched DNA methylation, and transcriptomic profiles, we generate a comprehensive and high-resolution map of human TE variation spanning both mobile and non-mobile TE polymorphisms.

We systematically characterize the genomic distribution, allele-frequency spectra, geographic differentiation, and predicted deleteriousness of TE variants, revealing an evolutionary continuum from newly inserted, highly polymorphic elements to long-fixed TEs with reduced population variation and accumulated internal sequence divergence. This genetic trajectory is paralleled by an epigenetic pattern: recently inserted TEs are predominantly hypermethylated, whereas older fixed elements progressively lose methylation over time. Polymorphic insertions additionally influence methylation levels in adjacent genomic regions, highlighting their contribution to local epigenetic heterogeneity.

At the transcriptomic level, approximately 50% of human transcripts contain TE-derived sequence. Among these, ~70% involve TE sequences contributing structural features such as transcriptional start sites, internal exons, or termination signals, and ~75% retain predicted coding potential. TE polymorphism further expands transcript diversity by altering isoform architecture, including the emergence of novel TE-containing isoforms and modification of pre-existing transcript structures. Notably, MEI-associated transcripts are predominantly chimeric, incorporating TE sequence into host gene architecture rather than forming autonomous canonical TE transcripts.

To evaluate regulatory impact, we map TE-associated expression effects at gene and transcript levels and partition underlying mechanisms. While disruption or creation of classical cis-regulatory elements represents one component, we find that transcript-structural remodeling and RNA-level processes, account for a substantial fraction of TE regulatory influence. Finally, we identify TE variants overlapping disease-relevant loci and detect population-differentiated elements consistent with contributions to local adaptation. Together, our study links high-resolution characterization of TE variation to its evolutionary dynamics, mechanistically diverse regulatory effects, and phenotypic consequences in human populations.

EXPLAINABLE MODELING OF LONG-RANGE REGULATORY INTERACTIONS FROM SEQUENCE

Junru Jin, Ruoyu Wang, Jian Zhou

University of Chicago, Department of Medicine, Chicago, IL

Long-range regulatory interactions are central to gene regulation and essential for interpreting noncoding genetic variants. Although sequence-based deep learning models have substantially advanced regulatory prediction, long-range dependencies remain challenging to capture and are often treated as implicit byproducts of end-to-end optimization, limiting mechanistic insight. We introduce a mechanism-driven and inherently explainable framework that explicitly models regulatory interactions directly from DNA sequence. By decomposing the modeling process into functional modules and explicitly parameterizing the interaction component, our architecture enables built-in explainability rather than relying on post-hoc interpretation methods. Our model achieves competitive performance on enhancer–target gene prediction benchmarks while generating biologically meaningful interaction maps. Case studies further demonstrate its ability to recover plausible regulatory relationships. Together, this work establishes a new paradigm for explainable sequence-based modeling of long-range gene regulation, bridging predictive accuracy and mechanistic understanding.

DECODING THE SEQUENCE BASIS OF POL II ELONGATION WITH DEEP LEARNING

Yijie Kang[1,2], Xin Zeng[1], Rebecca Hasset[1], Adam Siepel[1], Peter K Koo[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Stony Brook University, Graduate Program in Genetics, Stony Brook, NY

Understanding how DNA sequence governs RNA polymerase II (Pol II) elongation remains a central challenge in gene regulation. While initiation mechanisms are well-characterized, the sequence determinants of elongation dynamics across gene bodies are poorly understood. Precision Run-On Sequencing (PRO-seq) provides high-resolution snapshots of actively transcribing Pol II, yet prior studies have often focused on specific and predominant features—such as promoter-proximal pausing index—overlooking broader regulatory landscape across gene bodies. Unlike transcription initiation rates, elongation rates have proven difficult to connect to specific motifs except the pause button. Traditional statistical models have revealed linear correlations between sequence features and elongation-associated properties, but they often fail to capture the complex interactions and higher-order dependencies that also shape transcriptional regulation—limiting their ability to uncover mechanistic insight. Here we present an interpretable deep learning framework that predicts base-resolution, continuous elongation profiles from DNA sequence using time-course PRO-seq data. Our approach builds on a U-Net-based architecture with biologically grounded modules that capture multi-scale sequence features and their higher-order dependencies. The modular design of our framework enables the integration of epigenomic features alongside primary DNA sequence, facilitating interpretable analysis of the multifactorial determinants governing Pol II elongation dynamics. Importantly, the model treats time as a continuous input, enabling accurate predictions at arbitrary time points. Through post hoc model interpretation, we reveal the sequence basis of key motifs that drive transcriptional dynamics, including temporal patterns in pausing and elongation rate variation. This framework offers a powerful new approach for dissecting how DNA sequence and genetic variation shape Pol II elongation, including weak, dispersed signals across many bases that have been elusive for standard methods, providing insight into disease mechanisms and supporting the computational design of synthetic regulatory elements that control transcript output and isoform usage through elongation dynamics.

THE FIRST *MICROCEBUS* PANGENOME FOR EVOLUTIONARY GENOMICS RESEARCH

Hannah P Kania, J. Carolina Segami, Anne D Yoder

Duke University, Biology, Durham, NC

Pangenomes are powerful tools for identifying the entire genetic diversity within and among species. Their power is especially relevant for revealing rare variants that might have significant phenotypic effects. Here, we will present the first haplotype-resolved pangenome built from multiple mouse lemur species (genus *Microcebus*), including the first ever *Microcebus* Y-chromosome assembled using long-read sequencing and chromatin interaction data from a male *M. griseorufus*. Mouse lemurs have garnered recent attention for their potential to serve as a primate genetic model system, yet even though there are 19 recognized species, the current reference-quality genomic resources are limited to a single species, *M. murinus*. Our pangenome will serve as a powerful foundation for genomic investigations of the clade and beyond. We are leveraging the pangenome to explore genomic patterns of diversification within the speciose *Microcebus* clade with special attention to the distribution and characterization of transposable elements and other structural variants, with particular interest in their potential role as "speciation genes."

DEVELOPMENTAL GTEx DATA PROVIDES INSIGHT INTO GENE REGULATORY DYNAMICS WITH IMPLICATIONS FOR PEDIATRIC DISEASE

Rebecca Keener[1], Mingyuan Li[2], Marielle Bond[3], Winona Oliveros Diez[4], Jose Miguel Ramirez[5], Pau Clavell-Revelles[5], Laura Domenech[6], Kristin Ardlie[6], Deanne Taylor[7], Tuuli Lappalainen[3,4], Marta Mele[5], Alexis Battle[1]

[1]Johns Hopkins University, BME, Baltimore, MD, [2]Johns Hopkins University, CMDB, Baltimore, MD, [3]New York Genome Center, New York, NY, [4]SciLifeLab, KTH Royal Institute of Technology, Solna, Sweden, [5]Barcelona Supercomputing Center, Barcelona, Spain, [6]Broad Institute, Boston, MA, [7]CHOP, BHI, Philadelphia, PA

The developmental Genotype-Tissue Expression (dGTEx) project provides a data resource of pediatric (ages 0-18) multi-tissue gene expression, enabling analysis of gene regulatory dynamics related to childhood disease. The first data release includes 332 RNA-sequencing samples representing 24 organs and whole genome sequencing of 37 individuals. While this sample size prohibits quantitative trait locus (QTL) discovery, we found higher replication of GTEx expression QTLs (eQTLs) and allele specific expression (ASE) in matched tissues compared to unmatched tissues. In most tissues ASE replication was higher in adolescents than infants and we identified genes with changes in genetic effects across age.

Differential gene expression analysis identified 17,201 differentially expressed genes (DEGs) of which 2,897 DEGs were linked to childhood or developmental disease in Open Targets (Open Targets global score > 0.2). We observed an increasing proportion of DEGs linked to childhood diseases as the number of organs displaying differential expression increased. Differential splicing analysis displayed a similar enrichment of childhood disease connections. DEGs shared across six or more organs were overrepresented for diseases largely impacting connective tissue.

Drug development for pediatric disease suffers from the absence of a healthy pediatric reference. We examined expression patterns across age and tissue for *CD22*, the gene target for an FDA-approved CAR-T immunotherapy. Across GTEx and dGTEx, we identified 26 CD22 transcriptional isoforms that lack the genomic sequence encoding the drug interaction region of the protein. In GTEx these transcriptional isoforms are rare (maximum transcript ratio of 0.02). However, dGTEx immune organs highly express these transcripts (transcript ratio 0.32-0.4). In addition, other tissues had higher transcript ratios and displayed shifts in the transcript ratio with age. These patterns may underlie common side effects of CD22 CAR-T treatment.

In summary, the dGTEx dataset is an important resource with broad applications that captures dynamic expression patterns across organs during childhood development with novel insights into pediatric disease and complex traits.

SHORT- AND LONG-READ SINGLE-CELL RNA SEQUENCING REVEALS TRANSCRIPTOMIC AND ISOFORM DIVERSITY IN NATURAL INFECTIONS OF NEGLECTED HUMAN MALARIA PARASITES

Seri Kitada[1,2], Sunil Kumar Dogga[1], Jesse Rop[1,2], Yomna Gohar[1], Antoine Dara[3], Dinkorma Ouologuem[3], Sekou Sissoko[3], Arthur Talman[4], Abdoulaye Djimdé[3], Mara Lawniczak[1]

[1]Wellcome Sanger Institute, Tree of Life, Hinxton, United Kingdom, [2]University of Cambridge, Cambridge, United Kingdom, [3]Université des Sciences, des Techniques et des Technologies de Bamako, Malaria Research and Training Center, Bamako, Mali, [4]MiVEGEC, Evolution of Vectorial Systems, Montpellier, France

Malaria, caused by *Plasmodium* parasites, remains a major global health burden. While *P. falciparum* causes most malaria-related mortality, the burden of *P. malariae* and *P. ovale* is increasingly recognised. Despite their clinical relevance, the transcriptomic architecture underlying their distinct biology and disease phenotypes remains poorly characterised. To address this, we collected blood samples from individuals with natural infections of these species in Mali. Using long-read DNA sequencing with Hi-C scaffolding, we generated the first chromosome-level genome assemblies for both species, revealing extensive divergence across isolates within subtelomeric regions enriched for expanded antigenic multigene families. Using these assemblies and short- and long-read RNA sequencing, we constructed single-cell transcriptomic atlases comprising ~48,000 *P. malariae* and ~46,000 *P. ovale wallikeri* blood-stage parasites, representing 13 and 14 genotypically distinct strains from nine and eight participants, respectively. Across both species, we identified gene expression programmes underlying asexual development, including processes such as host cell invasion and remodelling, alongside sexual differentiation. Some genes, including those encoding vaccine candidate antigens, showed species-specific expression patterns. For example, *gamete antigen 27/25*, single-copy in *P. falciparum*, is expanded to >20 copies in *P. malariae*, several of which showed expression specific to particular asexual stages, suggesting functional diversification. In *P. ovale wallikeri*, strains within and across participants exhibited varied *pir* gene repertoires, an expanded subtelomeric antigenic family with potential relevance to immune evasion and chronic infection. Long-read analysis further revealed isoform usage differences across asexual, female, and male stages, highlighting regulatory complexity not captured by gene-level analyses alone. Cross-species analysis, including *P. falciparum*, revealed both conserved and divergent orthologue expression, including differences in the lifecycle stages at which ApiAP2 transcription factor family members are expressed. Together, these findings advance our understanding of the lifecycle and antigenic landscapes of these neglected human malaria parasites.

GENE-BY-LIFESTYLE INTERACTIONS CONTRIBUTE TO BLOOD PRESSURE VARIATION IN MULTI-ETHNIC POPULATIONS

Khushi Goda*[1,2], Noah Klimkowski Arango*[1,2], Francesco Tiezzi[1,3], Trudy Mackay[1,2], Fabio Morgante[1,2]

[1]Clemson University, Institute for Human Genetics, Greenwood, SC, [2]Clemson University, Department of Genetics and Biochemistry, Clemson, SC, [3]University of Florence, Department of Agriculture, Food, Environment and Forestry (DAGRI), Florence, Italy

* These authors contributed equally

Investigating the role of gene-by-environment (GxE) interactions remains a key step towards understanding the genetic architecture of complex human traits. While recent studies have shown that the contribution of GxE to phenotypic variation is non-negligible for many complex traits, this evidence mainly comes from the analysis of white individuals. Here, we hypothesized that GxE explain more variance in multi-ethnicity datasets because diverse populations exhibit greater heterogeneity in genetic background and environmental exposures than single-ethnicity datasets. We tested this hypothesis with three blood pressure (BP) traits – diastolic pressure (DP), systolic pressure (SP), and pulse pressure (PP) – and 23 lifestyle variables in a multi-ethnic subset of 24,000 individuals from the UK Biobank. Our results showed that, after carefully accounting for population structure, GxE explain 7% of the variance in DP, 4% of the variance in SP, and 3% of the variance in PP. Importantly, these estimates are larger than the estimates obtained using a subset of White British individuals of the same size as the multi-ethnic subset for DP (4%) and SP (3%), and similar for PP. In addition, including GxE in the model accounted for variance that would otherwise be unexplained and did not take variance away from other components in the model. Despite the increase in variance explained, including GxE effects did not improve prediction accuracy in cross-ethnicity or random cross-validation scenarios. We also sought to map individual interactions affecting blood pressure traits using GxE-GWAS. Although no interactions from our analysis achieved genome-wide significance, variant-lifestyle combinations exhibiting evidence for true signal have known connections to BP and other related traits such as cardiovascular and neuronal traits. Overall, our study highlights the importance of genetic diversity and environmental heterogeneity in detecting GxE effects contributing to complex trait architecture.

CONTEXT DEPENDENT EFFECTS OF NON-CODING NEUROPSYCHIATRIC VARIANTS IN HUMAN STEM CELL DERIVED NEURONS

Justin Koesterich[1,2,3,4], Sarah E Williams[5,6,7,8], Ratchell Sadovnik[9], Linda L Boshans[5], Kayla Townsley[4,10], Anat Kreimer[1,2,3,4], Kristen Brennand[11,12], Nan Yang[5,6,7,8]

[1]Rutgers University, Cell and Developmental Biology, Piscataway, NJ, [2]Rutgers University, Center for Advanced Biotechnology and Medicine, Piscataway, NJ, [3]Robert Wood Johnson Medical School, Biochemistry and Molecular Biology, Piscataway, NJ, [4]Rutgers University, Graduate Programs in Molecular Biosciences, Piscataway, NJ, [5]Friedman Brain Institute, Nash Family Department of Neuroscience, New York, NY, [6]Icahn School of Medicine at Mount Sinai, Institute of Regenerative Medicine, New York, NY, [7]Icahn School of Medicine at Mount Sinai, Alper Center for Neurodevelopment and Regeneration, New York, NY, [8]Icahn School of Medicine at Mount Sinai, The Graduate School of Biomedical Sciences, New York, NY, [9]Icahn School of Medicine at Mount Sinai, Stem Cell Biology And Regenerative Medicine, New York, NY, [10]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomics, New York, NY, [11]Yale University, Department of Psychiatry, New Haven, CT, [12]Yale University, Department of Genetics, New Haven, CT

Genome-wide association studies (GWAS) have identified ~98.5% of disease associated variants reside in the non-coding region of the DNA. Additionally, thousands of non-coding de novo variants (DNVs) have been identified in Autism spectrum disorder (ASD) probands and siblings. While the functional contribution of these non-coding variants to disease etiology remains unclear, we hypothesize they are disrupting context-dependent transcriptional regulatory mechanisms.

To investigate noncoding neuropsychiatric disorder associated variants, we utilized massively parallel reporter assays (MPRA) across five human stem cell-derived neuronal contexts. Furthermore, we performed activity-by-contact modelling to identify the genes regulated by enhancers containing disruptive variants.

Of the regulatory sequences tested via MPRA, approximately half showed activity in at least one cell-type with the context dependent transcriptional activity significantly influenced by epigenetic factors. 505 variants significantly altered activity across the 5 conditions. Of those, 85% were unique to one cellular state, highlighting the context specificity these variants. Gene-enhancer mapping revealed a subset of variant-containing enhancers predicted to regulate disorder associated gene genes. Furthermore, we identify significant differences in the transcriptional disruptions between proband ASD DNVs and healthy sibling DNVs. As well as significant differences between variants disrupting transcriptional activity within the NPC context versus the mature neuronal contexts.

We find that these variants have context specific effects on transcriptional activity and subsequent effects on the gene expression of disorder associated genes in human neurons. Together, these results emphasize the critical role of cellular context in interpreting the functional impact of non-coding regulatory variants, and their involvement with the associated disorder.

ELECTRONIC GENOME MAPPING FOR HIGH-THROUGHPUT ANALYSIS OF REPEAT EXPANSION DISORDERS

Syndi Koltz[1], Lindsay Schneider[1], Reger Mikaeel[1,2], Dong Zhang[1], Xu Tan[1], Shuk Shukor[3], Mike Kaiser[1], John Thompson[1]

[1]Nabsys 2.0 LLC, Providence, RI, [2]Washington University in St. Louis, Laboratory Genetics and Genomics, St. Louis, MO, [3]Hitachi Hi-Tech America, La Jolla, CA

Background: Many genetic disorders are caused by expansions of simple sequence repeats, including Fragile X syndrome (FXS) and Friedreich's ataxia (FA). FXS, the most common inherited form of intellectual disability, results from a CGG repeat expansion in the *FMR1* gene. FA, the most common early-onset inherited ataxia, results from a GAA repeat expansion in the *FXN* gene. Accurate sizing of these expansions is critical for classification into normal (< 45 repeats for FXS, < 66 repeats for FA), premutation (55–200 repeats for FXS), and full mutation (> 200 repeats for FXS, > 66 repeats for FA). However, conventional Polymerase chain reaction (PCR) and sequencing methods have limited sensitivity and are unreliable for long repeat regions. Electronic genome mapping (EGM) uses solid-state nanodetectors to survey long DNA molecules and construct high-density maps to detect structural variants (SVs), including those caused by repeat expansions and contractions. We evaluated the utility of EGM on the OhmX™ Platform for characterizing repeat expansion disorders such as FXS and FA.

Methods: We assessed EGM's ability to detect *FMR1* CGG repeat expansions and *FXN* GAA repeat expansions using Coriell cell lines. Data were analyzed using two bioinformatics pipelines: Human Chromosome Explorer®, which generates *de novo* whole genome assemblies, and RepX™, which maps individual molecules to specific variants and accurately calls repeat lengths.

Results: A single detector with a single sample loading generated sufficient data to call *FMR1* and *FXN* repeats using RepX. Multiple replicates of multiple samples were concordant with expectations. These findings confirm the accuracy and robustness of EGM for high-throughput characterization of repeat expansions that are challenging to analyze by standard techniques. We hypothesize that based on the success of the pipeline with detecting these repeat expansions, the pipeline should also be able to detect other repeat expansions. We recognize that some repeat expansion disorders, such as FA, result in mosaicism and we will discuss our progress with using OhmX to detect mosaicism. For those repeat expansions whose tagging patterns are not optimal for OhmX detection, the assay can be optimized using CRISPR to create or remove tag sites as needed.

Conclusion: EGM offers integrated workflows for high-resolution analysis of repeat regions in either whole-genome or gene-specific mode. This streamlined approach enables efficient, scalable detection of repeat expansions, supporting research applications in repeat expansion disorders.

CELL-TYPE-SPECIFIC PATTERNS OF SOMATIC MUTATIONS AND CONSEQUENCES FOR TRANSCRIPTIONAL HETEROGENEITY IN BRAIN AGING AND GLIOBLASTOMA

Andrea J Kriz*[1,2], Shulin Mao*[1,3,4], Diane D Shao[1,2], Daniel A Snellings[1,2], Rebecca E Andersen[1,2], Guanlan Dong[1,3], Luis E Guzman-Clavel[4], Hayley Cline[1], Chanthia C Ma[1], August Yue Huang[1,3], Eunjung Alice Lee[1,3], Christopher A Walsh[1,2,3]

[1]Boston Children's Hospital, Harvard Medical School, Division of Genetics and Genomics, Boston, MA, [2]Howard Hughes Medical Institute, Boston Children's Hospital, Harvard Medical School, Boston, MA, [3]Boston Children's Hospital, Manton Center for Orphan Disease Research, Boston, MA, [4]Harvard Medical School, Program in Biological and Biomedical Sciences, Boston, MA

*authors contributed equally

Somatic mutations arise in healthy tissues, cancer and aging. However, their patterns and functional impacts are highly cell type-specific, making them difficult to decipher using existing methods. Open “accessible” chromatin, including regulatory regions such as promoters and enhancers, are among the noncoding regions most prone to consequences of somatic mutation. Here, we developed Duplex-Multiome, incorporating duplex consensus sequencing to accurately discover somatic single-nucleotide variants (sSNV) from the same nucleus simultaneously analyzed for single-nucleus ATAC-seq (snATAC-seq) and RNA-seq (snRNA-seq). By strand-tagging the snATAC-seq libraries, duplex sequencing reduces sequencing error by >10,000-fold, eliminating artifactual mutational signatures. When applied to cell lines mixed in a 98:2 ratio, Duplex-Multiome identified sSNVs present in 2% of cells with 92% precision and accurately captured known sSNV mutational spectra, while revealing unexpected subclonal lineages. Duplex-Multiome of > 51,400 nuclei from non-malignant postmortem human brain captured sSNV burdens and spectra closely comparable to those determined by single-cell whole-genome sequencing (scWGS), and provided sSNV burdens and spectra in accessible chromatin regions for neuronal and subtypes that could not previously be studied. Whereas 2 sSNVs correlated with altered gene expression in non-malignant brain, in >17,400 nuclei from post-mortem glioblastoma brain, Duplex-Multiome identified 14 sSNVs with correlated changes in expression of nearby genes, including upregulation of PDGFRA and EGFR. Duplex-Multiome also discovered unexpected subsets of tumor cells with up to 4-fold higher mutational burdens than other tumor cells, correlating with increased mitotic activity or upregulation of synapse organization pathway genes, depending on the tumor. By directly bridging somatic mosaicism to phenotypic readouts in cell type-specific fashion, Duplex-Multiome reveals diverse burdens and functions of somatic mutations in normal brain and complex tumors like GBM.

POPULATION-FREE POLYGENIC RISK PREDICTION FROM ANCESTRAL RECOMBINATION GRAPHS

Nurdan Kuru[1], Shareef Khalid[1,2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Stony Brook University, Genetics, Stony Brook, NY

Polygenic Risk Scores (PRS) are widely used in precision medicine to predict disease risk from genetic variation, but most existing methods require individuals to be grouped into broad continental ancestries prior to analysis. This reliance on population labels limits accuracy and transferability, particularly in diverse and admixed populations.

We introduce a population-free approach to polygenic risk prediction that models individual genomes using Ancestral Recombination Graphs (ARGs), which encode fine-scale genealogical relationships along the genome. Rather than relying on predefined population labels, our method leverages local genealogies to directly capture genetic relatedness at the individual level.

Our framework applies phylogenetic linear mixed models with ARG-derived covariance kernels. Each local genealogy is converted into a tree-based kernel, and the contributions of multiple such kernels are jointly estimated within a penalized multi-kernel linear mixed model. This formulation allows genetic effects to be shared across ancestries while capturing individual-specific deviations through genealogical structure.

We evaluated our approach against PRS methods based on population-level trees, genome-wide averaged ARGs, and non-tree-based models. In simulations, the ARG-based framework improved predictive accuracy by up to 50 percent, with higher R2 and lower prediction error, with the largest gains observed in African and admixed populations. Analyses of height, LDL cholesterol, and type 2 diabetes in the All of Us cohort showed consistent improvements over existing methods.

By incorporating evolutionary history through ARGs, our framework provides a biologically grounded alternative to traditional PRS approaches, with the potential to improve genetic risk prediction across diverse populations and reduce health disparities.

GENE-AGE AND GENE-SEX INTERACTION PATTERNS ACROSS QUANTITATIVE PHENOTYPES IN UK BIOBANK

Yanina Kuzminich[1], Sri Gouri Rajaram[1], Sylvia Dai[1,2], Hakhamanesh Mostafavi[1,3]

[1]New York University Grossman School of Medicine, Center for Human Genetics and Genomics, New York, NY, [2]New York University, Division of Science, Abu Dhabi, United Arab Emirates, [3]New York University Grossman School of Medicine, Department of Population Health, New York, NY

Most complex traits change substantially with age and often in sex-specific manners. How these changes interact with genetic effects remains debated. Here, we systematically analyze gene-by-age and gene-by-sex interactions across a range of quantitative traits in the UK Biobank, focusing on GWAS loci while accounting for medication use, which can otherwise induce spurious interactions.

We observe pervasive interactions largely consistent with an amplification pattern – previously noted for gene-by-sex effects – where most loci show proportional changes in effect size with age or sex. For several traits, interactions are dominated by gene-by-sex amplification (e.g., systolic blood pressure, urate levels, and red blood cell distribution), whereas others show predominantly gene-by-age amplification (e.g., alkaline phosphatase).

Notably, LDL levels exhibit strong age-dependent amplification in a sex-specific manner: at younger ages, genetic effects are larger in men, and this difference diminishes with age. For example, the effect of rs58542926 decreases from about 0.2 at ages below 45 to about 0.1 at ages above 65 in men, with almost no change in women.

We further find that these interaction patterns are sensitive to phenotype scale, consistent with recent reports. However, this sensitivity is trait-specific, and no single transformation removes both age- and sex-dependent interactions, suggesting that changing phenotype scale may attenuate interactions with one exposure while accentuating those with another.

Finally, a subset of loci displays interaction patterns that deviate from the global amplification trend and are robust to phenotype scaling. For example, for urate levels, while most trait-increasing alleles have larger effects in women (e.g., rs113533538; effect about 29 vs. 15), a number of loci show the opposite pattern, with larger effects in men (e.g., rs75246752; effect about 16 vs. 9).

We conceptualize these patterns as reflecting two partially distinct phenomena: (i) sex and age entering the path from genotype to phenotype at different stages, and (ii) scale-dependent artifacts. Outlier loci may reflect pleiotropy with traits exhibiting different amplification dynamics or interactions arising at multiple biological steps. Altogether, these results sharpen the characterization of gene-by-age and gene-by-sex effects and, more broadly, offer novel ways to think about genetic interactions in complex traits.

A SCALABLE FRAMEWORK FOR EVIDENCE INTEGRATION AND GENE PRIORITIZATION IN POST-GWAS STUDIES

Fei Liu[1], Yuan Cao[1], Junbin Gao[1], Yao Ma[2], Boxiang Liu[1,3]

[1]National University of Singapore, Department of Pharmacy and Pharmaceutical Sciences, Faculty of Science, Singapore, [2]Xi'an Jiaotong University Second Affiliated Hospital, Department of Cardiology, Xi'an, China, [3]National University of Singapore, Department of Biomedical Informatics, Yong Loo Lin School of Medicine, Singapore

Genome-wide association studies (GWAS) have identified thousands of loci associated with complex traits, yet translating these signals into mechanistically grounded gene–disease relationships remains challenging, particularly for variants with modest effects or located in non-coding regions. Molecular quantitative trait loci (molQTLs) provide critical links between genetic variation and downstream molecular phenotypes and are widely used for gene implication. However, gene prioritization remains inconsistent across loci due to fragmented statistical evidence, sensitivity to analytical parameters, and the lack of a unified framework for integrating heterogeneous molecular and functional data, limiting reproducibility and interpretability in post-GWAS analyses.

Here, we present an upgraded version of LocusCompare2, a scalable and structured framework for post-GWAS gene prioritization. Beyond aggregating multiple methods, the framework formalizes evidence integration as a cross-method and cross-layer consistency problem, enabling systematic evaluation of concordant signals across heterogeneous data sources. The core platform incorporates six widely used gene implication methods and hundreds of QTL datasets, improving robustness to method choice and parameter variation through cross-method validation and flexible window settings. Building upon this foundation, the framework integrates multi-layer evidence across three dimensions: (i) molecular QTLs (eQTL, pQTL, sQTL, and caQTL), (ii) regulatory annotations, and (iii) functional perturbation evidence. These evidence types are harmonized to prioritize signals that are consistently supported across methods and biological layers.

Importantly, the framework explicitly models SNP-to-gene and gene-to-disease relationships, providing interpretable outputs that link genetic variants to candidate genes through multi-evidence support. This integrated design improves the stability and reproducibility of gene prioritization across loci and analytical settings, while maintaining transparent evidence structures that facilitate hypothesis generation and experimental follow-up. Together, this work provides a scalable and interpretable framework for transforming GWAS signals into coherent, multi-layered, and experimentally testable biological insights.

COLOCALIZATION OF TYPE 1 DIABETES RISK WITH GENE EXPRESSION REVEALS SEX-SPECIFIC GENE REGULATION

Benedict A Lenhart, Dominika Michalek, Wei-Min Chen, Stephen Rich, Suna Onengut-Gumuscu

University of Virginia, Department of Genome Sciences, Charlottesville, VA

Introduction and Objective: Type 1 Diabetes (T1D) is an autoimmune disease with heterogeneity in disease progression influenced by genetic and environmental factors. It is unclear what proportion of this variation in risk differs by sex, as well as the identity of gene regulatory pathways that contribute to initiation and progression to clinical disease. We provide initial data on the contribution of known T1D risk loci and effectors of gene regulation (eQTLs) through colocalization to identify distinct and common regions across the human genome.

Methods: We analyzed GWAS summary statistics from 61,427 individuals for T1D risk (Robertson et al., 2021), and performed colocalization with multiple eQTL datasets, including sex-stratified PBMC eQTLs (594 male and 671 female individuals). The genome was interrogated using coloc (v5.2.3) to identify variants/genes associated with both T1D risk and changes in gene expression, both shared and sex-specific.

Results: We identify several variants that are significantly enriched and shared across sex with respect to T1D risk and whole blood eQTLs. rs773653 in RPS26 is strongly colocalized across sex (male GWAS: $p = 4.51x10\text{-}5$, $\beta = -0.095$; female GWAS: $p = 3.48x10\text{-}5$, $\beta = -0.095$). rs7731626 in IL6ST has differential effect by sex (male GWAS: $p = 9.10x10\text{-}5$, $\beta = -0.09$), as well as rs1893592 in UBASH3A (female GWAS: $p = 4.09X10\text{-}12$, $\beta = -0.17$).

Conclusion: Future work will evaluate how genetic variation at these loci influences regulatory programs across different stages of T1D development. This work demonstrates sex-dependent patterns of gene regulation in T1D associated risk loci and offers new insight into diverse regulatory pathways that shape disease risk .

INVESTIGATING THE RELATIONSHIP BETWEEN RUNS OF HOMOZYGOSITY AND HEIGHT IN ANCIENT EURASIA.

Ana V Leon-Apodaca[1], George H Perry[2], Zachary A Szpiech[1]

[1]Pennsylvania State University, Department of Biology, University Park, PA, [2]Pennsylvania State University, Department of Anthropology, University Park, PA, [3]Pennsylvania State University, Department of Biology, University Park, PA

Demographic processes such as consanguinity, bottlenecks, and natural selection increase the likelihood of haplotypes being inherited identical-by-descent (IBD) from a recent common ancestor. When these identical haplotypes are inherited through both the maternal and paternal lineages, they manifest as runs of homozygosity (ROH). While ROH themselves are not directly heritable in the same way that genetic variants are, research has shown that long ROH harbors more deleterious variants, raising the question of whether elevated levels of homozygosity in the genome may increase the risk of inheriting deleterious recessive variants, and thus, increasing the risk for disease or presenting a trait. Research conducted on present-day human populations of European descent have found an association between genome-wide homozygosity and height, with more homozygosity resulting in lower height. Height is a highly heritable and polygenic trait influenced by both genetic and environmental factors. Archaeological evidence combined with genomic data spanning ~38,000 years have shown there was a marked reduction in human height that occurred during the Neolithic, as humans transitioned to agriculture, followed by an increase during the subsequent post-Neolithic periods of agricultural intensification. While the vast majority of genomic studies have been focused on understanding the relationship between genotype and phenotype at a single point in time, very few studies have looked at how the relationship between genotype and phenotype has changed over time. In this work, we investigate whether changes in ROH patterns over time are associated with changes in human height.

ISOFORM GAZER: AN INTERACTIVE WEBTOOL TO VISUALIZE ISOFORM DIVERSITY

Julia T Lewandowski[1], Megan D Schertzer[1,2], Keren Isaev[1,3], Stella H Park[1], David A Knowles[1,3,4]

[1]New York Genome Center, New York, NY, [2]University of Virginia, Molecular Physiology and Biological Physics, Charlottesville, VA, [3]Columbia University, Systems Biology, New York, NY, [4]Columbia University, Computer Science, New York, NY

Alternative RNA isoforms enable a single gene to produce distinct proteins. This diversity is fundamental to cellular function, underlies tissue-specific expression patterns, and contributes to disease phenotypes. RNA-sequencing technologies–including bulk and single cell, short and long read–reveal hundreds of thousands of novel junctions and isoforms across diverse cell types and disease states. As a result, a fundamental question persists: which isoforms are biologically meaningful? Reproducible detection across datasets and sequencing platforms increases confidence in biological relevance, but it remains challenging to integrate transcriptomic data across platforms and studies. Sequencing technologies provide complementary data: long-read sequencing (LRS) provides high-confidence transcript structures, short-read (SRS) offers quantitative junction-level resolution, and single-cell (scRNA-seq) outputs high-resolution cell-type specificity. Existing databases such as VastDB, MAJIQlopedia, and GTEx assemble transcript structures from SRS data and lack support for handling novel transcripts. These resources additionally do not support interactive filtering or figure customization.

To address these limitations, we developed Isoform Gazer, a web application that integrates multi-species transcriptomic data across sequencing technologies to provide a unified view of isoform diversity. It combines ENCODE4 PacBio LRS data (199,406 and 170,977 human and mouse transcripts in 55 and 9 cell types, respectively) with pseudobulked SRS data from Tabula Sapiens 2.0, Tabula Muris Senis, and Allen Brain Cell Atlas (123,383 and 89,831 human and mouse junctions each in 50 cell types). Isoform Gazer features three integrated panels: (1) a genome browser-style structure plot, which displays the hierarchical relationship between transcripts and junctions, integrating the LRS and SRS data (2) two expression clustergrams showing transcript abundance and junction percent spliced-in (PSI) values at both tissue and organ-level resolution, and (3) two summary tables capable of bidirectional filtering between transcript and junction information for all visualizations. To enable the search of novel isoforms, the webtool contains a function to convert GTF exon coordinates into unique, reproducible hash IDs. Isoform Gazer thus serves as a comprehensive resource for interactive exploration of isoform diversity across annotated and novel isoforms, sequencing technologies, species, and cell types.

NEURON-ASTROCYTE INTERACTIONS REPROGRAM THE EPIGENOME, GENE REGULATORY NETWORKS, AND CELLULAR FUNCTIONS THROUGH DISCRETE TRANSCRIPTIONAL AND EPIGENETIC EVENTS

Boxun Li[1,2], Kevin T Hagy[1,2], Alexias Safi[2,3], Michael A Beer[4], Alejandro Barrera[2], Sara Geraghty[1,2], Ruhi Rai[2], Alyssa N Pederson[1,2], Samuel J Reisman[2,5], Patrick F Sullivan[6], Cagla Eroglu[5,7,8], Gregory E Crawford*[2,3], Charles A Gersbach*[1,2,5]

[1]Duke University, Dept. of Biomedical Engineering, Durham, NC, [2]Duke University, Center for Advanced Genomic Technologies, Durham, NC, [3]Duke University Medical Center, Dept. of Pediatrics, Durham, NC, [4]Johns Hopkins University, Depts. of Biomedical Engineering and Genetic Medicine, Baltimore, MD, [5]Duke University Medical Center, Dept. of Cell Biology, Durham, NC, [6]University of North Carolina at Chapel Hill, Depts. of Genetics and Psychiatry, Chapel Hill, NC, [7]Duke University, Dept. of Psychology and Neuroscience and Howard Hughes Medical Institute, Durham, NC, [8]Duke University Medical Center, Dept. of Neurobiology, Durham, NC

*co-corresponding authors

Heterotypic cell-cell interactions are critical to governing cellular physiology, disease progression, and responses to the environment. For example, neurons and astrocytes engage in intricate interactions that are essential for brain development and function. However, the transformation of these extracellular signals into epigenomic regulation that governs cell function is poorly understood. Here, we report that weeks of neuron-astrocyte co-culture reprograms gene expression and chromatin landscape, affecting thousands of genes and open chromatin regions, including many transcription factors (TFs). These genes and regions are enriched for genes implicated in neuronal processes, schizophrenia, and autosomal dominant Alzheimer's Disease. Through CRISPR epigenetic repression and activation screens, we recapitulated hundreds of astrocyte-induced transcriptional and epigenetic events in mono-cultured neurons at both promoters and distal regulatory elements (REs) of TF genes. We discovered functional REs for ~50 astrocyte-responsive TF genes, providing a map of gene regulatory network control. Astrocyte-responsive TF genes fall into groups that exert independent or counter-balancing transcriptional effects, highlighting the complex coordination of the neuronal response to astrocytes. Functional effects of specific TFs, including POU3F2 and TFAP2E, on neurite morphology and neuronal electrophysiology are consistent with transcriptional effects, demonstrating the capacity of direct epigenetic control to mimic heterotypic cellular signals. This work illuminates the regulation of neurodevelopment- and disease-relevant genes by neuron-astrocyte interactions, and provides a blueprint for applying functional genomics to uncover the links between cell microenvironment and epigenomic programming.

GENETIC EPISTASIS OF PLASMA PROTEOME AND ITS IMPACT ON COMPLEX TRAITS

Jinghui Li[1], Xuanyao Liu[1,2]

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

Detecting and interpreting epistatic effects in the human genome is essential for a complete understanding of the genetic architecture of gene regulation and complex traits. Progress in this area, however, has been hindered by limited discovery power and an abundance of false positives. Here, leveraging large-scale plasma proteomic data from the UK Biobank Pharma Proteomics Project (UKB-PPP), we systematically map high-confidence cis-by-trans epistatic effects regulating protein expression. To reduce the multiple-testing burden, we focus on protein-level rather than variant-level interaction effects, and we employ a sandwich variance estimator to control false positives, as validated by extensive simulations. Using this framework, we identify seven significant cis-by-trans epistasis signals (FDR < 0.05), five of which involve ABO as the trans regulator and are immune-related proteins (MBL2, CD209, FCGR2B, NPTX1, and ICAM5). Those ABO-interaction partners exhibit distinct cis-regulatory effects and expression variance across different ABO alleles, consistent with a role for epistasis in expression buffering. We next asked whether these epistatic buffering effects are accompanied by signatures of long-term balancing selection, as the ABO locus itself is a canonical example of long-term balancing selection. Indeed, ABO-interaction partners have a 1.5-fold enrichment for elevated Tajima's D scores ($P < 0.001$), supporting the action of long-term balancing selection. We also identified 1,304 significant SNP-SNP interactions underlying the seven protein pairs. These epistatic SNPs tend to have larger additive effects but showed no strong enrichment in canonical regulatory chromatin states, such as enhancers or promoters. Finally, we demonstrate that despite the presence of large epistatic effects on protein regulation, naively incorporating interaction into protein-wide association study (PWAS) frameworks can lead to spurious protein-trait associations. Although individual epistatic effects can be large, they explain only a small proportion of total protein expression variance and therefore do not improve PWAS performance. Together, our results clarify both the biological relevance and the practical limitations of epistasis in human proteomic regulation.

REACTOME: STRUCTURED PATHWAY REPRESENTATION AND A NEXT-GENERATION PATHWAY BROWSER

Nancy T Li[1], Reactome Consortium[1,2,3,4]

[1]OICR, Computational Biology, Toronto, Canada, [2]NYU, New York, NY, [3]EBI, Hinxton, United Kingdom, [4]OHSU, Portland, OR

Reactome is an open, expert-curated knowledgebase that represents human biological processes as structured reaction networks defined by a formal data model of physical entities, molecular events, and their relationships. Each reaction is supported by primary literature and organized within a hierarchical pathway framework spanning DNA replication and repair, chromatin organization, transcription, cell cycle control, signaling, metabolism, and disease. Human pathways are computationally projected to additional species via orthology-based inference.

A major current development is the Pathway Browser Beta, a redesigned, high-performance interface for scalable visualization and analysis of pathway graphs. Built on a modern web architecture, the Beta supports multi-level navigation between global pathway overviews and detailed reaction diagrams, graph-based rendering of complexes and modifications, and dynamic overlay of user-supplied omics datasets for pathway enrichment and expression mapping. The updated implementation improves responsiveness and modularity while providing structured entity views and integrated analysis outputs.

The Pathway Browser Beta is under active development, and community feedback is essential to guide usability refinement, feature prioritization, and analytical workflows. Reactome provides RESTful APIs and standard data exports for integration with genome-scale analysis pipelines. As a FAIR-compliant, CoreTrustSeal-certified and ELIXIR-recognized resource, Reactome invites the genomics community to evaluate and help shape this next-generation interface.

USING THE HUMAN PANGENOME TO EFFICIENTLY DETECT COMPLEX GENETIC VARIATION AT SCALE

Linda Y Lin[1], Shuangjia Lu[1], Wen-Wei Liao[1], Nathan O Stitziel[2], Ira M Hall[1]

[1]Department of Genetics, Center for Genomic Health, Yale University School of Medicine, New Haven, CT, [2]Center for Cardiovascular Research, Department of Genetics, Washington University School of Medicine, Saint Louis, MO

Comprehensive variant detection will improve large-scale trait association studies that lend insight into the genetic basis of human disease. However, standard variant calling approaches rely on alignment to a linear reference genome that fails to represent human genetic diversity, leading to blind spots for many complex variants. Yet, due to their size and functional potential, complex variants can have large impacts. The Human Pangenome Reference Consortium (HPRC) is generating hundreds of high-quality genome assemblies from diverse global populations, and integrating them into pangenome graphs that represent all forms of genetic variation. A key unsolved problem is how to leverage this more representative pangenome reference to affordably genotype complex variants from short-read whole-genome sequencing (WGS) data at the scale of modern biobanks.

We are developing a pangenome-based method to efficiently detect previously overlooked genetic variation. In parallel work, we demonstrated via expression quantitative trait loci (eQTL) studies that aligning short-read WGS data to pangenome graphs improves trait association power relative to traditional methods. However, current graph aligners are computationally expensive, whereas alignment-free algorithms based on k-mers (sequences of length k) employ exact matches for significantly faster analysis. Our new method combines the rich information encoded in the pangenome graph with the efficiency of k-mer algorithms to enable more comprehensive and scalable complex variant detection. As proof-of-concept, a simple implementation achieves ≥ 0.8 correlation in genotyping with the best existing pangenomic approach at 92% of putative functional variants missed by standard methods (n = 9,962), with orders of magnitude faster runtime. In a pilot eQTL study, our method identifies just as many significant genes, with 85% overlap (n = 7,091). Our method therefore shows promise for enabling pangenome-based trait association at biobank scale, which will facilitate more comprehensive mapping of disease-relevant traits.

MORE IS MORE: SHARED PHENOTYPES AMONGST SEX CHROMOSOME TRISOMIES HINTS DOSAGE-SENSITIVE EFFECT OF PSEUDOAUTOSOMAL REGIONS IN GENETIC MALES AND FEMALES

Aoxing Liu[1,2,3], Yining Wang[1,3], Wenhan Lu[1,2], Zhili Zheng[1,2], Konrad Karczewski[1,2], Mark J Daly[1,2,3]

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, [3]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Genetic males and females differ biologically in many ways. Nearly all the time, comparisons searching for sex differences are made directly between XX females and XY males, with observed differences, either in behaviors, biological processes, or diseases, attributed to the apparent differences in sex gonadal hormones and genetically, the dosage effect of X chromosome or the testis-determining factor gene residing on the Y chromosome. Unlike autosomal aneuploidy, which usually causes spontaneous abortion, sex chromosome trisomies (SCTs) are relatively tolerated and have a considerable prevalence among general populations such as biobank participants. Introducing these trisomies into the analysis can disrupt the perfect correlation between genetic sex and sex chromosome karyotypes, thereby providing a unique perspective on the phenotypic consequences of sex chromosomes.

We recently published a large biobank-based survey of the prevalence and phenotypic consequences of SCTs (PMID: 40840450). Among the notable findings were a high prevalence of all three SCTs (47,XXY; 47,XYY; 47,XXX), an overwhelming majority of carriers (>85%) lacking a genetic or karyotypic diagnosis, and a surprising similarity of non-reproductive related phenotypes shared across all SCTs, including increased height and elevated risk of asthma and multiple vascular diseases. We further ask whether the origin of the extra sex chromosome would have any impact on the phenotypic consequence; we test it in 47,XXY and see no differences regarding both the parent of origin and the identical or homologous status of the two X chromosomes.

Collectively, these shared phenotypes suggest a common biological mechanism driven by gene dosage of the ~3 Mb pseudoautosomal regions (PAR) shared between the X and Y chromosomes. To test this hypothesis, we perform the first meta-pheWAS for PAR variants in FinnGen, UKB, and All of Us; among the >2000 phenotypes examined, height and asthma emerge with the strongest PAR associations (P<5.0e-8). The strongest PAR association to height is adjacent to SHOX - independent analysis of UKB shows loss-of-function variants in SHOX strongly reduce height (P<1.5e-52), consistent with increased height of SCT carriers. To expand these insights, we perform proteomic profiling of >100 SCT carriers using Olink in FinnGen and UKB-PPP identifies >100 proteins shared across SCTs (P<9.2e-06), with the strongest associations for PAR-encoded proteins IL3RA, CSF2RA, CD99, and XG. Many SCT-shared proteins are correlated and share pQTLs. Notably, a locus at 9q34.2 that increases circulating levels of the PAR-encoded protein IL3RA as well as the vascular integrity-related protein TIE1, is associated with increased risk of multiple SCT-shared vascular diseases. Our findings suggest a broad dosage-sensitive effect of PAR on traits such as height, asthma, and vascular diseases.

MULTI-CONTEXT, -OMICS, -METHOD TRANSCRIPTOME-WIDE ASSOCIATION STUDY RESOURCE ATLAS REVEALS PUTATIVE CAUSAL GENES IN ALZHEIMER'S DISEASE

C Liu[1], FunGen xQTL Consortium[2], G Wang[3], F Morgante[1,4]

[1]Institute for Human Genetics, Clemson University, Greenwood, SC, [2]National Institute on Aging, Division of Neuroscience, Bethesda, MD, [3]Department of Neurology, Columbia University Irving Medical Center, New York, NY, [4]Department of Genetics and Biochemistry, Clemson University, Clemson, SC

Complex traits are highly polygenic and often affected by many genetic variants with small individual effects. Genome-Wide Association Studies (GWAS) have identified large numbers of genetic variants associated with complex traits, most of which are located in non-coding regions. The regulatory mechanisms underlying these signals remain poorly understood. Transcriptome-Wide Association Studies (TWAS) address this gap by leveraging expression quantitative trait loci data to impute expression levels in GWAS samples. Yet, most TWAS efforts have focused on single context (i.e., tissue, cell type) or regulatory modality (i.e., gene expression, protein expression) and rely on a limited set of prediction models such as Elastic-net and Lasso.

Here, we present a comprehensive multi-context, multi-modality, multi-cohort, and multi-model TWAS resource for systematic investigation of the genetic architecture of aging brain related disorders. This TWAS resource integrates tissue-, cell-type-, and brain-region–specific imputation models across regulatory modalities including gene expression, protein abundance, glycoprotein abundance, and splicing regulation, spanning 4 bulk brain tissues, 6 pseudobulk cell types, 1 plasma cell type, and 4 brain regions. Expression reference panels were curated from large postmortem aging brain cohorts harmonized by the FunGen-xQTL Consortium. Imputation models were trained using multiple univariate and multivariate methods to flexibly capture molecular-trait (i.e., unique context-modality combination)-specific and shared regulatory effects.

To show the advantages of our resource, we integrated the imputation models for a subset of 11 molecular traits with a large-scale GWAS of Alzheimer's Disease. We generated prediction models for 16,976 genes, yielding over 73,873 imputable gene–molecular-trait pairs. Incorporating prediction models beyond Lasso and Elastic-net have yielded 28,014 additional imputable gene–molecular-trait pairs. We identified 327 significant TWAS associations, including 42 novel candidate genes (e.g., VGF, IGHG2, STX1B, ASPHD1). To distinguish putative causal genes from associations driven by linkage disequilibrium and co-regulation, we integrated causal TWAS fine-mapping, and further prioritized 146 plausible causal gene-molecular-trait pairs, such as microglia-specific causal signals for PICALM and PTK2B.

INTEGRATIVE MULTI-OMICS ANALYSIS OF NEURAL DIFFERENTIATION REVEALS REGULATORY ALTERATIONS AND NONCODING VARIANT ENRICHMENT ASSOCIATED WITH AUTISM SPECTRUM DISORDER RISK

Jiayi Liu[1,2], William DeGroat[2], Paul Matteson[2,3], James Millonig[2,3], Anat Kreimer[2,4]

[1]Rutgers University, Graduate Program in Cell & Developmental Biology, Piscataway, NJ, [2]Rutgers University, Center for Advanced Biotechnology and Medicine, Piscataway, NJ, [3]Rutgers University, Department of Neuroscience and Cell Biology, Rutgers Health, Piscataway, NJ, [4]Rutgers University, Department of Biochemistry and Molecular Biology, Rutgers, Piscataway, NJ

Autism spectrum disorder (ASD) is a neurodevelopmental disorder with extensive genetic heterogeneity, much of which involves noncoding regulatory variation. However, the cell-type-specific mechanisms by which these variants disrupt early neural development remain unclear. Here, we integrate ATAC-seq, H3K27ac ChIP-seq, and RNA-seq across induced pluripotent stem cells (iPSCs), neural progenitor cells (NPCs), and induced neurons (iNs) derived from neurotypical individuals and unaffected siblings of individuals with idiopathic ASD. Using the Activity-by-Contact (ABC) model, we reconstructed enhancer–promoter interaction (EPI) networks to map the regulatory architecture underlying human neurodevelopment.
Our analyses reveal distinct chromatin accessibility, enhancer activity, and transcriptional programs across developmental stages and sample origins. NPCs from ASD families exhibit altered accessibility and transcription at loci involved in immune signaling and neurogenesis. By overlaying ASD-associated de novo mutations and neuropsychiatric GWAS variants on EPI networks, we identify significant enrichment in NPC- and iN-specific enhancers. Disrupted transcription factor binding motifs in these enhancers implicate regulatory interference at key neurodevelopmental loci.
These findings define a mechanistic framework linking noncoding variation to cell-type-specific dysregulation in ASD and underscore the importance of enhancer–promoter architecture in early human neural development. Our study provides a foundation for the functional dissection of regulatory variants in human neuronal models of neurodevelopmental disorders.

EPITHELIAL-INTRINSIC ALTERATIONS AND MALADAPTATION TO LUMINAL METABOLITES UNDERLIE PERSISTENT CROHN'S DISEASE PATHOGENESIS

Jianqiao (Josh) Liu[1], Jason Koval[2], Peter Carbonetto[3], Candace M Cham[2], Ashley M Sidebottom[4], Matthew Stephens[3], Sebastian Pott[2], Eugene B Chang[2], Anindita Basu[2]

[1]University of Chicago, Department of Chemistry, Chicago, IL, [2]University of Chicago, Department of Medicine, Chicago, IL, [3]University of Chicago, Department of Human Genetics, Chicago, IL, [4]University of Chicago, Duchossois Family Institute, Chicago, IL

The noninflamed epithelium under Crohn's disease exhibits inflame-like transcriptional signatures that persist regardless of clinical remission, but whether this reflects sustained environmental triggers or epithelium's intrinsic immunological adaptions remains unclear. Therefore, we assayed the scRNA-seq and scATAC-seq of patient-matched intestinal tissues and organoids to deconvolute the intrinsic alterations that persist after isolation from local environment, followed by luminal metabolite stimulation to access the environmental contribution to disease signatures. We demonstrated organoid presented epithelial-autonomous disease changes despite of remissions whereas tissue additionally integrated signals from immune and microbial crosstalk. Exposing organoid to patient-derived luminal metabolites revealed a selective activation of disease-associated transcriptional alteration and environment-responsive inflammatory chromatin remodeling in Crohn's disease organoid but not in control. Overall, these results support that disease-associated epithelial states are contributed by both cell-intrinsic dysregulation and microenvironmental cues.

LONG-READ METHYLOME PROFILING IN THE HUMAN PANGENOME REVEALS ANCESTRY-ASSOCIATED METHYLATION STATES AND GENETIC-VARIANT-COUPLED REGULATORY EFFECTS

Tianjie Liu, Juan F Macias-Velasco, Xiaoyu Zhuo, Juan Jiang, Zheng Dong, Wenjin Zhang, Daofeng Li, Chad Tomlinson, Eddie Belter, Ting Wang

Washington University School of Medicine, Department of Genetics, St. Louis, MO

DNA methylation is fundamental to chromatin architecture, imprinting, transcriptional regulation, and disease, yet short-read approaches incompletely resolve complex loci and non-reference sequences. Here, we combine long-read methylation profiling with the human pangenome graph to define population-scale methylation diversity, CpG island polymorphism, and the local and cis-regulatory coupling between methylation and genetic variation.

Using a graph-projected harmonization framework anchored to the telomere-to-telomere CHM13 backbone, we enable direct comparison of methylation across 462 haplotypes, 5 populations, and structurally variable sequences, including previously inaccessible non-reference insertions. Graph-based CpG genotyping captures more than 40 million nonredundant CpGs and supports cross-haplotype analyses within variant-bearing regions. Across the genome, we identify ~11 Mb of highly variable methylated regions (HVMRs), enriched in distal regulatory elements and CpG island shores. Methylation variability shows a weak overall negative correlation with nucleotide and synteny diversity, with a particularly strong negative relationship in SVA elements. At a small subset of loci with extreme synteny diversity, however, methylation no longer conforms to canonical sequence-embedded patterns, instead, methylation states extend from variant sequence into adjacent flanks, reshaping local regulatory landscapes. We further construct a nonredundant pangenome-wide CpG island set and define its population polymorphism and saturation dynamics, providing a systematic view from CpG island sequence variation to methylation divergence.

Non-negative matrix factorization of HVMR methylation resolves ancestry-associated methylome states, with African-ancestry haplotypes showing the clearest separation from non-African haplotypes. Graph-based differential methylation analyses further reveal lower methylation in African-ancestry haplotypes at LTR-rich repeats and identify ancestry-specific differentially methylated regions. Finally, by integrating long-read Iso-Seq-defined novel transcripts, we perform haplotype-resolved, allele-specific xQTL analyses that link genetic variation to methylation, isoform expression, and alternative splicing. Together, these results establish a pangenome-enabled framework for dissecting how structural variation shapes human methylation and transcription, and reveal ancestry-associated epigenomic modules with broad implications for human biology and evolution.

ACE-OF-CLUST: ALIGNMENT, COMPARISON, AND EVALUATION OF OMICS FEATURES IN SINGLE-CELL CLUSTERING

Xiran Liu[1], Ritambhara Singh[1,2], Sohini Ramachandran[1,3]

[1]Brown University, Data Science Institute, Providence, RI, [2]Brown University, Department of Computer Science, Providence, RI, [3]Brown University, Department of Ecology, Evolution, and Organismal Biology, Providence, RI

Clustering is widely used to identify cell types in cellular-resolution transcriptomic data, including single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST). Mixed-membership clustering assigns fractional memberships across clusters to capture continuous variation beyond traditional hard clustering; however, integrating and interpreting results from either approach is complicated by the "clustering alignment problem," which arises from label switching, multi-modality (i.e., multiple alternative solutions), and differences in model settings (particularly the numbers of clusters).

We introduce *ACE-OF-Clust*, which enables a streamlined four-step workflow for single-cell clustering: multiple clustering, clustering alignment, model comparison, and identification of informative features. *ACE-OF-Clust* provides a framework for the direct comparison of clustering solutions, assessing consistency against annotations while leveraging feature-level mixed-membership clustering profiles to prioritize genes that discriminate between cell types. Its cross-model and cross-omic capabilities quantify shared clustering patterns across models or modalities, identifying the specific features—or feature pairs—that drive cell-group separation. We demonstrate its utility when applied to scRNA-seq, ST, and multi-omic single-cell datasets. *ACE-OF-Clust* enables benchmarking of existing models for cell-type identification by highlighting cluster assignments that remain robust or vary substantially across clustering runs and parameter settings. Notably, it quantifies cross-omic clustering variability and identifies putative cross-omic regulatory links. Overall, *ACE-OF-Clust* increases the interpretability, flexibility, and robustness of single-cell clustering, providing a scalable tool for studying cellular heterogeneity and gene expression dynamics.

COMPREHENSIVE GENE HERITABILITY ESTIMATION REVEALS THE ROLE OF RARE CODING VARIANTS IN HUMAN TRAITS AND DISEASES

Zhengtong Liu[1], Boyang Fu[2], Moonseong Jeong[1], Prateek Anand[1], Aakarsh Anand[1], Seon-Kyeong Jang[3], Aditya Gorla[4], Noah Zaitlen[3,5,6], Richard Border[7], Sriram Sankararaman[1,5,6]

[1]UCLA, Computer Science, Los Angeles, CA, [2]Harvard Medical School, Biomedical Informatics, Boston, MA, [3]UCLA, Neurology, Los Angeles, CA, [4]UCLA, Bioinformatics, Los Angeles, CA, [5]UCLA, Computational Medicine, Los Angeles, CA, [6]UCLA, Human Genetics, Los Angeles, CA, [7]CMU, Computational Biology, Pittsburgh, PA

Whole-exome sequencing (WES) enables direct interrogation of rare protein-coding variation, but quantifying how rare and ultra-rare variants shape complex traits remains challenging. Gene-level estimates can be inflated by local linkage disequilibrium (LD), while the sparsity of rare variants limits statistical precision. We present FLEX (Fast, LD-aware Estimation of eXome-wide and gene-level heritability), a scalable mixed-model framework that integrates coding variants across the allele-frequency spectrum and corrects LD by jointly modeling variants within each gene and its flanking regions. FLEX incorporates all coding variants—including singletons—into a unified model, supports both individual-level and summary-level data, and enables gene- and gene set-level inference at biobank scale.

We validated the calibration and power of FLEX through extensive simulations. Applying FLEX to 18,624 protein-coding genes across 20 quantitative traits in UK Biobank whole-exome data (N = 153,351), we observed that LD correction reduced the number of significant gene–trait associations by 36.7%. We identified 64 gene–trait pairs with genome-wide significant gene-level heritability ($p < 0.05/18{,}624$). While gene-level heritability was correlated with the number of GWAS associations per gene (Pearson $r = 0.48$), we also identified gene–trait pairs with significant gene-level heritability despite lacking genome-wide significant GWAS hits or rare-variant test signals (e.g., *GATA3*–BMI). Beyond identifying individual gene–trait associations, we sought to characterize gene-level polygenicity. We observed several genes with substantial contributions to trait heritability: *ALPL* accounted for 13.1% of the genome-wide gene-level heritability for alkaline phosphatase, and *APOE* explained 28.1% for LDL-C. To quantify gene-level polygenicity, we computed tau80, the minimum number of genes explaining 80% of cumulative gene-level heritability. Across traits, polygenicity was substantial (average tau80 = 295), with lipid traits least polygenic (tau80 ≈ 247) and anthropometric traits most polygenic (tau80 ≈ 385).

At the exome scale, incorporating rare and ultra-rare coding variants increased heritability by 24.8% on average relative to common-variant imputed SNP analyses, with substantially larger per-allele effects at rare variants (~18× on average). Rare and ultra-rare variants explained 38% of gene-level heritability on average. Functional stratification showed that missense variants accounted for most rare coding heritability (76.3%), while per-allele effects tended to be larger for rare predicted loss-of-function (pLoF) than for missense variants, on average. Together, FLEX provides gene-resolved heritability estimates from WES, complementing GWAS and rare-variant tests to prioritize genes and refine coding-variant genetic architecture.

S2F: A PACKAGE FOR DEEP AND MECHANISTIC SEQUENCE TO FUNCTION MODELING

Zhihan Liu, Justin Kinney

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

Learning quantitative relationships between genotype and phenotype from multiplexed assays of variant effects (MAVEs) is central to interpreting genetic variants and designing biomolecules. We present S2F, a modular neural-network-based package for interpretable sequence-to-function modeling of MAVE datasets. S2F models one or more latent phenotypes as sequence-dependent quantities that explain the observed measurements while accounting for nonlinearities and noise in the experimental process. To capture sequence-to-latent-phenotype mappings of varying complexity, S2F supports a range of models from linear and deep learning to explicit thermodynamic formulations. Compared to previous methods, S2F provides (1) joint modeling across datasets and loci to learn consensus regulatory models, (2) multidimensional latent phenotypes, (3) support for arbitrary-order epistasis, and (4) a modular thermodynamic layer that constructs biophysical models from high-level specifications.
Applied to MPRA data for the E. coli lacZ promoter, S2F recovers known thermodynamic regulatory architecture, including transcription factor interaction energies. S2F also highlights sequence contexts poorly explained by the initial biophysical assumptions, enabling iterative refinement of the underlying model. The refined model remains mechanistically interpretable while outperforming deep learning approaches with far fewer parameters.

GENOME EXPANSION DRIVEN BY TRANSPOSABLE ELEMENTS AND POTENTIAL SYMBIONT-TO-HOST HORIZONTAL GENE TRANSFER IN LUCINID BIVALVES

Alejandro Llanos-Lizcano, Lisa Wybranitz, Thomas Rattei, Jillian Petersen

University of Vienna, Center for Microbiology and Environmental Systems Science, Vienna, Austria

Reliance on beneficial microbial symbionts is a conserved feature of all plant and animal life. Lucinidae is a diverse, ancient, widespread, and ecologically important family of marine bivalves and is also a prominent example of specific and highly intimate host-microbe symbiosis. All known species in this family rely on symbiosis with a specific family of gammaproteobacterial for nutrition. So far, studies of lucinid genomes were limited to the use of a few genetic markers for phylogeographic analyses. The Aquatic Symbiosis Genomics Project recently released chromosomal-level assemblies of many species including several lucinids, providing a unique opportunity to investigate how this signature trait of eukaryotic life, species evolving and living together for mutual benefit, has shaped the genomes of the eukaryotic hosts. Here, we present the first comparative genomics study of 13 lucinid species. Genome size in this family is surprisingly dynamic: the average is 1.8 Gb, but some species have undergone drastic genome reduction (e.g., Rugalucina vietnamica; 1.1 Gb), or expansion (e.g., Lucinisca nasulla; 2.9 Gb). Selfish DNA elements may explain part of their dynamic nature, with transposable elements (TEs) making up between 44% and 77% of the total (Median; 63%). The TEs responsible for genome expansion are predominantly class II DNA transposons from the families PIF-Harbinger, hAT and TcMar. We identified 53 TE types occurring only within lucinid genomes when compared with six outgroup mollusc species. Lucinid chromosomes (18 ± 1) are typically conserved with rare inter-chromosomal mixing events, however, two major chromosomal reorganization events have occurred; one lineage that includes many members of one subfamily has undergone a chromosomal fusion-with-mixing event. The other identified event (chromosomal breakage) affects one so-far sequenced species. Lucinid genomes encode, on average, 33,000 proteins with a core 10.7k orthogroups in all species. Massive transfer of symbiont genes to the nuclear genome was a feature of organellar evolution, however, in other studied symbiotic systems, this is mostly restricted to one or a few symbiont genes. Here we identify 60 high-confidence genes in the host genome that were likely acquired from (relatives of) the intracellular symbionts. Our study highlights how large-scale collaborative (eukaryotic) genome sequencing efforts are enabling new insights into the genomic mechanisms underpinning such fundamental processes as co-evolution between hosts and symbionts in non-model organisms that play key roles in global ecosystems.

A GEOMETRIC THEORY OF PARAMETER IDENTIFIABILITY IN THERMODYNAMIC STATE MODELS

Kaiser Loell, Justin Kinney

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

Thermodynamic state models are a highly interpretable class of sequence-function models, with parameters that have explicit mechanistic interpretations. However, these models are frequently difficult to fit and are limited in the types of data on which they can be trained. To identify the sources of these challenges, we analyze the geometry of thermodynamic energy-activity functions to develop a theory to predict which thermodynamic parameters are identifiable under which circumstances. We then validate this theory by systematically simulating sequence-function datasets based on minimal thermodynamic models of small numbers of transcription factor binding sites and evaluating the ability of models with the same architecture to recover the ground-truth parameters when trained on the simulated datasets. These results suggest library design strategies for generating datasets that allow for the training of more accurate biophysical models, as well as reparameterization strategies to extract metrics that can be accurately inferred across a broad range of assumptions, which we evaluate through simulation studies.

MOM DOES IT BEST: HOW HORMONE-GATED ENHANCERS RECONFIGURE NEURONAL CIRCUITS FOR PARENTING

Brandon L Logeman

University of Kentucky, Molecular and Cellular Biochemistry, Lexington, KY

Across many animal species, including lab mice, systemic hormones bias infant-directed behavior: lactating females exhibit robust maternal care, whereas virgin males frequently attack and kill infants. How these circulating cues engage cell type–specific regulatory elements to tune defined neuronal circuit nodes and brain activity remains unclear. Here I combine behavioral analysis, intersectional genetics, in vivo activity monitoring, chemogenetic perturbations, and genetically targeted single-nucleus RNA-seq and ATAC-seq across infanticidal virgin males, partially parental fathers and virgin females, and highly parental lactating mothers to map state-dependent transcription and chromatin remodeling within a defined hypothalamic circuit that controls infant-directed behaviors.

In a neuronal population selectively engaged during maternal pup care, chemogenetic activation in virgin females increased parental behavior. Multiomic profiling identified a mother-specific, prolactin-dependent gene regulatory program characterized by increased Stat5b motif activity and induction of the oxytocin receptor gene. Acute-slice calcium imaging showed that, in mothers, these neurons acquire sensitivity to oxytocin, linking regulatory remodeling to functional excitability. Consistent with this model, conditional disruption of prolactin receptor–to–Stat5b signaling reduced parenting in mothers, whereas expressing a constitutively active prolactin receptor was sufficient to enhance caregiving in virgins.

A second, downstream neuronal population functions as a bidirectional hub for parenting versus infant attack. Chemogenetic inhibition biased behavior toward infant-directed attack, whereas activation increased parental behavior. Chromatin accessibility analysis revealed a strongly male-biased regulatory landscape enriched for androgen receptor (Ar) motifs, with Ar-linked candidate enhancers associated with coordinated changes in ion-channel genes that parallel sex-dependent intrinsic electrophysiological properties. Removing Ar in this population abolished pup attack in males, while introducing a constitutively active Ar allele into females suppressed pup care.

Together, these results support a hormone-to-enhancer-to-excitability framework in which endocrine cues engage cell type–specific cis-regulatory elements to recalibrate hypothalamic circuit dynamics and determine whether infant cues evoke nurturing or aggression.

AGE MODIFIES METHYLATION QTL ACROSS TISSUES IN A FREE-RANGE POPULATION OF RHESUS MACAQUES

Amy Longtin[1], Rachel M Petersen[1], Baptiste Sadoughi[2], Christina E Costa[3], Cayo Biobank Research Unit[4], Angelina V Ruiz Lambides[5], Amanda D Melin[6], Michael L Platt[4], Michael J Montague[4], James P Higham[3], Noah Snyder-Mackler[2], Amanda J Lea[1]

[1]Vanderbilt University, Biological Sciences, Nasvhille, TN, [2]Arizona State University, Life Sciences, Tempe, AZ, [3]New York University, Anthropology, NY, NY, [4]University of Pennsylvania, Neuroscience, Philadelphia, PA, [5]University of Puerto Rico, CPRC, San Juan, PR, [6]University of Calgary, Anthropology, Calgary, Canada

DNA methylation (DNAm) is an epigenetic gene regulatory mechanism that plays a critical role in biological processes such as development, aging, and tissue differentiation. Genetic variation can also strongly impact DNAm, and thus a major area of interest has been understanding how genotype interacts with these processes to shape interindividual variation. Previous work, for example the GTEx Project, has mapped genetic effects on DNAm levels through cis methylation quantitative trait loci (meQTL) in humans in nine tissues, but a wider survey of tissue-dependent genetic effects on DNAm and their interaction with demographic processes is needed. In this study, we aimed to test the hypothesis that loss of methylation fidelity with age dampens genetic effects in older individuals. Here, we use multi-tissue (n=14) DNAm data from a population of rhesus macaques (n=237 individuals; 2,485 total samples) to map tissue-dependent meQTLs as well as interactive effects of genotype and age on DNAm. We tested for cis meQTL at 1.57 million SNPs and an average of 145k CpG regions per tissue using a binomial mixed model (PQLseq2) followed by an empirical Bayes approach (mashR). We identified 561,622 meQTL significant in >1 tissue, with 94,222 (17%) being specific to a single tissue and 183,803 (33%) being fully tissue-shared. Tissue-specific meQTLs are consistently enriched in enhancer chromatin states in the focal tissue ($p < 0.05$), supporting the gene regulatory role of DNAm in shaping tissue-specific phenotypes. Tissue-specific meQTLs are also enriched for motifs of tissue-relevant transcription factors–e.g., liver-specific meQTLs are enriched for motifs of TFs (NR1D1/2), which is associated with lipid metabolism regulation and bile acid synthesis. Consistent with our hypotheses, we found more (>2X) meQTL in younger compared to older individuals, along with age-associated differences in effect sizes and functional enrichment. Together, this study provides a novel, powerful approach to understand the effects of genetic variants on DNAm, as well as how these genetic effects interact with aging processes to vary across the lifespan.

META-ANALYSIS OF RARE VARIANT ASSOCIATION RESULTS FOR 222 TRAITS ACROSS 786,871 SAMPLES ENHANCES GENETIC DISCOVERY AND IDENTIFIES PLEIOTROPIC EFFECTS AMONG COMPLEX DISEASES

Wenhan Lu[1], Robert J Carroll[2], Matthew Solomonson[1], Dan M Rodan[3], Benjamin M Neale[1], Konrad J Karczewski[1]

[1]Broad Institute, MPG, Cambridge, MA, [2]Vanderbilt University Medical Center, Vanderbilt Institute for Clinical and Translational Research, Nashville, TN, [3]Vanderbilt University Medical Center, Personalized Medicine, Nashville, TN

Large-scale genome-wide association studies (GWAS) and rare variant association studies (RVAS) provide powerful resources for gene discovery in complex traits. Combining global biobanks increases power further across a broad range of phenotypes, enabling identification of potential therapeutic targets.

To maximize statistical power for gene discovery, we integrated phenome-wide genetic association results from two biobanks: the All by All resource from All of Us (AoU) Research Program v8, comprising 392,030 participants from six genetic groups and 3,602 phenotypes, and the Genebass dataset based on UK Biobank, including 394,841 European participants and 4,529 phenotypes. After harmonizing phenotypic definitions across these resources, we performed cross-biobank meta-analyses across 222 phenotypes with up to 786,871 individuals. This joint analysis identified 244 gene-phenotype associations through predicted loss-of-function (pLoF) variants that are not significant in either biobank alone. To further prioritize potentially novel gene associations, we developed an AI-based literature review approach to systematically evaluate prior evidence and identify 29 novel pLoF associations, including rare pLoF variants in NAA15 associated with type II diabetes. Moreover, the large sample size and broad phenotype spectrum empower systematic investigation into the pleiotropic effects of rare protein-coding variants. For example, pLoF variants in TIMD4 show coherent effects on key metabolic traits, such as LDL, triglycerides, and hyperlipidemia, which are only found through either All by All meta-analysis or cross-biobank meta-analysis. Finally, we investigate the effect of having a rare pLoF variant in a highly constrained gene across common diseases and complex traits, and find associations with hundreds of phenotypes, suggesting a broader phenotypic impact than previously observed.

We also present a public interactive browser for rapid querying of the All by All summary statistics, along with support documentation, including a Featured Workspace in the AoU Researcher Workbench. We release the full set of summary statistics on the Controlled Tier of the AoU Researcher Workbench, where researchers can query or download the results to investigate the functional effects of genetic components on complex diseases with unprecedented resolution and power.

EXPLORE GENE×ENVIRONMENT INTERACTIONS IN REGULATING THE CIRCULATING LEVELS OF POLYUNSATURATED FATTY ACIDS

Yueqi Lu[1], Kaixiong Ye[1,2]

[1]University of Georgia, Department of Genetics, Athens, GA, [2]University of Georgia, Institute of Bioinformatics, Athens, GA

Polyunsaturated fatty acids (PUFAs) are crucial dietary components for human nutrition. The levels of circulating PUFAs (cPUFAs) are results of genetic backgrounds, environmental exposures (e.g., diet and lifestyle), and their interactions (i.e., G×E). Most existing genome-wide G×E studies of cPUFAs focused on specific environmental exposures and were limited by statistical power due to insufficient sample sizes and the statistical burden of multiple testing. Variance quantitative trait locus (vQTL) analysis offers a powerful method to detect candidate GxE loci without explicit environmental information. vQTLs could then be tested for GxE with specific environmental exposures, enhancing statistical power.

The longitudinal cohort UK Biobank (UKB) measured metabolic biomarkers using high-throughput nuclear magnetic resonance (NMR) spectroscopy from EDTA plasma samples at baseline. We included genetically determined European participants with 12 cPUFA-related biomarkers and genotype data for vQTL analysis. We explored shared and trait-specific vQTLs across these 12 traits by both the single-trait method with median-based Levene's test and the multiple-trait method with the Cauchy combination test. We then compared the cPUFA-associated loci from previous genome-wide association studies (GWAS) with our vQTLs to evaluate the degree of their overlap. Functional and epigenetic annotations will be applied to investigate the molecular mechanism of vQTLs in regulating cPUFA levels.

In this study, a total of 354,466 participants with a mean age of 56.8 years (53.8% females) from the UKB were included. We identified 3-34 significant independent loci for 12 cPUFA-related traits by the single-trait method. Then we found 22 additional vQTLs through multiple-trait analysis. By comparing GWAS and vQTL results, we seek to assess how strongly genetic influences on cPUFA levels are shaped by environmental context. Finally, multiple annotations will elucidate the potential molecular process of how vQTLs regulate target gene expression through epigenetic mechanisms and subsequently influence cPUFA levels.

In summary, these results will offer insights into the complex mechanisms underlying cPUFA regulation and may inform the development of advanced precision nutrition strategies.

DEVELOPMENTAL DYNAMICS OF 3D GENOME ORGANIZATION IN THE MALARIA MOSQUITO *ANOPHELES COLUZZII*

Varvara Lukyanchikova[1], Vitaly Dravgelis[2], Igor Sharakhov[1]

[1]Virginia Polytechnic and State University, Department of Entomology, Blacksburg, VA, [2]ITMO University, Saint Petersburg, Russia

Long-range communication between genomic regulatory elements is particularly critical for fine-tuning gene expression during development and cell differentiation. Recent studies demonstrate that long-range chromatin interactions in neurons of diverse organisms play an important role in sensory perception and signal integration from sensory organs. In mosquitoes, sensory perception underlies key biological behaviors relevant to disease transmission, including host-seeking, foraging, oviposition, mating, threat avoidance, and heat and cold avoidance.

In this study, we characterized the developmental dynamics of 3D genome organization in several tissues in *Anopheles coluzzii* and identified sets of developmentally conserved and tissue-specific long-range chromatin interactions. Using genome-wide and Capture Hi-C approaches across embryonic, larval, and adult stages, as well as in multiple adult tissues, we detected widespread long-distance chromatin interactions on both autosomes and the X chromosome. We identified multiple extremely large, multi-megabase chromatin loops specific to somatic tissues, present in adult heads and thoraxes but absent in ovaries and testes. Among adult tissues, head samples displayed the most pronounced contact enrichment and the greatest abundance of giant loops, particularly in eye and brain tissues. In contrast, antennae and maxillary palps exhibited substantially fewer long-range interactions.

Our genomic analysis of chromatin loop anchors revealed that genes located at these anchors in antennae and eye/brain samples have specific functions in the nervous system and include several long non-coding RNA (lncRNA) genes. This suggests that they may play a role in regulating neuronal differentiation and sensory perception. In addition, we uncovered a distinct set of smaller head-specific loops (120–2,000 kb) located within intercalary heterochromatin and anchoring neural-cadherin loci. Integrating these findings with RNA-seq datasets revealed that developmental reconfiguration of chromatin loops often coincides with altered transcription of anchor-associated genes.

Collectively, our findings show that the dynamic nature of long-distance interactions, together with the presence of nervous system-related and lncRNA genes at loop anchors, suggests their regulatory importance in mosquitoes. Deciphering of molecular mechanisms and functional consequences of dynamic chromatin interactions will provide important insights into mosquito biology that can ultimately be used for vector control.

INTEGRATING MULTI-OMICS AND MULTI-CONTEXT QTL DATA WITH GWAS REVEALS THE GENETIC ARCHITECTURE OF COMPLEX TRAITS AND IMPROVES THE DISCOVERY OF RISK GENES

Sheng Qian*[1], Kaixuan Luo*[1], Xiaotong Sun*[1], Wesley Crouse[1], Jing Gu[1], Lifan Liang[1], Siming Zhao[3], Matthew Stephens[1,2], Xin He[1]

[1]University of Chicago, Department of Human Genetics, Chicago, IL, [2]University of Chicago, Department of Statistics, Chicago, IL, [3]Dartmouth College, Department of Biomedical Data Science, Dartmouth Cancer Center, Hanover, NH

*These authors contributed equally.

Expression QTLs (eQTLs) are often used to interpret GWAS findings. However, recent studies showed limited colocalization between eQTLs and GWAS signals, and eQTLs explain a small fraction of complex trait heritability. Incorporating molecular QTLs beyond gene expression across diverse tissue and cellular contexts may help close this gap. Yet, integrating heterogeneous QTLs is analytically challenging, as molecular traits often share QTLs or have QTLs in high LD, complicating the attribution of GWAS signals to specific molecular traits. To address this challenge, we leveraged our recently developed method, causal-TWAS (cTWAS, Nature Genetics, 2024), which jointly model variants and gene expression in GWAS regions, leading to better control of false discoveries than commonly used methods. Building on cTWAS, we developed multi-group cTWAS (M-cTWAS), a scalable framework for integrating multi-omics QTLs across contexts to fine-map causal molecular traits. M-cTWAS identifies the molecular modalities and contexts through which genetic variations act on the phenotype, while combining information across molecular traits targeting the same genes to increase power.

Applying M-cTWAS to GWAS of common traits with expression, splicing, RNA stability QTLs across tissues from GTEx, we found that single-tissue eQTLs explain 3-20% of trait heritability, joint analysis increases this to 10-50%. Heritability contributions from splicing and RNA stability are comparable to expression. M-cTWAS identified many genes missed by single-tissue eQTL analysis. In most loci, M-cTWAS prioritized a single gene, whereas colocalization and TWAS implicated ≥2 genes in over half of loci. Those findings were often driven by multiple molecular traits in different genes with QTLs in high LD, highlighting the challenge of standard colocalization analysis and TWAS. Applying M-cTWAS to psychiatric traits using eQTLs and epigenetic QTLs (chromatin accessibility and DNA methylation) from postmortem brains, we found epiQTLs explained substantial heritability not captured by eQTLs, suggesting epiQTLs capture regulatory effects during early development missed by eQTLs from adult samples.

Overall, our results highlight the promise of integrating multi-modal molecular QTLs across diverse tissue/cellular contexts to dissect complex trait genetics. M-cTWAS provides a powerful framework for this integrative analysis to enable causal gene discovery.

SINGLE-CELL eQTL DATASET OF LUNG TISSUES FROM ASIAN NEVER-SMOKERS HIGHLIGHT THE ROLES OF ALVEOLAR EPITHELIAL CELLS IN LUNG CANCER ETIOLOGY

Thong Luong*[1], Jinhu Yin*[1], Bolun Li[1], Ju Hye Shin[2], Elelta Sisay[1], Sama Mikhail[1], Fei Qin[1], Samuel Anyaso-Samuel[1], Christopher Amos[3], Qing Lan[1], Kai Yu[1], Tongwu Zhang[1], Erping Long[4], Jianxin Shi[1], Jin Gu Lee#[5], Eun Young Kim#[2], Jiyeon Choi#[1]

[1]National Cancer Institute, National Institutes of Health, Division of Cancer Epidemiology and Genetics, Bethesda, MD, [2]Yonsei University College of Medicine, Department of Internal Medicine, Seoul, South Korea, [3]University of New Mexico, Department of Internal Medicine, Albuquerque, NM, [4]Chinese Academy of Medical Sciences and Peking Union Medical College, Institute of Basic Medical Sciences, Beijing, China, [5]Yonsei University College of Medicine, Department of Thoracic and Cardiovascular Surgery, Seoul, South Korea

*Equal contributions, #co-corresponding

Single-cell expression quantitative trait loci (sc-eQTL) analyses are powerful in identifying context-specific susceptibility genes from genome-wide association studies (GWAS) loci. However, few studies have comprehensively investigated cells of lung cancer origin in non-European populations. Here, we built a lung sc-eQTL dataset from 129 Korean women never-smokers with epithelial cell enrichment. eQTL mapping identified 2,229 genes with an eQTL (eGenes) in 33 cell types, including East Asian-specific findings when compared to predominantly European datasets. Integration with single-cell chromatin accessibility data demonstrated an enrichment of cell-type-specific eQTLs in cell-type matched candidate enhancers, while shared eQTLs were more frequently found near promoters. Colocalization and transcriptome-wide association study (TWAS) unveiled 36 susceptibility genes from 22 cell types in 22 lung cancer loci, including 10 loci not achieving genome-wide significance in prior GWAS. Around 47% of these genes were from cells of the alveoli, underscoring their importance, especially in lung adenocarcinoma (LUAD) susceptibility. Focusing on the trajectory of alveolar epithelial cell regeneration, we detected 785 cell-state-interacting QTLs, which overlapped with 28% (10) of the identified susceptibility genes. Finally, we experimentally validated variant-to-gene connections for *ROS1* (6q22.1) and *TCF7L2* (10q25.2) from LUAD loci initially identified in East Asian never-smokers, susceptibility genes that were not evident with bulk or European eQTL datasets. Our data highlighted context-specific susceptibility genes, especially from alveolar cells of lung, contributing to lung cancer etiology.

SYSTEMATIC DISCOVERY OF NON-CODING DRIVER MUTATIONS IN EVOLUTIONARILY CONSTRAINED REGIONS: A PAN-CANCER ANALYSIS.

Firoj Mahmud[1,2], Suvi Mäkeläinen#[1,2,8], Raphaela Pensch#[1,2], Sergey V Kozyrev[1,2], Anna Darlene van der Heiden[1,2], Ananya Roy[2,3], Eric Pederson[1,2], Åsa Karlsson[1,2], Sharadha Sakthikumar[4], Mats Pettersson[1,2], Eric S Lander[5], Maja-Louise Arendt[1,2,6], Karin Forsberg-Nilsson%[2,3,7], Kerstin Lindblad-Toh%[1,2,5]

[1]Uppsala University, Department of Medical Biochemistry and Microbiology, 75123, Uppsala , Sweden, [2]Uppsala University, Science for Life Laboratory, 75223, Uppsala, Sweden, [3]Uppsala University, Department of Immunology, Genetics and Pathology, 75185, Uppsala, Sweden, [4]The Translational Genomics Research Institute (TGen), Phoenix, AZ, [5]Broad Institute of MIT and Harvard, Cambridge, 02142, MA, [6]University of Copenhagen, Department of Veterinary Clinical Sciences, Copenhagen, Denmark, [7]University of Nottingham Biodiscovery Institute, Division of Cancer and Stem Cells, Nottingham NG72RD, United Kingdom, [8]Swedish University of Agricultural Sciences, Department of Animal Biosciences, 750 07, Uppsala, Sweden

# Contributed equally.
% Contributed equally.

Cancer is a collective term for approximately 200 diseases that arise from genetic alterations promoting uncontrolled cell proliferation and metastatic capability. Large-scale genomic analyses have uncovered numerous driver genes harbouring protein-coding mutations that contribute to malignant transformation. In contrast, the non-coding genome has received comparatively less attention, despite strong indications that it also contains mutations with driver potential. We analyzed data from 2,539 cancer genomes representing 16 tumor types and 33 subtypes from the International Cancer Genome Consortium (ICGC). We intersected the non-coding regions of the genome with phyloP scores from the Zoonomia 240 mammals to identify non-coding constraint mutations (NCCMs) with regulatory potential.

We developed a statistical framework modeling NCCM enrichment while controlling for background mutation rate and other confounders. Using complementary gene-centric and 2-kb sliding-window analyses, we identified 56 pan-cancer genes, including FOXA1 and TRIM27, as well as 132 cancer-type–specific genes, such as MYC and CARD11. Genes with high NCCMs burden generally contained few protein-coding driver mutations, suggesting selective pressure to preserve protein integrity while altering regulatory control. Clustering of recurrent NCCMs revealed 176 regulatory hotspots, including both previously described and novel loci, such as regions near RMRP, FSHR, ALPK2, and MIR122HG. Functional assays demonstrated that selected NCCMs alter regulatory activity and gene expression. Together, these results identify conserved non-coding regions as an important and distinct class of cancer driver elements and demonstrate the utility of evolutionary constraint for systematic discovery of regulatory mutations in cancer.

MASSIVELY PARALLEL CHARACTERIZATION OF ADOLESCENT IDIOPATHIC SCOLIOSIS RISK VARIANTS

Darius Ramkhalawan[1], Justin Koesterich[2], Fahim Tasin[1], Carlos Cuna[1], Anat Kreimer[2], Nadja Makki[1]

[1]University of Florida, Department of Physiology and Aging, Gainesville, FL, [2]Rutgers The State University of New Jersey, Department of Biochemistry and Molecular Biology, Piscataway, NJ

Adolescent idiopathic scoliosis (AIS) is a common pediatric musculoskeletal disorder characterized by lateral spinal curvature, often leading to chronic pain and deformity. While a significant genetic component to AIS is recognized, the functional impact of most associated genetic variants, particularly those in non-coding regions, remains largely unknown. This study aims to identify functionally relevant AIS-associated non-coding variants, focusing on their regulatory activity in chondrocytes, a major cell type implicated in AIS pathogenesis.

Using massively parallel reporter assays (MPRAs), we characterized 1,664 variant positions in linkage disequilibrium with 26 AIS lead variants identified by genome-wide association studies (GWAS). We designed a library of 7,173 candidate regulatory sequences (CRSs), comparing the 1,664 reference alleles against 4,708 alternate alleles, and performed MPRAs in two human chondrocyte cell lines (TC28a2 and SW1353).

Our analysis identified 234 variants that exhibited significant differential regulatory activity between their reference and alternate alleles. Of these, 201 are predicted to disrupt transcription factor binding sites (TFBSs), often correlating with their observed regulatory effect. Notably, we validated rs9496392, a single-nucleotide variant near the *ADGRG6* locus, which showed consistent differential regulatory activity in both cell lines. *ADGRG6* is a key regulator of cartilage homeostasis, and its cartilage-specific knockout in mice results in a scoliosis-like phenotype. The AIS risk allele of rs9496392 (T) is predicted to strongly disrupt several TFBSs, including SP1, which is known to enhance *COL2A1* expression.

This study provides a foundational catalog of functional AIS-associated regulatory variants active in chondrocytes, offering crucial insights into the perturbed gene regulatory networks in AIS. These findings lay the groundwork for identifying biomarkers and potential therapeutic targets for this complex childhood disease.

PRINCIPLES AND FUNCTIONAL CONSEQUENCES OF PLASMID CHROMATINIZATION IN MAMMALIAN CELLS

Benjamin J Mallory[1,2], Thomas W Tullius[3], Carina G Biar[1], Conor P Herlihy[1], Jonas A Gustafson[4,5], Stephanie C Bohaczuk[6], Danilo Dubocanin[7], Brian J Beliveau[1,2,8], Devin K Schweppe[1,2,8], Lea M Starita[1,2], Andrew B Stergachis[1,2,6]

[1]University of Washington, Department of Genome Sciences, Seattle, WA, [2]University of Washington, Brotman Baty Institute for Precision Medicine, Seattle, WA, [3]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [4]University of Washington, Department of Pediatrics, Seattle, WA, [5]University of Washington, Molecular and Cellular Biology Program, Seattle, WA, [6]University of Washington, Division of Medical Genetics, Seattle, WA, [7]Stanford University, Department of Genetics, Stanford, CA, [8]University of Washington, Institute for Stem Cell and Regenerative Medicine, Seattle, WA

Plasmids have been a foundational tool for deciphering the regulatory genome across the tree of life for over 50 years. Yet despite their widespread adoption, plasmid-based assays do not measure the overlying chromatin architecture, a key variable that dictates regulatory element activity in the endogenous context. In fact, it remains poorly understood how, or whether, regulatory sequences are chromatinized when placed within a plasmid context, and how this chromatin architecture compares to the element's endogenous chromatin state. This knowledge gap limits the ability to confidently translate plasmid-based measurements of regulatory activity to conclusions about how these regulatory sequences function in their endogenous biological context.

To address this limitation, we developed plasmid Fiber-seq, a single-molecule chromatin fiber sequencing method that maps chromatin architecture along individual full-length transfected plasmid molecules at near single-nucleotide resolution. By directly comparing chromatin architectures between plasmid and endogenous genomic contexts for the same regulatory sequences, we demonstrate that plasmids are indeed capable of faithfully recapitulating nuclear genome-encoded chromatin architectures. However, the fidelity of this recapitulation varies between elements and is largely dependent on the specific genomic sequence context provided in the plasmid setting. Importantly, plasmid Fiber-seq can readily detect differences in chromatin architecture between plasmid and endogenous contexts, enabling direct assessment of whether a given plasmid construct design accurately reflects the endogenous chromatin architecture of the regulatory element under study.

Furthermore, by combining plasmid Fiber-seq with massively parallel reporter assays, we can resolve the molecular mechanisms underlying pathogenic non-coding variants. Where traditional MPRAs measure only changes in transcript abundance, plasmid Fiber-seq can detect how variants alter the overlying protein occupancy of regulatory elements, including the gain or loss of transcriptional activators and repressors. Additionally, plasmid Fiber-seq can identify variants whose effects on measured activity may reflect altered RNA stability rather than changes in transcriptional regulation. Together, these findings expand the power of plasmid-based assays to decipher regulatory grammar and the molecular mechanisms of disease-associated non-coding variation.

UNCOVERING GENES INVOLVED IN UN(DER)-STUDIED FUNCTIONS AND PHENOTYPES ACROSS SPECIES USING GRAPH LEARNING OF MOLECULAR NETWORKS AND BIOMEDICAL ONTOLOGIES

Keenan Manpearl, Alexander McKim, Arjun Krishnan

University of Colorado, Anschutz Medical Campus, Department of Biomedical Informatics, Aurora, CO

Despite decades of experiments in both humans and model organisms, our functional understanding of the human genome remains both incomplete and systematically biased with a small number of genes and biological contexts (such as processes, phenotypes, and diseases) that are over-represented in literature and annotation databases, while most remain largely un(der)-studied and un(der)-characterized. Machine learning (ML) methods have been successful at predicting novel links between genes and biological contexts. However, as most ML approaches require labeled training data for each context, they are not suitable for contexts for which no genes/proteins are already known to be associated (i.e. zero-shot prediction). Recent methods aim to overcome this limitation by incorporating representations of biological contexts derived from biomedical ontologies into the prediction workflow to enable information sharing across contexts, and/or by incorporating training data from multiple species to improve annotation coverage. However, most methods incorporate novel genes by using representations derived from their primary protein sequences and consistently underperform for prediction of biological processes, which requires information about how a gene works with other genes, compared to, say, molecular function which is mostly intrinsic to the gene/protein. Methods that use representations derived from molecular networks overcome this limitation by incorporating interactions between genes/proteins, but, to our knowledge, no network-based method exists that uses both zero-shot techniques and multi-species features. We present a novel ML framework for making and evaluating few- and zero-shot predictions using a multi-species feature space that captures both within-species interactions and cross-species orthology. Through a rigorous evaluation scheme, we show that our approach can accurately predict annotations in the Gene Ontology Biological Process Branch and Unified Phenotype Ontology, even when both the gene and term were unseen during training (i.e. true zero-shot generalizability). In addition, by creating a multi-species feature space, we provide a framework for designing informed model organism studies based on shared or unique biology, and for transferring the results of model organism studies back to human health and disease, even in the absence of one-to-one orthologs. Taken together, we present a novel multi-species network-based framework for gene classification that improves functional annotation in understudied genes and terms.

DISSECTING THE CIS-REGULATORY CODE BEYOND MOTIF SYNTAX

Pablo J Mantilla Puccetti[1], Peter K Koo[2]

[1]School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Precise regulation of gene expression underlies development, cellular differentiation, and disease. The cis-regulatory code, which governs how transcription factors (TFs) read regulatory DNA, is characterized by recurrent sequence patterns, or motifs, that serve as recognition sites for TF families. These motifs appear to form a combinatorial syntax in which their arrangements and spacings can influence TF interactions and downstream regulatory processes such as nucleosome remodeling, histone modification, co-factor recruitment, and ultimately gene expression.

Deep Neural Networks (DNNs) learn the mapping from regulatory sequence to functional outcome, allowing pre-trained DNNs to serve as virtual experimental platforms. In this paradigm, DNNs have been shown to learn established features of cis-regulatory code, including motif distance, orientation, affinity, and flanking sequence effects, highlighting the power of in-silico experiments to dissect regulatory logic. However, most virtual experiments remain motif-centric, focusing on perturbations of known motifs while largely ignoring regulatory information embedded in surrounding sequence context. This gap has proved a major limitation in our ability to design regulatory elements with desired functional outcomes simply from known motif grammars.

To address this gap, we designed in-silico experiments to systematically quantify sequence context contributions to model predictions. Leveraging SEAM (Systematic Explanation of Attribution-based Mechanisms), we identify attribution signals robust to partial random mutagenesis, enabling us to disentangle motif syntax from sequence context. Building on this capability, we develop motif-context swap experiments to quantify the extent to which motif syntax and sequence context drive predictions. In fly enhancers, we find that motif syntax drives substantial changes to model predictions whereas sequence context sets a strong baseline activity level. In human promoters, we systematically characterize oncogene and Hox gene transcriptional start sites to understand the context-dependent regulatory logic at disease-relevant loci. Overall, this context-aware dissection of the cis-regulatory code moves us closer to a comprehensive understanding of how regulatory DNA encodes gene expression programs beyond motif syntax.

PRIMATE GENOME COMPLEXITY AND ITS EVOLUTIONARY INSIGHTS

Yafei Mao

Shanghai Jiao Tong University, Bio-X Institutes, Shanghai, China

Genetic variation provides the foundation for phenotypic diversity, adaptation, and human disease susceptibility. Despite its importance, the evolutionary dynamics and biological consequences of structural variants (SVs) and complex genomic alterations—such as structurally divergent regions (SDRs)—remain poorly understood in primates. This gap is largely due to historically incomplete reference genomes and the absence of scalable, cross-species frameworks for accurate SV detection.

Here, we develop and apply novel computational tools that enable near-perfect, complete-level assembly of primate genomes and systematic characterization of lineage-specific and recurrent SVs and SDRs across primate evolution. Leveraging these high-resolution genomic resources, we uncover distinct evolutionary regimes of SVs, showing that insertions evolve ~ threefold faster than deletions among small SVs (<10 kbp), while exhibiting a threefold slower evolutionary rate than deletions in large SVs (>=10kbp). Functionally, we identify hundreds of genes disrupted by SVs that are linked to lineage-specific adaptive traits. By integrating single-cell transcriptomic and epigenomic data, we further show that SV-driven regulatory divergence leads to widespread, cell-type-specific gene expression changes affecting more than four hundred genes between humans and macaques. In addition, we identify twenty SDRs in which human-specific segmental duplications generate unique genomic architectures that predispose to recurrent microdeletions and microduplications underlying rare human genetic disorders.

Together, these results highlight the central role of SVs in shaping primate genome evolution and provide new insights into the genetic mechanisms underlying species-specific traits and human disease risk.

THE CAUSAL EPIGENETIC DRIVERS OF CANINE AGING

Blaise L Mariner[1], Brianah M McCoy[1], Benjamin R Harrison[2], The Dog Aging Project Consortium[3], Joshua M Akey[4], Elhanan Borenstein[5], Daniel Promislow[6], Noah Snyder-Mackler[1]

[1]Arizona State University, Tempe, AZ, [2]University of Washington, Seattle, WA, [3]Dog Aging Institute, Jamaica Plain, MA, [4]University of Princeton, Princeton, NJ, [5]Princeton University, Princeton, NJ, [6]Tufts University, Boston, MA

To identify healthspan-extending interventions and pinpoint the mechanisms of aging, we must distinguish the molecular processes that drive aging from those that merely accompany it. Changes in DNA methylation (DNAm) track predictably with age, but it is unclear whether these shifts drive aging or simply record its progress. The companion dog offers a unique opportunity to resolve this causality, given their shorter lifespan and substantial genetic and phenotypic diversity. Leveraging the Dog Aging Project, we first generated genome-wide DNAm data for 899 dogs sampled over three years and found that age, size, and the interaction of age and size significantly reshape the DNAm landscape, reflecting accelerated aging in shorter-lived, larger dogs. Building on these findings, we generated the first genome-wide map of the genetic architecture underlying DNAm, identifying 326,282 methylation quantitative trait loci (meQTL; FDR<0.05). Using these meQTLs in Mendelian randomization, we identified 445 CpG regions with strong causal evidence (FDR<0.05) for DNAm driving an age-related phenotype (mobility). These causally deleterious regions were enriched for autonomous, but not non-autonomous, transposons (OR = 1.22, p = 0.03) that have age-associated demethylation. This suggests that the loss of epigenetic suppression of transposons is causal to the age-related decline in mobility. By integrating genome-wide epigenetic mapping with causal inference, site-specific DNAm at autonomous transposons emerges as a primary pathway driving age-related deterioration, suggesting that therapeutics that quench transposon activation may help mitigate age-related decline.

ANCESTRY INFERENCE FROM PANGENOMES

Franco Marsico, Silvia Buonaiuto, Laura Pignata, Farnaz Salehi, Robert W Williams, Erik Garrison, Vincenza Colonna

University of Tennessee, Dept of Genetics, Genomics and Informatics, Memphis, TN

Traditional identical-by-descent (IBD) detection and local ancestry inference rely on variant calling against a reference genome, introducing reference bias and loss of information on large variants. Here, we present impop$_k$, a toolkit that performs IBD and local ancestry inference directly from pangenome alignments. Our approach builds over pairwise sequence similarity obtained from implicit pangenomes, then applies Hidden Markov Models (HMM), a 2-state Gaussian HMM for IBD and an N-state softmax-emission HMM for ancestry, to call segments from the identity signal.
We validated impop$_k$ across multiple systems and species. To validate IBD inference, we compared impop$_k$ against hap-ibd on the HPRCv2 human pangenome. In all tested regions, every hap-ibd–confirmed IBD pair ranked in the top 10% of pairwise identity scores, with over half ranked first. The main differences were detected in segments surrounding centromeres. The Platinum Pedigree provided instead ground-truth validation of IBD transmission across up to four generations. impop$_k$ correctly resolved haplotype inheritance and pinpointed recombination breakpoints at expected positions in a parent–child trio, with ongoing multi-generational extension.
To validate local ancestry inference, we compared impop$_k$ against RFMix v2 on 4-way inference in a subset of admixed individuals from HPRC. Across validated regions, we achieved 98.4% concordance with RFMix with zero wrong-ancestry calls, and per-population F1 scores of 99.8% (EUR), 99.2% (AMR), and 97.8% (AFR). We extended validation to recombinant inbred mice and bats, confirming cross-species generalizability of the pangenome identity signal for relatedness detection. Both local ancestry and IBD signals were used to perform whole genome natural selection screening in the studied systems. Preliminary results show ancestry specific IBD enrichment in LCT and EDAR, two well-documented examples of recent selection in humans.
impop$_k$ provides a reference-free based IBD and ancestry inference applicable across species as pangenome references become standard for population and ecological genomics.

ACTIVITY-DEPENDENT GENE REGULATION IN AN OCTOPUS LEARNING AND MEMORY CIRCUIT

Matthew McCoy[1,2], Ernie Hwaun[3,4], Chew Chai[5], János Szabadics[3,6], Gergely Szabo[3], Keyue Shi[7], Andrew Fire[2,8], William Gilly[9], Bo Wang[5,10], Ivan Soltesz[3,4]

[1]University of Chicago, Organismal Biology & Anatomy, Chicago, IL, [2]Stanford University, Pathology, Stanford, CA, [3]Stanford University, Neurosurgery, Stanford, CA, [4]Stanford University, Wu Tsai Neurosciences Institute, Stanford, CA, [5]Stanford University, Bioengineering, Stanford, CA, [6]HUN-REN Institute of Experimental Medicine, Laboratory of Cellular Neuropharmacology, Budapest, Hungary, [7]Stanford University, Biology and Howard Hughes Medical Institute, Stanford, CA, [8]Stanford University, Genetics, Stanford, CA, [9]Stanford University, Oceans Department and Hopkins Marine Station, Pacific Grove, CA, [10]Stanford University, Developmental Biology, Stanford, CA

The deep evolutionary divergence between cephalopods and vertebrates offers a rare opportunity to study how complex neural circuits are built from conserved and lineage-specific molecular components. We investigated the octopus superior frontal–vertical lobe (SFL-VL) system, an anatomically compact yet highly neuron-dense circuit critical for associative learning and memory-guided behavior. Utilizing chromosome-scale genomes and advanced deep-learning methodologies, we performed single-nucleus profiling of gene expression and chromatin accessibility in tandem with single-cell electrophysiology. This comprehensive approach allowed us to identify distinct neuronal classes and demonstrated that neuronal activity activates conserved transcriptional programs alongside cephalopod-specific zinc-finger regulators. Notably, we found that activity-dependent gene induction is closely linked to alterations in sodium currents, establishing a connection between transcriptional regulation and neuronal function. Our findings suggest that an ancient framework of activity-dependent transcriptional plasticity with lineage-specific elaborations contributed to the evolution of advanced cognitive capabilities in cephalopods.

EXPLORING THE ARCHAIC INTROGRESSION LANDSCAPE OF ADMIXED POPULATIONS THROUGH JOINT ANCESTRY INFERENCE

Jazeps Medina Tretmanis[1], Maria C Avila-Arcos[2], Flora Jay[3], Emilia Huerta-Sanchez[1]

[1]Brown University, CCMB, Providence, RI, [2]UNAM, LIIGH, Queretaro, Mexico, [3]Universite Paris-Saclay, LISN, Paris, France

Local Ancestry Inference (LAI) has been useful to study the evolution of recently admixed populations in the Americas. Similarly, detecting archaic introgressed tracts in Eurasian populations has provided insights into our interactions with archaic humans. Archaic introgression in recently admixed populations remains understudied, leaving questions about how recent admixture has altered patterns of archaic ancestry. This is partly due to a lack of methods that detect LAI and archaic ancestry simultaneously. Here, we present the first deep learning method capable of jointly inferring continental and archaic introgressed regions, even when trained on a mixture of real and synthetic data.

We show that we can train our model on a mixture of the 1KG dataset and synthetic data, and achieve high inference accuracies on the Simons Genome Diversity Project dataset (94.1% and 93.3% for continental and archaic ancestry classification). Our method is also better at detecting continental ancestry from old admixture events (>100 generations), especially when the admixture proportion from one of the donors is small (~10%), where we demonstrate an 8% increase in accuracy for LAI compared to the next best available method.

Applying our method to American populations from the 1KG dataset, we find that Peruvian individuals harbor a large part of their archaic ancestry in Native American tracts (~50%), while other American populations like Puerto Ricans harbor most of their archaic ancestry tracts on European tracts (~70%). These results suggest that recent admixture has shifted patterns of archaic ancestry in different geographical regions due to local differences in their history of admixture. We also infer the joint ancestry of modern individuals from the Mexican BioBank dataset, and use these results to illustrate how our method can find candidates of adaptive introgression in admixed individuals, through the identification of introgression hotspots where local ancestry patterns differ from global ancestry proportions.

EARLY CHROMATIN ACCESSIBILITY LANDSCAPE OF PERIPHERAL BLOOD CD4$^+$ T CELLS IN CHILDREN PROGRESSING TO TYPE 1 DIABETES

Gopika J Menon[1,2], Sini Junttila[1,2], Mohd M Moin Khan[1], Meraj Hasan Khan[1], Niklas Paulin[1], Omid Rasool [1,2], Mikael Knip[1,3], Laura Elo[1,2,4], Riitta Lahesmaa[1,2,4], Ubaid Ullah Kalim[1,2]

[1]Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland, Interdisciplinary Research Centre, Turku, Finland, [2]InFLAMES Research Flagship Center, University of Turku, Turku, Finland, Faculty of Medicine, Turku, Finland, [3]Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, FInland, Faculty of Medicine, Helsinki, Finland, [4]Institute of Biomedicine, University of Turku, Finland, Institute of Biomedicine, Turku, Finland

The disruption of immune homeostasis alters the delicate balance between inflammatory effector and suppressive regulatory T cell programs, predisposing to either immune-mediated tissue destruction or tumour immune escape. Exacerbated CD4$^+$ T cell responses contribute to the development of autoimmune diseases, including type 1 diabetes by promoting inflammation. While several transcriptional and DNA methylation changes have been reported in peripheral blood samples from children who later developed beta cell autoimmunity and clinical disease, the upstream regulatory mechanisms driving these changes remain poorly understood.

Here we studied changes in chromatin accessibility in CD4$^+$ T cells associated with later development of T1D to explore underlying regulatory mechanisms that drive gene expression changes before beta cell autoimmunity. To this end we performed Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) analyses on longitudinal peripheral blood samples from 13 children genetically at risk for T1D and their age-, sex-, and HLA-risk matched controls. Our ongoing analysis aims to uncover early alterations in chromatin accessibility that precede immune activation and may contribute to the initiation of autoimmunity.

Understanding these early epigenetic changes could provide novel insights into the molecular events leading to T1D and identify potential targets for early intervention or prevention.

EXPLORING THE EVOLUTIONARY DYNAMICS AND MITOCHONDRIAL LOCALIZATION OF C/EBPβ ISOFORMS

Gavirel Minor[1], Daria Arakelova[1], Gilad Barshad[1,2], Dan Mishmar[1]

[1]Ben-Gurion University of the Negev, Department of Life Sciences, Beer Sheva, Israel, [2]Technion-Israel Institute of Technology, Department of Genetics and Developmental Biology, Haifa, Israel

It is currently accepted that regulation of mitochondrial DNA (mtDNA) transcription occurs via a dedicated set of transcription factors, suggesting a separate regulatory mechanism from nuclear DNA-encoded genes. However, accumulating evidence in human cells revealed in vivo mtDNA binding by several known regulators of nuclear gene transcription (TFs), thus suggesting direct mito-nuclear co-regulation. We previously showed that one of these TFs, CCAAT Enhancer binding protein beta (C/EBPβ), binds a negatively selected mtDNA site in human cells, thus supporting its functional importance. C/EBPβ silencing in human cells led to reduced mtDNA copy number, yet increased mtDNA gene expression supporting a negative regulatory role. Our experiments in human cells indicate that two of its three translation isoforms (LAP1, LAP2) are in the cell nucleus, whereas the third and shortest isoform (LIP) localizes in both the mitochondria and in nucleus, implying its involvement in mito-nuclear co-regulation. As a first step to investigate the evolutionary trajectory of this phenomenon, we predicted (TRISTAN) the translation initiation sites of aligned C/EBPβ orthologs from 657 vertebrate species. Our findings indicate that LAP1 is highly conserved within Tetrapoda and in most fish, yet is nearly absent in ray-finned fish (Actinoptery). LAP2 is present in all tested classes, yet LIP conservation was reduced within Amphibians and cartilaginous fish (Chondrichthyes). Secondly, we noticed that in all species in which LIP translation was predicted, its mitochondrial localization was also predicted. This suggests that the emergence of LIP associates with the involvement of C/EBPβ in mito-nuclear co-regulation. This interpretation is currently investigated.

GENOME-WIDE CRISPR KNOCKOUT AND KNOCKDOWN SCREENING TO IDENTIFY KEY HOST FACTORS IN MEDIATING VIRAL PATHOGENESIS OF ALPHAVIRUSES

Tyler Dao[1,2,3,4], Sergio Triana[1,2,3], Ruthie Mitchell[3], Cheyanne L Bemis[6], Lisa Hensley[5], Christopher J Neufeldt[6], Alex Shalek[1,2,3], Pardis Sabeti[3,7,8,9]

[1]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA, [2]Institute for Medical Engineering and Science (IMES), Department of Chemistry, and Koch Institute for Integrative Cancer, Massachusetts Institute of Technology, Cambridge, Massachusetts, Cambridge, MA, [3]Broad Institute of MIT and Harvard, Cambridge, MA, [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, [5]Zoonotic and Emerging Disease Research Unit, National Bio- and Agro-defense Facility, Agricultural Research Service, United States Department of Agriculture, Manhattan, KS, [6]Department of Microbiology and Immunology, Emory University School of Medicine, Atlanta, GA, [7]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, [8]Howard Hughes Medical Institute, Chevy Chase, MD, [9]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA

Recent epidemics of Ebola and Zika and the SARS-CoV-2 pandemic are stark reminders of the public health threat posed by emerging viruses and the urgent need for transformative vaccines and antivirals. Mayaro virus (MAYV) and chikungunya virus (CHIKV), mosquito-borne viruses of the *Togaviridae* family and *Alphavirus* genus, are emerging pathogens with pandemic potential. Both cause rheumatic disease, and CHIKV can cause neurological symptoms. MAYV is endemic in South and Central America and the Caribbean, and CHIKV circulates in Africa and Asia with outbreaks in Europe and the Americas. No antivirals exist for either virus, and the most effective action against both viruses is to avoid human contact with mosquitos. Limited understanding of pathogenic mechanisms remains a major barrier to therapeutic and vaccine development for these viruses. We implemented genome-wide CRISPR-Cas9 knockout and knockdown survival screens in human cell lines during viral infection to identify host factors as potential therapeutic targets. This hypothesis-generating workflow is an unbiased, scalable approach to identify host factors without prior knowledge of a virus. We performed the screens in Huh-7 and A549 cell lines infected with MAYV or an attenuated CHIKV strain. Our screens identified host genes important during infection, including the vacuolar ATPase, previously implicated in replication of the alphavirus Sindbis virus. These shared host factors are promising targets for broad-spectrum antivirals and vaccines against *Togaviridae* pathogens, a major step toward pandemic preparedness for these emerging viruses.

GENETIC ADAPTATION OF BALTIC HERRING TO LOW SALINITY TARGETS REPRODUCTION AND EARLY DEVELOPMENT

Fahime Mohamadnejad Sangdehi*[1], Cheng Ma*[1], Mari Kawaguchi**[2], Kaori Sano**[3], Svenja V Dannenberg**[4], Mats E Pettersson[1], Andreas Wallberg[1], Joshua L Wort[5], Yumeng Yan[4], Sergei Moshkovskii[4,6], Florian Berg[7], Arild Folkvord[7,8], Christof Lenz[4,6], Henning Urlaub[4,6], U. Benjamin Kaupp[5,9], Shigeki Yasumasu[2], Leif Andersson[1,10]

[1]Uppsala University, Uppsala, Sweden, [2]Sophia University, Tokyo, Japan, [3]Josai University, Saitama, Japan, [4]Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany, [5]University of Bonn, Bonn, Germany, [6]University Medical Center Göttingen, Göttingen, Germany, [7]Institute of Marine Research, Bergen, Norway, [8]University of Bergen, Bergen, Norway, [9]Max-Planck-Institute for Multidisciplinary Sciences, Göttingen, Germany, [10]Texas A&M University, College Station, TX

Understanding how species genetically adapt to new environments is a central question in evolutionary biology. Atlantic herring colonized the brackish Baltic Sea within the last ~8,000 years, where salinity is in the range 2–12 parts per thousand (ppt), compared to 34–35 ppt in the North Atlantic. Here, we combine whole-genome population genetics with long-read assemblies and functional analyses, including proteomics, RNA-seq, and enzymatic assays, to dissect the molecular mechanisms underlying adaptation to low salinity. To distinguish salinity-driven adaptation from geographic effects, we used herring from Ringkøbing Fjord — an Atlantic population spawning in brackish water — as a key reference population. We focus particularly on reproduction and early development, the life stages directly exposed to environmental conditions in species with external fertilization. Genes involved in sperm, egg, and embryo function emerge as primary targets of natural selection, highlighting a critical life-history window when gametes and early embryos encounter osmotic stress directly. We identify four unlinked loci showing near-complete fixation of variant alleles in Baltic herring: a sperm-specific volume-regulated anion channel (VRAC, encoded by *LRRC8C2*), a zona pellucida egg envelope protein (*ZPBA1*), a cluster of three fish transglutaminase genes (*FTG1-3*), and a copy number expansion of fish hatching enzyme genes (*HE1C*). Adaptation involved both amino acid sequence changes and structural variation, including gene duplication and copy number variation. Notably, alleles at two of these loci have been introgressed from Pacific herring. Three of the four genes, located on different chromosomes, are functionally interconnected. Modified ZPBA1 sequence and enhanced FTG1-3 enzyme activity at low salinity together produce a harder egg envelope that resists swelling in brackish water. The massive copy number expansion at the *HE1C* locus in Baltic herring drives high hatching enzyme expression and activity, enabling larvae to digest this reinforced envelope at hatch. The sperm-specific VRAC encoded by *LRRC8C2* likely protects sperm from osmotic swelling during spawning in low salinity. Our results demonstrate how coordinated changes in multiple unlinked genes enabled a marine species to adapt to a low-salinity environment.

WIN SOME, NOT LOSE SOME: DEEP TRANSCRIPTOME ANALYSIS EXPANDS GENETIC DISCOVERY IN BULK AND SINGLE-CELL DATA

Daniel Munro[1,2], Yan Hao[1], Alexander Gusev[3], Abraham Palmer[2], Pejman Mohammadi[1,4]

[1]Seattle Children's Research Institute, Center for Immunity and Immunotherapies, Seattle, WA, [2]UC San Diego, Department of Psychiatry, La Jolla, CA, [3]Dana-Farber Cancer Institute and Harvard Medical School, Division of Population Sciences, Boston, MA, [4]University of Washington School of Medicine, Department of Pediatrics, Seattle, WA

Transcriptomic diversity across individuals reflects multiple modes of RNA regulation and is widely used to interpret GWAS signals via quantitative trait locus (QTL) mapping. Yet most studies focus on total expression, and sometimes intron excision rates, leaving additional regulatory variation underused. We introduce **Pan transcriptome phenotyping (Pantry)**, a systematic pipeline for multimodal RNA phenotyping beyond total expression, incorporating isoform ratios, splice junction usage, alternative TSS and polyA usage, and RNA stability. Applying Pantry to GTEx data and 114 GWAS traits, Pantry yields a **67% increase** in independent QTL across six modes of transcriptome regulation and an **87% increase** in unique genes with colocalized associations with GWAS signals compared to conventional eQTL analysis.

We next adapt Pantry to single-cell RNAseq from the OneK1K cohort (1.27M cells from 982 donors). Despite 3-prime tag sequencing, **single-cell Pantry identifies 53% more independent xQTLs and 63% more unique colocalized associations** with GWAS signals than the expression-only analysis used in the original study by Yazar et al., Science (2022), while capturing cell-type-specific regulatory signals. Improvements are consistent across regulatory modes, except that intron excision rates are not adequately quantifiable in 3-prime tag single-cell RNA seq.

To further expand discovery, we present **Latent Data-Driven RNA phenotyping (LaDDR)**, a mechanism-agnostic framework that learns orthogonal modes of transcriptome variation across individuals, enabling xQTL discovery and GWAS integration without requiring complete gene annotations. Applying LaDDR to GTEx, we identify on average 95% more independent QTLs per tissue than the six modes implemented in Pantry. Residualizing the six Pantry modalities prior to LaDDR yields, on average, **41% more QTLs** per tissue beyond the six-mode analysis while retaining interpretability of knowledge-driven signals. Integrating LaDDR-derived phenotypes with GWAS data uncovers on average **45% more unique colocalized gene-trait pairs** per GTEx tissue compared with six-mode Pantry analysis.

Together, Pantry and LaDDR provide a unified, scalable framework for multimodal and data-driven RNA phenotyping that increases genomic discovery across transcriptome cohorts without sacrificing interpretability.

TRANSPOSABLE ELEMENTS SHAPE OLAPARIB RESPONSE ACCORDING TO BRCA1 STATUS IN TRIPLE-NEGATIVE BREAST CANCER

Daniela Moreira Mombach[1,2], Carlos Mendez-Dorantes[3,4,5], Rafael L V Mercuri[2], Suelen C Soares Baal[6], Maria A Poersch[6], Kathleen H Burns[3,4,5], Jaqueline Carvalho de Oliveira[6], Elgion L S Loreto[7], Pedro A F Galante[2]

[1]Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Genética e Biologia Molecular, Porto Alegre, Brazil, [2]Hospital Sírio-Libanês, São Paulo, Brazil, [3]Dana-Farber Cancer Institute, Department of Pathology, Boston, MA, [4]Harvard Medical School, Department of Pathology, Boston, MA, [5]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, [6]Universidade Federal do Paraná, Departamento de Genética, Curitiba, Brazil, [7]Universidade Federal de Santa Maria, Departamento de Bioquímica e Biologia Molecular, Santa Maria, Brazil

Triple-negative breast cancer (TNBC) is an aggressive subtype with limited therapeutic options. While PARP inhibitors, such as olaparib, show promise in BRCA1-deficient TNBC through synthetic lethality, up to 50% of patients fail to respond, highlighting the need to understand the molecular mechanisms underlying PARP inhibitors efficacy. Transposable elements (TEs), particularly LINE-1 elements, are increasingly recognized as modulators of genomic instability associated with DNA repair processes and potential key players in synthetic lethality. Here, we investigate the functional relationship between TE activity and olaparib treatment in TNBC with distinct BRCA1 functional status. We performed comprehensive multi-OMICs analysis of four TNBC cell lines (two BRCA1-deficient and two BRCA1-proficient) treated with olaparib. We analyzed expression and differential expression of protein-coding genes, TEs, and gene-TE chimeric transcripts. Long-read whole-genome sequencing was employed to detect de novo TE insertions, complemented by a functional assay to quantify LINE-1 retrotransposition activity in olaparib-treated cells. Olaparib treatment induces extensive transcriptomic and genomic disorganization mediated by TEs, especially LINE-1, exclusively in BRCA1-deficient cells. Consistently, orthogonal assays confirmed LINE-1 retrotransposition in BRCA1-deficient cells following olaparib exposure. Our findings demonstrate that olaparib treatment induces TE activation especially in BRCA1-deficient cells, a novel mechanism that may underlie synthetic lethality in TNBC. This TE activation triggers immune responses and genomic instability, providing new therapeutic opportunities through immunotherapy combinations and suggesting that TE activity may serve as a potential biomarker for treatment stratification of TNBC.

CHARACTERIZING 5-HYDROXYMETHYLATION IN MOUSE TISSUE WITH NANOPORE SEQUENCING

Luke B Morina[1], Jessica Hosea[1], Sheridan Cavalier[1], Paul Hook[1], Winston Timp[1,2]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, Molecular Biology and Genetics, Baltimore, MD

The epigenome is pivotal in mammalian gene regulation and often becomes dysregulated in disease. However, it is still not understood how epigenetic patterns are regulated—especially how methylation is lost at specific loci. Most methylation studies in mammals have focused on 5-methylcytosine (5mC), while 5-hydroxymethylcytosine (5hmC) remains poorly characterized because it is less abundant and more difficult to detect. We reasoned that tissues from well-established inbred mouse models offer the closest approximation to a "Genome in a Bottle" because, unlike dividing cultured cells, they maintain realistic levels of 5hmC. Here we present the most comprehensive profiling of 5mC and 5hmC to date, combining short- (NEB EMseq/E5hmCseq) and nanopore sequencing to generate high-coverage (>50X) datasets across four mouse tissues with distinct epigenetic landscapes: cortex, cerebellum, liver, and heart.

Our analysis revealed strong concordance (Pearson r > 0.9) between short-read and nanopore results. We identified two classes of differentially modified regions between tissues: (1) differentially modified regions (DMRs), reflecting total cytosine modification and serving as a proxy for bisulfite-based assays; and (2) differentially hydroxymethylated regions (DhMRs), specific to 5-hydroxymethylcytosine. Genome-wide comparison revealed ~75% of DhMRs were undetectable when considering total cytosine modification alone. Furthermore, most DhMRs were identified in tissue-specific genes with higher gene expression reflecting higher 5hmC frequency. This suggests a connection between active regulation/transcription and hydroxymethylation, even in tissues with low global 5hmC abundance.

Our results highlight the importance of 5hmC in different tissues for gene regulation and epigenetic control, while also providing a benchmark for future method development.

GENE REGULATORY SIGNATURES OF ARCHAIC INTROGRESSION ACROSS HUMAN TISSUES AND CELL TYPES

Kitty B Murphy, Laurits Skov

Globe Institute, Molecular Ecology and Evolution, Copenhagen, Denmark

Archaic humans, including Neanderthals and Denisovans, contributed to modern human genomes multiple times through a process known as archaic introgression. These introduced archaic DNA fragments have shaped human evolution and genetic diversity, and with the increasing availability of omics datasets, there is now an opportunity to infer the regulatory impacts of these fragments across many tissues and cell types. The existence of introgression deserts, regions strongly depleted of archaic DNA, suggests that some archaic genetic material is incompatible with modern human biology and potentially contributes to human uniqueness. Furthermore, as ancient DNA is preserved mainly from remains in colder climates, we may never be able to study the molecular profiles of archaic populations from warmer regions like South Asia. Here, we leveraged the large-scale sequencing efforts of GTEx (n=800) and UK Biobank (n=500,000) to generate maps of archaic introgression and introgression deserts. Using a reference-free method for detecting archaic introgression, we identify all archaic groups that contributed DNA to modern genomes. Using gene expression data for >50 tissues, scRNA-seq, methylation patterns, and H3K27ac data, we explored whether introgressed DNA and introgression deserts are associated with gene regulatory changes across specific tissues, including skin, testes, and brain regions, and cell types, such as T cells of the immune system and melanocytes of the skin.

REPRODUCIBLE AND RESONSIBLE USE OF AGENTIC AI WITH GALAXY FOR GENOMIC DATA ANALYSIS

Dannon Baker[1], Danielle Callan[2], Marius Van Den Beek[3], Junhao Qiu[4], David Rogers[5], Aysam Guerler[1], John Chilton[3], Hiram Clawson[6], Scott Cain[3], Teresa O'Meara[7], Kelsey Beavers[8], Michael Schatz[1], Maximilian Haeussler[6], Bjorn Gruning[9], Jeremy Goecks[4], Sergei Kosakovsky Pond[2], Anton Nekrutenko[3]

[1]Johns Hopkins University, Dept. of Biology, Baltimore, MD, [2]Temple University, Dept. of Biology, Philadelphia, PA, [3]The Pennsylvania State University, Dept. of Biochemistry and Molecular Biology, University Park, PA, [4]Moffitt Cancer Center, Genomics, Tampa, FL, [5]Clever Canary, LLC, Santa Cruz, CA, [6]University of California, Santa Cruz, Baskin School of Engineering, Santa Cruz, CA, [7]University of Michigan, Dept. of Microbiology and Immunology, Ann Arbor, MI, [8]The University of Texas, Texas Advanced Computing Center, Austin, TX, [9]Albert-Ludwigs-University Freiburg, Dept. of Bioinformatics Freiburg, Germany

AI agents that autonomously write and execute code are the most transformative tools to hit biological research in years. They are also a reproducibility disaster waiting to happen. The community is racing to build new models and tout their capabilities while paying almost no attention to how these tools are actually used. Conversations vanish. Intermediate results disappear. The same prompt yields different analyses on different days. If we do not act now, agentic AI will unleash a fresh wave of irreproducibility on a field that is already struggling with it.

Power and reproducibility do not have to be incompatible. We integrate agentic AI with Galaxy (https://usegalaxy.org)—an open platform with 20 years of provenance tracking and over 12,000 analysis tools running on free public infrastructure. Galaxy handles computationally intensive primary analysis through standardized workflows; AI agents interpret secondary datasets through iterative refinement captured in versioned notebooks. Agents connect to Galaxy via the Galaxy MCP server and Galaxy skills—structured instructions that enforce best practices and prevent common mistakes.

We validated this with three studies spanning variant interpretation, custom tool creation, and full experiment reproduction. An agent characterized variants across 5,000+ Candidozyma auris surveillance samples and tracked clade dynamics over 3.5 years. It wrote a custom annotation tool for overlapping coding regions in measles virus—a problem no existing predictor handles. It reproduced a published RNA-seq experiment end-to-end, independently discovering experimental design from scattered metadata and matching published results.

Every artifact—plans, notebooks, histories—is public. We propose concrete guidelines: detailed plans over vague prompts, file-based interaction over chat, rerunnable notebooks, and full provenance preservation. The tools are here. The question is whether we use them responsibly.

CHARACTERIZING THE IMPACT OF INDUSTRIALIZATION ON HOST GENETIC-MICROBIOME INTERACTIONS IN HUMAN INTESTINAL ORGANOIDS

Shreya Nirmalan[1], Sabrina Arif[2], Adnan Alazizi[1], Gabrielle Garlicki[3], Henriette Mair-Meijers[1], Mathilde Poyet[4,5,6], Mathieu Groussin[4,5,7], Roger Pique-Regi[1,8], Ran Blekhman[2], Francesca Luca[3]

[1]Wayne State University, Molecular Medicine and Genetics, Detroit, MI, [2]University of Chicago, Genetic Medicine, Chicago, IL, [3]University of Chicago, Human Genetics, Chicago, IL, [4]Global Microbiome Conservancy, microbiomeconservancy.org, Cambridge, MA, [5]Massachusetts Institute of Technology, Biological Engineering, Cambridge, MA, [6]Kiel University, Institute of Experimental Medicine, Kiel, Germany, [7]Kiel University, Institute of Clinical Molecular Biology, Kiel, Germany, [8]Wayne State University, Obstetrics and Gynecology, Detroit, MI

Industrialization is associated with shifts in gut microbiome composition and increased burden of non-communicable diseases. Prior studies identified genotype-by-microbiome (GxM) interactions in intestinal gene expression in vivo and differential colonic epithelial transcriptional responses to urban versus rural microbiomes in vitro. However, whether lifestyle-associated microbial communities modulate host gene regulation and thus disease risk remains unresolved. We developed a live microbiome-colonic organoid co-culture system to interrogate interactions between host genetic variation and the microbiome. Primary human colonic epithelial organoid lines (colonoids) from five healthy donors were co-cultured with eight live microbial communities representative of urban and rural populations from Ghana and Rwanda,(59 samples including controls), followed by bulk RNA-sequencing. We identified 823 host genes differentially expressed in response to microbiome exposure (FDR=10%), which were enriched for type I interferon, antiviral response, and extracellular structure organization. Responses varied substantially across colonoid lines (57 to 3054 differentially expressed genes per line). In addition, we found that urban microbiomes elicited a stronger transcriptional response compared to rural microbiomes. We performed allele specific expression (ASE) analysis using QuASAR with multivariate adaptive shrinkage (mash) and identified 14,682 SNPs in 4,106 genes with shared ASE (lfsr < 0.05) and 1,360 SNPs in 1,043 genes that exhibit conditional ASE in response to microbiome exposure indicating microbiome-dependent allelic effects on gene expression. Together, these results extend prior in vivo and in vitro findings by demonstrating that lifestyle-associated differences in microbiome composition shape host intestinal transcriptional response and modulate host allelic effects on transcription in human colonic organoids.

INTERACTIVE VISUALIZATION OF WHOLE GENOME ALIGNMENTS AND PANGENOMES USING NCBI CGV AND MCGV

Dong-Ha Oh, Dmitry Rudnev, Sanjida H Rangwala, Andrea Asztalos, Evgeny Borodin, Vadim Lotov, Marina Omelchenko, Joël Virothaisakun, Vamsi Kodali

National Center for Biotechnology Information, Information Engineering Branch, Bethesda, MD

NCBI offers a set of interactive browsers to visualize and explore whole genome alignments and pangenomes derived from across the eukaryotic tree of life.
The Multiple Comparative Genome Viewer (MCGV; https://ncbi.nlm.nih.gov/mcgv) allows users to browse multiple genome alignments and pangenomes as linear alignment tracks. Genomes in the alignment and pangenome are anchored to a reference, with an option to switch the anchor among a set of selected references. The “Sequence conservation” panel summarizes the degree of conservation for each region or position in the anchor assembly. The “Assembly alignments” panel shows the synteny blocks or aligned nucleotide sequences depending on the zoom level. Users can color the synteny blocks by either the average sequence identity compared to the anchor, the sequence orientation, or the chromosome aligned. These options facilitate exploration of both sequence divergence and structural variation. Gene models, including exon and intron structures, are displayed for all annotated genomes, providing functional context to the alignment data.
To illustrate a variety of use cases, MCGV currently hosts five diverse alignment sets. Alignments range from the default set of 33 eutherian mammals (Ensembl EPO) to 8 telomere-to-telomere primate assemblies (NHGRI), collections of fungal pathogens (*Candidozyma* and *Cryptococcus*), and 38 insect species from NCBI RefSeq. As we incorporate alignments from the Vertebrate Genomes Project (VGP), we continue to invite users to propose additional datasets for inclusion in MCGV.
For research questions requiring more detailed comparison of a pair of genomes, the Comparative Genome Viewer (CGV; https://ncbi.nlm.nih.gov/cgv) offers interactive browsing of >1450 pairwise alignments across >650 species, the vast majority of which were added at user request. CGV is interconnected with the Genome Data Viewer (GDV; https://ncbi.nlm.nih.gov/gdv), NCBI’s genome browser, which can display any number of pairwise genome alignment tracks in detail at sequence resolution, along with other tracks such as gene expression, variation, and user-provided data. We encourage users to submit requests to include additional genomes in CGV using the “Feedback” button on the webpage.

This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health (NIH). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

VARIATION IN CARDIOTONIC STEROID RESISTANCE IN *D. MELANOGASTER* AND ITS IMPLICATIONS

Naima Okami*, Flora Borne*, Julia Holder, Miyoung Jang, Arya Rao, Peter Andolfatto

Columbia University, Department of Biological Sciences, New York, NY

*Contributed equally

The repeated evolution of toxin resistance is arguably among the most compelling and informative examples of adaptive evolution. However, such systems have largely been studied as monogenic traits, focusing on adaptation of the toxin's molecular target. A well-studied instance of this phenomenon is the repeated evolution of resistance to dietary cardiotonic steroids (CTS) in animals. Work in this system has highlighted the role of convergent adaptive amino acid substitutions and neofunctionalization of sodium–potassium ATPase (NKA), the target of CTS inhibition. While CTS resistance is often depicted as a simple trait determined by NKA sequence, CTS resistance—at the level of whole-organism fitness—likely involves a complex interplay of cell biology, physiology, and behavior, each involving many genes and substantial standing genetic variation. *Does this "background" variation for CTS resistance confound our ability to relate adaptive evolution of the NKA to resistance in vivo? To what extent does background variation contribute to CTS resistance evolution at the level of whole organisms?*

Despite a lack of ecological exposure to CTSs, we find that wild-derived *D. melanogaster* strains display near-continuous variation in CTS resistance, potentially consistent with a highly polygenic basis. Indeed, while this variation in CTS resistance is highly heritable, we find no evidence for common large-effect variants driving resistance. A differential gene expression analysis highlights the potential contribution of several mechanisms to CTS resistance variation, including known detoxification pathways and barriers to toxin ingestion and absorption. Finally, CTS resistance has previously been studied by engineering NKA substitutions into isogenic *Drosophila* lines without considering the effects of genetic background. We show that genetic background can indeed alter the inferred effect size of NKA variants *in vivo*. Together, these findings highlight the potential importance of background variation both in interpreting putative adaptive amino acid substitutions and in the expected dynamics of these variants during trait evolution.

MULTIOMIC ANALYSIS OF CIRCADIAN AND SEASONAL BIOMARKERS

Lea Urpa[1], Nasa Sinnott-Armstrong[2], Finngen FInnGen[1], Hanna M Ollila[1,2,3,4]

[1]University of Helsinki, Institute for Molecular Medicine Finland, FIMM, HiLIFE, Helsinki, Finland, [2]Fred Hutchinson Cancer Center, Herbold Computational Biology Program, Seattle, WA, [3]Massachusetts General Hospital, Department of Anesthesiology, Boston, MA, [4]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA

Biological processes exhibit dynamic regulation over time, particularly across circadian (daily) and seasonal cycles. However, the genetic and temporal architecture of such oscillatory processes in human clinical biomarkers remains poorly characterized. Leveraging the large-scale FinnGen cohort (N = 500,314) integrated with the national digital health infrastructure, for in- and outpatient diagnoses, primary care, laboratory measurements and demographic information, we systematically investigated time-of-day and time-of-year effects on routine clinical laboratory measurements and proteomics across over 200 biomarkers. Furthermore, we validate the associations in clinical data from the UK Biobank (N = 420,000) and MassGeneralBrigham Biobanks (N = 52,000).

For each biomarker measured in at least 1,000 individuals, we compared two nested machine learning models and classical epidemiological multivariable regression models: one accounting for covariates (age, sex, disease endpoints), and a second model additionally incorporating time-of-day and time-of-year of sampling. We discovered that 167 (44.1%) biomarkers showed significant temporal variation, including glucose, albumin, calcium, lipid levels, creatinine, and leukocyte counts. In several cases, time-of-sampling explained up to 5.48% of observed variance, suggesting biologically meaningful temporal regulation. Furthermore, genome-wide interaction analyses identified genetic variants whose effects on biomarker levels varied over time. Notably, we identified significant time-dependent genetic effects in the melatonin receptor gene MTNR1B (rs10830963, p=1e-22) and the circadian regulator CRY2 (rs12419690, p=1e-16), where variant effects were modulated by time-of-day of sampling.

These findings demonstrate that routine clinical biomarkers exhibit widespread and biologically meaningful temporal variation, which can obscure or enhance genetic effects depending on sampling time. By integrating the dimension of time into genetic analyses, we uncover novel gene-by-time interactions and identify circadian-linked loci that modulate biomarker levels. Our results highlight the importance of accounting for temporal context in biobank-scale studies and provide new insights into how circadian biology intersects with human physiology, disease risk, and personalized medicine.

GRAMENE: ADVANCING PLANT PAN-GENOME RESOURCES AND COMMUNITY STANDARDS

Andrew Olson[1], Sunita Kumari[1], Xuehong Wei[1], Kapeel Chougule[1], Zhenyuan Lu[1], Peter Van Buren[1], Audra Olson[1], Suyun Kim[1], Janeen Braynen[1], Lifang Zhang[1], Nicholas Gladman[1,2], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]USDA-ARS, Robert Holley Center, Ithaca, NY

Since 2001, Gramene has served as a major resource for comparative plant genomics, supporting research in genetics, breeding, systems biology, and evolution. With over 1,500 citations and a global user base spanning more than 100 countries, Gramene provides integrated access to genome annotations, genetic variation, pathways, regulatory networks, environment–variant associations, and gene expression data through coordinated platforms including Ensembl, Plant Reactome, CLIMtools, EMBL-EBI Expression Atlas, and BAR.

Rebranded as Gramene Plants to reflect its expanded taxonomic scope, Release 69 includes 233 reference genomes, curated pathways for 139 species, expression data from 1,026 studies across 27 species, and variation data mapped to 27 genomes representing 19 species. Recent updates include integrated expression visualization tools, a literature-curated catalog of gene functions, and a Germplasm tab linking accessions with loss-of-function alleles to public seed repositories.

A major recent focus has been the expansion of crop pan-genome resources, including GrameneMaize, GrameneOryza, GrameneGrapevine, and SorghumBase. These portals enable exploration of structural variation, gene presence/absence variation, gene expression and gene family evolution across diverse germplasm. Integration of standard reference SNP identifiers (rsIDs) further enhances reproducibility, interoperability, and alignment with FAIR data principles.

In addition to data integration, Gramene actively supports the research community through workshops, webinars, and training initiatives in FAIR data practices, biocuration, and data management. Together, Gramene Plants and its crop-specific pan-genome portals provide a scalable, community-driven infrastructure for advancing plant genomics research.
This work is supported by USDA-ARS grant 8062-21000-051-000D.

EPIGENETIC REGULATION OF GENE COPY-NUMBER VARIATION IN STICKLEBACK GENOMES

Michael J Olufemi[1], Sarah L Chang[2], Trevor J Krabbenhoft[2], Frédéric J. J Chain[1]

[1]University of Massachusetts, Biological Sciences, Lowell, MA, [2]University at Buffalo, Biological Sciences, Buffalo, NY

Gene duplication can facilitate adaptation to changing environments. However, most new duplicate genes are expected to disrupt stoichiometric relationships via gene dosage changes, resulting in their loss over time through purifying selection. How some genes that are not immediately adaptive might overcome these dosage effects and become preserved in the genome remains unclear. It has been observed in several species that many new gene duplications initially display a decrease in expression levels, seemingly reducing detrimental dosage effects. This decreased expression can be facilitated by epigenetic modifications like DNA methylation, which may rapidly balance dosage and allow new duplicates to persist as gene CNVs in a population. In this study, we leveraged high-quality genome assemblies and nanopore sequencing data from stickleback ecotypes to test whether genes CNVs are transcriptionally repressed by DNA methylation. We quantified promoter and gene-body DNA methylation across gills, gonads, muscle, and spleen and found that gene CNVs show tissue-dependent methylation, context-specific copy-number–methylation relationships, and frequent asymmetry in methylation between duplicate copies. By revealing differential methylation of gene duplications among individuals from ecotypes adapted to distinct environments, our findings support a model in which epigenetic buffering stabilizes duplicate genes against deleterious dosage effects while maintaining their capacity for subsequent adaptive divergence. These results expand our understanding of the evolutionary and tissue-specific dynamics that govern gene duplication and their potential to influence adaptation and speciation.

HIGH-FIDELITY LONG-READ SEQUENCING UNCOVERS TANDEM REPEAT VARIATION ASSOCIATED WITH VIRAL VIRULENCE

Alejandro Ortigas-Vasquez[1], Christopher D Bowen[1], Daniel W Renner[1], Moriah L Szpara[1,2], Anton Bankevich[3]

[1]The Pennsylvania State University, Department of Biology, State College, PA, [2]The Pennsylvania State University, Department of Biochemistry and Molecular Biology, State College, PA, [3]The Pennsylvania State University, School of Electrical Engineering and Computer Science, State College, PA

Herpesviruses like herpes simplex virus type 1 (HSV-1) are among the most ubiquitous pathogens in the world, infecting two out of three people. Despite their prevalence, there are no available vaccines for eight of the nine herpesvirus species that primarily infect humans. The most effective strategy to date has been the use of live-attenuated vaccines, which consist of a "weakened" form of the virus. However, the inability of single-nucleotide variants (SNVs) to fully explain phenotypic differences between attenuated and virulent herpesvirus strains greatly complicates the rational design of live-attenuated herpesvirus vaccines. In recent years, tandem repeats have become increasingly recognized as a major source of genomic variation, but one that is poorly resolved due to limitations of short-read sequencing technologies. In our recent work, we used ultra-deep (> 1,000x) high-fidelity long-read (PacBio HiFi) sequencing to characterize tandem repeat variation in five strains of HSV-1. To enable high-throughput assessments of tandem repeat diversity within each viral sample, we developed TRIDENT, a novel tool that leverages profile hidden Markov models to resolve tandem repeat architectures across thousands of individual PacBio HiFi reads. In addition to enabling comprehensive characterizations of HSV-1 tandem repeat loci, the combination of TRIDENT and long read sequencing uncovered a kaleidoscope of repeat architectures in HSV-1 genomes, with individual HSV-1 strains exhibiting markedly distinct variation patterns in several tandem repeat loci. These patterns not only varied in length and composition, but also in the extent of variation present within each viral sample. In two of the sequenced strains, these patterns represented the only DNA-level differences, potentially explaining their different virulence levels. We also found that many of the non-repetitive regions flanking HSV-1 tandem repeat loci were highly variable, showcasing the potential impact that repetitive DNA can have on adjacent regions over time. These results highlight tandem repeats as important contributors to viral virulence, and suggest that repeat-resolved genomics may provide new avenues for understanding herpesvirus biology and help guide rational vaccine design.

PANGENOME-BASED GENOTYPING OF STRUCTURAL VARIANTS IN MEDICAL COHORTS

Chiara Paleni[1], Davide Bolognini[1], Andrea Guarracino[2,3], Thomas S Dudley[1], Alessandro Raveane[1], Peter H Sudmant[4], Erik Garrison[3], Nicole Soranzo[1,5,6]

[1]Human Technopole, Population & Medical Genomics Research Centre, Milan, Italy, [2]The Translational Genomics Research Institute, Bioinnovation and Genome Sciences, Phoenix, AZ, [3]University of Tennessee Health Science Center, Department of Genetics, Genomics and Informatics, Memphis, TN, [4]University of California Berkeley, Department of Integrative Biology, Berkeley, CA, [5]University of Cambridge, Cambridge, United Kingdom, [6]Wellcome Sanger Institute, Hinxton, United Kingdom

Pangenome graphs built from long-read assemblies enable comprehensive variant discovery, particularly at highly complex and structurally variable loci such as the HLA. However, most large-scale biobanks rely on short-read sequencing. Moreover, translating pangenome-based genotyping to population and medical genomics applications, such as linking genotypes to molecular traits, remains challenging. We present an application of pangenomes for structural variant (SV) genotyping with short-read data in a Southern-Italian medical cohort, representative of an understudied Mediterranean population.

Our method, COSIGT, assigns haplotype sequences from pangenome graphs to short-read samples by comparing read coverage and haplotype copy-number vectors. We extended COSIGT to convert haplotype assignments into structural variant genotypes by decomposing the pangenome graph at variable regions, enabling downstream analysis with standard variant analysis tools. We applied this approach to the Moli-sani cohort, comprised of >6,000 short-read samples. Using the HPRCy2 pangenome reference, we benchmarked COSIGT on the hyper-variable HLA genes. We then genotyped structural variants across challenging, medically relevant genes (CMRGs).

For HLA typing, COSIGT shows comparable accuracy to specifically designed tools despite being a general-purpose approach. Over 47 CMRGs, we can genotype hundreds of structural variants in the Moli-sani cohort, demonstrating scalability to population-level datasets. Additionally, a large fraction of SVs are not strongly tagged by neighbouring SNPs, demonstrating that this approach can potentially discover new associations and help explain known ones.

Pangenome-based genotyping with COSIGT bridges the gap between pangenome resources and large-scale medical cohorts, enabling structural variant analysis and association studies at complex loci using widely available short-read sequencing data.

COMPREHENSIVE MAP OF MEDIATOR COMPLEX INTERACTOME ACROSS HUMAN CELL MODELS - ADDING THE PROTEIN LAYER TO GENE REGULATION

Petra Páleníková[1,2], Xuening He[1,3], Justus F Gräf[1,4], Travis Botts[1,2], Glen Munson[1], Makayla Martorana[1,2], Daya Mena[1,2], Danzel Rebelo[1,2], Judhajeet Ray[1], Paulina Strzelecka[1,2], Yu-Han Hsu[1,2], Greta Pintacuda[1,2], Robin Andersson[1,3], Elisa Donnard[1], Jesse M Engreitz[1,5], Kasper Lage[1,2]

[1]Broad Institute of MIT and Harvard, Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Cambridge, MA, [2]Broad Institute of MIT and Harvard, Stanley Center for Psychiatric Research, Cambridge, MA, [3]University of Copenhagen, Department of Biology, Copenhagen, Denmark, [4]University of Copenhagen, Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, Copenhagen, Denmark, [5]Stanford University, Department of Genetics and BASE Initiative, Stanford, CA

Gene expression in specific cell types is highly orchestrated and dependent on combinations of transcription factors (TFs) and cofactors that link enhancers and promoters to specific gene programs. However, the dynamic time- and cell-type-specific composition of the TF-cofactor protein complexes central to this process is poorly understood. To study general and cell-type-specific gene regulation, we built a protein-protein interaction (PPI) map of Mediator, a multi-subunit ubiquitously expressed cofactor complex. First, we identified proteins interacting with the Mediator complex by immunoprecipitations followed by mass spectrometry in K562, iPSC, neuronal progenitors, excitatory neurons and hepatic progenitor cells. We used a total of 11 Mediator subunits as baits, spanning head, middle, tail, and kinase modules, with two subunits analyzed across all cell models. The resulting PPI network contained more than 1000 proteins, most of which have not been reported as Mediator interactors previously. The network included several TFs and cofactors, some of which interacted in cell-type-specific manner. We validated functional relationships of proteins in the PPI networks by integration with predicted TF-TF cooperativity and published genome-wide Perturb-seq data. Nine out of 11 baits showed significantly higher correlation (FDR < 5%) with the transcriptional response to perturbation of their network proteins compared to perturbation of other expressed proteins. This orthogonal data integration allows us to better understand the combinatorial protein interaction logic of cofactors and TFs that orchestrate tightly regulated gene expression programs.

THE INFLUENCE OF DEMOGRAPHIC HISTORY AND GENETIC ARCHITECTURE ON COMPLEX TRAITS VIA RUNS OF HOMOZYGOSITY

Mingzuyu Pan, Zachary A Szpiech

Penn State University, Department of Biology, State College, PA

Runs of homozygosity (ROH) are contiguous genomic regions where all sites are homozygous, inherited from identical haplotypes due to shared ancestry. The number and length of ROH in individuals varies based on population history and sociocultural behaviors. Although often discussed in the context of inbreeding, ROH are ubiquitous in putatively outbred human populations, and their prevalence are associated with multiple complex traits, including height and measures of lung function. Importantly, ROH have been shown to be enriched for deleterious alleles, suggesting a mechanism by which ROH prevalence can influence traits. Here we employ realistic forward-in-time population genetic simulations and a flexible quantitative model of a generic complex phenotype to explore how population history and genetic architecture influence ROH associations with a generic quantitative phenotype. We show that ROH are important for all simulated demographic histories and genetic archi-tectures but especially when phenotypes have a recessive component. This is even more prominent when the rare-allele contribution to the phenotype is upweighted and in high-diversity populations (e.g. African). For a fully recessive phenotype, ROH can account for 25-45% of an individual's total phenotype score, depending on demographic history and rare-allele weight. Our results emphasize the utility of ROH in helping to explain phenotype variation across different population histories and genetic architectures.

REGENSEQ: AN ECOSYSTEM FOR HIGH-THROUGHPUT HIGH-CONTENT IMAGING USING DECOMMISSIONED SEQUENCERS.

Kunal Pandit[1], Craig Fouts[1,2,3,4], Sarah Rodwin[1], Karan Dhingra[1], Silas Maniatis[1], Jagjit Singh[2,3], Hemali Phatnani[1,6], Bianca Dumitrascu[3,5], Sanja Vickovic[1,2,3,4]

[1]New York Genome Center, Technology Innovation, New York , NY, [2]Columbia University, Biomedical Engineering, New York, NY, [3]Herbert Irving Institute of Cancer Dynamics, Columbia University, New York, NY, [4]Uppsala University, Immunology, Genetics and Pathology, Uppsala, Sweden, [5]Columbia University, Statistics, New York, NY, [6]Columbia University Irving Medical Center, Neurology, New York, NY

The analysis of spatially resolved proteomic data is critical for linking protein localization to cellular function and disease mechanism, yet current iterative multiplexed imaging platforms are often cost-prohibitive and lack scalable analytical tools for complex high content datasets. We developed ReGenSeq, an open-source high-content imaging ecosystem, by repurposing decommissioned Illumina sequencers to make high-content imaging accessible to any research laboratories. For a proof-of-concept, we automated ImmunoSABER, a multiplexed immunofluorescence assay utilizing DNA conjugated antibody to study the progression of amyotrophic lateral sclerosis (ALS) in a SOD1-G93A transgenic mouse model. An antibody panel targeting 29 early ALS biomarkers and relevant spinal cord cell types was developed to characterize nearly 100 spinal cord sections from transgenic and wild type animals across 3 time points. To process all the images we developed a Snakemake pipeline which segments single cells, and measures interpretable features such as antibody intensity, antibody texture, and cellular morphology. Interpretable features can also optionally be augmented with embeddings of cells from a vision transformer. With a combined set of over 10,000 features, we discovered spatial microenvironments using a single-cell embedded latent Dirichlet allocation (sceLDA) model that combines the feature scalability of deep generative models with the interpretability of topic modeling. The sceLDA model efficiently categorizes cells and assigns them into anatomical regions to determine the underlying cell type mixtures associated with disease states. Using sceLDA we discovered pre-symptomatic ALS cellular interaction niches characterized by molecular and morphological features. High-content imaging assays like ImmunoSABER are easily adaptable to the ReGenSeq ecosystem, which offers a complete solution for assay automation, image processing, and spatial microenvironment discovery.

DEVELOPMENT AND IMPLEMENTATION OF MOSASAUR: A NOVEL TOOL FOR THE ANALYSIS OF OXFORD NANOPORE LONG-READ SEQUENCING MODIFICATION DATA.

Lauren E Patterson[1], Kip D Zimmerman[2,3]

[1]Wake Forest University School of Medicine, Biomedical Research Master of Science Program, Winston-Salem, NC, [2]Wake Forest University School of Medicine, Biostatistics and Data Science, Winston-Salem, NC, [3]Wake Forest University School of Medicine, Internal Medicine, Winston-Salem, NC

Oxford Nanopore long-read sequencing confers great advantages over short-read methods for the assembly of complete genomes, haplotype phasing, and the detection of epigenetic modifiers such as methylation and hydroxymethylation. Current tools for the analysis of long-read methylation data lack robustness, and there is no single unified framework for performing statistical analyses. There are no current tools for the analysis of hydroxymethylation, particularly for the joint analysis of hydroxymethylation and methylation. We developed Mosasaur, an R-based tool for the analysis of Oxford Nanopore long-read modification data. Mosasaur is a flexible, user-friendly R-package that enables users to perform simulations of realistic CpG count data, calculate power for proposed studies, and perform statistical analyses with Oxford Nanopore long-read modification data. As we generate larger and more diverse datasets, we will continue to maintain and expand the package.

ON COALESCENT-BASED INTROGRESSION INFERENCE: THEORY, BIASES, AND SOLUTIONS

David Peede[1,2,3], Jazeps Medina Tretmanis[2], Léo Planche[4], Marco Rosario Capodiferro[5], Diego Ortega-Del Vecchyo[6], Emilia Huerta-Sánchez[1,2,5,7]

[1]Brown University, Department of Ecology, Evolution, and Organismal Biology, Providence, RI, [2]Brown University, Center for Computational Molecular Biology, Providence, RI, [3]Brown University, Institute at Brown for Environment and Society, Providence, RI, [4]Université Paris-Saclay, Interdisciplinary Laboratory of Digital Sciences, Orsay, France, [5]Trinity College Dublin, Smurfit Institute of Genetics, Dublin, Ireland, [6]Universidad Nacional Autónoma de México, Laboratorio Internacional de Investigación sobre el Genoma Humano, Juriquilla, Mexico, [7]Brown University, Data Science Institute, Providence, RI

The evolution of species has traditionally been viewed as a bifurcating process; however, genomic studies over the past 15 years have revealed that introgression is pervasive and a powerful force shaping genetic variation across the Tree of Life. Introgression is typically inferred from genomic sequence data using statistics that quantify discordance between marginal genealogies and an assumed species tree, often using site patterns or genetic distances.

Here, we use coalescent theory to derive the expected branch lengths for all six site patterns, along with expected pairwise coalescence times, under a single-pulse model with variable population sizes. These results allow us to derive and analytically study the expectations of previously proposed introgression statistics, including those proposed by verbal arguments and simulations. We show that all existing statistics are inherently demographically biased by unavoidable contributions from incomplete lineage sorting (ILS), and therefore cannot provide unbiased estimates of the introgression proportion. Motivated by this result, we introduce two new statistics. First D^*, the only statistic that is fully agnostic to ILS, provides a robust test for the presence of introgression. Second, f^*, dynamically corrects the bias in estimating the introgression proportion by approximating a demographic nuisance parameter, enabling more accurate inferences across diverse evolutionary scenarios in the absence of prior demographic knowledge.

We benchmark existing and new statistics for detecting and quantifying introgression at genome-wide and local scales using empirical human- and canine-like demographic models that incorporate heterogeneity in recombination and mutation rates. We find that D^*, consistently outperforms existing approaches for detecting introgression, while f^* substantially improves estimates of introgression proportions, especially when the demographic bias is severe. Finally, we provide practical guidelines for selecting appropriate methods based on prior demographic knowledge and the target of inference, further empowering studies of introgression across the Tree of Life.

AGE AND EARLY LIFE ADVERSITY SHAPE HETEROGENEITY OF THE EPIGENOME ACROSS TISSUES IN MACAQUES

R M Petersen*[1], B Sadoughi*[2], S K Patterson[3,4], M M Watowich[1], C R Kelsey[2], E A Goldman[5], Cayo Biobank Research Unit[6], A R DeCasien[7], K L Chiou[8], A V Ruiz Lambides[9], A D Melin[10], LJ N Brent[11], J P Higham[3], M J Montague[6], M L Platt[6], N Snyder-Mackler[2], A J Lea[1]

[1]Vanderbilt University, Biological Sciences, Nashville, TN, [2]Arizona State University, School of Life Sciences, Tempe, AZ, [3]NYU, Anthropology, New York, NY, [4]Notre Dame, Anthropology, Notre Dame, IN, [5]Oregon Health & Science University, Knight Cancer Institute, Portland, OR, [6]UPenn, Department of Neuroscience, Philadelphia, PA, [7]NIH, NIA, Bethesda, MD, [8]UA Birmingham, Biology, Birmingham, AL, [9]University of Puerto Rico, CPRC, Punta Santiago, PR, [10]U Calgary, Anthropology, Calgary, Canada, [11]U Exeter, Centre for Research in Animal Behaviour, Exeter, United Kingdom

Aging is universal, yet its pace varies among individuals and across organ systems within the same individual. Early-life adversity (ELA) is linked to age-related disease and reduced lifespan in humans, but its molecular contributions to aging heterogeneity remain unclear. We generated DNA methylation (DNAm) profiles across 14 tissues from 237 semi-free-ranging rhesus macaques (n = 2,485 total samples) with corresponding information about six naturally occurring measures of ELA. Age-associated DNAm differences varied in direction and magnitude across tissues, with many differences being unique to a single tissue (42%) and few being shared across all tissues (1.3%). DNAm clocks accurately estimated chronological age and revealed moderate, within-individual consistency across tissues, suggesting a coordinated age-related epigenetic state (19% variance explained by individual identity). ELA was associated with DNAm variation at 7,533 regions (198,740 CpGs), with maternal loss and adipose tissue showing the strongest effects. For a given ELA, DNAm differences were broadly shared across tissues, whereas different ELA measures targeted largely distinct CpGs, indicating pathway-specific responses with cross-tissue coordination. Tissue-dependent ELA effects were enriched near transcription start sites (FDR < 0.001), and were strongest in tissues with long-lived cell types, and immune and endocrine tissues ($p < 10\text{-}4$), suggesting functional relevance. Although age and ELA overlapped at many genomic regions, ELA did not consistently accelerate epigenetic age (FDR = 0.67–0.75). Nevertheless, both age- and ELA-associated regions were enriched for loci linked to human aging and mortality (FDR < 0.05), highlighting organism-level relevance. Together, these findings advance our understanding of how early environments sculpt the molecular foundations of aging and underscore the need to integrate developmental context to explain aging heterogeneity.

*Equal contribution as first author

LEVERAGING POPULATION GENETICS TO IMPROVE RARE VARIANT INTERPRETATION IN dbSNP

Lon Phan, Qiang Wang

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

The dbSNP database (Build 157) catalogs >1.2 billion reference SNPs (rsIDs) and now represents a comprehensive archive of common and rare human variation. Interpreting rare variants at scale remains challenging due to technical artifacts, complex population structure, and limited functional context. We developed a population-genetic framework to characterize the functional landscape of the human genome using allele frequency data from >409,000 individuals in NCBI ALFA (R4) integrated with dbSNP.

Our multi-phase workflow begins with stringent quality control by blacklisting artifact-prone genomic regions (e.g., MHC, IGH) to reduce false signals from Hardy–Weinberg disequilibrium in structurally complex loci. We then performed genome-wide locus characterization across genetically distant population pairs using two complementary metrics: allele frequency differentiation (Fst) and genotype proportion differentiation (absolute difference in Fis). High-confidence “functionally constrained” loci were defined as 100 kb windows ranking in the top 0.1% for both metrics, identifying regions under strong purifying selection.

Pathogenic ClinVar variants are enriched within these constrained loci and display Fis distributions consistent with evolutionary intolerance, whereas Benign and VUS variants largely reflect neutral population structure. These results demonstrate that pathogenic alleles preferentially reside in constrained genomic regions and provide a data-driven strategy for prioritizing novel rare variants discovered in sequencing studies.

We further show that population-specific baseline homozygosity and structure influence variant interpretation and model transferability. Rare homozygotes in populations with elevated background homozygosity have different prior probabilities of pathogenicity than in more outbred populations. Network analyses of genetic proximity illustrate why polygenic risk score (PRS) models trained in one ancestry group often underperform in genetically distant populations.

Together, this work establishes a scalable framework for generating a genome-wide functional constraint map from population-scale variation. By integrating evolutionary signals with rigorous QC, we provide a principled method to prioritize rare variants, identify ClinVar assertions discordant with population-genetic evidence, and support development of population-aware disease risk models.

REPRODUCIBLE AUTOSOMAL GENE EXPRESSION CHANGES WITH LOSS OF TYPICAL X AND Y COMPLEMENT ACROSS TUMOR TYPES

Seema B Plaisier[1], Robert Phavong[2], Mason Farmwald[2], Teagan Allen[2], Malli Swamy[2], Ilsa Rodriguez[2], MacKenzie Wells[2], Nadia Phaneuf[2], Susan C Massey[2], Jared Del Rosario[2], Juvelyn Hart[2], Alexander Magelsdorf[2], Martin Van Der Jagt[2], Alex R DeCasien[3], Kenneth H Buetow[2], Melissa A Wilson[1]

[1]National Institutes of Health, National Human Genome Research Institute, Bethesda, MD, [2]Arizona State University, School of Life Sciences, Tempe, AZ, [3]National Institutes of Health, National Institute of Aging, Bethesda, MD

Although there are known sex differences in cancer incidence, severity, and treatment, the sex chromosomes are typically excluded from genomic analyses due to technical challenges assessing their copy number, sequence variation, and expression. Using sex chromosome gene expression and copy number in primary human tumor tissues from The Cancer Genome Atlas, we examined the gene expression profile of tumors across tissue types that showed atypical sex chromosome complements including loss of chromosome X (LOX) and reactivation of the inactive X chromosome (XaXa) in tumors from female patients and loss of chromosome Y (LOY) from male patients. Across several tissue types, tumors from male patients with LOY and tumors from female patients who have LOX (both X0 genotype) have more similar gene expression profiles than their typical sex chromosome complement counterparts. LOX and LOY largely reduces the number of differentially expressed genes between tumors from different patient sexes, affecting sex chromosomal and autosomal gene expression. We identified genes differentially expressed consistently between tissue types for each sex chromosome alteration compared to the karyotypic complement. XaXa had the most genes consistently differentially expressed; the profile included both sex-linked and autosomal genes across the genome, including genes known to be involved in the hallmarks of cancer, druggable genes, and genes with molecular functions relevant to cancer signaling, such as kinase activity. Going forward, considering patient sex as well as the entire genome, including the sex chromosomes, will provide additional insights into personalized tumor etiology, progression, treatment, and patient outcome.

DISSECTING GENETIC EFFECTS ON GENE REGULATORY MECHANISMS WITH SINGLE-MOLECULE FOOTPRINTING

Kaixuan Luo*[2], Ayelen Lizarraga*[1], Xiaotong Sun[2], Diana Vera Cruz[1], Xin He[2], Sebastian Pott[1]

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

*Contributed equally

Gene regulation is a highly dynamic process tightly controlled in time and space, mediated through interactions of transcription factors (TFs), nucleosomes, and DNA methylation. Effects of non-coding genetic variants on these gene regulatory features and how they contribute to disease traits have been studied extensively. However, the precise molecular mechanisms by which these variants affect gene expression remain poorly understood in most cases. To address this gap, we used single-molecule footprinting (SMF) to map effects of genetic variants on chromatin organization of individual molecules. SMF combines in vitro DNA methyltransferase treatment and direct long-read sequencing to simultaneously capture chromatin accessibility, DNA methylation, and TF binding across intact 10-20 kb molecules. SMF data capture multimodal, coordinated gene regulatory mechanisms and are thus uniquely suited to identify genetic variants affecting multiple regulatory modalities. We generated SMF data in a cohort of 30 human lymphoblastoid cell lines with a combined genome-wide coverage >700-fold. Using these data, we identified 6,655 regulatory elements exhibiting allele-specific accessibility, with most variants located directly within the regulatory region. Notably, in most cases these allele-specific effects corresponded to changes in the fraction of accessible molecules between the two alleles. To test whether variants directly affected binding of TFs, we identified footprints of TF binding (sized 10-30bp) within single molecule accessible regions and found that 1,657 of these showed allele-specific differences.
Our analysis demonstrates that SMF detects genetic effects on multiple regulatory features simultaneously. Motivated by this observation, we are extending this framework to identify multimodal quantitative trait loci (QTLs) and to detect genetic variants associated with novel regulatory modes including accessibility variability, nucleosome positioning patterns, and co-accessibility between distant elements. These single-molecule chromatin QTLs have the potential to reveal regulatory mechanisms previously inaccessible with other experimental approaches.

PREDICTING HOSPITAL-ACQUIRED INFECTION RISK THROUGH MULTI-OMIC INTEGRATION OF ELECTRONIC HEALTH RECORDS, GUT MICROBIOME AND METABOLOME

Sambhawa Priya[1], Ashwin Chetty[1], Christopher Lehmann[1], Matthew Odenwald[1], Dinanath Sulakhe[2], Bhakti Patel[1], Brett K Beaulieu-Jones[1], Eric Pamer[1,2], Ran Blekhman[1]

[1]University of Chicago, Department of Medicine, Chicago, IL, [2]University of Chicago, Duchossois Family Institute, Chicago, IL

Hospital-acquired infections are a major cause of morbidity, mortality, and healthcare costs worldwide. Yet existing prediction models, which predominantly rely on clinical risk factors derived from electronic health records, show inconsistent performance and limited biological insight. The gut microbiome is increasingly recognized as a key determinant of infection susceptibility, mediating colonization resistance and immune modulation, and undergoing alterations during hospitalization. Here, we leverage longitudinal gut metagenomics, metabolomics, and electronic health record (EHR) data from 3,335 fecal samples and 180,394 patient visits across 1,275 hospitalized patients at the University of Chicago Medical Center. We developed a machine learning framework integrating multi-omics factor analysis, sparse survival learning, and deep learning to jointly model gut microbiome composition, functional pathways, and metabolites alongside longitudinal EHR data, including diagnoses, medications, lab values, vital signs, and procedures. We identified latent factors capturing shared variation between microbiome features and clinical variables, characterized by loss of gut microbial diversity and depletion of anaerobic carbohydrate degraders, such as *Bacteroides thetaiotaomicron* and *Bacteroides ovatus*, alongside functional pathways that maintain gut barrier integrity and immune regulation. These microbial features co-varied with clinical markers of systemic metabolic stress, including glucose dysregulation, reduced renal function, and prolonged hospitalization. We then evaluated how microbiome and clinical features jointly predict 30-day hospital-acquired bloodstream infection risk. Clinical features associated with elevated risk included laboratory markers of hepatic and renal dysfunction (elevated bilirubin and blood urea nitrogen), frequent inpatient encounters, and invasive procedures such as mechanical ventilation. Microbial features further refined risk stratification, with opportunistic pathogens including *Pseudomonas aeruginosa* and *Enterococcus faecium* associated with increased risk. Conversely, commensal microbes linked to short-chain fatty acid production and colonization resistance, such as *Bifidobacterium longum* and *Akkermansia muciniphila*, showed protective effects. Overall, our results demonstrate microbiome–clinical interactions during hospitalization and identify microbial taxa and pathways that may modulate susceptibility to infection.

MAPPING DRUG RESPONSE AND TOXICITY ACROSS HUMAN CELL TYPES USING HETEROGENEOUS DIFFERENTIATING CULTURES

Henry W Raeder[1], Katherine Rhodes[2], Hae Kyung Im[1,2], Yoav Gilad[1,2]

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

Patient responses to pharmaceutical interventions are highly heterogeneous: while some individuals derive substantial therapeutic benefit with minimal side effects, others experience mild to severe adverse reactions. Severe adverse events pose a significant risk to susceptible individuals and, when identified, can derail clinical trials or lead to post-market drug withdrawals. These outcomes not only endanger patients but also deny effective therapies to non-susceptible populations and result in billions of dollars in lost resources. The ability to predict individual responses to drug exposure *a priori* would improve patient safety and treatment efficacy while streamlining the drug development pipeline.

Gene expression has emerged as a powerful predictor of drug response and can provide insight into underlying mechanisms of action. However, many pharmacogenomic studies rely on immortalized human cell lines, which poorly recapitulate normal human tissue; model organisms such as mice, which often fail to translate to human outcomes; or primary human tissues, which are difficult to scale across diverse cell types and drug perturbations.

To address these limitations, we leverage heterogeneous differentiating cultures (HDCs), iPSC-derived organoid systems that model diverse human cell types within a single experimental framework. We generated single-cell gene expression profiles from HDCs comprising 39 cell types following exposure to 86 compounds with well-characterized toxicity profiles, yielding a dataset of approximately 4.1 million cells. These cell types span all three germ layers and a continuum from pluripotent to fully differentiated states, enabling a holistic interrogation of drug response and toxicity across human tissues.

Initial analyses show that differential expression between untreated and drug-treated cells captures both global and cell type–specific responses, including generalized stress signatures and drug-specific transcriptional programs. For example, cardiac cell types exposed to rofecoxib exhibit downregulation of genes involved in prostaglandin synthesis and extracellular matrix maintenance, consistent with known COX-2 inhibition and proposed cardiotoxic mechanisms.

Together, these results suggest that HDCs can provide mechanistic insight into drug action and support predictive modeling of toxicity. This framework may complement existing preclinical models and offers a scalable, human-centered platform for studying both pharmaceutical and non-pharmaceutical exposures, including common environmental chemicals and metabolic stressors.

SCALABLE AND INTERPRETABLE MPRA-BASED PREDICTION OF REGULATORY VARIANT EFFECTS

Mahmudur Rahman Hera[1], Jiayi Liu[1,2], Anat Kreimer[1,3]

[1]Rutgers, the State University of New Jersey, Center for Advanced Biotechnology and Medicine, Piscataway, NJ, [2]Rutgers, the State University of New Jersey, Graduate Program in Cell and Developmental Biology, Piscataway, NJ, [3]Rutgers, the State University of New Jersey, Department of Biochemistry and Molecular Biology, Piscataway, NJ

Characterizing the functional consequences of noncoding genetic variation is a central problem in genomics, with applications in fine-mapping disease loci, understanding molecular evolution, and prioritizing regulatory variants for experimental follow-up. High-throughput experiments such as massively parallel reporter assays (MPRAs) offer a scalable way to systematically measure cis-regulatory function across thousands of candidate elements and variants in relevant cellular contexts. However, MPRA experiments cannot exhaustively investigate all variants across all cell types and environments, motivating the development of predictive models that generalize to unseen data.

Existing MPRA-driven variant-effect prediction approaches fall into two broad categories. First, ``meta-analysis/feature engineering'' approaches construct large sets of sequence- or annotation-derived predictors and fit classical machine-learning models. Such models have been used in predicting variant effects in the perturbation MPRA setting, where changes in regulatory activity are assayed by systematically altering sequences (e.g.\ altering transcription factor binding sites) within putative regulatory regions. Models built on differences between wild-type and perturbed sequences have been shown to be useful for both classification and regression for such settings, yet there is room for improvement in the performance of these feature-driven models in the case of SNP-scale allelic effect prediction, where reference and alternate alleles are separated by a small edit distance, and both occur in natural populations.

The second category of approaches tries to characterize non-coding variant effects using deep learning. These models typically rely solely on the sequence and use CNN-based deep networks to predict variant effects. These deep learning models can achieve strong generalization, but they often require substantial training data and compute, and their learned representations can be difficult to map to interpretable biological features.

Here, we present an efficient framework for SNP-scale variant effect prediction from MPRA using an ensemble of gradient-boosted models. By leveraging biologically grounded feature sets derived from paired reference/alternate sequences, our approach achieves performance comparable to deep models, yet requires significantly less training data and compute. In addition to reduced training requirements, our approach is more robust when projected to new cell types where MPRA data are scarce, and it yields direct feature attributions to facilitate biological interpretation and hypothesis generation.

LOSS-OF-FUNCTION VARIANTS IN KEY GENES ATTENUATES POLYGENIC EFFECTS ON LDL CHOLESTEROL

Gouri Rajaram[1], Yanina Kuzminich[1], Sylvia Dai[1,3], Hakhamanesh Mostafavi[1,2]

[1]New York University School of Medicine, Center for Human Genetics and Genomics, New York, NY, [2]New York University School of Medicine, Department of Population Health, New York, NY, [3]New York University Abu Dhabi, Division of Science, Abu Dhabi, United Arab Emirates

Characterizing gene-by-gene (GxG) and gene-by-environment (GxE) interactions is critical for understanding complex trait variation and improving disease prediction and treatment. However, identifying such interactions remains challenging, in part due to difficulties in quantifying environmental exposures and the multiple testing burden arising from limited prior knowledge of relevant genotypes or exposures.

We propose a model-driven approach to study GxG interactions. Specifically, we consider a model in which a small number of key genes mediate all effects on complex traits, such that inactivation of these genes, for example due to loss-of-function (LoF) mutations, nullifies the effects of their upstream regulators, including polygenic background and environmental factors.

As a proof of concept, we studied low-density lipoprotein (LDL) cholesterol levels in the UK Biobank. Consistent with the model, we observed that the effect of polygenic background on LDL is systematically reduced in individuals carrying LoF variants in LDL-regulating genes such as APOB and LDLR. In addition, applying our interaction test genome-wide, we identify other genes with similar behavior including ABCA10, SETD1B, CREB3L1, and UGT2B10, which are genes with both established and previously unrecognized roles in LDL metabolism.

We plan to build on this framework by developing kinetic models of key physiological processes involved in LDL homeostasis, in order to explore which classes of mutations in key genes lead to interaction effects. More broadly, this study highlights the value of studying well-characterized biological systems to understand how interactions arise, which may ultimately guide the search for novel biology through GxG and GxE interactions.

MULTI-ANCESTRY MAPPING OF GENETIC EFFECTS ON SPLICING IN 10,000 HUMAN BRAIN SAMPLES REVEALS NOVEL MEDIATORS OF NEUROLOGICAL DISEASE RISK

Aline Réal[1,2], Kailash BP[3], Winston H Dredge[3], Derek Lamb[4], Benjamin Z Muller[3], Beomjin Jang[3], Alex Tokolyi[1,2], Hong-Hee Won[5], Brielin Brown[4], Jack Humphrey[3], Towfique Raj[3], David A Knowles[1,2]

[1]New York Genome Center, New York Genome Center, New York, NY, [2]Columbia University, Department of Computer Science, New York, NY, [3]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences, New York, NY, [4]University of Pennsylvania, Philadelphia, Department of Biostatistics, Epidemiology & Informatics, Philadelphia, PA, [5]Samsung Genome Institute, Samsung Medical Center, Seoul, South Korea

Alternative splicing shapes isoform diversity and gene dosage, yet how genetic variation impacts splicing in brain disease remains incompletely characterized. We assembled BigBrain, a multi-ancestry resource of 10,725 bulk RNA-seq profiles with matched genotypes from 4,656 individuals across 43 tissue–cohort pairs, and mapped 52,696 cis-sQTLs affecting 10,833 genes using random-effects meta-analysis. By integrating signals across ancestries while accounting for cohort-specific heterogeneity, this framework increases power to detect shared regulatory effects while preserving ancestry differences.

Using SuSiE, we fine-mapped over half of these sQTLs into 95% credible sets, frequently to a single variant near splice sites. Fine-mapping leveraged linkage disequilibrium structure to refine candidate causal variants enriched near canonical splice donor and acceptor sites. We further annotated variants predicted to alter dosage through frameshifts, nonsense-mediated decay, or disruption of protein domains, highlighting widespread isoform-level functional consequences beyond total gene expression changes.

Colocalization with seven neurodegenerative and psychiatric GWAS highlighted 97 loci where alternative splicing appears to mediate genetic risk. Among sQTL–eQTL pairs with colocalization probability ≥ 0.8 (posterior probability of a shared causal variant), half shared credible-set variants, showing that splicing can complement or act independently of expression. Mechanistic examples include *CAMLG* (Parkinson's), *ZDHHC2* (Schizophrenia), *CLU* and *TREM2* (Alzheimer's). Together, these results establish alternative splicing as a key mediator of genetic risk in neurodegenerative and psychiatric disorders. Ongoing analyses extend fine-mapping and colocalization across ancestries to further refine shared and ancestry-specific regulatory effects.

TARGETED INTERCHROMOSOMAL MEGABASE-SCALE GENOME AND EPIGENOME COPYING IN HUMAN STEM CELLS

Martin Lackner[1], Svante Pääbo[1,2], Stephan Riesenberg[1]

[1]Max Planck Institute for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany, [2]Okinawa Institute of Science and Technology, Human Evolutionary Genomics Unit, Onna, Japan

Current experimental tools are struggling to generate and study genomic changes spanning from hundreds of kilobases to megabases. This limits our ability to investigate large-scale genetic and epigenetic variation. Here, we present a method that enables the copying of megabase-scale genomic regions, alongside parent-specific DNA methylation patterns, from the homologous chromosome. This method is based on CRISPR targeting and modulation of the DNA repair pathway. Targeted copying efficiencies range from 1% to 68%. We used this method to repair a 1.9 Mb deletion responsible for Sotos syndrome in patient-derived human induced pluripotent stem cells, as well as to generate model systems for Prader-Willi and Angelman syndrome imprinting diseases. By generating deletions of a few megabases as a first step, we can subsequently direct interchromosomal copying to copy until the end of the chromosome, which allowed us to copy up to 80 Mb - almost an entire chromosome arm. The ability to induce megabase-scale interchromosomal copying has the potential to facilitate large-scale genome modification, the investigation of the functional effects of entire haplotypes, as well as therapeutic genome editing.

STEDD: RESOURCE-EFFICIENT ENSEMBLE DISTILLATION FOR UNCERTAINTY-AWARE GENOMIC DEEP LEARNING

Kaeli Rizzo, Peter Koo

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

Deep learning models have revolutionized genomic sequence analysis, but their deployment in high-stakes applications requires reliable uncertainty quantification. Deep ensembles provide gold-standard uncertainty estimates through model disagreement, yet their computational cost scales linearly with ensemble size, creating barriers for both training and inference. While ensemble distillation can compress these models into efficient single-network predictors, existing methods assume simultaneous access to all ensemble members during training, a constraint that becomes prohibitive as models and ensembles scale.

We introduce STEDD (Stochastic Teacher-sampling for Ensemble Distribution Distillation), a framework that enables uncertainty-aware ensemble distillation when only limited teacher queries are available per training example. Rather than requiring predictions from all M ensemble members, STEDD methods estimate ensemble statistics from as few as k=1 stochastically sampled teachers per input. STEDD encompasses multiple complementary strategies optimized for different computational constraints, providing practitioners with flexible tools matched to their specific resource limitations.

We validate STEDD on genomic regulatory element prediction tasks, demonstrating that students trained with dramatically reduced teacher access preserve both the predictive performance and uncertainty estimates of full ensemble distillation. Our results show that uncertainty-aware distillation remains effective even in extreme resource-constrained regimes. By democratizing access to ensemble-quality uncertainty quantification, STEDD enables practical deployment of trustworthy genomic AI in settings where computational budgets are limited.

BENCHMARKING METHODS FOR INFERRING BIOLOGICAL RELATEDNESS IN ANCIENT DNA

Xavier Roca-Rada[1,2], David Peede[1,2], Linda Ongaro[3], Mayra M Bañuelos[1,2], Laura Carrillo-Olivas[4], Flora Jay[5], María C Ávila-Arcos*[4], Emilia Huerta-Sanchez*[1,2,3]

[1]Brown University, Center for Computational Molecular Biology, Providence, RI, [2]Brown University, Department of Ecology, Evolution and Organismal Biology, Providence, RI, [3]Trinity College Dublin, Smurfit Institute of Genetics, Dublin, Ireland, [4]National Autonomous University of Mexico, International Laboratory for Human Genome Research, Querétaro, Mexico, [5]Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS, INRIA, Paris, France

* These authors contributed equally.

The analysis of human ancient DNA (aDNA) has revolutionized our understanding of past populations, with kinship inference providing direct insight into burial practices, family structure, social organization, and population demography when integrated with archaeological and osteological evidence. However, aDNA poses inherent challenges for relatedness estimation: datasets are often low coverage, affected by post-mortem damage, and lack appropriate reference panels, complicating the interpretation of genetic relationships.

Despite the widespread application of kinship analyses in archaeogenomics, the performance of commonly used methods across different degrees of relatedness remains poorly characterized. Here, we systematically evaluate nine widely used kinship inference tools representing diverse methodological approaches: pairwise mismatch rates (READv2, BREADR, Kennett 2017, GRUPS-rs), identity-by-descent (IBD) probability given identity-by-state (IBS) (NgsRelateV2, lcMLkin), IBS sharing (TKGWV2), and IBD segment detection (KIN, ancIBD).

We apply these methods to a well-characterized Early Neolithic dataset from Great Britain comprising 62 individuals, including nine from a previously reported extended pedigree at the site of Hazleton North. This context provides a robust framework for assessing performance across varying degrees of relatedness.

Across approaches, close relationships (e.g., first-degree) are consistently identified, whereas inferences of more distant relationships show substantial methodological disagreement. In particular, genotype-likelihood–based methods yield less consistent estimates of extended relatedness than pseudohaploid-based approaches, underscoring key differences in how kinship is inferred under typical aDNA constraints.

By defining the strengths and limitations of current tools, this study provides a guideline for rigorous kinship analysis and identifies key priorities for future methodological development.

TIMING AND DEVELOPMENTAL ORIGINS OF SINGLE BASE MUTATIONS IN RHESUS MACAQUES AND ASSOCIATED PLACENTAL SAMPLES

Jeffrey Rogers[1], Yadira Pena-Garcia[2], Richard Wang[2], Muthuswamy Raveendran[1], R.Alan Harris[1], Jenna Schmidt[3], Matthew W Hahn[2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Indiana University, Dept.of Biology, Bloomington, IN, [3]Wisconsin National Primate Research Ctr., Madison, WI

Understanding when and where germline and somatic mutations arise is essential for revealing the mechanisms that shape genetic variation and influence risk of disease. At present, most of what we know about mutational processes comes from studies focused on the germline, using trio-based analysis. The frequency and timing of early embryonic and tissue-specific mutation profiles, particularly in primate models of human biology, is unclear. In this study, we analyzed whole genome sequences from 24 rhesus macaque “quartets”, each consisting of the blood of both parents and their offspring, along with the associated placenta from each offspring. Our goal was to investigate the developmental origins of *de novo* germline and somatic mutations. Whole genome sequence data was generated using Nova-Seq X methods (~35x coverage) and analyzed using a conservative variant-calling approach. We identified mutations separately in offspring blood and placenta, two tissues that separate at 6-7 days post-fertilization. We distinguish true germline mutations from those arising shortly after fertilization by counting the haplotypes present at genomic locations showing *de novo* mutations. We identified an average of 248.9 mutations in offspring blood samples and 259.7 in placental samples. On average, 89.2 mutations were shared between offspring blood and placenta, indicating that most mutations occurred after the separation of those developing tissues. Haplotype-counting analysis based on 989 haplotype-informative mutations shows that 88.4% of mutations shared by infant blood and placenta are true germline mutations and 11.6% occurred during early post-zygotic development. Offspring blood (sampled at mean 1.53 years of age) had an average of 159.7 mutations not found in placental tissues; placental samples had an average 170.5 tissue-specific mutations. Our results show that ~12% of apparently *de novo* germline mutations originated at a very early embryonic stage, and not in the parental germline. Interestingly, we found that blood from male offspring and male-associated placenta carry more mutations than female samples by 18% and 16%, respectively. This difference was not observed in early post-zygotic mutations, which may indicate that male biased accumulation, previously observed in germline mutations, begins only after sex differentiation has begun. These findings provide new insights into the timing and pattern of somatic mutations across tissues and sexes.

PAN-CANCER LANDSCAPE OF ALTERNATIVE LENGTHENING OF TELOMERES REVEALED BY MACHINE LEARNING ANALYSIS OF LARGE-SCALE CLINICAL SEQUENCING DATA

Harshit Sahay, Bill Diplas, Oluchi Ezekwenna, Divya Koyyalagunta, Simran Chhabria, Madison Darmofal, Quaid Morris, Agnel Sfeir

Memorial Sloan Kettering Cancer Center, New York, NY

Replicative immortality is a hallmark of cancer, enabled by telomere maintenance mechanisms (TMMs), which prevent telomere shortening and senescence. While most tumors achieve this via telomerase reactivation, 10-15% rely on alternative lengthening of telomeres (ALT), a recombination-driven TMM. ALT is strongly associated with inactivating mutations in the chromatin remodeling genes ATRX and DAXX. However, many ALT tumors lack these alterations, and ATRX loss alone is insufficient to trigger ALT in vitro. This indicates that the genetic basis of ALT remains incompletely defined. Determining TMM status also has high clinical relevance, since ALT+ve tumors show distinct prognoses and therapeutic vulnerabilities. However, genetic and clinical characterization of ALT has been limited to small cohorts and select tumor types due to the labor-intensive nature of ALT detection assays.

To address this, we leveraged the MSK-IMPACT clinical sequencing cohort, which consists of >100,000 tumors across 80+ cancer types profiled with an FDA-approved targeted sequencing panel. We found that typically discarded off-target reads contained telomeric sequences, enabling us to quantify telomere content and repeat composition. For 700 patient samples, we assessed the presence of ALT experimentally using the C-circle assay (CCA). We then used this to train an ensemble of Random Forest classifiers to predict ALT-status, using stratified five-fold CV and Platt calibration. This yielded robust predictive performance (mean ROC-AUC = 0.84; PRC-AUC = 0.76), enabling predictions of ALT-status for 78,704 IMPACT-profiled tumors.

Model predictions corresponded well with known patterns. Highest ALT-prevalence was seen in sarcomas and gliomas, particularly in ATRX/DAXX-mutant tumors, and low/no prevalence in TERT-altered (telomerase reactivating) tumors. Strikingly, the model predicted several ATRX/DAXX wild-type tumors in ALT-relevant histologies as ALT+ve, validated by additional CCAs. Integrating MSK-IMPACT annotations identified novel genetic factors associated with ALT. Notably, a tumor harboring concurrent ATRX and TERTp mutations with high predicted ALT-probability was also validated by CCA, suggesting co-occurrence of ALT and telomerase-based TMMs. Predicted ALT-status associated with differential survival with tissue-specific patterns.

Overall, our framework enables scalable ALT detection in clinical cohorts, and our resource of >78,000 tumors provides an expanded view of ALT prevalence, genetics, and clinical outcomes across diverse cancers.

CROSS-PRIMATE DGTEX MAPS EARLY-LIFE GENE PROGRAM DYNAMICS AND THEIR SELECTIVE CONSTRAINT

Irepan Salvador-Martínez[1], Jose M Ramirez[1,2], Pau Clavell-Revelles[1], Winona Oliveros[3], Zhiwei Wang[4], Laura Colbran[5], Kristin G Ardlie[6], Ziyue Gao[5], Lin S Chen[4], Tuuli Lappalainen[2,3], Marta Melé[1], and the dGTEx Consortium[6]

[1]Barcelona Supercomputing Center, Life Sciences Department, Barcelona, Spain, [2]New York Genome Center, New York City, NY, [3]SciLifeLab, KTH Royal Institute of Technology, Department of Gene Technology, Solna, Sweden, [4]The University of Chicago, Department of Public Health Sciences, Chicago, IL, [5]University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, [6]Broad Institute, Harvard and MIT, Boston, MA

Early-life gene regulation in primates remains poorly characterized, largely due to limited transcriptomic resources spanning birth to adulthood. Here, we leverage the developmental Genotype-Tissue Expression project (dGTEx), a multi-tissue human pediatric dataset (0–18 years), and the complementary non-human primate dGTEx (NHP-dGTEx), which profiles matched developmental stages—including prenatal time points—in *Macaca mulatta* (rhesus macaque) and *Callithrix jacchus* (common marmoset).

Across species, we identify hundreds of age-differentially expressed genes (age-DEGs) in multiple tissues. In humans, we uncover hundreds of genes that are switched off during early life and would be missed by adult-only resources. In macaque, dozens of prenatally expressed genes are silenced after birth, highlighting regulatory transitions that current human datasets likely miss due to limited prenatal sampling.

Using a cross-species topic modeling approach we identify 63 primate gene programs. Broadly expressed programs are enriched for core cellular functions and show reduced interspecies expression divergence. These programs are also preferentially down-regulated with age in humans, matching the pattern that down-regulated age-DEGs are broadly shared across tissues, whereas up-regulated age-DEGs are more tissue-specific. Together, these results suggest progressive silencing of a shared developmental program alongside tissue-specific transcriptional activation that drives organ maturation.

Finally, we link developmental regulation to selective constraint and gene evolutionary age. Down-regulated genes are generally more constrained across organs, while muscle up-regulated genes show unexpectedly high constraint, consistent with strong purifying selection on late-muscle programs. Transcriptome age index analyses show dynamic evolutionary patterning, with testis and muscle having evolutionary younger and older transcriptomes in adolescents, respectively.

Overall, dGTEx and NHP-dGTEx provide an unprecedented primate resource for defining early-life transcriptional programs across organs and interpreting human developmental regulation in a comparative evolutionary context.

NETWORK AND PATHWAY ANALYSIS OF TIME-DEPENDENT TRANSCRIPTOMIC RESPONSES TO SENOLYTIC THERAPY IN NONHUMAN PRIMATES

McKinley Santiago[1,2], Darla DeStephanis[1], Kylie Kavanagh[1]

[1]Wake Forest University School of Medicine, Comparative Medicine, Pathology, Winston-Salem, NC, [2]Johns Hopkins Krieger School of Arts and Sciences, Advanced Academic Programs, Baltimore, MD

Cellular senescence in adipose tissue is a state in which cells stop dividing but remain metabolically active. These cells frequently secrete senescence-associated secretory phenotype (SASP), which involves the release of pro-inflammatory cytokines, chemokines, and matrix-remodeling proteases that affect the adipose microenvironment. Senescent cell accumulation drives metabolic dysfunction and chronic inflammation in aging tissues. Senolytic therapies selectively reduce senescent cell burden, but the transcriptional programs affected in adipose tissue are not well characterized. Bulk RNA-seq method was chosen for this tissue type because, in preclinical and human pilot studies, senescent cells preferentially accumulate in adipose tissue and are removed by dasatinib plus quercetin (D+Q). Gene expression profiles were measured in SQ-adipose biopsies from middle-aged cynomolgus macaques, treated monthly for 5 months with D+Q (n=9) or vehicle (n=5). Differential gene expression (DGE) profiles were made between the treatment groups and between baseline and 5 months post-treatment using DESeq2. Ingenuity Pathway Analysis (IPA) was used to contextualize these DGEs within signaling pathways, biological functions, and molecular interaction networks. IPA provides an inferred systems-level overview of transcriptional programs affected by senolytic treatment in adipose tissue. DESeq2, network, and pathway analysis indicate coordinated, treatment-associated changes in immune and stromal gene regulation. Senolytic treatment resulted in decreased expression of cytokine-associated inflammatory signaling and immune cell trafficking networks. These results provide functional genomic context for the observation that senolytic intervention alters adipose immune and tissue-remodeling programs in a translational model of aging.

CONVERGENT EVOLUTION AND GENETICS OF HETERANTHERY IN *SOLANUM*

Miguel Santo Domingo[1,2], Srividya Ramakrishnan[3], Joyce Van Eck[4], Michael C Schatz[3], Zachary B Lippman[1,2]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [3]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [4]Boyce Thompson Institute, Ithaca, NY

Solanum is a large and diverse genus in the Solanaceae family, including major crops such as tomato, eggplant, and potato. To better understand this diversity, recently developed genomic and genetic resources across diverse *Solanum* lineages are enabling deeper insight into the clade's morphological variation. Flower morphology is particularly interesting, as symmetry has independently transitioned from radial to bilateral at least 14 times. Radial flowers, as in tomato, present equal-length stamens, whereas bilateral flowers exhibit heteromorphic stamens, a trait known as heteranthery. This trait is also found in other plant clades, where it has evolved independently numerous times. Understanding how heteranthery is genetically controlled in *Solanum* can reveal how selective pressures and genetic architecture drive the evolution of novel morphological traits.

To address this, we are building an integrated framework that combines comparative genomics and later functional validation in a newly established experimental system. First, we are generating high-quality genomes for a panel of species covering multiple repeated gains of heteranthery and their sister taxa. These new assemblies, together with those previously generated in our group, will comprise an enriched *Solanum* pangenome designed to capture genetic diversity associated with heteranthery and overall floral symmetry. Using independent origins as natural replicates, we are testing for shared signals across heterantherous lineages, which we will also integrate with existing genetic and transcriptomic resources in *Solanum*.

In parallel, we are performing genetic mapping in biparental populations in different clades. This complementary approach will allow us to directly identify the genetic architecture and the genomics regions associated with the independent gains of heteranthery in those clades.

Finally, we are establishing *S. citrullifolium* (heterantherous) and *S. sisymbriifolium* (non-heterantherous sister species) as a tractable model pair to functionally test candidate genes emerging from all discovery modes (mapping populations, mutants, and phylogenetic mapping). We have developed a gene-editing protocol in both species and are targeting homologs of genes implicated in floral symmetry in other taxa, together with a set of heteranthery-related genes identified in a mutagenesis experiment. These comparative perturbations in a shared experimental system will enable us to translate repeated evolutionary signals into causal developmental mechanisms.

This work will clarify how developmental programs are repeatedly rewired to generate heteranthery in *Solanum* and will provide a general strategy for connecting convergent phenotypes to genomic variation in plants.

OPENOMICS: BUILDING BEST-PRACTICES BIOINFORMATICS PIPELINES THROUGH COMMUNITY-DRIVEN SNAKEMAKE WORKFLOWS

Ryan Routsong[1], Paul Schaughency[1], Vicky Chen[1], Tovah Markowitz[1], Keyur Talsania[2], Thomas Hill[1], Yue Zhang[1], Oladele Oluwayiose[1], Neelam Redekar[1], Katherine Hornick[1], Subrata Paul[1], Cihan Oguz[1], Elisabeth Meyer[1], Sofia Roitman[1], Justin Lack[1], Skyler Kuhn[1]

[1]National Institute of Allergy and Infectious Diseases, Integrated Data Sciences Section, Bethesda, MD, [2]Frederick National Laboratory for Cancer Research, Advanced Biomedical Computational Science, Frederick, MD

The Integrated Data Science Section (IDSS) part of the Research and Technologies Branch (RTB) of the National Institutes of Allergy and Infectious Disease (NIAID) is a bioinformatics collective of scientists aimed at supporting research at NIAID as well as other institutes. With an exponential demand for complex and multi-omic research project design, we have strived to create a repository of reproducible community driven snakemake workflows following the current best practices for a broad variety of modalities. OpenOmics currently contains workflows for common methods like RNAseq (gene/isoform/ERV expression, gene fusion, alternative splicing, isoform switching, comprehensive QC, differential expression), Single-cell (GEX, VDJ, CITE, MULTI, ATAC Multiome), Spatial transcriptomics, ChIP-seq, ATAC-seq, WGS/WES/Amplicon (germline, somatic, CNV, SV, HLA), long-read (ONT and PacBio) transcriptomics, and Bisulphite sequencing. We also host repositories for more specialized workflows focused on viral metagenomics (short and long read), metagenomics and metatranscriptomics, cfChIP-seq, alternative splicing, genome assembly and annotation, miR-seq, cell-free DNA fragmentomics, HiChIP and others. These workflows, together, allow the end user to focus less on the primary data analysis and more on the integration and interpretation. The goal of OpenOmics is to promote open and transparent sharing of pipelines to further the cause of reproducibility and best practices as data science projects get increasingly complex.

GENE-BY-ENVIRONMENT INTERACTIONS IN ENDOTHELIAL CELLS REVEAL GENETIC MODULATION OF VASCULAR RESPONSES TO BPA AND PHTHALATE EXPOSURE

Madyson Scherr[1], Carly Boye[2], David B Witonsky[1], Gabrielle Garlicki[1], Adnan Alazizi[2], Mikhail Y Salnikov[2], Xiaoquan Wen[3], Roger Pique-Regi[2], Francesca Luca[1]

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

Cardiovascular disease is the leading global cause of death, driven by complex interactions between genetic and environmental risk factors. Endothelial cells form the inner lining of blood vessels and play central roles in vascular homeostasis and atherosclerotic plaque development. Bisphenol A (BPA) and phthalates are ubiquitous endocrine-disrupting chemicals in consumer plastics and personal care products that disrupt endothelial function and induce cell death, processes driving atherosclerotic disease. Epidemiological studies have linked these exposures to increased risk of atherosclerosis and coronary artery disease, yet how genetic variation modifies molecular responses to these chemicals and contributes to individual differences in disease susceptibility remains largely unknown.

To systematically identify gene-environment interactions in endothelial cells, primary human umbilical vein endothelial cells (HUVECs) from 100 genetically diverse donors were treated with BPA and mono-n-butyl phthalate (MBP) for 6 and 24 hours, followed by bulk RNA-seq. We identified 3,619 differentially expressed genes (DEGs) following BPA exposure at 6 hours. These genes were enriched for biological processes including autophagy and intracellular transport. Only a small number of genes were differentially expressed after 24 hours. MBP induced 5,273 and 2,740 DEGs at 6 and 24 hours, respectively, with early responses dominated by cell cycle regulation, and late response enriched for genes in the cell motility and migration processes. Expression quantitative trait locus (eQTL) mapping across all samples identified 3,984-4,945 eGenes per condition at 10% FDR. Mashr was applied to identify variants with condition-specific regulatory effects; response eQTLs (reQTLs) were defined between treatment-control pairs as variants with local false sign rate (lfsr) < 0.05 in at least one condition and either opposing effect direction or at least two-fold difference in effect magnitude. This analysis identified 87 genes with reQTLs for BPA at 6 hours and 30 at 24 hours, and 87 genes with reQTLs for MBP at 6 hours and 63 at 24 hours. Notably, response eGenes showed minimal overlap with treatment-induced DEGs, indicating that genetic variation modulates endothelial response through distinct pathways from the primary toxicological mechanisms and may capture individual-specific susceptibility to endocrine disruptor exposure.

PERSONALIZED VARIANT EFFECT PREDICTION WITH GENOMIC AI REVEALS WIDESPREAD SEQUENCE CONTEXT DEPENDENCE

Brian M Schilder[1], Zihan Liu[1], Jack Desmarais[1], David Laub[2], Fahimeh Rahimi[3], Palash Sethi[3], Lucas Pereira[3], Mengyi Sun[1], Justin B Kinney[1], David McCandlish[1], Juannan Zhou[3], Peter Koo[1]

[1]Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY, [2]University of California San Diego, Division of Biomedical Informatics, La Jolla, CA, [3]University of Florida, Department of Biology, Gainesville, FL

Computationally predicting the consequences of genetic variants remains a major goal in genomics. Conventional approaches assess single-nucleotide variants in the context of a single reference genome, but this assumption fails to capture human genetic diversity that can modulate variant effects. Here we introduce a computational framework that predicts variant effects across thousands of personalized genomes from globally diverse populations. Using sequence-based deep learning models, we quantify how genetic background influences predicted effects of clinical variants on DNA regulation, RNA processing, and protein structure. We observe widespread heterogeneity in predicted variant effects, with the same clinical variant predicted to be pathogenic in some individuals and benign in others. We find that interactions between variants and local haplotypes shape predicted three-dimensional protein contacts and alternative splicing outcomes, suggesting plausible molecular mechanisms for background-dependent effects. Together, these findings advance population-aware variant interpretation and support progress toward personalized genomic medicine.

FLUCTUATION STRUCTURE PREDICTS GENOME-WIDE PERTURBATION OUTCOMES.

Ben Kuznets-Speck[1,2], Jaekwon Jung[1,2], Leon Schwartz [1,2], Jacob L Schlamowitz[1,2], Yogesh Goyal [1,2,3]

[1]Northwestern University Feinberg School of Medicine, Cell and Developmental Biology, Chicago, IL, [2]Northwestern University, Center for Synthetic Biology, Chicago, IL, [3]Chan-Zuckerberg Biohub Chicago, LLC, Chicago, IL

Pooled single-cell perturbation screens represent powerful experimental platforms for functional genomics, yet interpreting these rich datasets for meaningful biological conclusions remains challenging. Most current methods fall at one of two extremes: either opaque deep learning models that obscure biological meaning, or simplified frameworks that treat genes as isolated units. As such, these approaches overlook a crucial insight: gene co-fluctuations in unperturbed cellular states can be harnessed to model perturbation responses. Here we propose a new conceptual framework leveraging linear response theory from statistical physics to predict transcriptome-wide perturbation outcomes using gene co-fluctuations in unperturbed cells. We validated our approach on synthetic regulatory networks before applying it to 11 large-scale single-cell perturbation datasets covering 4,234 perturbations and over 1.36M cells. Our work robustly recapitulated genome-wide responses to single and double perturbations by exploiting baseline gene covariance structure. Importantly, eliminating gene-gene covariances, while retaining gene-intrinsic variances, reduced model performance by 11-fold, demonstrating the rich information stored within baseline fluctuation structures. Moreover, gene-gene correlations transferred successfully across independent studies of the same cell type, revealing stereotypic fluctuation structures. Furthermore, our framework outperformed conventional differential expression metrics in identifying true driver perturbations while providing uncertainty-aware effect size estimates through Bayesian inference. Additionally, our framework allows for reverse inferences of drug perturbed cells allowing us to determine which drug was applied from the induced transcriptomic signature. Finally, most genome-wide responses propagated through the covariance matrix along approximately three independent and global gene modules. Our study underscores the importance of theoretically-grounded models in capturing complex biological responses, highlighting fundamental design principles encoded in cellular fluctuation patterns.

LEARNING TRANSFERABLE PHENOTYPE-TO-GENOTYPE MAPPINGS VIA MULTIMODAL CONTRASTIVE MODELING

Leon Schwartz[1], Ben Kuznets-Speck[1], Jaekwon Jung[1], Jacob Schlamowitz[1], Auinash Kalsotra[2,3], Ekta Prashnani[4], Carsten Marr[5], Yogesh Goyal[1,3]

[1]Northwestern University, Cell and Developmental Biology, Chicago, IL, [2]University of Illinois, Department of Biochemistry, Urbana-Champaign, IL, [3]Chan-Zuckerberg Biohub, LLC, Chicago, IL, [4]NVIDIA, NVIDIA, Santa Clara, CA, [5]Helmholtz, Institute of AI for Health, Munich, Germany

Linking genetic perturbations to cellular phenotypes has been central to biology since Mendel's early experiments and remains a defining challenge in the CRISPR era. Technologies such as Perturb-seq and CROPseq now enable massively parallel genotype-to-phenotype screening by coupling CRISPR-based perturbations with single-cell transcriptomic readouts. Since such experiments remain costly and limited to cell lines, perturbation-response prediction models have emerged to complement experimental discovery and improve generalization. However, these models are typically trained and evaluated within individual datasets and exhibit limited ability to generalize across experimental contexts, cell types, and unseen perturbations. Moreover, modeling the full forward transcriptomic response to a perturbation is inherently noisy and difficult to evaluate at scale. While most existing approaches focus on predicting gene expression changes given a perturbation, we deliberately turn this paradigm on its head. Instead of modeling genotype-to-phenotype mappings, we address the inverse problem: learning transferable phenotype-to-genotype mappings that infer the underlying genetic drivers from observed transcriptional states.

Here, we present ExPert, a contrastive Variational Autoencoder framework designed explicitly to learn context-invariant representations of perturbation effects across heterogeneous datasets. Largest to date, ExPert integrates 30 Perturb-seq studies comprising over 30 million cells and learns a harmonized latent space that disentangles perturbation-driven transcriptional signatures from context-specific variation. Crucially, we adopt a multimodal alignment strategy that projects cellular transcriptomic representations into a shared embedding space with gene representations derived from detailed gene descriptions encoded by pretrained large language models. This multimodal formulation allows ExPert to leverage semantic structure captured in natural language to inform genotype inference, enabling scalable prediction across thousands of genetic targets and supporting zero-shot generalization to unseen perturbations. Across different perturbation settings, ExPert consistently achieves approximately two-fold higher macro F1-scores than existing methods, while maintaining performance in entirely unseen cellular contexts. Importantly, performance remains robust with over 1000 perturbations, while baseline methods fall below 0.1 macro F1 at this scale. Beyond supervised classification, the model enables zero-shot perturbation inference to genes related to observed modules by leveraging a shared gene-embedding structure. Together, these results establish phenotype-to-genotype mapping as a scalable, transferable, and context-agnostic framework for causal gene discovery at scale.

EVOLUTIONARY CONSEQUENCES OF CHROMOSOMAL FISSION FOR CENTROMERE EVOLUTION AND SPECIATION IN GELADAS (*THEROPITHECUS GELADA*)

Brooklynn R Scott[1], Jacinta C Beehner[2], India A Schneider-Crease[3], Amy Lu[4], Thore J Bergman[2], Kenneth L Chiou [5], Andrea Guarracino[6], Noah Snyder-Mackler[1]

[1]Arizona State University , School of Life Sciences, Tempe, AZ, [2]University of Michigan , Department of Psychology, Ann Arbor, MI, [3]Arizona State University, School of Human Evolution and Social Change, Tempe, AZ, [4]Stony Brook University, Department of Anthropology, Stony Brook, NY, [5]University of Alabama at Birmingham, Department of Biology, Birmingham, AL, [6]Translational Genomics Institute, Phoenix, AZ

Chromosomal rearrangements are recognized drivers of reproductive isolation, yet the molecular mechanisms linking structural change to divergence — particularly neocentromere formation and pericentromeric restructuring — remain poorly understood, in part because such events are rarely observed in recent evolutionary history. Geladas (Theropithecus gelada), the only surviving species in their genus, provide a compelling primate model to study the effects of chromosomal rearrangements and their role in speciation in a primate lineage that typically resists karyotypic change. A recently discovered novel karyotype (2n=44) in the northern gelada population results from a Robertsonian fission of chromosome 7. This karyotype differs from the ancestral, evolutionarily conserved karyotype (2n=42) found in the central gelada population and the rest of the Papionini clade (macaques, baboons, and mangabeys), which has a crown origin approximately 10-11 million years ago. In this study, we use population resequencing data from northern and central geladas to estimate that the populations diverged ~175,000 years ago, with limited gene flow between them, suggesting that the fission emerged around this time and may act as a reproductive barrier contributing to incipient speciation. To characterize centromere evolution and pericentromeric reorganization following this recent fission, we are generating haplotype-resolved long-read genome assemblies from a northern (23x ONT, 58x PacBio HiFi), a central (33x ONT), and a zoo-born hybrid gelada (53x ONT, 42x PacBio HiFi,) with 2n=43 chromosomes. By comparing the fissioned chromosomes to their intact homologs, we aim to characterize centromere structure and pericentromeric restructuring that emerged after the Robertsonian fission. The hybrid genome will further allow us to investigate how chromosomal rearrangements shape chromosome pairing, recombination patterns, and genome architecture — mechanisms through which chromosomal fissions can drive reproductive isolation. Together, these analyses will illuminate the functional consequences of recent chromosomal rearrangements and their potential role in primate genomic divergence and speciation.

ACCURATE AND PERSONALIZED ALZHEIMER'S DISEASE RISK ASSESSMENT FOR INDIVIDUALS OF AFRICAN ANCESTRY DEMONSTRATED FOR SUBJECTS IN THE ALL OF US RESEARCH PROGRAM.

Janan Semseddin[1], Jianhua Zhang[1], Harrison McNabb[1], Dayo Shittu[2], Shaojian Gao[3], Huan Mo[2], William F Simonds[1]

[1]NIH/NIDDK, Metabolic Diseases Branch, Bethesda, MD, [2]NIH/NHGRI, Cohorts Data Analytics Core, Bethesda, MD, [3]NIH/NEI, Division of Extramural Science Programs, Bethesda, MD

Accurate assessment of an individual's genetic burden for developing Alzheimer's disease (AD) could enable early intervention in high-risk individuals. In this study, we perform a *de novo*, unbiased search for genes that distinguish subjects with AD from healthy populations. A gene-function-centered analysis of whole-exome sequencing data from individuals with African ancestry enrolled in the NIH All of Us Research Program and comparing it to healthy controls from the gnomAD database, allowed us to identify 72 potential AD-risk genes that mapped primarily to Toll-like receptor and innate immune system signaling pathways. Profiling based on the frequency of disruptive variants in these 72 genes proved able to separate Blacks with AD from healthy Blacks by unsupervised clustering analysis. Applied blindly to a cohort of 260 newly recruited Blacks naïve to prior analysis, this method assigned the AD phenotype with an accuracy that exceeded 95%. The gene-function-centered analytic approach we employed identifies potentially relevant AD-risk genes and may allow the preclinical estimation of the genetic burden of AD risk in Black individuals. We are now extending this approach beyond Black populations to evaluate variants across additional ethnic groups in an ancestry-specific fashion. Ultimately, we hope this work will enable clinicians to identify high-risk individuals earlier and initiate preventive interventions sooner, well before the appearance of AD symptoms or biomarkers.

MAPPING NON-CODING VARIANT EFFECTS TO CELL STATES VIA PREDICTIVE MODELING OF VARIANT-TO-GENE LINKS AND PERTURB-SEQ

Rintsen N Sherpa[1], Weizhou Qian[1], Elysia Chou[1], Maureen A Sartor[1,2], Joshua D Welch[1,3], Alan P Boyle[1,4]

[1]University of Michigan, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI, [3]University of Michigan, Department of Electrical Engineering and Computer Science, Ann Arbor, MI, [4]University of Michigan, Department of Human Genetics, Ann Arbor, MI

Much work has been done to understand the role of non-coding regions of the human genome. CRISPR screens have enabled cell-type-specific mapping of enhancers to genes and with them accurate predictive models have been developed. However, understanding the molecular mechanisms through which non-coding variation impacts higher-order cellular phenotypes remains elusive. This is a complex problem and requires a way to map non-coding variant effects to genes and causally probe these genes for their effects on cell states. Perturb-seq has emerged as a useful assay to directly observe the effects of gene overexpression or inhibition at the single-cell level, but no assay currently exists that can measure individual variant effects on cell states. To predict how non-coding variants shift cell states at scale, we integrated variant-to-gene (V2G) predictions with generative models of Perturb-seq data to define an integrative variant-to-cell-state (V2CS) score. We treated the overall effect of a variant on cell state as the sum of effects of the genes that variant regulates, scaled by the regulatory effect between each variant-gene pair. We trained PerturbNet, a state-of-the-art generative perturbation model, on existing K562, HepG2, and Jurkat Perturb-seq datasets to predict perturbation effects on untested genes. We measured mean shift of perturbation effect in the generated latent space and quantified the change in state distribution using energy distance. We demonstrated the utility of this method for mapping V2CS for blood, liver, and immune trait-associated variants in relevant cell lines. We showed that the V2CS score imposes sparsity on functional variant scores and prioritizes trait-relevant genes. We evaluated the integrative score for heritability enrichment in these traits against baseline functional variant predictions and on concordance with CRISPR base editing data in HepG2 and PBMC cell-state abundance QTLs. V2CS showcases a novel approach to move beyond associative methods for functional variant prioritization and suggests mechanistically driven hypotheses about trait-associated variants. This work is part of a broader collaborative effort through the Impact of Genomic Variation on Function (IGVF) Consortium to characterize and model functional variants and disseminate the findings to the scientific community.

THE ASSOCIATION OF GENETIC ANCESTRY AND EGFR DRIVER MUTATIONS IN A COHORT OF 131,000 NON-SMALL CELL LUNG CANCER PATIENTS

Alaina Shumate[1,2,3,4], Owen Hirschi[1,2,3], Dexter Jin[4], Garrett Frampton[4], Matthew Meyerson[1,2,3]

[1]Dana Farber Cancer Institute, Medical Oncology, Boston, MA, [2]Harvard Medical School, Genetics, Boston, MA, [3]Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, [4]Foundation Medicine Inc., Boston, MA

Lung cancer is the leading cause of cancer-related death worldwide, accounting for over 1.8 million deaths annually. Although commonly associated with smoking, up to 20% of lung cancer deaths occur in patients who have never smoked. The majority of lung cancer cases are classified as non-small cell lung cancer (NSCLC), which frequently harbors somatic driver mutations in the epidermal growth factor receptor (*EGFR*) gene. These mutations are of particular interest because they can be targeted with tyrosine kinase inhibitors. It is well known that the frequency of *EGFR* mutations varies significantly across populations ranging from 40-50% in patients of East Asian ancestry to ~10% in patients of European and/or African ancestry. Previous studies suggest germline genetic factors contribute to this discrepancy, but specific loci have not been identified. In this study, we investigate the relationship between genetic ancestry and somatic *EGFR* mutations in a multi-ancestry cohort of over 131,000 NSCLC patients in the United States. The sample size of our population is far larger than previous studies providing substantial statistical power to detect associations between genetic ancestry and *EGFR* mutation status. We examine both the association of global ancestry (the overall fraction of the genome derived from different ancestral populations) and local ancestry (the genetic ancestry at a specific locus). Proportions of European, East Asian, African, admixed American, and South Asian ancestry were calculated for each sample. Over 7,900 samples have detectable East Asian ancestry spanning a wide range of proportions across both admixed and predominantly East Asian individuals. We assessed the association of East Asian ancestry proportion and somatic *EGFR* mutation status, adjusting for sex, age, and smoking status. We find that *EGFR* mutations are significantly associated with the proportion of East Asian ancestry with an odds ratio of 4.90 (95% CI: 4.66-5.18). We then investigated specific mutations in *EGFR* and found significant associations were overwhelmingly concentrated in the kinase domain. These include the canonical L858R point mutation and exon19 deletions, which showed odds ratios of 7.01 (95% CI: 6.53-7.52) and 4.54 (95% CI: 4.24-4.86), respectively. To identify specific loci underlying this association, we are currently inferring local ancestry in admixed individuals and performing admixture mapping analyses. Elucidating the role of germline genetics in lung cancer susceptibility is essential for improving prevention efforts and risk stratification, especially in non-smokers who are often perceived to be low-risk and are not routinely screened.

SINGLE-STRANDED AND NON-CANONICAL DNA FORMATION IN HUMAN AND OTHER APE CELLS WITH TELOMERE-TO-TELOMERE GENOMES

Jacob Sieg[1], Huiqing Zeng[1], Hana Pálová[1], Saswat Mohanty[1], Linnéa Smeds[1], Angelika Lahnsteiner[2], Francesca Chiaromonte[1,3,4,5], Kateryna Makova[1,4]

[1]Penn State University, Department of Biology, University Park, PA, [2]Paris Lodron University Salzburg, Department of Biosciences and Medical Biology, Salzburg, Austria, [3]Penn State University, Department of Statistics, University Park, PA, [4]Penn State University, Center for Medical Genomics, University Park, PA, [5]Sant'Anna School of Advanced Studies, L'EMbeDS, Pisa, Italy

Non-canonical (non-B) DNA secondary structures, such as G-quadruplexes, Z-DNA, cruciforms, and triplex DNA, are mutation hotspots and genome regulators that contribute to evolution and disease. Yet they remain uncharacterized in vivo in complete genomes. Here, we exploited the fact that many non-B DNA structures form single-stranded DNA (ssDNA). Using permanganate/S1 footprinting across 14 cell lines, we generated and analyzed ssDNA profiles for human and six non-human ape telomere-to-telomere (T2T) genomes. Our analyses revealed three major findings. First, Hidden Markov Models applied to our multispecies ssDNA data demonstrated that approximately 11% of ape genomes exist in a state with a high ssDNA level that is conserved across species. These conserved ssDNA loci correspond to genomic domains with specific functions—e.g., replication, transcription, and recombination—each enriched in distinct non-B DNA types. This finding implicates non-B DNA conformations in the regulation of these fundamental genomic processes. Second, in human embryonic and cancer cells, ssDNA was decreased at promoters and enhancers but increased at transposable elements, indicating altered structural regulation during development and cancer. Third, elevated ssDNA levels were found at satellite arrays (ribosomal DNA, centromeres, HSat3, and multiple species-specific satellites), which vary widely among species and were unresolved prior to T2T assemblies. Non-B DNA was enriched at many satellite arrays, and we confirmed its formation using orthogonal biophysical methods for select arrays. Thus, non-B DNA can contribute to satellite expansion and function. Taken together, our ssDNA analyses across ape T2T genomes uncovered conserved and species-specific DNA structural dynamics central to genome regulation.

SMOKING AND ENVIRONMENTAL-EXPOSURE RELATED CHROMATIN INTERACTION OF LUNG CELLS IDENTIFIES TARGET GENES OF LUNG CANCER-ASSOCIATED VARIANTS

Elelta Sisay[1], Thong Luong[1], Maryam Vaziripour[2], Chia Han Lee[1], Mai Xu[1], Bolun Li[1], Jinhu Yin[1], Kevin Brown[1], Jinyoung Byun[3], Nathaniel Rothman[1], Qing Lan[1], Christopher Amos[3], Jianxin Shi[1], Jun Xia[2], Jiyeon Choi[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, [2]Center for Genomic and Precision Medicine, Texas A&M Health, Houston, TX, [3]University of New Mexico Comprehensive Cancer Center, Albuquerque, NM

Genome-wide association studies (GWAS) of lung cancer have identified over 50 genomic loci, including those dependent on smoking exposure. The genes and mechanisms promoting the risk remain poorly understood, as most candidate causal variants (CCVs) are non-protein-coding and might regulate gene expression in a context-dependent manner. To identify target genes of lung cancer CCVs and assess the effect of smoking-related exposures on these genes, we profiled chromatin interaction of lung cells by incorporating relevant exposures.

We performed HiChIP to capture interactions anchored to active chromatin regions (marked by H3K27ac) in a normal bronchial epithelial cell line, BEAS-2B, with exposures to benzo[a]pyrene (BaP), a tobacco smoke and environmental carcinogen. Cells were treated under acute (48-hour) and chronic (10-day) BaP protocols at two dosages (0.25μM and 0.50μM or 0.10μM and 0.25μM, respectively) or a DMSO control in triplicate. Sequenced HiChIP libraries (n=18) were processed and analyzed with the HiCPro and FitHiChIP pipelines to identify interacting regions. RNA-seq was conducted on matching samples to profile expression-level changes.

We identified a median of 62,020 significant interactions anchored at H3K27ac peaks ($q < 0.05$) across the samples. We then profiled the effects of BaP-exposure with differential loop and differentially expressed gene (DEG) analyses to assess both loop-level and expression-level changes, respectively. A greater number of differential loops and DEGs were detected with higher dosage and exposure time, with a consistent upregulation of BaP metabolizers and inflammatory markers. Lung cancer CCVs from diverse populations were overlaid onto loop anchors, linking 1,004 CCVs from 35 loci (69%) to a target gene based on CCV-interaction to a promoter, direct promoter overlap, or both. 311 candidate causal genes were nominated, and 186 CCGs were prioritized based on functional and regulatory scoring of their associated CCVs and loops, including 43 BaP-sensitive CCGs such as *FUBP1*. Context-specific susceptibility genes identified in this study highlight the interplay between genetic predisposition and exposures in lung cancer risk.

A PLATFORM FOR LARGE-SCALE EXPERIMENTAL MUTAGENESIS OF INTEGRAL MEMBRANE PROTEINS

Oliver B Smith[1,2], Ben Lehner[1,3,4,5]

[1]Wellcome Sanger Institute, Hinxton, United Kingdom, [2]University of Cambridge, Cambridge, United Kingdom, [3]Centre for Genomic Regulation (CRG), Barcelona Institute for Science and Technology (BIST), Barcelona, Spain, [4]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Integral membrane proteins such as ion channels, solute carriers and G protein-coupled receptors are an important class of proteins targeted by over 50% of FDA-approved drugs. They are frequently mutated in genetic disease: ClinVar reports pathogenic coding variants in over 700 integral membrane proteins and variants of uncertain significance (VUS) in more than 1200. Understanding the effect of protein sequence variation on integral membrane proteins would allow for mechanistic classification of VUS in ClinVar and facilitate the development of models for prediction of destabilising variants that could be targeted using pharmacological chaperones or other drugs. This has not yet been possible because of a lack of appropriate highly scalable selection systems. Here, we have developed a scalable Membrane Protein Fragment Complementation (mPCA) assay for deep mutational scanning of integral membrane proteins, enabling us to systematically map the relationship between protein sequence and abundance at cellular membranes. In our pilot, we perform pooled site saturation mutagenesis library to study the effects of 23,680 mutations on the membrane abundance of 19 single-pass integral membrane proteins. We establish a highly reproducible molecular construct setup that captures biologically meaningful effects and is an order of magnitude more scalable than any existing integral membrane protein deep mutational scanning technology. We have progressed to the characterisation of large multipass integral membrane proteins from human, yeast and plant species using mPCA to describe the relationship between sequence and abundance for diverse integral membrane proteins in Eukaryota. We use this data to train simple predictive models of integral membrane protein abundance and systematically study the role of abundance in mutant pathogenicity, with the aim of identifying universal mechanisms of loss-of-abundance that could be pharmacologically targetable. This work will provide a foundational dataset that describes the genetic architecture of stability in integral membrane proteins.

S-LiDER: EXPLOITING LINKAGE DISEQUILIBRIUM GEOMETRY TO REFINE FUNCTIONAL HERITABILITY ESTIMATES

Hannah Snell[1], Dhruv Raghavan[2], Sohini Ramachandran[1,3], Ritambhara Singh[1,2]

[1]Brown University, Center for Computational Molecular Biology, Providence, RI, [2]Brown University, Computer Science, Providence, RI, [3]Brown University, Ecology, Evolution, and Organismal Biology, Providence, RI

Among the complexities of understanding the heritability of diseases and traits in humans, the variation of linkage disequilibrium (LD) across the human genome and between populations remains underexplored. Previous methods, such as LD Score Regression (LDSC, Bulik-Sullivan et al., 2015) and its functionally stratified version (S-LDSC, Finucane et al., 2015), use summary statistics from genome-wide association studies and LD to estimate heritability enrichment in various traits. Both methods, however, collapse information from the LD matrix so that only the amount of LD is modeled per SNP, thus missing other LD-related patterns in the full LD matrix that may explain variation in trait heritability (such as SNP correlation with many weak independent signals versus one strong signal). As an extension of these methods, LD Eigenvalue Regression (LDER, Song et al., 2022) uses eigenvalue decomposition of the LD matrix to represent LD across spatial genomic regions. Building on these approaches, we present S-LiDER, a conceptual advance to LDER that incorporates functional information to refine heritability estimates. S-LiDER has two modes to represent LD structure: eigenvalue decomposition and a graph-based latent representation derived from a nonlinear model. We validate S-LiDER on quantitative traits and molecular biomarkers represented in the UK Biobank. In the future, we will assess its performance on more diverse genetic backgrounds and developmental insights at the single-cell level.

ORIGIN AND EVOLUTION OF ACROCENTRIC CHROMOSOMES IN HUMAN AND GREAT APES

Steven J Solar[1,2,3], Prajna Hebbar[4], Leonardo G de Lima[5], Alex Sweeten[1], Arang Rhie[1], Tamara Potapova[5], Luciana de Gennaro[6], Andrea Guarracino[7], Juhyun Kim[1], Brandon D Pickett[1], Benedict Paten[4], Melissa A Wilson[8], Sergey Koren[1], Erik Garrison[9], Evan E Eichler[10,11], Mario Ventura[6], Jennifer L Gerton[5], Adam M Phillippy[1]

[1]Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [2]Harvard Medical School, Harvard University, Boston, MA, [3]Harvard-MIT Division of Health Science and Technology, MIT, Cambridge, MA, [4]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, [5]Stowers Institute for Medical Research, Kansas City, MO, [6]Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Bari, Italy, [7]Bioinnovation and Genome Sciences Division, Translational Genomics Research Institute, Phoenix, AZ, [8]Comparative Genomics and Reproductive Health Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [9]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, [10]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, [11]Howard Hughes Medical Institute, University of Washington, Seattle, WA

The short arms of human acrocentric chromosomes are characterized by nucleolar organizer regions essential for ribosome biogenesis, but their highly repetitive nature has hindered genomic analysis. Leveraging the recently completed telomere-to-telomere (T2T) genome assemblies of all major ape lineages, we identified recurrent features of their acrocentrics, including enriched repeat classes, centromere repositioning by whole-arm inversion, interchromosomal sequence exchange, and birth-and-death evolution of multiple gene families. Together, these processes have enabled the repeated amplification and diversification of the FRG1 gene family over 25 million years of ape evolution, and, in gorilla, the formation and amplification of a novel IGSF3-GGT fusion gene under positive selection. Similar evolutionary events also explain the distribution of segmental duplications and heterochromatin in the modern human genome, predisposing it to karyotypic abnormalities such as Robertsonian translocations. Our findings highlight acrocentric chromosomes as key drivers of evolution in the great apes, with implications for speciation, adaptation, and clinical genomics.

WHY NEANDERTALS WERE HOTTER THAN US: INCREASED THERMOGENESIS VIA ELEVATED IRISIN LEVELS

Volker Soltys[1], Hugo Zeberg[1,2]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]Karolinska Institutet, Department of Physiology and Pharmacology, Stockholm, Sweden

Irisin is a peptide hormone which induces non-shivering thermogenesis via fat browning. It is cleaved off of FNDC5, a transmembrane protein produced in skeletal muscle. Here, we identified a Neandertal haplotype overlapping the *FNDC5* locus with allele frequencies of 7-25% outside Africa, which doubles *FNDC5* expression. We measured the relative level of Irisin in blood plasma and show that this increased expression results in significantly higher hormone levels as well as elevated levels of proteins directly implicated in thermogenesis. Carriers of the haplotype show a reduced risk of several lipid metabolism–related diseases, including type 2 diabetes, and have a favorable profile across all components of the metabolic syndrome. Using a combined approach of statistical finemapping, genome editing and *in vitro* differentiation, we are identifying the likely causal DNA variant on the haplotype responsible for the increased *FNDC5* gene expression. Lastly, we find that allele frequencies are strongly correlated with latitude and mean annual surface temperature. This haplotype thus may have helped some modern humans adapt to colder climates when expanding out of Africa.

CHARACTERIZATION OF A HUMAN-SPECIFIC VNTR ASSOCIATED WITH NEUROPSYCHIATRIC DISEASE RISK

Janet Song[1], Fikri Birey[2], Tzu-Chiao Hung[3], Vivien Zhao[1], Nicola Hall[4], Catherine A Guenther[3], Xiaoyu Chen[5], Ibrahim Alkuraya[1], Elizabeth M Tunbridge[4], Wilfried Haerty[4,6,7], Sergiu P Pasca[5], David M Kingsley[3,8]

[1]Harvard University, Human Evolutionary Biology, Cambridge, MA, [2]Emory University, Human Genetics, Atlanta, GA, [3]Stanford University, Developmental Biology, Stanford, CA, [4]University of Oxford, Psychiatry, Oxford, United Kingdom, [5]Stanford University, Psychiatry and Behavioral Sciences, Stanford, CA, [6]Earlham Institute, Norwich, United Kingdom, [7]University of East Anglia, Biology, Norwich, United Kingdom, [8]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA

The recent development of long-read sequencing has made it possible to catalog variable number tandem repeats (VNTRs) in the human genome. However, little is known about their functional consequences or evolutionary history. Here, we characterized TRACT, a human-specific VNTR that is composed of 100-1000+ 30bp repeats and is intronic to the calcium channel gene *CACNA1C*. Sequence variation in TRACT is tightly linked to nearby SNPs that have been strongly and consistently associated with bipolar disorder and schizophrenia in genome-wide association studies. By modeling TRACT in mice and human brain organoids, we found that TRACT affects the neuronal response to stimulation and downstream transcriptional processes, suggesting that TRACT contributed to changes in neuronal maturation during human evolution and is the causative allele at this neuropsychiatric disease risk locus. To determine how TRACT sequence and length variation evolved, we next examined multiple human cohorts and found that TRACT alleles are strikingly bimodal in both length and sequence. Short alleles ($TRACT^{S}$, ~6 kb, 95% of alleles) and long alleles ($TRACT^{L,}$ ~24 kb, 5% of alleles) have distinct sequence compositions and are found on separate haplotypes that arose prior to the human migration out of Africa. Our data suggest that these ancient alleles expanded via perfect repeat tracts that were disrupted by accumulated mutations to result in relative length stability in extant humans, where there is no evidence for overt germline or somatic instability. Differences between $TRACT^{S}$ and $TRACT^{L,}$ allele lengths likely arose due to differences in the propensity of specific 30-bp variants to expand and mutate. Together, these findings reveal how ancient sequence divergence at a neuropsychiatric disease risk locus can drive VNTR length and sequence variation to shape neural traits, and motivate the study of VNTRs as a genetic source of phenotypic variation in both evolution and disease.

VECTOR2VARIANT: DISCOVERY OF GENETIC ASSOCIATIONS FROM ML DERIVED REPRESENTATIONS WITHOUT PHENOTYPE ENGINEERING

Ramprakash Srinivasan*, Matt Sooknah*, Sivaramakrishnan Sankarapandian*, Zhenghao Chen, Jun Xu

Calico Life Sciences, Applied Machine Learning, South San Francisco, CA

* authors contributed equally

Genome-wide association studies (GWAS) have transformed our understanding of human biology, yet their scope remains constrained by our ability to define and quantify phenotypes. While modern cohorts like the UK Biobank provide high-dimensional data (including multi-organ MRI, ECG, and omics), traditional analyses typically reduce these to handcrafted features, discarding a significant fraction of the information contained in the raw measurements. Machine learning (ML) methods can extract information rich, high-dimensional representations that capture the intrinsic structure of the data, but how to best leverage these representations for genetic discovery remains an open problem.

We introduce a modality-agnostic framework that identifies genetic associations directly from high-dimensional representations (such as machine learning embeddings) eliminating the need for manual feature engineering. For each genetic variant, our method trains a classifier to identify the axis in representation space that maximizes genotype separation. This axis defines a de novo "projection phenotype" optimized to capture the specific biological perturbation of that variant. These continuous phenotypes quantify an individual's alignment with expected genetic effects, facilitating biological validation through PheWAS against clinical labels. Our approach is computationally efficient at a genome-wide scale, providing both robust association statistics and interpretable phenotypic axes.

We applied this framework across diverse modalities in the UK Biobank, including abdominal, brain, cardiac MRI, ECG, spine DEXA, retinal imaging, proteomics, and metabolomics. In addition to uncovering novel gene-organ associations not previously reported, we demonstrate that our method recovers established biology from raw data alone, effectively capturing associations that had previously required manual feature engineering or additional clinical assays to discover: CASP9-renal failure association from kidney MRI without eGFR measurement, HFE-hemochromatosis from liver imaging without iron contrast, and IL11-hyperostosis from spine DEXA without pathology scoring. In proteomics, the multivariate framework disentangles complex interactions, correctly capturing the regulatory feedback governing the PCSK9-LDLR interaction and protective effect of PCSK9 LoF variants. We further identify pleiotropic effects of genes where variants in SH2B3 and SLC39A8 exhibit distinct phenotypic signatures across organ systems.

Our method provides a unified, modality-agnostic framework to identify genetic associations from ML representations and other high-dimensional data without the need for manual phenotype engineering, and subsequently generate interpretable clinical associations to aid in target discovery.

REGULATORY LANDSCAPE OF ESSENTIAL GENES IN AGE RELATED DISORDERS

Jaya Srivastava, Ivan Ovcharenko

National Institutes of Health, National Library of Medicine, Bethesda, MD

Essential genes (EGs) are indispensable for organismal viability and are therefore subject to strong purifying selection. This results in a marked depletion of loss-of-function (LoF) variants and an enrichment of rare pathogenic variants associated with both Mendelian and complex disorders, particularly developmental disorders. While numerous forward and reverse genetic studies have interrogated the coding regions of essential genes to define their biological functions and disease contributions, their non-coding regulatory loci remain comparatively under explored. This represents a critical gap, as more than 95% of disease-associated variants reside in non-coding regions. To address this, we investigated the regulatory landscapes of 3,230 LoF-depleted EGs and compared them to those of non-essential genes (NEGs). We find that enhancers of EGs exhibit greater redundancy, are 1.6-fold more evolutionarily conserved, and are more likely to harbor disease-associated variants. Variants linked to Alzheimer's disease (AD), Parkinson's disease, and dementia show significant enrichment within regulatory loci of essential genes. Notably, for AD, variants located within EG enhancers demonstrate 1.3-fold higher effect sizes that are more likely to be rare. This observation contrasts with the expectation that regulatory regions of EGs would be depleted of large-effect variants due to purifying selection. To further dissect these patterns, we applied a deep learning framework capable of prioritizing functional regulatory variants to compare the contributions of essential versus non-essential gene loci to complex disease risk. Variants targeting EGs converge on inflammatory and protein degradation pathways, processes strongly implicated in neurodegenerative disorders (NDDs). Among the top candidates enriched for deleterious variants in EG loci are inflammatory transcription factors such as STAT1 and NFKB2, central mediators of the JAK/STAT and NF-κB signaling pathways, which play key roles in inflammaging. We further assessed age-associated variance in gene expression using GTEx eQTL data comparing younger and older individuals and found that EGs exhibit significantly greater age-associated expression variance than NEGs, indicating increased regulatory instability with aging. This pattern may reflect reduced fitness constraints after reproductive age and suggests that age-dependent dysregulation of EGs contributes to pathological trajectories during aging. These findings may reflect either a direct role for EGs as risk genes or, more plausibly, their function as central nodes within regulatory networks that influence downstream disease-relevant genes through non-coding risk loci that incur lower fitness costs. We will elucidate these mechanisms using computational approaches to further define how regulatory loci of EGs contribute to age-related neurodegeneration.

SCALABLE BAYESIAN PHYLOGENETIC INFERENCE FOR SINGLE-CELL LINEAGE TRACING ANALYSIS

Stephen Staklinski, Rebecca Hassett, Adam Siepel

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

Recent advances in cell lineage tracing now make it possible to reconstruct evolutionary histories for thousands of single cells, yet the large size and shallow divergence of these phylogenies lead to substantial uncertainty in tree structure. In earlier work, we addressed these challenges in metastatic cancer by developing Bayesian Evolutionary Analysis of Metastasis (BEAM), a Bayesian Markov chain Monte Carlo (MCMC) framework that jointly models cell lineage evolution and tissue migration, leading to improved inference of cancer metastasis histories by explicitly accounting for phylogenetic and migratory uncertainty. While this approach substantially improved performance over existing methods, its reliance on MCMC limited scalability to the largest lineage-tracing datasets. Motivated by these constraints, we subsequently re-examined Bayesian phylogenetic inference through a variational lens and developed Variational Inference with Node Embeddings (VINE), a scalable variational framework that replaces MCMC sampling with differentiable optimization while retaining flexible posterior approximations. We show that VINE achieves comparable accuracy to MCMC-based methods at a fraction of the computational cost. VINE infers phylogenies from both traditional DNA sequence datasets and recent CRISPR cell-lineage tracing datasets, and we have extended it to model tissue migration at scales previously infeasible with BEAM.

HAPLOTYPE-RESOLVED STRUCTURAL VARIATION AND FUNCTIONAL CONSEQUENCES ACROSS GLOBALLY DIVERSE HUMAN POPULATIONS

Margaret R Starostik[1], Jonas A Gustafson[2,3], Katherine M Munson[4], Hope Eden[5], Rebecca Martin[6], Kaitlyn Sun[4], Zev Kronenberg[7], Stacy L Musone[7], Elizabeth Tseng[7], 1000 Genomes Project Long-read Sequencing Consortium[10], Rob Patro[8], Chia-Lin Wei[4], Winston Timp[5], Rajiv C McCoy[1], Evan E Eichler[4,9], Danny E Miller[10]

[1]Johns Hopkins University, Biology, Baltimore, MD, [2]University of Washington, Div. Genetic Medicine, Dept. Pediatrics, Seattle, WA, [3]University of Washington, Molecular and Cellular Biology Program, Seattle, WA, [4]University of Washington, Genome Sciences, Seattle, WA, [5]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [6]Seattle Children's Research Institute, Seattle, WA, [7]PacBio, Menlo Park, CA, [8]University of Maryland, Computer Science, College Park, MD, [9]University of Washington, HHMI, Seattle, WA, [10]University of Washington and Seattle Children's Hospital, Div. Medical Genetics, Dept. Pediatrics, and Dept. Laboratory Medicine and Pathology, Seattle, WA

Structural variants (SVs; ≥50 bp) contribute substantially to functional and phenotypic diversity, but technical limitations of short-read DNA sequencing have hindered efforts to characterize their biological impact. Moreover, limited ancestry representation in human genomic datasets has obscured knowledge of the global extent of SV diversity and evolution across populations. Although recent advances in long-read sequencing and improved alignment and variant calling tools resolve previously inaccessible SVs, large-scale, high-coverage long-read genomic datasets from diverse populations remain scarce.

To address this gap, as part of the 1000 Genomes Long-read Sequencing Consortium, we performed high-coverage PacBio HiFi long-read DNA sequencing for 126 individuals from the 1000 Genomes Project (mean 35× coverage; mean N50 19.3 kb), representing 26 populations across five continental groups. Samples were sequenced using two chemistries, SPRQ and the forthcoming SPRQ-Nx, allowing performance comparisons. We complemented these data with Kinnex full-length RNA sequencing (mean full-length non-chimeric read length >2 kb), enabling integrated analysis of SVs and their transcriptional consequences.

Using long-read DNA assemblies, we show that personalized references improve transcriptomic analysis compared to even high-quality references based on distinct individuals (GRCh38 and T2T-CHM13), particularly at segmental duplications and SVs. Integrating transcriptomes with DNA methylation profiles, we map associations between SVs, isoform-level expression, and DNA methylation, providing insight into their functional and epigenetic effects across populations.

Together, this globally diverse, matched long-read DNA and RNA dataset expands current SV catalogs and establishes a high-resolution resource for understanding how SVs shape human genomes and their regulation across populations. These data are publicly available on Amazon Web Services.

WHAT COUNTS AS A SPATIAL PATTERN AND HOW TO RELIABLY DETECT ONE?

Jiayu Su

Columbia University, Systems Biology, New York, NY

Since Robert Hooke's first microscopic glimpses of "cellular" structure in 1665, scientific progress has relied on recognizing spatial patterns. The rapid rise of spatial omics has given this task new urgency: platforms now measure tens of thousands of molecular features simultaneously, making manual inspection impossible and spurring an explosion of computational methods to find spatially variable genes (SVGs) and beyond. Yet these tools often disagree with each other, lack clear theoretical grounding, and struggle to handle the million-cell datasets that new technologies now produce.

In this work, we present a theoretical unification to the fragmented landscape and provide algorithmic accelerations to move it forward. We show that virtually all major spatial pattern detection approaches—including Moran's I, Gaussian processes, and non-parametric tests—are mathematically equivalent instances of a single quadratic-form test (*Q-test*). The only real difference is their choice of *kernel* , which defines spatial similarity. This insight lets us to evaluate methods based on their "spectral spectrum," revealing the general rules that explain when and why methods succeed or fail.

Through this lens, we expose a critical flaw in Moran's I, the field's most widely used statistic. We prove that Moran's I is mathematically inconsistent as a spatial variability metric due to "spectral cancellation". That is, when biological signals aligned with opposite spatial modes, they can cancel each other out, causing the metric to miss real patterns. In practical applications, we observe measurable power loss and false negatives especially for sparse features. In tumor lineage tracing data where we consider variation along the phylogenetic tree, Moran's I and its bivariate counterpart systematically fail to detect heritable markers of rare subclones and co-expression modules.

To fix this, we propose a consistent alternative using an inverse-Laplacian (CAR) kernel. This approach guarantees reliable pattern detection while staying computationally efficient. Motivated by the convolutional connection of the Q-test, we further accelerate it using Fast Fourier Transforms, reducing runtime and memory requirements from cubic to near-linear. The resulting framework enables genome-wide spatial pattern detection across millions of locations in seconds.

Together, our work provide the necessary mathematical foundation and computational infrastructure for reliable spatial pattern detection at the million-cell scale.

Preprint: https://arxiv.org/pdf/2602.02825

A DISPERSION-BASED FRAMEWORK FOR EVALUATING CLUSTERING RESOLUTION IN SINGLE-CELL RNA-SEQ DATA

Michelle Sun, Brendan Jamison, Yoav Gilad

University of Chicago, Department of Medicine, Chicago, IL

Single-cell RNA-seq studies increasingly aim to resolve subtle cell states, particularly those that emerge in response to environmental perturbations, drug exposure, or disease. In these settings, accurate unsupervised identification of cell populations is central to biological interpretation. A persistent challenge in this process is the choice of clustering resolution (K). Standard approaches use distance-based metrics to select a single global K, which is then applied across the dataset. However, cellular heterogeneity is not uniform: some populations are well separated and risk being over-partitioned, while others contain meaningful substructure that can only be resolved at higher resolutions. As a result, a single global K may obscure biologically relevant states or artificially fragment coherent ones.

Here, we present a data-driven framework that leverages gene-level dispersion estimates to evaluate clustering resolution. Our approach is motivated by the observation that well-defined clusters exhibit low gene-level dispersion within clusters and high gene-level dispersion between clusters. We begin with a low-resolution partition and derive mean-adjusted dispersion estimates for each gene, computing cluster-level summaries based on the most dispersed genes. By tracking changes in the ratio of within- and between-cluster dispersion as K increases, our framework identifies resolutions that preserve coherent populations while revealing meaningful substructure.

This approach retrieves clusters representative of distinct cell types while guarding against distortions introduced by dimension reduction. In addition, correlations of dispersed genes between clusters enable identification of cell states within cell types and evaluation along continuous differentiation trajectories. By leveraging variability intrinsic to single-cell data, this dispersion-based framework provides a principled alternative to one-size-fits-all resolution selection strategies. The approach enables simultaneous interrogation of broad cell type differences and subtle cell state variation, improving our ability to investigate molecular mechanisms underlying disease, drug response, and gene–environment interactions.

A NEW METHOD FOR POLYGENIC PREDICTION INTEGRATING ADDITIVE AND DOMINANCE EFFECTS

Yuxuan Sun[1,2], Fabio Morgante*[1,3], Trudy F Mackay*[1,3]

[1]Clemson University, Institute for Human Genetics, Greenwood, SC, [2]Clemson University, School of Computing, Clemson, SC, [3]Clemson University, Department of Genetics and Biochemistry, Clemson, SC

*Corresponding authors: Fabio Morgante, Trudy F. Mackay

Polygenic scores (PGS) predict individual genetic predisposition by aggregating effect size estimates (β) from genome-wide association studies (GWAS). However, standard GWAS are typically conducted under a purely additive model, where β approximates the average effect of allele substitution. This approach ignores dominance effects that can be important components of the genetic architecture of complex traits. To address this, we developed a PGS method ($PGS_{a,d}$) that incorporates both additive (a) and dominance (d) genetic effects.
In this method, a and d are estimated from phenotype means and integrated into polygenic prediction according to individual genotypes. We evaluate the method using large-scale simulations based on UK Biobank genotypes under complete dominance, incomplete dominance, and purely additive genetic architectures. In settings with non-additive effects, $PGS_{a,d}$ improved phenotype prediction accuracy, showing higher cross-validated R^2 and lower Root Mean Squared Error (RMSE), relative to standard PGS (PGS_{β}). Under purely additive architectures,$PGS_{a,d}$ performed comparably to PGS_{β}, indicating robustness when dominance is absent. We further applied the framework to 16 blood cell traits in 273,795 UK Biobank participants and identified variants with significant dominance effects for multiple traits after Bonferroni correction. While $PGS_{a,d}$ achieved prediction R^2 comparable to PGS_{β} for all these traits, $PGS_{a,d}$ showed lower prediction RMSE for 9 out of the 16 traits, indicating more accurate individual predictions. Additionally, the identification of variants with significant dominance effect provides a more nuanced map of the genetic basis. Together, these results demonstrate that the proposed method provides a scalable approach for incorporating dominance effects into polygenic prediction, extending standard PGS methods to accommodate a broader range of genetic architectures in large-scale data.

MODELING INDIRECT MOLECULAR QUANTITATIVE TRAITS

Maha Syed, Hannah V Meyer

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

Genome wide association studies (GWAS) have identified many risk variants associated with diseases. To better understand their mechanistic function, molecular quantitative trait mapping aids in identifying risk conferring variants associated with changes such as differences in gene expression levels or splicing in relevant cell types. Here, we hypothesize that variants do not only act by directly impacting a cell type of interest but might confer their effect indirectly. Indirect genetic effects have been examined on an organismal level, such as modeling the effect on phenotype from a mouse's cage mate genotypes, or on at cellular level, where cell abundances are associated with genetic variants expressed another cell type. However, there remains a lack of understanding of how to model indirect genetic associations on molecular level. For example, in the autoimmunity disorder APS-1, a mutation in the AIRE gene is expressed in the "educator" of T cells, which in turn impairs T cell maturation and results in dysfunctional T cells. We developed a framework to model indirect genetic effects with a "molecular risk score." To examine the calibration and power of this new framework, we test it on independently simulated matched gene expression and genotype data with known indirect genetic effects modeled with three distinct biological hypotheses. We apply our method to a public dataset of two interacting immune cell types, expanding the level at which indirect genetic effects can be explored to the molecular state.

A PANGENOME REFERENCE OF THE SUBTELOMERES REVEALS EXTENSIVE SEQUENCE VARIATION AT HUMAN CHROMOSOMAL ARMS

Kar-Tong Tan[1,2,3,4], Ryan Jun Xiang Ong[1,2], Russell Ker Han Yap[1,2], Brandon Bing Rui Kee[1,2], Alicia Jun Ting Ng[1,2], Max Garrity-Janger[4,6,7], Qiyu Lin[1,2], Cin Thet Kyi[1,2], Matthew Meyerson[4,6,7,8], Heng Li[3,5]

[1]National University of Singapore, Department of Pharmacy and Pharmaceutical Sciences, Singapore, [2]National University of Singapore, Department of Biomedical Informatics, Singapore, [3]Dana-Farber Cancer Institute, Department of Data Sciences, Boston, MA, [4]Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, [5]Harvard Medical School, Department of Biomedical Informatics, Boston, MA, [6]Broad Institute of MIT and Harvard, Cancer Program, Cambridge, MA, [7]Harvard Medical School, Department of Genetics, Boston, MA, [8]Dana-Farber Cancer Institute, Center for Cancer Genomics, Boston, MA

Subtelomeres are regions directly adjacent to the telomeres and may play important regulatory roles in telomere maintenance due to their proximity to chromosomal ends. However, as these regions were poorly represented in earlier versions of the reference genome, it was not previously possible to determine the degree of subtelomeric sequence variation, or its impact on arm-specific telomere length and human diseases. Here, we present a collection of n=860 unique human genome assemblies of the subtelomeres across all chromosomal arms, spanning six major ancestral groups (Africans, Americans, East Asians, Middle Easterners, South Asians, and Europeans). Remarkably, the comparison of subtelomeric sequences of each chromosomal arm revealed large complex structural variations (>50-100kb) in 54% (21/39) of all non-acrocentric autosomal arms, which are characterized by a mosaic patchwork of homologous blocks shared across chromosomal arms. These variations also vary in frequency across ancestral populations, with chromosomal arms like 11p, 6p, 16q, and 9q, exhibiting the highest degree of haplotypic diversity within the human population. Notably, our subtelomeric reference collection increases the accuracy of assigning telomeric reads for arm-level telomere length determination from 62.8% (using CHM13 alone) to 93.6%. Leveraging long-read sequencing data, we also observed haplotype specific differences in telomere length at chromosomal arms. Furthermore, while the Facioscapulohumeral Muscular Dystrophy-associated *DUX4* gene is typically localized to a repeat array on the 4q subtelomere, we identified it in 0.6% (3/490) of 10q assemblies. This suggests inter-chromosomal recombination and demonstrates the utility of our pangenome for resolving complex, medically relevant loci. Overall, this collection reveals extensive sequence diversity in a poorly represented region of the human genome, providing a critical reference for arm-specific telomere biology. More broadly, we anticipate that this resource will enhance our ability to investigate how subtelomeric variation impacts telomere function and human diseases.

KERNELIZED GENE PRIORITIZATION APPROACH ENABLES INTERPRETABLE GENE PREDICTIONS

Taotao Tan, Md. Abul Hassan Samee

Baylor College of Medicine, Department of Integrative Physiology, Houston, TX

Genome-wide association studies (GWAS) have linked thousands of genetic variants to complex traits, motivating gene prioritization methods that integrate functional genomic annotations to rank putative effector genes. These approaches enable the systematic nomination of candidate genes for experimental follow-up. Polygenic Priority Score (PoPS), for instance, predicts gene-level association statistics from functional annotations and outputs a ranked gene list. Despite its effectiveness, PoPS lacks interpretable units and is unable to quantify the confidence of a prediction. To address these limitations, we introduced Kernelized-Polygenic Priority Score (K-PoPS), an intuitive dual-space reformulation of PoPS using kernel ridge regression. K-PoPS enhances interpretability by decomposing the prediction score into contributions from each training gene. The learned contribution scores enable the assessment of prediction reliability and the nomination of new candidate genes. Empirical testing of K-PoPS on a variety of complex traits, including cancers and metabolic traits, highlights its utility by identifying potential false-positive and false-negative genes. In summary, K-PoPS addresses a critical gap in explainable gene prioritization and will improve the transparency of gene-trait association predictions.

IDENTIFYING GENETIC RISK FACTORS FOR VASCULAR CALCIFICATION

Fahim Rejanur Tasin[1], Justin Koesterich[2], Anat Kreimer[2], Nadja Makki[1]

[1]University of Florida, Physiology & Aging, Gainesville, FL, [2]Rutgers University, Biochemistry & Molecular Biology, Piscataway, NJ

Vascular calcification (VC) is a pathological deposition of calcium minerals in arterial walls, leading to vessel rigidity and significant cardiovascular risk. While genome-wide association studies (GWAS) have identified numerous loci linked to VC, the causal single nucleotide polymorphisms (SNPs) often remain obscured by linkage disequilibrium (LD). Given that the majority of SNPs reside in non-coding regions, we hypothesize that these variants disrupt distal regulatory elements, such as enhancers, to alter target gene expression.

In this study, we expanded 140 GWAS lead SNPs into a candidate pool of ~5000 variants based on strong LD ($r^2 \geq 0.8$). Using the novel E-P-INAnalyzer platform, we integrated cell-specific multiomics datasets - including RNA-seq, H3K27ac ChIP-seq, ATAC-seq, and Hi-C - from primary human coronary artery endothelial cells (HCAECs) and smooth muscle cells (HCASMCs). Our computational pipeline identified 31 high-priority SNPs predicted to alter the function of enhancers regulating target genes critical to vascular health.

To validate these findings, we are performing in vitro luciferase reporter assays to characterize the allele-specific effects of these SNPs on enhancer activity. By mapping these functional variants and their target genes, this work identifies potential predictive biomarkers and provides mechanistic insights into the genetic architecture of vascular calcification, paving the way for earlier clinical intervention.

LINKING GENETIC VARIATION TO PHENOTYPE VIA COMPUTATIONAL SATURATION MUTAGENESIS AND FUNCTIONAL GENOMICS

Shaolei Teng

Howard University, Biology, Washington DC, DC

Interpreting the functional consequences of genetic variation remains a central challenge in genomics. We present an integrative framework that combines large-scale computational saturation mutagenesis with genome editing and transcriptomic profiling to systematically evaluate the impact of missense mutations in Curly suppressor (cysu/dMPO), the Drosophila ortholog of human myeloperoxidase. We computationally modeled 11,191 possible missense variants and identified a predominance of destabilizing mutations, including G378W and W621R, which target evolutionarily conserved residues. CRISPR-Cas9–engineered mutant flies exhibited developmental defects, reduced lifespan, and widespread transcriptional reprogramming affecting metabolic and immune pathways. These results directly link sequence-level variation to protein structural stability, organismal phenotypes, and genome-wide expression changes. Comparative analyses with human peroxidases reveal conserved mutational vulnerabilities associated with disease. Our study provides a scalable strategy for functional annotation of coding variants, bridging structural bioinformatics with experimental genomics and systems-level analysis. This approach advances efforts to interpret genetic variation across genomes and to connect genotype with phenotype at multiple biological scales.

THE RAPID EVOLUTION OF CENTROMERIC SATELLITE SEQUENCES IN GEOGRAPHICALLY ISOLATED HOUSE MOUSE LINEAGES

Keenan Wiggins, Jitendra Thakur

Emory University, Biology, Atlanta, GA

Centromeres are chromosomal sites where spindle fibers attach via the kinetochore to enable chromosome segregation. Despite conserved function, centromeric DNA sequences evolve rapidly across eukaryotes. Yeast centromeres are short, sequence-dependent regions (~100–400 bp), whereas mammalian centromeres comprise long tandem satellite repeats (up to several megabases). The centromere drive model explains this paradoxical rapid sequence evolution: centromeric variants that strengthen centromeres preferentially segregate to the egg pole during asymmetric female meiosis, while centromeric proteins subsequently evolve to suppress these deleterious effects. Because established mammalian species exhibit substantial centromeric sequence variation, identifying conserved motifs that define centromere identity remains challenging, especially given the highly repetitive nature of mammalian centromeres. To investigate how DNA sequences contribute to centromere identity, we examined centromeric variants in house mouse (*Mus musculus*) lineages that diverged 0.5 million years ago into Eastern European (*M.m. musculus*) and Western European (*M.m. domesticus*) populations. Despite recent divergence, these strains exhibit functional molecular differences at centromeres. We used PacBio and Nanopore long-read sequencing to generate high-resolution assemblies of centromeric regions across several Mus lineages. Our findings reveal the complex organization of satellite DNA, the prevalence of Robertsonian translocations and inversions, and the diversification of specific motifs within lineage groups. Interestingly, we observed lineage-specific enrichment of transposable elements within centromeres, suggesting a dynamic interplay between repetitive elements and centromeric repeats. These findings establish a framework for understanding centromere evolution within geographically isolated house mouse lineages, providing insights into rapidly evolving genomic elements and reproductive isolation.

MAPPING CELL-TYPE-SPECIFIC HOST-MICROBIOME ASSOCIATIONS IN THE DISTAL LUNG

Polina Tikhonova[1,2], Hanh Tran[2,3], Nicholas E Banovich[4], Emily R Davenport[2,5]

[1]The Pennsylvania State University, Bioinformatics and Genomics Graduate Program, University Park, PA, [2]The Pennsylvania State University, Department of Biology, University Park, PA, [3]The Pennsylvania State University, Molecular, Cellular, and Integrative Biosciences Program, University Park, PA, [4]Translational Genomics Research Institute, Phoenix, AZ, [5]The Pennsylvania State University, The Huck Institutes of the Life Sciences, University Park, PA

Our lungs, like other organs, function alongside a microbial community. Although microbial biomass in the lung is low, microbes can exert meaningful effects on host cell homeostasis. However, we lack a systematic understanding of how microbial composition relates to host cell types, gene expression programs, and biological processes in the distal lung. We hypothesized that variation in lung microbial communities is linked to specific host cell lineages and coordinated gene expression programs rather than uniform effects across cell types. To address this gap, we conducted a comprehensive multi-omics study integrating paired single-cell RNA sequencing and 16S rRNA profiling from distal lung tissue of 25 Idiopathic Pulmonary Fibrosis or Interstitial Lung Disease patients and 26 donor controls. At the community level, microbial diversity was significantly associated with the abundances of multiple host cell types, including adventitial fibroblasts, a fibrotic proinflammatory cell population (Shannon index, $p < 0.05$). Moving beyond broad measures of diversity, we identified lineage-specific associations between individual microbial taxa and host cell type abundances, including *Halotalea* with epithelial cell types and *Sphingomonas* with immune cell types (zero-inflated models, adjusted $p < 0.05$). To resolve these lineage-specific relationships at the gene expression level, we identified 349 gene co-expression network modules across cell types using hdWCGNA and tested microbe-cell-type-specific associations using zero-inflated linear models. Together, these results reveal a broad spectrum of associations between lung microbial variation and coordinated, lineage-specific host gene expression programs spanning all major cell-type lineages, including endothelial, epithelial, immune, and mesenchymal cells, and implicating diverse biological processes such as immune and metabolic functions. Notably, these patterns suggest that microbial community structure and composition in the lung are non-random with respect to host cell lineage. This study provides a cell-type resolved framework for linking microbial communities and host cell composition and gene expression in the distal lung, offering new insight into host-microbiome interactions in health and disease.

RAPID CENTROMERE TURNOVER AND THE ADAPTIVE RADIATION OF LEMURS

Mihir Trivedi[1,2], Luciana de Gennaro[3], Francesca Gianfrate[3], Marcelo Ayllon[1], Katherine M Munson[1], Kendra Hoekzema[1], Erin E Ehmke[5], Anne Yoder[6], Stephen Chang[4], Mark Krasnow[4], Mario Ventura[3], Evan E Eichler[1,2]

[1]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [2]Howard Hughes Medical Institute, University of Washington, Seattle, WA, [3]University of Bari Aldo Moro, Department of Biosciences, Biotechnology and Environment, Bari, Italy, [4]Stanford University School of Medicine, Department of Biochemistry, Stanford, CA, [5]Duke Lemur Center, Duke University, Durham, NC, [6]Duke University, Department of Biology, Durham, NC

Centromeres represent essential chromosomal structures required for faithful chromosome segregation during cell division. Despite their fundamental importance to genomic stability and inheritance, they are paradoxically hypermutable, leading to centromere drive and reproductive isolation in closely related species. The recent advent of high-fidelity long-read sequencing technologies has finally enabled complete assembly and characterization of these critical yet previously inaccessible genomic territories. Here, we generate nearly complete genomes from eight Strepsirrhini lemur species (haploid genome sizes ranging from 2.1-2.5 Gbp) and characterize the sequence, epigenetic and cytogenetic structure of centromeres. The basal position and diversity of the Strepsirrhini suborder provide an alternative primate perspective of centromere evolution since they diverged from Old World and New World monkeys ~70 million years ago. We show that none of the fully sequenced 223 centromeres in these eight lemur species consist of α-satellite DNA that typifies the haplorrhine primates. Instead, each species evolved its own distinct higher-order centromeric repeat sequence, varying substantially in both monomer length (ranging from 41-1405 bp, with mean of 368 bp) and primary sequence composition (GC percentages 28.7-67.9, mean = 44.2%). Despite this diversity, analysis of DNA methylation landscapes across these complete centromeric regions shows the presence of centromeric dip regions ranging in length from 110-300 kbp as candidates for kinetochore attachment. Remarkably, comparison of the predominant monomer sequence shows no apparent sequence homology among lemur genera, even for species separated by <15 million years (Lemur and Eulemur). This diversity suggests rapid turnover and independent evolutionary trajectories of centromeric DNA across strepsirrhine lineages over short periods. Even within species, we observe unexpected complexity: four species exhibit centromeric heterogeneity, i.e., two or more distinct monomer types capable of forming functional centromeres on different chromosomes within the same genome. We suggest that the high rates of diversification and bursts of speciation characteristic of lemuriform primates,all occupying the Madagascar island, were facilitated by the extraordinary turnover of centromeres, providing a stasipatric barrier during evolution.

COMPREHENSIVE CHARACTERIZATION OF INVERSIONS ACROSS THE HUMAN POPULATION USING POOLED STRAND-SEQ AND LONG-READ SEQUENCING

Vasiliki Tsapalou[1], Thomas Weber[1], Tiffany Leung[2], Daniel Chan[2], David Porubsky[1,3], Evan E Eichler[3,4], Peter Lansdorp[2,5], Jan O Korbel[1]

[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, [2]BC Cancer Agency, Terry Fox Laboratory, Vancouver, Canada, [3]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, [5]University of British Columbia, Department of Medical Genetics, Vancouver, Canada

Inversions are major contributors to human genomic diversity, yet significant knowledge gaps persist. These copy-number–neutral structural variants are often flanked by highly identical repeat sequences and associated with disease-linked microdeletions and microduplications, making them challenging to detect using conventional genomic technologies and necessitating integrated approaches.

We generated data from all 1000 Genomes Project cohorts, enabling systematic population-scale characterization of inversion frequencies and functional associations. We integrated single-cell template strand sequencing (Strand-seq) from 728 genomes with paired Oxford Nanopore Technologies (ONT) data generated as part of the IMP/MARVL study. The dataset maximizes global genetic diversity through balanced representation across five continental ancestries and 26 populations, with emphasis on African individuals. Strand-seq captures DNA strand directionality, enabling sensitive detection and genotyping of inversions >50 kbp even in repetitive regions. To scale sequencing while preserving accurate inversion detection, we coupled lymphoblastoid cell line pooling with Strand-seq. In parallel, ONT reads provided orthogonal validation and resolved additional structural complexity.

We identified 466 inversion regions, including 247 previously unreported loci and 44 novel multi-megabase events. Notably, we discovered a complex 57 Mb pericentromeric inversion in a Yoruba individual overlapping neurodevelopmental disorder loci. Population genetic analyses revealed pronounced stratification, including population-specific variants consistent with heterogeneous evolutionary histories and up to 3-fold enrichment of certain inversions, particularly in African and admixed American ancestries. Breakpoints intersect genes and clinically annotated variants implicated in neurodevelopment, immune regulation, and cancer susceptibility. Ongoing analyses aim to quantify inversion recurrence and evaluate the functional impact of highly recurrent loci. Together, this study delivers the most comprehensive population-scale inversion callset to date, advancing understanding of inversion polymorphisms in human genome diversity and evolution.

TRANSFORMER-BASED DEEP LEARNING FRAMEWORK FOR GENE REGULATORY NETWORK INFERENCE FROM SINGLE-CELL MULTIOME DATA

Eric Moeller, Karamveer Karamveer, Hannah Valensi, Yasin Uzun

Penn State College of Medicine, Pediatrics, Hershey, PA

Mapping transcription factor (TF) to target gene interactions is central to understanding cellular identity, regulatory control, and disease mechanisms. However, accurate inference of gene regulatory networks (GRNs) remains challenging due to the complex interplay of transcriptional and epigenetic regulation. Although single-cell multiomic technologies now enable cell-type-specific GRN inference, the absence of standardized benchmarking has hindered objective evaluation and comparison of existing methods.

To address this limitation, we developed **SC-MO-GRN-DB**, a publicly available resource comprising more than 22 million experimentally validated regulatory interactions paired with matched **single-cell multiomic datasets** across human and mouse tissues. Leveraging this resource, we created **BEAR-GRN**, a benchmarking pipeline for systematic evaluation of GRN inference algorithms. Using four independent single-cell multiomic datasets spanning diverse cell types, BEAR-GRN revealed that current methods exhibit low accuracy and poor reproducibility, underscoring the need for improved integrative modeling approaches.

We therefore developed a deep learning framework that integrates chromatin accessibility and gene expression at single-cell resolution using a **transformer-based architecture**. In this model, transcription factors and target genes are represented by learned identity embeddings, while accessible chromatin regions are encoded as genomic windows with positional information. A transformer encoder captures long-range chromatin dependencies, and bi-directional cross-attention enables explicit modeling of TF-chromatin and chromatin-gene relationships, incorporating distance-aware regulatory biases. Gene expression is predicted from these learned representations, linking chromatin state, TF activity, and transcriptional output. To mitigate data sparsity, the model is pretrained on pseudobulk data, stabilized using elastic weight consolidation, and subsequently fine-tuned on single-cell data.

From the trained model, interpretable features including attention weights, predicted expression, distance effects, and gradient-based attributions are extracted and used to score TF-target interactions. Evaluation on a mouse embryonic stem cell multiomic dataset demonstrated that this framework outperforms state-of-the-art GRN inference methods based on ROC and precision-recall analyses. Together, this work provides a scalable and interpretable approach for reconstructing gene regulatory networks and identifying candidate transcriptional drivers in complex biological systems.

EMBRYORADAR – A MACHINE LEARNING MODEL TO UNCOVER THE IMPACT OF EARLY EMBRYONIC TRANSCRIPTIONAL REAWAKENING IN CANCER

Tongtong Wang[1,2], Benjamin HernandezRodriguez[1,2], Janith A Seneviratne[1,2], Alicia Oshlack[1,2,3], Melanie A Eckersley-Maslin[1,2,4]

[1]Peter MacCallum Cancer Centre, Department of Laboratory Research, Melbourne, Australia, [2]The University of Melbourne, Sir Peter MacCallum Department of Oncology, Melbourne, Australia, [3]The University of Melbourne, School of Mathematics & Statistics, Melbourne, Australia, [4]The University of Melbourne, Department of Anatomy and Physiology, Melbourne, Australia

Cancers and embryos share many similarities, including heightened cellular plasticity. During development, this plasticity enables the generation of diverse cell types, whereas in cancer it has been linked to tumour aggressiveness, metastasis, therapy resistance, and relapse. Consistently, transcriptional programs associated with embryonic stem cells, foetal development, and adult stem cells are frequently reactivated in cancers and are generally associated with poor patient outcomes. However, these studies have largely focused on later developmental stages or in vitro–adapted embryonic stem cell states. In contrast, the preimplantation embryo represents the highest developmental potency in vivo, yet its systematic contribution to cancer biology has not been systematically examined.

To address this gap, we developed EmbryoRadar, a suite of generalisable machine-learning models that detect transcriptional similarity to distinct preimplantation embryonic states in non developmental contexts. Applying EmbryoRadar across multiple human cancer types revealed widespread reactivation of preimplantation embryonic transcriptional programs, with tumour type specific associations with patient outcomes. Importantly, different embryonic states exhibited divergent prognostic effects, challenging the prevailing assumption that embryonic program reactivation is uniformly detrimental.

By adapting EmbryoRadar to single-cell and spatial datasets, we further resolved the cellular and microenvironmental contexts underpinning these signals, uncovering distinct tumour-intrinsic and stromal associations across different types of cancers. EmbryoRadar offers a scalable and generalisable approach to quantify developmental transcriptional reawakening in cancer and highlights the context-dependent nature of embryonic plasticity in tumour progression.

SINGLE-NUCLEUS RNA-SEQ OF ASIAN SKELETAL MUSCLE REVEALS ANCESTRY- AND LIFESTYLE-DEPENDENT REGULATORY PROGRAMS ACROSS OBESITY AND WEIGHT LOSS

Wenjing Wang[1], Yihan Tong[1], Wei Lin Liew[2], Chi Tian[1], Zixian Zhao[1], Yuntian Zhang[2], E Shyong Tai[2,3,4], Mei Hui Liu [5], Boxiang Liu[1,6,7,8]

[1]National University of Singapore, Faculty of Science, Pharmacy and Pharmaceutical Sciences, Singapore, [2]National University of Singapore, Yong Loo Lin School of Medicine, Medicine, Singapore, Singapore, [3]Duke-National University of Singapore Medical School, Medicine, Singapore, [4]National University Health System, Division of Endocrinology, Medicine, Singapore, Singapore, [5]National University of Singapore, Food Science and Technology, Singapore, [6]National University of Singapore, Yong Loo Lin School of Medicine, Biomedical Informatics, Singapore, [7]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A∗STAR), Singapore, [8]National University of Singapore, Yong Loo Lin School of Medicine, Precision Medicine Translational Research Programme, Singapore

Skeletal muscle plays a vital role in maintaining the body's metabolic balance. However, the genetic and transcriptional factors that influence obesity and responses to weight loss remain poorly understood, particularly in Asian populations that are underrepresented in large-scale molecular studies. To address this gap, we constructed a comprehensive single-nucleus RNA sequencing (snRNA-seq) atlas from 313 human skeletal muscle biopsies in the Singapore Adult Metabolism Study (SAMS), including 205 cross-sectional samples from lean and obese Chinese, Malay, and Indian adults, alongside 108 longitudinal paired samples collected before and after a structured 16-week weight-loss intervention. Each participant is deeply phenotyped with over 200 metabolic, biochemical, and anthropometric measurements. We generated a high-depth snRNA-seq profile across these samples, comprising more than one million high-quality nuclei and identifying more than 20 distinct cell types. This dataset captures substantial cellular heterogeneity across ancestry, adiposity, and metabolic states, enabling systematic investigation of cell-type-specific transcriptional programs in human skeletal muscle. Integration with genotype information and quantitative metabolic phenotypes provides a foundation for identifying regulatory variations that influence gene expression and cellular composition, particularly in pathways linked to insulin sensitivity, mitochondrial function, and type 2 diabetes risk. Longitudinal analyses further enable the dissection of transcriptional remodeling in response to weight loss, revealing dynamic regulatory programs within myonuclear, stromal, and muscle microenvironment interactions. . Ongoing work extends this work by generating spatial transcriptomic profiles from dozens of paired pre- and post-intervention muscle samples, enabling the spatial localization of regulatory variation and the characterization of microenvironment-specific adaptations to weight loss. Together, this study establishes the first longitudinal and ancestry-diverse single-cell and spatial regulatory reference for Asian human skeletal muscle and provides a foundation for understanding how genetic background and lifestyle perturbations jointly shape metabolic tissue organization at cellular resolution.

INDUSTRIALIZATION INFLUENCES MOLECULAR MECHANISMS OF AGING IN IMMUNE CELLS IN THREE NON-INDUSTRIAL POPULATIONS

Marina Watowich[1], Julien Ayroles[2], Alexander Bick[3], Michael Gurven[4], Hillard Kaplan[5], Thomas Kraft[6], Yvonne Lim[7], Amy Longtin[1], Sospeter Njeru[8], Yash Pershad[3], Benjamin Trumble[9], Vivek Venkataraman[10], Ian Wallace[11], Amanda Lea[1]

[1]Vanderbilt University, Biological Sciences, NVE, TN, [2]Univ of California, Integrative Biology, BKL, CA, [3]Vanderbilt University Medical Center, Genetic Medicine, NVE, TN, [4]Univ of California, Anthropology, SB, CA, [5]Chapman University, Economic Science, ORG, CA, [6]Univ of Utah, Anthropology, SL, UT, [7]Universiti Malaya, Parasitology, KL, Malaysia, [8]Kenya Medical Research Institute, Community Research, NBI, Kenya, [9]Arizona State University, Evolution and Medicine, TPE, AZ, [10]Univ of Calgary, Anthropology, CAL, Canada, [11]Univ of New Mexico, Anthropology, ABQ, NM

The Geroscience Hypothesis posits that molecular aging processes, including changes to DNA methylation (DNAm), contribute to non-communicable age-related diseases (NCDs). Yet, molecular aging has been studied almost exclusively in Western, industrialized populations where NCDs are common; in contrast, non-industrial populations have low rates of NCDs and lifestyle features (e.g., high physical activity) that may slow disease-generating aging mechanisms. To develop true mechanistic targets of the aging process and understand whether industrialized lifestyles exacerbate molecular aging and disease, we quantified DNAm in blood from four populations spanning non-industrial to industrialized contexts. We generated comparable data for Tsimane horticulturalists in Bolivia (2060 samples from 1252 individuals, including 727 longitudinal samples from 219 individuals), Orang Asli hunter-gatherers in Malaysia (650 individuals), and Turkana pastoralists in Kenya (1155 individuals). We also drew on a separate DNAm dataset from urban, US residents (372 individuals).

We find that DNAm undergoes extensive remodeling with age in the 3 non-industrial groups (70% of tested CpGs are age-associated; n=1.3M; LFSR<5%), and that 40% of these age effects are consistent in magnitude and direction across populations (<2 fold difference in effect size). These shared effects fall into two general categories: hypermethylation linked to transcriptional repression in bivalent regions and developmental genes and hypomethylation in quiescent and non-coding regions. We find 13,998 CpGs for which age effects are shared in the non-industrial populations but amplified in the US (>2 fold difference); these are enriched in promoters and enhancers and overlapped EWAS loci for cardiometabolic disease and inflammation. Within the 3 non-industrial groups, population-specific age effects (>2 fold difference in 1 population vs others; n CpGs=129,834) are enriched in active regulatory regions, immune genes (e.g., BTLA), and loci from previous EWAS studies of mortality and cardiovascular diseases. Population-specific sites exhibit higher within-individual stability with age in our longitudinal data, compared to sites with shared age effects, which exhibit evidence for stochastic epigenetic drift. Together, our findings suggest that some epigenomic decay is unavoidable across human aging and proceeds in consistent ways across diverse environmental contexts; in contrast, age effects that are slowed or accelerated by lifestyle may be the most relevant to non-communicable disease.

STANDARDIZED rsID PROPAGATION AND COMMUNITY-DRIVEN SNP GENOTYPING: AN INTEGRATED, FAIR FRAMEWORK FOR CROP PAN-GENOMICS AND MOLECULAR BREEDING

Sharon Wei[1], Kapeel Chougule[1], Suyun Kim[1], Andrew Olson[1], Zhenyuan Lu[1], Doreen Ware[1,2]

[1]Cold Spring Harbor Laboratory, Ware Lab, Cold Spring Harbor, NY, [2]USDA ARS NAA, Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY

The establishment of stable and interoperable variant identifiers is critical for advancing crop genomics and accelerating molecular breeding. Inspired by the success of reference SNP cluster IDs (rsIDs) in human genetics, we integrated over 193 million standardized rsIDs from the European Variation Archive (EVA) into Gramene's crop pan-genome databases, including sorghum, rice, maize, and grape. Because rsIDs are initially assigned only to single reference assemblies, we pioneered a strategy to propagate these identifiers across multiple assemblies and pan-genomes using EVA's Ensembl Variant Remapping pipeline. Validation in sorghum demonstrated high remapping accuracy (~98% between reference versions and ~87% across pan-genomes), enabling scalable implementation in rice and maize. These stable identifiers are accessible through Gramene's genome browser as searchable variant tracks and gene-level annotations, supporting cross-assembly comparisons, trait association studies, and translational research. Complementing this informatics framework, we developed and validated a community-driven, mid-density sorghum SNP genotyping array using the PlexSeq™ NGS platform. The array comprises 2,421 SNPs distributed across all ten Sorghum bicolor chromosomes, including trait-associated and quality-control markers prioritized by stakeholders. Genotyping 2,726 diverse accessions achieved high call rates, low missing data, and population structure resolution consistent with whole-genome datasets. Importantly, genomic prediction accuracy for key agronomic traits matched high-density genotyping-by-sequencing platforms. Together, standardized rsID propagation and an accessible, targeted genotyping platform provide a stable, FAIR, and cost-effective infrastructure to support germplasm management, genomic prediction, and breeding applications in sorghum and other crops.

INFERENCE OF POSITIVE SELECTION USING ANCESTRAL RECOMBINATION GRAPHS.

Xinzhu (April) Wei

Cornell University, Computational Biology, Ithaca, NY

Inference of positive selection is a central theme in population genetics. Recent advances in ancestral recombination graph (ARG) inference offer new opportunities to tackle these long-standing questions. Speidel et al. proposed a test for partial selective sweeps, which leverages marginal coalescent tree topology within ARGs. Here, we proposed a simple generalization for this statistic, tested its performance, and applied it to conduct a selection scan in East Asians. Moreover, we introduced PAC, a novel neutrality test that uses coalescence times from each marginal tree or subtree to identify genomic regions under selection. We found that PAC achieves high power and sensitivity across a range of selective scenarios, including hard, soft, partial, and ancient sweeps. In addition, PAC improves localization of causal variants. While the topology-based statistic tends to benefit more from having larger sample sizes, PAC performs well even with tens of samples. We also investigated the best practices for applying these two tests to ARGs inferred from real human genomes, such as genomes from short-read sequencing and long-read assembly, and ARGs inferred by different methods. Finally, we discuss the possibilities in combining the two statistics with complementary information. Together, these approaches and results demonstrate how ARGs can be inferred for real data and utilized for the purpose of detecting positive selection.

A PATHWAY FOR DE-EXTINCTION OF BLACK-FOOTED FERRET LOCI BY GENOME WRITING

Jordan M Welker[1], Antonio Vela Gartner[1], Aleksandra M Wudzinska[1], Henrique v Figueiró[2], Klaus-Peter Koepfli[2], Jef D Boeke[1]

[1]NYU Grossman School of Medicine, Institute for Systems Genetics, New York, NY, [2]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA

As more species begin to face extinction due to habitat loss and climate change, one problem that will persist even if the species can be saved from extinction is a severe decrease in genetic diversity. The Black-footed ferret (BFF) faces such a crisis today, nearly 40 years since 7 founder animals were collected into a captive breeding program. Today's BFFs are the survivors of an extreme population bottleneck, with limited options to increase the living genetic diversity. Here we describe a genome writing method for replacing whole loci up to 120kb in common ferret cells produced from synthetic copies of genes from a deceased BFF. This process provides a pathway to de-extinct alleles from museum specimens collected from extinct populations of BFFs that never interacted with the surviving population, and thus greatly expands the potential genetic diversity that can be restored to this endangered species. We also describe attempts to derive animals from such genome written ferret cells as an avenue to restore lost genetic diversity to the living population of BFFs.

GENETIC INFLUENCE ON BLOOD PRESSURE TRAJECTORY DURING PREGNANCY

Prabhavi Wijesiriwardhana[1], Guisong Wang[2], Tesfa D Habtewold[1], Kunal Kathuria[1], Fasil Fasil Tekola-Ayele[1]

[1]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Epidemiology Branch, Division of Population Health Research, Division of Intramural Research, Bethesda, MD, [2]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, The Prospective Group, contractor for Division of Population Health Research, Division of Intramural Research, Bethesda, MD

Maternal blood pressure (BP) has a unique trajectory during pregnancy, with levels typically dropping until mid-pregnancy and then rising until delivery. The change in BP during gestation signals maternal circulatory adaptive response, and has been linked to pregnancy complications including preeclampsia, placental abruption, preterm birth, and low birth weight. Genome-wide association studies have identified hundreds of genetic loci associated with BP in adults; however, the genetic underpinnings of BP trajectory during pregnancy are unknown. We aimed to identify maternal genetic loci associated with the prenatal trajectory of four BP traits (systolic BP, diastolic BP, mean arterial pressure, and pulse pressure). Data included pregnant woman with at least 3 measurements of systolic and diastolic BP in two pregnancy cohorts (total N=7,646; 36,868 BP measurements). Pulse pressure and mean arterial pressure were calculated from systolic and diastolic BP using standard methods. To identify genetic variants that shift the mean BP trait trajectory and the intra-individual variability in BP trait trajectory, genome-wide analyses were performed in each cohort using a linear mixed effects-based model adjusted for maternal age, parity, and genotype principal components, with varying slope for gestational week at BP measurement. Results were combined by meta-analyses. We identified 41 loci associated with variability in trajectory of diastolic BP, 38 loci for systolic BP, 91 loci for mean arterial pressure, and 1 locus for pulse pressure ($p < 5x10-8$). At least eleven loci were shared by two or more BP traits. Genes linked to variability in diastolic BP and mean arterial pressure trajectories showed marked upregulation in heart left ventricle, while systolic BP–associated genes were downregulated in tibial artery, suggesting tissue-specific regulation of blood pressure components. Several loci have previously been associated with traits related to blood pressure, lipid traits, and fetal growth. It is possible that genetic loci regulate intra-individual variability in BP during pregnancy through their influence on women's cardiovascular and other physiological adaptations across gestation. These findings would help in identifying new biological pathways in blood pressure regulation with potential for improved cardiovascular disease prevention for the mother and child.

A BLUEPRINT FOR USE OF A SINGLE-CELL ATLAS IN N-OF-1 INTERPRETATION OF A CASE OF MULTIPLE CHORANGIOMA SYNDROME.

Brandon M Wilk, Manavalan Gajapathy, Elizabeth Worthey

University of Alabama at Birmingham Center for Computational Genomics and Data Science, Genetics, Birmingham, AL

The human placenta is a complex and essential, yet temporary organ characterized by a heterogenous, mosaic genomic landscape. Abnormal placental development is associated with severe complications, including fetal hydrops, stillbirth, and preeclampsia. Chorangiomas, benign capillary lesions of the placenta, can result in significant hemodynamic alterations when large or numerous. Although often attributed to late gestation localized hypoxia, emerging evidence suggests an earlier developmental origin. The molecular drivers and mechanisms in these lesions and the rare Multiple Chorangioma Syndrome (MCS), remain largely undefined. We aimed to study these mechanisms by defining central regulatory hubs across placental development and defining how disruption contributes to MCS.

Access to placental tissue across developmental stages is limited. We therefore integrated existing scRNA-Seq datasets from 5 to 41 weeks gestation to generate a single-cell reference atlas. We applied a standardized workflow using nf-core fetchngs and scrnaseq pipelines for technology-agnostic processing with FIRM for batch-aware integration across platforms. Using this integrated dataset, we generated seven stage-specific gene regulatory networks using SCENIC to define developmental regulatory programs and identify hubs of importance.

This data was then used to aid interpretation of a rare MCS case where we previously noted germline loss-of-function variation in the key hypoxia-sensing transcription factor EPAS1(HIF-2α) and somatic mutation in the PI3K-AKT-mTORC2 pathway. Mapping these alterations onto the healthy developmental gene regulatory networks supported a model in which early germline EPAS1 variation disrupts placental hypoxia responses during villous development.

This work provides a plausible molecular mechanism for a maternal and fetal life-threatening rare condition. More broadly, this framework provides a blueprint for the construction of developmentally resolved single-cell reference atlases and demonstrates how such atlases can support n-of-1 genomic interpretation even in rare, data-sparse disorders.

EXERCISE CONSTRAINS STRESS-RESPONSIVE ENHANCER ACTIVATION DURING CARDIAC AGING

Jack Clarke[1], Fujian Wu[1], Vaibhoa Janbandhu[1], Alvaro Gonzalez-Rajal[1], David Zheng[1], Xueqian Zhuang[2], HoorE Maksura[1], Robert Shearer[1], Alex Pinto[3], Lee Jones[4], Tuomas Tammela[2], Richard Harvey[1], Emily Wong[1]

[1]Victor Chang Cardiac Research Institute, Division of Molecular, Computational and Structural Biology, Sydney, Australia, [2]MSKCC, SKI, New York, NY, [3]Baker Institute, Department of Cardiometabolic Health, Melbourne, Australia, [4]City of Hope, Department of Population Sciences, Duarte, CA

Aging is accompanied by chronic inflammatory signaling that progressively reshapes tissue gene regulation, yet how it influences the enhancer landscape in vivo remains unclear. In the mouse heart, aging activated an IL-1–associated signaling axis, it expanded inflammatory macrophage states, and drove widespread AP-1 engagement at distal regulatory elements across cardiac lineages. These age-associated AP-1 sites were enriched within low-connectivity, lineage-restricted enhancers and preferentially linked to stress and extracellular matrix gene programs. Long-term exercise broadly attenuated IL-1 receptor expression, reduced AP-1 chromatin engagement, suppressed inflammatory signaling strength, and shifted transcriptomic age toward a younger state.

To test whether dynamically engaged AP-1 sites encode intrinsic regulatory differences, we interrogated their sequence logic. While nucleotide composition predicted baseline enhancer activity, sites that gained or lost AP-1 engagement with age or exercise were markedly more sensitive to full sequence disruption than stably bound sites. Thus, dynamic and stable enhancers represent fundamentally distinct regulatory architectures: dynamically bound sites depend on intact sequence organization beyond the core motif, whereas constitutively bound sites function through simpler, composition-driven mechanisms. This distinction indicates that condition-dependent enhancer activation requires coordinated regulatory context rather than modulation of canonical motif strength alone.

Together, these findings identify a stress-responsive, sequence-sensitive enhancer class that is activated during aging and restrained by exercise. They demonstrate that aging and exercise engage distinct enhancer classes rather than uniformly modulating AP-1–containing elements.

INTEGRATING LONG-READ RNA SEQUENCING WITH GENOMICS AND PHENOMICS TO DISCOVER NOVEL DISEASE-RELEVANT SPLICE-ALTERING GENETIC VARIANTS

David Wu[1,2], Feng Wang[1], Quan Sun[1,3,4], Xinjun Ji[1], Robert Wang[1], Joseph Park[1], Ryan Park[1], Stacy Woyciechowski[5], Lan Lin[1,6], William Gaynor[5], Yi Xing[1,3,4]

[1]Center for Computational and Genomic Medicine, CHOP, Philadelphia, PA, [2]MD/PhD Program, UPenn, Philadelphia, PA, [3]Department of Pathology and Laboratory Medicine, UPenn, Philadelphia, PA, [4]Department of Biomedical and Health Informatics, CHOP, Philadelphia, PA, [5]Division of Cardiothoracic Surgery, CHOP, Philadelphia, PA, [6]Department of Pathology and Laboratory Medicine, CHOP, Philadelphia, PA

Splicing variants represent 15-60% of disease variants, but they are poorly cataloged. A key driver of this gap is the use of short-read RNA sequencing (srRNA-seq), which incompletely characterizes splicing. Long-read RNA-seq (lrRNA-seq) overcomes the limitations of srRNA-seq, but current lrRNA-seq studies are small and profile healthy cohorts.

To address this limitation, we generated whole genome sequencing, phenotyping, and Nanopore lrRNA-seq on 91 CHOP Birth Defects Biorepository patients. With a median of 27.8 million reads/sample, our cohort is one of the largest, most deeply sequenced lrRNA-seq studies and the only one to profile birth defects. We use this dataset to discover splicing variants behind complex disease risk and Mendelian conditions.

To uncover common splicing variants, we performed splicing QTL (sQTL) mapping, discovering 280,979 significant sQTL-gene relationships at 0.05 FDR. Notably, 111,378 (39.6%) of these relationships were not detected in any GTEx tissue, despite GTEx having ~10X more samples. GWAS colocalization revealed 1,278 associations across 166 complex diseases. Importantly, we find a novel CD2AP sQTL, rs10676828, that activates an unannotated transposon-derived poison exon and increases Alzheimer's disease risk.

To uncover rare splicing variants, we performed allele-specific splicing analyses to identify haplotype-specific splicing events and their proximal causal mutations. We discover 57,689 splicing variants across 6,979 genes at 0.05 FDR. Notably, we find that the rare synonymous variant rs138344913 causes CARD9 exon 11 skipping. Importantly, rs138344913 protects against inflammatory bowel disease and our analysis reveals the mechanism behind this effect.

Lastly, we developed a pipeline to discover rare splicing variants that cause Mendelian diseases. Our method finds splicing outliers, links them with causal mutations, and uses machine learning to prioritize mutations that explain the patient’s phenotypes. Using this approach, we find Mendelian diagnoses for two undiagnosed patients. In one patient, we made a new diagnosis of TARP Syndrome by identifying that a synonymous RBM10 VUS causes pathogenic intron retention.

In summary, we establish a comprehensive lrRNA-seq resource to reveal novel disease-associated splicing variants.

EXPLAINABLE SEQUENCE-BASED MODEL REVEALS DIVERGENT TRANSCRIPTION INITIATION RULES IN *DROSOPHILA* AND HUMAN

Ruoxuan Wu[1], Kseniia Dudnyk[2], Jian Zhou[1]

[1]University of Chicago, Genetic Medicine, Chicago, IL, [2]UT Southwestern Medical Center, Biomedical Engineering, Dallas, TX

While sequence-based rules for human transcription initiation have been well studied through deep learning, the extent to which these rules generalize to *Drosophila* remains poorly understood. In this study, we applied an explainable sequence-based model named Puffin-Fly, to *Drosophila melanogaster* transcription initiation data and conducted a comparative analysis of promoters between flies and humans. Our model identifies key sequence patterns, including motifs, and their corresponding position- and strand-specific effects, thereby providing mechanistic hypotheses for *Drosophila* transcription initiation. Motif contribution analysis reveals clear differences between human and *Drosophila* promoters, specifically in the synergistic interplay among the Initiator (Inr) element, the Downstream Promoter Region (DPR), and the TATA box. More generally, this model could uncover how promoter sequences have been conserved or diverged across species, revealing evolutionary rules of transcription initiation across diverse taxa.

SEX-STRATIFIED SINGLE-CELL TRANSCRIPTOMIC ANALYSIS REVEALS MOLECULAR AND CELLULAR SIGNATURES ACROSS MULTIPLE PSYCHIATRIC DISORDERS

Yan Xia, Ro Malik, Zhongzheng Mao, Nancy Fang, Declan Clark, Mark Gerstein

Yale University, MBB, New Haven, CT

Sex differences are pervasive across multiple psychiatric disorders, influencing prevalence, symptomatology, and treatment response. However, the molecular mechanisms underlying these differences remain incompletely understood. Here, we present a comprehensive sex-stratified single-nucleus RNA-seq analysis of 3.3 million nuclei from 636 postmortem brains (PsychENCODE Phase III) spanning schizophrenia, bipolar disorder, autism spectrum disorder, major depressive disorder, post-traumatic stress disorder, and controls. We examined four contrasts: male versus female within cases and controls, and case versus control within males and females. Pseudobulk differential expression was performed using DESeq2, Dreamlet, and glmGamPoi to ensure robustness. Across disorders, we identified numerous sex-specific disease-associated differentially expressed genes (DEGs). Intriguingly, in each disorder, the sex with lower disease prevalence exhibited stronger transcriptional perturbations. These sex-biased DEGs were enriched for synaptic and glutamatergic signaling pathways, particularly within GABAergic and layer 6 excitatory neurons. Moving beyond single-gene analyses, network-level approaches revealed broader mechanistic insights. Key differences emerged in GABAergic signaling networks—including altered NRXN–NLGN interactions—and in gene modules regulated by estrogen-responsive transcription factors such as JUN/FOS. Integration with summary-based Mendelian randomization analyses further linked sex-biased expression to putative causal loci and known drug targets (e.g., GABRA5, SNCA), revealing sex-specific therapeutic opportunities. Together, these findings highlight the necessity of incorporating sex as a core biological variable and illustrate how integrative, systems-level approaches can illuminate the cellular and regulatory architecture underlying sex differences in psychiatric disorders.

VOUS: VARIATIONAL ORNSTEIN-UHLENBECK STOCHASTICS LINKING SINGLE-CELL LINEAGE TRACING WITH DYNAMIC GENE EXPRESSION

Jiawei Xing, Stephen Staklinski, Adam Siepel

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

Single-cell gene expression evolves dynamically along cell division histories. However, most existing single-cell methods treat cells as static snapshots, neglecting the rich information encoded in their underlying lineage structures. Recent advances in single-cell lineage tracing now enable the reconstruction of high-resolution lineage phylogenies, providing a natural framework pinpoint exactly when and where transcriptional changes occur. This capability is fundamental to decoding the dynamics of development, differentiation, and disease progression. To fully leverage this lineage information, we present VOUS (Variational Ornstein-Uhlenbeck Stochastics), a flexible probabilistic framework that models stochastic single-cell gene expression over inferred cell lineage trees. By grounding gene expression analysis in explicit cell lineage phylogenies with topology and branch lengths, VOUS enables the inference of continuous expression dynamics, despite the high sparsity and low coverage of sequencing data. We applied VOUS to scRNA-seq data from metastatic lung cancers, identifying gene programs associated with metastasis and potential therapeutic targets. By providing a rigorous foundation for modeling sparse count data on latent tree structures, VOUS establishes a generalizable framework that naturally extends to multi-gene programs, lineage uncertainty, and multi-modal integration, paving the way for a comprehensive atlas of single-cell stochastic dynamics.

ESCAPE FROM X INACTIVATION DRIVES SEX DIFFERENCES IN GENE EXPRESSION

Carrie Zhu[1,2], Liaoyi Xu[1,2], Arbel Harpak[1,2]

[1]University of Texas at Austin, Department of Integrative Biology, Austin, TX, [2]University of Texas at Austin, Department of Population Health, Austin, TX

X chromosome inactivation (XCI) partially balances gene dosage between sexes, yet many genes are expressed from the inactive X (Xi) to a variable degree. In this study, we investigate whether variation in Xi expression among genes predicts transcriptional and phenotypic consequences of X-linked variation. We find that Xi expression levels are a strong linear predictor of female-male expression differences, suggesting that other compensatory or regulatory mechanisms play a more minor role in sex differences in X-linked gene expression. Among females, we identify traits—including BMI, estradiol, and testosterone levels—for which higher Xi expression correlates with the strength of evidence for either additive or dominance effects on the trait. We hypothesize that an underappreciated mechanism could generate dominance effects of X-linked variants on a trait—specifically when the variant influences skew in X inactivation. This work establishes Xi expression as important for understanding transcriptional sex differences and physiological variation among females.

GENETIC REGULATION OF CELL TYPE–SPECIFIC CHROMATIN ACCESSIBILITY SHAPES IMMUNE FUNCTION AND DISEASE RISK

Angli Xue[1,2], Jianan Fan[1,2], Oscar Dong[1], Hao Huang[1,2], Peter Allen[1], Eleanor Spenceley[1,2], Anna Cuomo[1,2,3,4], Albert Henry[1,2], Ling Chen[5], Elizabeth Dorans[6,7], Kyle K Farh[5], Wei Zhou[7], Alkes L Price[6,7], Gemma A Figtree[8], Alex W Hewitt[9], Daniel G MacArthur[2,3,4], Joseph E Powell[1,2]

[1]Garvan Institute of Medical Research, Translational Genomics Program, Sydney, Australia, [2]University of New South Wales, Faculty of Medicine and Health, Sydney, Australia, [3]Garvan Institute of Medical Research, Centre for Population Genomics, Sydney, Australia, [4]Murdoch Children's Research Institute, Centre for Population Genomics, Melbourne, Australia, [5]Illumina, Illumina Artificial Intelligence Laboratory, San Diego, CA, [6]Harvard T.H. Chan School of Public Health, Department of Epidemiology, Boston, MA, [7]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [8]University of Sydney, Charles Perkins Centre, Sydney, Australia, [9]University of Tasmania, Menzies Institute for Medical ResearchHobart, Australia

Deciphering how genetic variation influences gene regulation is crucial for elucidating complex disease mechanisms. However, limited large-scale single-cell data have constrained our understanding of how variants regulate cell type-specific expression. Here we present chromatin accessibility profiles from 3.5M PBMCs across 1,042 donors using scATAC-seq and multiome sequencing with matched whole-genome sequencing in the TenK10K program.

We characterized 440,996 chromatin peaks across 28 cell types and mapped 243,272 chromatin accessibility quantitative trait loci (caQTLs), of which 60% are cell type-specific and 27,927 harbor cumulative rare variant signals (MAF<5%). Integrating caQTL with eQTL from TenK10K scRNA-seq data (5.4M PBMCs) identified 31,688 candidate *cis*-regulatory elements by colocalization. Incorporating caQTL with GWAS for 16 diseases and 44 blood traits revealed 9.8-30% more colocalized signals, which would be missed if using eQTL, largely due to multiple causal variants. Using a graph neural network integrating caQTL+eQTL signals with unpaired multiome data, we inferred peak-gene links with up to 80% higher accuracy than paired data without QTL and replicated 68% of CRISPR-validated links. This further enhanced gene regulatory network inference by identifying 128 additional transcription factor–target gene pairs (a 22% increase), some with drug repurposing potential like *IKZF1*.
This study provides a comprehensive single-cell map of chromatin accessibility and genetic variation in human circulating immune cells, advancing our knowledge of cell type-specific regulation for complex diseases. We are expanding TenK10K to Phase 2, targeting 50M cells from 10,000 samples by 2027, with increased diversity in disease backgrounds and ancestry.

MULTI-LAYER OMICS STUDIES TO UNDERSTAND HUMAN IMMUNE SYSTEM

Kazuhiko Yamamoto[1], Rintaro Fujimoto[1,2], Hiroki Kitaoka[1,2], Saya Hisano[1,2], Akari Suzuki[1], Yasuhiko Murakawa[1], Koshi Imami[1], Makoto Arita[1], Yosuke Isobe[1], Hiroshi Ohno[1], Shohei Asami[1], Shin-ichiro Fujii[1], Takeya Kasukawa[1], Jun Seita[1], Yukinori Okada[1]

[1]RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan, [2]The University of Tokyo, Genome Informatics, Tokyo, Japan

Genome-wide association studies (GWAS) have been used to detect genetic variants in diseases and traits involving multiple factors, such as the immune system. It was previously thought that 80-90% of genetic variations associated with diseases or traits were expression quantitative trait loci (eQTLs) related to gene expression. Furthermore, the importance of chromatin accessibility QTLs (caQTLs), which affect the degree of open chromatin upstream as a trait, is increasingly being reported. Cell-specific caQTLs and eQTLs are considered crucial. The mechanism is thought to involve the cell-specific epigenome and transcription factors, where gene expression levels are determined via enhancers and promoters, ultimately leading to traits and diseases.

Against this backdrop, we are advancing multi-omics analyses of immune cell subsets in healthy individuals, 13 steady-state and 7 stimulated states, to elucidate the mechanisms of the human immune system and to understand and control diseases. Particularly regarding subset-specific caQTLs and eQTLs related to genetic variants, we note that many colocalize with transcription factor binding sites, enhancers, and promoters in non-coding regions as well as GWAS risk variants. Therefore, we are conducting analyses using not only ATAC-seq and histone modifications to detect caQTLs, but also the CAGE method for precise measurement of 5'-end transcription RNA, and the ReapTEC method (Oguchi A et al. Science 2024). Additionally, we are analyzing subset-specific proteomes, lipidomes, and metabolomes as intermediate traits closer to the final trait, including their relationship with genetic variants suggesting causality. Furthermore, we are incorporating gut microbiota as an environmental factor and the metabolome as effector factors produced by them, ultimately aiming to construct a dataset of 400 individuals. Data analysis employs statistical genetics aimed at elucidating mechanisms and causality, alongside AI analysis focused on improving functional prediction. Utilizing this constructed dataset is expected to significantly advance research into the mechanisms of the human immune system. For instance, integrating single-cell analysis data from small volumes of peripheral blood to this data set could enable more detailed prediction of the individual immune function, contributing to the analysis of numerous human immune diseases and traits.

GENOMIC MOSAICISM REVEALS DEVELOPMENTAL ORGANIZATION OF SENSORY AND SYMPATHETIC GANGLIA

Xiaoxu Yang

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

The neural crest (NC) generates a broad spectrum of cell types that migrate across the body plan to populate multiple tissues. However, the relationship between lineages of NC derivatives remains unclear, and the extent to which NC cells delaminated from the neural tube has specified fates remains debated. Lineage reconstruction of these cells requires novel genome technologies for both model organisms and humans. Here, leveraging CRISPR barcoding in mice and mosaic variant barcode analysis in humans, we demonstrate robust bilateral progenitor clonal spread of NC progenitors along the rostrocaudal axis but limited clonal overlap between sensory and sympathetic lineages. Computational modeling of mosaic variants suggests that most NC cells show strong fate restriction before delamination. Real-time imaging of quail embryos with fluorescent proteins further reveals an FGF-dependent rostrocaudal dispersion of NC cells across multiple axial levels. These findings support a model in which NC fate bias predominantly emerges within the neural tube, while only a minor subset of delaminated progenitors retaining multipotency to generate both sensory and sympathetic derivatives.

PROFILING OF INTERNAL VARIATION OF SINE-VNTR-ALU ELEMENTS IN THE ALL OF US LONG-READ COHORT

Alex Yenkin[1,2,3], Yulia Mostovoy[2,3], Karan Jaisingh[2,3], Xuefang Zhao[2,3,4], Yongqing Huang[2,3], Fabio Cunial[5], Samuel Lee[5], Kiran Garimella[2,5], Michael Talkowski[2,3,4]

[1]Harvard University, Division of Medical Sciences, Boston, MA, [2]Massachusetts General Hospital, Center for Genomic Medicine, Boston, MA, [3]Broad Institute of MIT and Harvard, Medical & Population Genomics, Cambridge, MA, [4]Massachusetts General Hospital and Harvard University, Department of Neurology, Boston, MA, [5]Broad Institute of MIT & Harvard, Data Sciences Platform, Cambridge, MA

SINE-VNTR-Alu (SVA) elements are the youngest class of active human retrotransposon and have been implicated in a number of Mendelian diseases and common disease risk through transcriptomic modulation. They contain two tandem repeats: a hexamer repeat on their 5' end and an internal CG-rich VNTR sequence. Previous studies have found that variation within these repeats has strong associations with human health, including affecting Alzheimer's risk and age of onset for X-linked Dystonia-Parkinsonism, a rare neurodegenerative disease. However, their repetitive content and rarity compared to other classes of mobile element insertions (MEIs) has prevented large-scale study of these elements' internal variation and their consequences on human biology.

We have leveraged the unprecedented scale of long-read sequencing and multi-omics data from the All of Us (AoU) program to comprehensively profile MEIs, their internal variation, and their impacts on RNA transcription, protein levels, and disease phenotypes. Following structural variant discovery in this cohort of 13,662 individuals, we annotated Alu, LINE-1, and SVA elements within the variant dataset. Using a novel HMM-based approach validated using data from the Human Genome Structural Variation Consortium, we profiled the internal variation of SVA elements present in the reference genome and elements in the MEI dataset to identify the composition of the hexamer and VNTR sequences. We found substantial variation in these tandem repeats, with 53.4% of reference SVA hexamer sequences and 85.3% of reference VNTR sequences showing variation between individuals.

To discover SVA internal variation significantly associated with human biology, we performed association testing on the wide array of disease phenotype data available in AoU and the multi-omic data available for these individuals, including RNA-seq (n=9,043) and O-link proteomics data (n=10,043). For each phenotype and multi-omic signature, we identified significant loci by comparing previously calculated SNP-only models with models including internal SVA variation, harnessing the power of large-scale long-read-sequencing to deeply interrogate this understudied source of human genetic variation.

FUNCTIONAL DISSECTION OF CIRCULATING FATTY ACIDS-ASSOCIATED LOCI USING CRISPR-BASED GENETIC PERTURBATIONS

Ke Yi[1], Huifang Xu[1], Haifeng Zhang[1,2], Pengpeng Bi[1,2], Kaixiong Ye[1,3]

[1]University of Georgia, Department of Genetics, Athens, GA, [2]University of Georgia, Center for Molecular Medicine, Athens, GA, [3]University of Georgia, Institute of Bioinformatics, Athens, GA

Fatty acids (FA) have diverse biological functions and are associated with complex diseases, such as type 2 diabetes and cardiovascular diseases. To identify novel genetic loci that are associated with 19 human circulating fatty acid traits such as total FAs, five polyunsaturated fatty acid (PUFA) absolute concentrations and their relative percentages in total FAs, our lab conducted a genome-wide association study (GWAS) using UK Biobank data of European ancestry (n=239,268) and five other ancestries (n=508-4,663). After fine-mapping and multi-omics integrative analysis, we identified over 400 genetic loci that show significant association, but most of them reside in noncoding regions, making it difficult to define their target genes and regulatory functions. To address this challenge, we designed a gRNA library targeting these variant-containing genomic regions and performed a large-scale single-cell CRISPR screen in a human liver cancer cell line HepG2, enabling the identification of 239 cis-target genes for 238 GWAS loci. From these results, we prioritized the top 16 candidate variant-gene pairs and plan to apply CRISPR-based genetic perturbation approaches for functional validation in HepG2 cells and further explore the regulatory mechanism of these variants. Specifically, we will perform CRISPR interference to assess the regulatory relationships between variants and candidate target genes, followed by prime editing to introduce precise single nucleotide substitutions to directly validate the causal effects of these variants. In parallel, we will integrate multi-omics data such as histone modification and chromatin accessibility, and combine prediction results obtained from AI-based tools (e.g. DeepSEA and AlphaGenome) to help further assess their potential regulatory mechanism. Taken together, this study aims to bridge the gap between GWAS-identified variants and their regulatory function, thereby advancing our understanding of how genetic variants shape human lipid metabolism.

SATELLITE DNA FRAGILITY, EPIGENETIC DISRUPTION, AND EXTRACHROMOSOMAL AMPLIFICATION CONVERGE TO DRIVE STRUCTURAL GENOME INSTABILITY IN CANINE OSTEOSARCOMAS

Feyza Yilmaz[1], Wonyoung Kang[1], Sabriya A Syed[1], Francis H O'Neill[1], Jody T Lombardi[1], Patrick Kwok Shing Ng[1,2], Ching C Lau[1,2,3], Charles Lee[1,2]

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT, [2]University of Connecticut School of Medicine, Farmington, CT, [3]Connecticut Children's Medical Center, Hartford, CT

Osteosarcoma (OS) exhibits extreme instability characterized by genome shattering and complex chromosomal rearrangements. Yet despite abundant catalogs of structural variation (SV) in OS, the initiating lesions that precipitate catastrophic genome remodeling in these tumors remain unclear. Studying these mechanisms in humans is challenging because these tumors are rare, and short read sequencing cannot reliably resolve breakpoints in repetitive DNA, where chromosome fragility likely originates. Spontaneous OS in dogs offers a strong alternative model. Canine OS is similar to human OS and provides high-quality matched germline samples, making it ideal for long read sequencing to study somatic structural genomic instability.

We used PacBio HiFi long-read whole genome sequencing (51x-96x) on matched tumor and blood samples from four purebred dogs with OS, enabling us to detect SVs and CpG methylation patterns at base pair resolution across the genome.

Tumor genomes had a high number of somatic SVs (184-738 SVs per tumor) and included complex chromosomal rearrangement patterns consistent with chromothripsis and chromoplexy. SV breakpoints showed enrichment at satellite repeats, particularly the canine-specific SAT1_CF family (17.8-44.1-fold) identifying satellite DNA as recurrent sites of chromosome breakage. Integrated methylation analysis revealed that, among SVs overlapping differentially methylated regions, 93% (69 out of 74) had focal CpG hypomethylation, at or near the breakpoint, connecting local methylation loss to somatic structural disruption. TP53 was the only gene recurrently mutated across all tumors, consistent with near-universal TP53 loss in both canine and human OS.

Long read sequencing further uncovered another aspect of genome structural instability: extrachromosomal DNAs (ecDNAs). We found ecDNA in three of four tumors, ranging in size from 103 kb to 3.07 Mb. These ecDNAs often contained genes associated with OS, such as MDM2, CTNNB1, TGFBR2, and FGF7, matching amplification patterns observed in human OS and suggesting ecDNA as a mechanism by which gene dosage drives cancer in spontaneous canine disease.

Together, these findings integrate structural and epigenetic observations into a unified mechanistic framework. The strong directional coupling of satellite hypomethylation with SV breakpoints suggests that focal methylation loss at repetitive regions may predispose to chromosomal fragility. Further, ecDNA formation provides a mechanism for rapid oncogene amplification through high-copy circular structures. The use of a naturally occurring animal model and long read sequencing to simultaneously resolve breakpoints and methylation status provides a foundation for targeted cross-species investigations of satellite fragility and epigenetic instability as drivers of catastrophic rearrangements and oncogene amplification in OS.

A 200 MILLION CELL GENOME-WIDE PERTURB-SEQ ATLAS WITH CRISPRi, CRISPRa, AND siRNA

Kwontae You, Alejandro Mendez Mancilla, Dulguun Amgalan, Eyal Ben David, Hong Gao, Jiang Zhu, Doyeon Kim, Emily Laubscher, Jonatan Perez, Ling Chen, Lenka Dohnalova, Marcos Nascimento, Sebastian Pineda, Zala Sekne, Wenhe Lin, Martijn Vochteloo, Lauren Varanese, Kyle Kai-How Farh

Illumina, Artificial Intelligence Lab, San Diego, CA

To study molecular pathways across the diversity of human cell types, we performed genome-wide perturb-seq in 12 different cell lines comprising over 200 million single cells, including both IPSC-derived and cancer cell lines. Single cell data were generated using droplet emulsion PIP-seq, averaging 2 million single cells per reaction, with median on-target knockdown of 80% with CRISPRi perturb-seq, and median 2.8-fold overexpression with CRISPRa perturb-seq. For each cell type, we generated paired genome-wide CRISPRi and CRISPRa perturb-seq datasets, identifying genes and pathways where CRISPRi and CRISPRa produce symmetric or asymmetric effects on downstream gene expression. We performed orthogonal genome-wide perturbation experiments using a 60,000-element siRNA library to validate downstream targets from CRISPRi perturb-seq. Despite the two perturbation technologies acting at different stages of transcriptional regulation, we observed strong agreement in the resulting downstream gene expression programs, while also identifying off targets specific to each perturbation platform.

EXPANDING THE READABLE GENOME: A NOVEL APPROACH FOR ANALYZING MONONUCLEOTIDE C REPEATS

Zhezhen Yu[1,2], Inessa Hakker[1], Antoine Gruet[1], Asya Stepansky[1], Jude Kendall[1], Joan Alexander[1], Zihua Wang[1], Michael Wigler[1], Dan Levy[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]Stony Brook University, Department of Molecular and Cell Biology, Stony Brook, NY

Microsatellites are short, repetitive sequences scattered throughout the genome. They are among the most dynamic regions of the genome, providing important information about genomic stability, patterns of inheritance, evolution, and disease risk. Unfortunately, many microsatellite loci remain poorly characterized due to sequencing and alignment challenges. Because of slippage or stutter during PCR, mononucleotide repeats are notoriously difficult to sequence, especially mononucleotide C (mono-C) repeats.

Here, we present an integrated experimental and computational approach that significantly improves microsatellite analysis and reveals previously inaccessible variations. Our method builds on the muSeq protocol, which introduces random cytosine deamination via partial bisulfite conversion. Introducing random mutations into an otherwise perfect repeat prevents stutter during both PCR amplification and sequencing, enabling accurate sequencing of mono-C repeats.

Standard flank matching algorithms often fail when aligning reads around mono-C repeats due to numerous mutational variants in these dynamic regions. To obtain accurate genotypes for mono-C loci, we built a specialized analysis pipeline that (i) refines repeat annotations using Tandem Repeat Finder, (ii) constructs adaptive reference sequences that account for extended repeat structures, and (iii) employs a modified Needleman-Wunsch algorithm optimized for bisulfite-treated sequences and invariant to the repeat length. Then, by integrating haplotype-aware variant calling, we precisely genotype the microsatellite alleles, capturing disruptions in the repeat and identifying flanking sequence polymorphisms. We applied this framework to a dataset of 630 mono-C loci in 100 individuals. We identified many mono-C repeat alleles that are unresolvable with standard WGS. We reported the distribution of alleles observed over the population, revealing variations in both length and sequence content in microsatellites. By distinguishing the two alleles at each locus, we obtained an accurate measure of somatic variation. We found that the length of the uninterrupted repeat is the key indicator of genetic and somatic instability. By extending the range of accurately measurable sequences, our work provides unprecedented resolution into the mutational dynamics of microsatellites, paving the way for deeper insights into genetic diversity, evolutionary processes, and disease mechanisms.

FITNESS IN HUMAN POPULATIONS FOR NON-CODING GENOMIC REGIONS INFORMED BY GENOME LANGUAGE MODELS

Aziz Zafar[1], Guojie Zhong[2,6], Audrey Kris[3], Jingyi Han[4], Wendy K Chung[5], Yufeng Shen[1,6]

[1]Columbia University Irving Medical Center, Biomedical Informatics, New York, NY, [2]New York Genome Center, New York Genome Center, New York, NY, [3]Colgate University, Neuroscience, Hamilton, NY, [4]Peddie High School, Peddie High School, Hightstown, NJ, [5]Boston Children's Hospital, Pediatrics, Boston, MA, [6]Columbia University Irving Medical Center, Systems Biology, New York, NY

Non-coding variants can modulate expression levels of dosage sensitive genes and play a key role in human diseases and traits. We aim to quantify genetic constraint of non-coding regions based on fitness impacts of rare non-coding variants in human populations. For this, a primary source of information comes from allele count distributions in human populations. While allele counts allow for computing a likelihood under a population genetics model, given *s* (selection coefficient) and mutation rate for a site, a very high sample size of whole-genome sequences (WGS) would be required for well-powered and precise estimation due to the generally small effect sizes of non-coding variants. Therefore, we need to leverage strong and informative priors, for which we can utilize conservation information from multiple-sequence alignments (MSAs) and functional genomics data, such as chromatin accessibility and histone modifications. We perform a feasibility analysis using simulations and gnomAD WGS data. We first investigated if scores from genome language models (gLM) pre-trained on epigenomic and functional genomics data are correlated with regional genomic constraint metrics (Gnocchi). Using a black-box variational inference (VI) model named NoncoBayes, with Enformer and GPN-MSA embeddings as input and Gnocchi's observed-expected ratio as the label, we learned that language model embeddings can strongly augment Gnocchi's constraint metric, improving its zero-shot performance on downstream tasks. We then performed simulations to test the feasibility of integrating these sources of information in a probabilistic model. We simulated allele counts for sites in windows of various sizes using *q*, a latent parameter measuring the proportion of sites under strong selection (*s*=0.1), as opposed to neutral selection (*s*=0.0001). Then we used the Poisson-inverse-gaussian (PIG) model to obtain a likelihood for an observed value for allele count and performed maximum a-posteriori estimation for *q*. We found that estimation of *q* was accurate with informative priors and robust even in the absence of informative or correct priors, for simulated regions under strong selection. This work supports the feasibility of use a probabilistic model, in conjunction with gLMs, to estimate non-coding constraint using human population data.

POLYGLYCYLATION: RETAINED BY NEANDERTHALS, DENISOVANS, AND VIRTUALLY ALL ANIMALS BUT A LOST TRAIT IN MODERN HUMANS

Tomislav Maricic[1], Sabina Kelly-Falke[1,2], Miriam Berrieter[2], Ziqi Zhao[1], Svante Pääbo[1], Wieland B Huttner[3], Carsten Janke[4], Hugo Zeberg[1,2]

[1]Max Planck Institute for Evolutionary Anthropology, Dept. Evolutionary Genetics, Leipzig, Germany, [2]Karolinska Institutet, Dept. Physiology and Pharmacology, Stockholm, Sweden, [3]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany, [4]Institut Curie, Genome Integrity, RNA and Cancer, Paris, France

Polyglycylation is the addition of glycine chains to tubulins, a protein modification widespread in the animal kingdom and present even in single-celled protists such as *Paramecium*. It is particularly prevalent in beating cilia and flagella. In contrast, modern humans lack this capability due to two inactivating amino acid substitutions in the enzyme TTLL10. Here, we demonstrate that Neanderthals and Denisovans retained a functional *TTLL10*, enabling polyglycylation. We show that virtually all modern humans carry an inactive *TTLL10* gene. Additionally, gorillas have independently acquired a loss-of-function allele in *TTLL10*. A current hypothesis is that the absence of sperm competition in both modern humans and gorillas has led to relaxed purifying selection on this gene. To investigate the biological implications of this loss of function in humans, we are utilizing gene-edited brain organoids, transgenic zebrafish and mice, and in vitro assays. *TTLL10* represents the first example of a non-gradual functional genetic change distinguishing modern and archaic genomes.

GWAS HIGHLIGHTS THE NEURONAL CONTRIBUTION TO MULTIPLE SCLEROSIS SUSCEPTIBILITY

Lu Zeng*[1], Atlas Khan*[2], Kathryn Fitzgerald[3], Tsering Lama[1], Jessy Chen[1], the International Multiple Sclerosis Genetics Consortium[4], Tanuja Chitnis[5], Quentin Le Grand[6,7], Stéphanie Debette[6,8], Gao Wang[9], Mariko Taga[1], Krzysztof Kiryluk[2], Philip De Jager[1]

[1]Columbia University Irving Medical Center, Center for Translational and Computational Neuroimmunology & Columbia Multiple Sclerosis Center, Department of Neurology, New York, NY, [2]Columbia University Irving Medical Center, Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, New York, NY, [3]Johns Hopkins University School of Medicine, Department of Neurology, Baltimore, MD, [4]the International Multiple Sclerosis Genetics Consortium, N/A, International, NY, [5]Brigham & Women's Hospital, Anne Romney Center for Neurologic Diseases and Brigham Multiple Sclerosis Center, Department of Neurology, Boston, MA, [6]University of Bordeaux, INSERM, Bordeaux Population Health research center, Bordeaux, France, [7]Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany, [8]Bordeaux University Hospital, Department of Neurology, Institute for Neurodegenerative Diseases, Bordeaux, France, [9]Columbia University Irving Medical Center, The Gertrude H. Sergievsky Center and the Department of Neurology, New York, NY

Multiple Sclerosis (MS) is a chronic inflammatory and neurodegenerative disease. Genetic studies have identified susceptibility risk loci that primarily impact immune cells and microglia. Here, a multi-ancestry genome-wide association study with 20,831 MS and 729,220 control participants identified 236 susceptibility variants outside the Major Histocompatibility Complex, including four novel genomic loci. We derived a polygenic score for MS; optimized for European ancestry, it is informative for African-American and Latino participants. Integrating single-cell data from blood and brain tissue, we identified 76 candidate causal genes. Notably, inhibitory neurons emerged as a key target cell type for MS variants, with 7 loci such as *STAT3* displaying altered expression only in inhibitory neurons. The *STAT3* variant is also associated with cognition and white matter integrity in non-MS individuals and greater sNfL levels in MS participants, suggesting that MS susceptibility may involve altered brain development and a CNS less resilient to inflammatory challenges.

SEQUENCE-BASED REGULATORY CODE FOR HETEROGENEOUS AND DYNAMIC CHROMATIN

Ruoyu Wang, Junru Jin, Jian Zhou

University of Chicago, Section of Genetic Medicine, Department of Medicine, Chicago, IL

Much of our understanding of the regulatory code has been built upon the view of chromatin as a single, sequence-determined, state, typically measured through population average across cells. Single-molecule techniques provide the opportunity to push measurement and modeling closer to the molecular mechanisms, by capturing the distribution of heterogeneous chromatin states at near basepair resolution. Here we present a sequence-based generative modeling approach that faithfully captures the distribution of single-molecule chromatin states, including co-accessibility patterns not visible from bulk data. This sequence-to-distribution framework allows us to uncover distinct modes of transcription factor effects on nucleosomes, including roles of motifs in positioning the +1 nucleosome in transcription start sites, the dependencies between accessibility and methylation across transcription factor motifs, and identifying factors that mediate long-range regulatory coupling. Extending beyond static snapshots, we generated single-molecule simulations of pseudo-dynamic chromatin trajectories as a step toward a true dynamic model of chromatin. Furthermore, our models enable interpretation of the genetic variant effects at single-molecule resolution. Together, this work establishes a sequence-to-distribution framework that effectively transforms single-molecule regulatory measurements into a virtual chromatin model, providing a foundation for decoding how genomic sequence encodes heterogeneous and dynamic chromatin behavior and how genetic variation perturbs it, one molecule at a time.

LEARNING CONTEXT SPECIFIC DISEASE MECHANISMS FROM SINGLE CELL DATA

Alexander Gusev

Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts

Genome-Wide Association Studies have now identified hundreds of thousands of disease-associated loci, but most associations still cannot be linked to specific gene functions: a “missing mechanism" problem. Some missing mechanisms may reside in specific contexts, such as individual cell types, cell states, or environmental/cellular exposures. I will describe new methods for identifying context-specific mechanisms using single cell data and connecting them to complex disease. We propose a method for quantifying cell-type/state interaction heritability of gene expression using population scale sing-cell data. Applying the method to single-nucleus RNA-seq from the ROSMAP brain study, we find that cell-type and cell-state specific heritability is sizable and comparable to the heritability of main effects. Consistent with prior work, we find that genes under stronger evolutionary constraint tend to have lower heritability (and, by extension, fewer eQTL). Surprisingly, we also find that the same genes are likely have higher cell-state specific heritability, suggesting that common cell-state (but not cell-type) specific effects may be particularly relevant to functional mechanisms. To validate these cell-state specific effects, we propose a method for inferring cell-type/state interaction effects from genome-wide perturbational screens by leveraging sparse, low rank representations of the data. We apply this approach to perturb-multiome data across hematopoietic differentiation and show that our approach can recover substantially more cell-type specific interactions than conventional methods in both scATAC-seq and scRNA-seq modalities. Overall, our findings demonstrate that single-cell data provides important new insights into the missing mechanisms of disease.

POPULATION GENETIC STRUCTURE THROUGH TIME USING LATENT SPACE MODELS

John Novembre[1,2]

[1]University of Chicago, Human Genetics, Chicago, IL, [2]University of Chicago, Ecology and Evolution, Chicago, IL

The patterns of biological variation we observe today, from deep phylogenetic splits to isolation by distance, arise from how genetic ancestry is shared among individuals. Recent progress in the inference of ancestral recombination graphs (ARGs) provides new opportunities to understand genetic ancestry with unprecedented detail; but it remains challenging to uncover the layers of genetic structure that ARGs can in principle reveal. In this talk, I will share recent progress from my group in this area, including a new conceptual framework that provides a flexible, individual-based, time-varying view of genetic ancestry. A feature of this approach is that it emphasizes variation in the rates at which genetic ancestors coalesce in a way that reflects the consequences of population size change and migration without requiring pre-defined demographic models. Using simulations, I will demonstrate the approach's utility and subtleties in interpreting its outputs. Finally, I will share applications of the method to chimpanzees, bonobos, and humans, illustrating how the method can provide new resolution on time-varying genetic ancestry in these species.

NOTES

NOTES

NOTES

Participant List

Elnaz Abdollahzadeh
University of California, Irvine
elnaza@uci.edu

Dr. Mohamed Abuelanin
University of California, Davis
mabuelanin@gmail.com

Prof. Alexej Abyzov
Mayo Clinic
abyzov.alexej@mayo.edu

Mr. Sandesh Acharya
University of Calgary
sandesh.acharya@ucalgary.ca

Mr. Temidayo Adeluwa
The University of Chicago
temi@uchicago.edu

Dr. Quadri Adewale
Beth Israel Deaconess Medical Center
qadewale@bidmc.harvard.edu

Mr. Jeremy Aguilar
Duke University School of Medicine
jeremy.aguilar@duke.edu

Dr. Kennedy Agwamba
Stanford University
agwamba@stanford.edu

Dr. Nirmala Akula
National Institutes of Health
akulan@mail.nih.gov

Valeria Anorve Garibay
Brown University
valeria_anorve_garibay@brown.edu

Dr. Peter Arndt
Max Planck Institute for Molecular Genetics
arndt@molgen.mpg.de

Peter Audano
The Jackson Laboratory
paudano@gmail.com

Mr. Kailash Babu Panneerselvam
Icahn School of Medicine at Mount Sinai
kailashbp10@gmail.com

Ms. Gali Bai
University of California, Santa Cruz
gabai@ucsc.edu

Dr. Floris Barthel
TGen
fbarthel@tgen.org

Dr. Michelle Bartolo
Mass Eye and Ear
mbartolo@meei.harvard.edu

Dr. Anindita Basu
University of Chicago
onibasu@uchicago.edu

Dr. Alexis Battle
Johns Hopkins University
ajbattle@jhu.edu

Dr. Christine Beck
The Jackson Laboratory and UConn Health
Christine.Beck@jax.org

Dr. Andres Bendesky
Columbia University
a.bendesky@columbia.edu

Dr. Jan Bergmann
Atrandi Biosciences Inc.
jan.bergmann@atrandi.com

Dr. Rameen Beroukhim
Dana Farber Cancer Institute/Harvard Medical School
Rameen_Beroukhim@dfci.harvard.edu

Mr. Sarang Bhutada
University College Dublin
sarang.bhutada@ucdconnect.ie

Mr. Dmitry Biba
Cold Spring Harbor Laboratory
dmitriy.biba@gmail.com

Prof. Minou Bina
Purdue University
Bina@Purdue.edu

Ms. Kate Blackwell
Stony Brook University
Kate.Blackwell@stonybrook.edu

Dr. Malgorzata Borczyk
Maj Institute of Pharmacology PAS
malgorzata.m.borczyk@gmail.com

Dr. Alan Boyle
University of Michigan
apboyle@umich.edu

Layla Brassington
Vanderbilt University
layla.brassington@vanderbilt.edu

Dr. Sean Bresnahan
The University of Texas MD Anderson Cancer Center
stbresnahan@mdanderson.org

Kelly Brewer
Icahn School of Medicine at Mount Sinai
kelly.brewer@mssm.edu

Dr. Joanne Bujnoski
Woodlands Radiation Oncology
jbujnoski@gmail.com

Dr. Ben Calverley
Scripps Research
bcalverley@scripps.edu

Wenjia Cao
NIAID/NIH
caow3@nih.gov

Maya Caskey
California Institute of Technology
mcaskey@caltech.edu

Ms. Sophie Chapelle
Pensieve Health
sophiechapelle@pensievehealth.com

Prof. Taosheng Chen
St. Jude Children's Research Hospital
taosheng.chen@stjude.org

Dr. Xiyue Chen
CD Genomics
genome@cd-genomics.com

Alexander Chen
University of Chicago
awchen55@uchicago.edu

Xingyi Chen
Johns Hopkins University
xchen274@jh.edu

Yixuan Chen
University of Chicago
yixuanc@uchicago.edu

Kapeel Chougule
Cold Spring Harbour Laboratory
kchougul@cshl.edu

Trevor Christensen
Cold Spring Harbor Laboratory
christen@cshl.edu

Ms. Zoe Clarke
University of Cambridge
zoe.clarke@utoronto.ca

Brian Cleary
Boston University
bcleary@bu.edu

Ms. Celeste Cohen
Wellcome Sanger Institute
cc53@sanger.ac.uk

Dr. Leonardo Collado Torres
Lieber Institute for Brain Development; JHBSPH
leo.collado@libd.org

Dr. Vincenza Colonna
University of Tennessee
enza.colonna@gmail.com

Anna Cormack
University of Chicago
cormack@uchicago.edu

Dr. John Costella
Personal interest
john.costella@gmail.com

Sylvia Dai
NYU Grossman School of Medicine
yd2812@nyu.edu

Dr. Simona Dalin
Broad Institute & Dana Farber Cancer Institute
sdalin@broadinstitute.org

Dr. Aidan Daly
New York Genome Center
adaly@nygenome.org

Prof. Mehdi Damaghi
Stony Brook University
mehdi.damaghi@stonybrookmedicine.edu

Dr. Jyoti Dayal
National Institutes of Health
jyotig@nih.gov

Mr. William DeGroat
Rutgers, The State University of New Jersey
williambdegroat@gmail.com

Brenda Delamonica
Stony Brook University / IRACDA
brenda.delamonica@stonybrook.edu

Dr. Jennifer DeLeon
Genome Research, Senior Assistant Editor
deleon@cshl.edu

Kira Delmore
Columbia University
ked2195@columbia.edu

Dr. Ahmet Denli
CSHL Press
denli@cshl.edu

Ms. Astrid Deschenes
Cold Spring Harbor Laboratory
deschene@cshl.edu

Ross DeVito
University of California San Diego
rdevito@ucsd.edu

Dr. Kushal Dey
Memorial Sloan Kettering Cancer Center
deyk@mskcc.org

Mr. Alex Diaz-Papkovich
Brown University
alex_diaz-papkovich@brown.edu

Dr. Laura Domenech Salgado
Broad Institute of MIT and Harvard
ldomenec@broadinstitute.org

Dr. Zheng Dong
Washington University in St. Louis
zdong@wustl.edu

Dr. Huw Dorkins
University of Oxford
huw.dorkins@spc.ox.ac.uk

Dr. Julie Dragon
University of Vermont
julie.dragon@med.uvm.edu

Ms. Ava Dreiband
American Society of Human Genetics (ASHG)
aldreiband@gmail.com

Laura Duntsch
Livestock Improvement Corporation
laura.duntsch@lic.co.nz

Hope Eden
Johns Hopkins University
heden3@jh.edu

Mohamed Yousry ElSadec
Boston University
myousry@bu.edu

Prof. Barbara Engelhardt
Stanford University/Gladstone Institutes
barbara.e.engelhardt@gmail.com

Dr. Tiago Faial
Nature Genetics
tiago.faial@us.nature.com

Tim Fessenden
Life Science Alliance
t.fessenden@life-science-alliance.org

Ms. Ingrid Flaspohler
University of Michigan
imf@umich.edu

Willard Ford
University of Pennsylvania
willardf@pennmedicine.upenn.edu

Rienna Franks
University of Missouri
riennafranks@missouri.edu

Dr. Tawfiq Froukh
Philadelphia University in Jordan
tfroukh@philadelphia.edu.jo

Dr. Naoko Fujito
Niigata University
naokofujito@gmail.com

Dr. Manavalan Gajapathy
University of Alabama at Birmingham
magajapathy@uabmc.edu

Dr. Pedro Galante
Hospital Sirio Libanes
pgalante@mochsl.org.br

Jake Galvin
Johns Hopkins University
jgalvin6@jh.edu

Dr. Hong Gao
Illumina Inc.
hgao@illumina.com

Dr. Mateusz Garbulowski
Uppsala University
mateusz.garbulowski@igp.uu.se

Dr. David Garfield
Genentech
garfield@gene.com

Dr. Erik Garrison
University of Tennessee Health Science Center
erik.garrison@gmail.com

Dr. Audrey Gasch
University of Wisconsin-Madison
agasch@wisc.edu

Deepa Gayadin
NHGRI
gayadind2@nih.gov

Dr. Alejandro Gil Gomez
Stony Brook University
alegigo@gmail.com

Dr. Yoav Gilad
University of Chicago
gilad@uchicago.edu

Camila Goclowski
University of Utah
camila.goclowski@genetics.utah.edu

Dr. Fernando Goes
Johns Hopkins
fgoes1@jhmi.edu

Dr. David Gokhman
Weizmann Institute of Science
david.gokhman@weizmann.ac.il

Dr. Michael Goldberg
University of Utah
megoldberg2@gmail.com

Prof. Amy Goldberg
University of California, Los Angeles
amygoldberg@mednet.ucla.edu

Huanfa Gong
Zhejiang University
gonghuanfa@zju.edu.cn

Dr. Kevin Gori
University of Cambridge
kcg25@cam.ac.uk

Dr. Yogesh Goyal
Northwestern University
yogesh.goyal@northwestern.edu

Gabriel Griffin
Dana-Farber Cancer Institute
gabriel_griffin@dfci.harvard.edu

Dr. Jeremy Grushcow
Juniper Genomics
jeremy@junipergenomics.com

Andy Gu
Yale University
andy.gu@yale.edu

Dr. Rodrigo Gularte Merida
Memorial Sloan Kettering Cancer Center
gularter@mskcc.org

Dr. Qi Guo
Insmed Innovation UK
qi.guo@insmed.com

Mr. Hersh Gupta
Albert Einstein College of Medicine
hersh.gupta@einsteinmed.edu

Dr. Gabriela Gurria
Wellcome Sanger Institute
gg7@sanger.ac.uk

Dr. Alexander Gusev
Dana-Farber Cancer Institute
Alexander_Gusev@DFCI.HARVARD.EDU

Dr. Tesfa Habtewold
National Institutes of Health
tesfa.habtewold@nih.gov

Dr. Maximilian Haeussler
University of California, Santa Cruz
mhaeussl@ucsc.edu

Dr. Christopher Harbort
Max Planck Institute for Infection Biology
harbort@mpiib-berlin.mpg.de

Dr. Arbel Harpak
The University of Texas at Austin
arbelharpak@utexas.edu

Taylor Head
The University of Texas MD Anderson Cancer Center
sthead@mdanderson.org

Ms. Prajna Hebbar
University of California Santa Cruz
pnhebbar@ucsc.edu

Mr. Jakob Heinz
Harvard University
jheinz@g.harvard.edu

Dr. Amy Herbert
University of Chicago
herbert6@uchicago.edu

Dr. Rachael Herman
Stony Brook University
rachael.herman@stonybrook.edu

Dr. Alli Hickman
EpiCypher, Inc
ahickman@epicypher.com

Dr. Paul Hook
The American Journal of Human Genetics
phook@ashg.org

Dr. Wolfram Hops
Radboud University Medical Center
wolfram.hops@radboudumc.nl

Kathleen Houlahan
McMaster University
houlahke@mcmaster.ca

Jingqing Hu
Dana Farber Cancer Institute
jasonhuusa@gmail.com

Dr. Xiaoqin Huang
National Institutes of Health
xiaoqin.huang@nih.gov

Dr. Jaan Huik
University of Tartu
jaanmarten.huik@kliinikum.ee

Dr. Tobias Hunt
EMBL-EBI
toby@ebi.ac.uk

Dr. Dam Kim Tuyen Huynh
University of South Carolina
huynhdam@email.sc.edu

Dr. Taeyoung Hwang
Lieber Institute for Brain Development
taeyoung.hwang@libd.org

Dr. Florian Iser
KiTZ Heidelberg
f.iser@kitz-heidelberg.de

Dr. Pritesh Jain
LKCMedicine, Nanyang Technological University
priteshrajesh.jain@ntu.edu.sg

Brendan Jamison
university of chicago
bvjamis2@uchicago.edu

Dr. Minal Jamsandekar
Cold Spring Harbor Laboratory
jamsande@cshl.edu

Dr. Peilin Jia
Beijing Institute of Genomes
pjia@cncb.ac.cn

Mr. Yunzhe Jiang
Yale University
yunzhe.jiang@yale.edu

Dr. Juan Jiang
Washington University
juanj@wustl.edu

Mr. Junru Jin
University of Chicago
junruj@uchicago.edu

Dr. Granton Jindal
University of California San Diego
gjindal@ucsd.edu

Dr. Leia Judge
Nature Communications
leia.judge@nature.com

Dr. Morten Kallberg
Novo Nordisk
qmok@novonordisk.com

Yijie Kang
Stony Brook University
ykang@cshl.edu

Hannah Kania
Duke University
kania.hannah@duke.edu

Dr. Charikleia Karageorgiou
University at Buffalo
charikle@buffalo.edu

Dr. Dave Kaufman
NHGRI
dave.kaufman@nih.gov

Dr. Rebecca Keener
Johns Hopkins University
rkeener@jhmi.edu

Mr. Cameron Kelsey
Arizona State University
ckelsey4@asu.edu

Dr. Janet Kelso
Max Planck Institute for Evolutionary Anthropology
kelso@eva.mpg.de

Dr. Alex Kentsis
Memorial Sloan Kettering Cancer Center
kentsisresearchgroup@gmail.com

Shareef Khalid
Cold Spring Harbor
khalid@cshl.edu

Dr. Aziz Khan
MBZUAI
aziz.khan@mbzuai.ac.ae

Dr. Seri Kitada
Wellcome Sanger Institute, University of Cambridge
sk25@sanger.ac.uk

Noah Klimkowski Arango
Clemson University
nklimko@clemson.edu

Mr. Justin Koesterich
Rutgers University
jhk148@scarletmail.rutgers.edu

Dr. Syndi Koltz
Nabsys
koltz@nabsys.com

Dr. Eli Kopel
Dexoligo by Dexcel Pharma
eli.kopel@dexcel.com

Dr. Kanako Koyanagi
Hokkaido University
kkoyanag@ist.hokudai.ac.jp

Dr. Anat Kreimer
Rutgers University
kreimer@cabm.rutgers.edu

Dr. Andrea Kriz
Boston Children's Hospital
andrea.kriz@childrens.harvard.edu

Dr. Nurdan Kuru
Cold Spring Harbor Laboratory
kuru@cshl.edu

Yanina Kuzminich
New York University Grossman School of Medicine
yanina.kuzminich@nyulangone.org

Dr. Anna Lagani
NYU Grossman School of Medicine
anna.berenson@nyulangone.org

Dr. Eric Lander
Broad Institute of MIT and Harvard
esloffice@broadinstitute.org

Ms. Anastasiia Latypova
Moscow Institute of Physics and Technology
ana.a.latypova@gmail.com

Dr. Amanda Lea
Vanderbilt University
amanda.j.lea@vanderbilt.edu

Ms. Charlotte LeMay
University of Texas at Austin
cmlemay@cs.utexas.edu

Dr. Adam Lenhart
University of Virginia
Benedictlenhart@gmail.com

Ana Leon-Apodaca
Penn State
avl5902@psu.edu

Julia Lewandowski
New York Genome Center
jlewandowski@nygenome.org

Dr. Yining Li
Columbia University
yl4437@cumc.columbia.edu

Dr. Jinghui Li
University of Chicago
jhuli@uchicago.edu

Dr. Boxun Li
Duke University
boxun.li@duke.edu

Dr. Nancy Li
Ontario Institute for Cancer Research
n2li@oicr.on.ca

Prof. Zhikai Liang
North Dakota State University
zhikai.liang@ndsu.edu

Dr. Lifan Liang
University of Chicago
xinhe@uchicago.edu

Linda Lin
Yale University
linda.yq.lin@yale.edu

Dr. Tianjie Liu
WashU Medicine
tianjie@wustl.edu

Fei Liu
National University of Singapore
phoebefl530720@gmail.com

Ms. Chunming Liu
Clemson University
chunmil@g.clemson.edu

Dr. Zhihan Liu
Cold Spring Harbor Laboratory
zhliu@cshl.edu

Mr. Josh Liu
University of Chicago - Chicago, IL
ljq@uchicago.edu

Dr. Aoxing Liu
Broad Institute
liuaoxin@broadinstitute.org

Dr. Xiran Liu
Brown University
xiran_liu1@brown.edu

Zhengtong Liu
UCLA
ericliu2023@g.ucla.edu

Ms. Jiayi Liu
Rutgers University
jl2791@scarletmail.rutgers.edu

Prof. Boxiang Liu
National University of Singapore
boxiangliu@nus.edu.sg

Ms. Jasmine Liu
Brown University
jasmine_c_liu@brown.edu

Mr. Alejandro Llanos-Lizcano
University of Vienna
alejandro.llanos.lizcano@univie.ac.at

Dr. Kaiser Loell
Cold Spring Harbor Laboratory
loell@cshl.edu

Dr. Brandon Logeman
University of Kentucky
brandon.logeman@uky.edu

Dr. Glennis Logsdon
University of Pennsylvania
glogsdon@pennmedicine.upenn.edu

Amy Longtin
Vanderbilt University
amy.l.longtin@vanderbilt.edu

Ms. Wenhan Lu
Broad Institute
wlu@broadinstitute.org

Mr. Zhenyuan Lu
CSHL
luj@cshl.edu

Ms. Yueqi Lu
University of Georgia
yueqi.lu@uga.edu

Dr. Varvara Lukyanchikova
Virginia Tech
lukva@vt.edu

Dr. Kaixuan Luo
University of Chicago
kevinlkx@gmail.com

Thong Luong
National Cancer Institute
thong.luong@nih.gov

Mr. Luiz Machado
Cold Spring Harbor Laboratory
lmachado@cshl.edu

Dr. A K M Firoj Mahmud
Uppsala University
mahmud.firoj@imbim.uu.se

Ms. Nadja Makki
University of Florida
nadja.makki@ufl.edu

Mr. Benjamin Mallory
University of Washington
bmallo@uw.edu

Ms. Keenan Manpearl
University of Colorado, Anschutz Medical Campus
keenan.manpearl@cuanschutz.edu

Pablo Mantilla Puccetti
Cold Spring Harbor Laboratory
pmantill@cshl.edu

Prof. Yafei Mao
Shanghai Jiao Tong University
yafmao@sjtu.edu.cn

Dr. Blaise Mariner
Arizona State University
bmarine2@asu.edu

Dr. Franco Marsico
University of Tennessee
franco.lmarsico@gmail.com

Dr. William McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Dr. Matthew McCoy
University of Chicago
mjmccoy@uchicago.edu

Mr. Jazeps Medina Tretmanis
Brown University
jazeps_medina_tretmanis@brown.edu

Arnav Mehta
Stanford University
arnav@thecolumngroup.com

Ms. Gopika Menon
University of Turku j
gopika.g.jayanmenon@utu.fi

Prof. Laurent Mesnard
INSERM UMR1155
laurent.mesnard@aphp.fr

Dr. Matthew Meyerson
Dana-Farber Cancer Institute
matthew_meyerson@dfci.harvard.edu

Dr. Stephen Meyn
University of Wisconsin - Madison
stephen.meyn@wisc.edu

Dr. Dan Mishmar
Ben-Gurion University of the Negev
dmishmar@bgu.ac.il

Ruthie Mitchell
Broad Institute
cmitchel@broadinstitute.org

Dr. Fahime Mohamadnejad Sangdehi
Uppsala University
fhmohamadnejad@gmail.com

Dr. Pejman Mohammadi
University of Washington
pejmanm@uw.edu

Dr. Nilah Monnier Ioannidis
University of California, Berkeley
nilah@berkeley.edu

Dr. Alison Monroe
Springer Nature
alison.monroe@springernature.com

Dr. Carolina Montano
Children's Hosp of Phil & Univ of Pennsylvania
cmontan2@gmail.com

Dr. Daniela Moreira Mombach
Hospital Sirio-Libanes
danielamombach@gmail.com

Mr. Luke Morina
Johns Hopkins University
lmorina2@jhmi.edu

Dr. Stephanie Morris
John Templeton Foundation
smorris@templeton.org

Prof. Ali Mortazavi
University of California, Irvine
ali.mortazavi@uci.edu

Dr. Hakhamanesh Mostafavi
NYU School of Medicine
hakhamanesh.mostafavi@nyulangone.org

Dr. Kousuke Mouri
The Jackson Laboratory
kousuke.mouri@jax.org

Dr. Elizabeth Murchison
University of Cambridge
epm27@cam.ac.uk

Dr. Alan Murphy
Cold Spring Harbor Labs
amurphy@cshl.edu

Dr. Kitty Murphy
Globe Institute, University of Copenhagen
kitty.murphy@sund.ku.dk

Ms. Shloka Negi
University of California, Santa Cruz
shnegi@ucsc.edu

Dr. Anton Nekrutenko
Penn State / galaxyproject.org
anton@nekrut.org

An Nguyen
Stony Brook University
an.nguyen.3@stonybrook.edu

Ms. Carol Nguyen
National Institute on Aging
carol.nguyen@nih.gov

Ms. Shreya Nirmalan
Wayne State University
go7535@wayne.edu

John Novembre
University of Chicago
jnovembre@uchicago.edu

Dr. Yevgeniya Nusinovich
AAAS/Science
ynusinov@aaas.org

Dr. Dong-Ha Oh
National Center for Biotechnology Information
dongha.oh@nih.gov

Naima Okami
Columbia University
neo2118@columbia.edu

Dr. Hanna Ollila
Institute for Molecular Medicine Finland
hanna.m.ollila@helsinki.fi

Mr. Andrew Olson
Cold Spring Harbor Laboratory
olson@cshl.edu

Mr. Michael Olufemi
University of Massachusetts Lowell
michael_olufemi@student.uml.edu

Mr. Alejandro Ortigas-Vasquez
The Pennsylvania State University
amo5740@psu.edu

Prof. Svante Paabo
Max Planck Institute for Evolutionary Anthropology
paabo@eva.mpg.de

Dr. Chiara Paleni
Human Technopole
chiara.paleni@fht.org

Dr. Petra Palenikova
Broad Institute
ppalenik@broadinstitute.org

Mingzuyu Pan
Penn state University
mzp5919@psu.edu

Dr. Kunal Pandit
New York Genome Center
kpandit@nygenome.org

Dr. Bogdan Pasaniuc
Perelman School of Medicine, UPenn
bgp@upenn.edu

Lauren Patterson
Wake Forest University School of Medicine
lauren.e.patterson@wfusm.edu

Katarina Pavlovic
University of Michigan
katrinp@umich.edu

Mr. David Peede
Brown University
david_peede@brown.edu

Dr. Stacey Pereira
Baylor College of Medicine
spereira@bcm.edu

Dr. Silvia Perez-Lluch
Center for Genomic Regulation
silvia.perez@crg.cat

Dr. Rachel Petersen
Vanderbilt University
rachel.m.petersen@vanderbilt.edu

Dr. Lon Phan
NIH
lonphan@ncbi.nlm.nih.gov

Dr. Seema Plaisier
National Institutes of Health
seema.plaisier@nih.gov

Lizzie Plender
University of Washington
plendere@uw.edu

Sebastian Pott
University of Chicago
spott@uchicago.edu

Gabriel Preising
Stanford University
gpreisin@stanford.edu

Dr. Jonathan Pritchard
Stanford University
pritch@stanford.edu

Dr. Sambhawa Priya
University of Chicago
priyas@uchicago.edu

Dr. Zhen Qiao
Garvan Institute of Medical Research
z.qiao@garvan.org.au

Dr. Yijian (Evan) Qiu
Cold Spring Harbor Laboratory
qiu@cshl.edu

Dr. Aaron Quinlan
University of Utah
aquinlan@genetics.utah.edu

Mr. Henry Raeder
The University of Chicago
hraeder@uchicago.edu

Dr. Mahmudur Rahman Hera
Rutgers, NJ
mahmudur.r@rutgers.edu

Dr. Sri Raj
Albert Einstein College of Medicine
srilakshmi.raj@einsteinmed.edu

Dr. Anil Raj
Calico Life Sciences
anil@calicolabs.com

Ms. Sri Gouri Rajaram
NYU Grossman School of Medicine
srigouri.rajaram@nyulangone.org

Prof. Sohini Ramachandran
Brown University
sramachandran@brown.edu

Dr. Erin Ramos
National Institutes of Health
ramoser@nih.gov

Dr. Aline Real
New York Genome Center
areal@nygenome.org

Dr. Arang Rhie
NHGRI
arrhie@gmail.com

Dr. Stephan Riesenberg
Max Planck Institute for Evolutionary Anthropology
stephan_riesenberg@eva.mpg.de

Kaeli Rizzo
Cold Spring Harbor Laboratory
rizzo@cshl.edu

Dr. Xavier Roca-Rada
Brown University
xavier_roca_rada@brown.edu

Dr. Jeffrey Rogers
Baylor College of Medicine
jr13@bcm.edu

Dr. Jeff Ross-Ibarra
University of California Davis
rossibarra@ucdavis.edu

Mr. Harshit Sahay
Memorial Sloan Kettering Cancer Center
sahayh1@mskcc.org

Dr. Irepan Salvador-Martinez
Barcelona Supercomputing Center
irepan.salvador@bsc.es

Dr. Mark Sanda
Stony Brook University
mark.sanda@stonybrook.edu

McKinley Santiago
Johns Hopkins University
msanti19@jh.edu

Dr. Miguel Santo Domingo
Cold Spring Harbor Laboratory
santodom@cshl.edu

Dr. Michael Schatz
Johns Hopkins University
mschatz@cs.jhu.edu

Dr. Paul Schaughency
NIAID/NIH/Axle Informatics
schaughencypm@nih.gov

Madyson Scherr
University of Chicago
mscherr@uchicago.edu

Dr. Megan Schertzer
New York Genome Center
mschertzer@nygenome.org

Dr. Brian Schilder
Cold Spring Harbor Laboratory
schilder@cshl.edu

Jacob Schlamowitz
Northwestern University School of Medicine
jacob.schlamowitz@nortrhwestern.edu

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Mr. Leon Schwartz
Northwestern University
leon.schwartz@northwestern.edu

Ms. Brooklynn Scott
Arizona State University
brscott4@asu.edu

Dr. Carolina Segami Marzal
Duke University
carolina.segami@duke.edu

Ms. Janan Semseddin
NIH/NIDDK
janan.semseddin@nih.gov

Vivaswat Shastry
AbbVie
vivaswat.shastry@abbvie.com

Rintsen Sherpa
University of Michigan
rintsen@umich.edu

Dr. Alaina Shumate
Dana Farber Cancer Insistute
alainashumate@gmail.com

Dr. Jacob Sieg
Penn State University
jus841@psu.edu

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Param Singh
University of California, San Francisco
param.singh@ucsf.edu

Ms. Elelta Sisay
National Cancer Institute (NCI/NIH)
easb2018@mymail.pomona.edu

Mr. Oliver Smith
Wellcome Sanger Instituite
os10@sanger.ac.uk

Hannah Snell
Brown University
hannah_snell@brown.edu

Dr. Noah Snyder-Mackler
Arizona State University
nsnyderm@asu.edu

Mr. Steven Solar
Harvard University | MIT
steven.j.solar@gmail.com

Dr. Volker Soltys
Max Planck Institute for Evolutionary Biology
volker_soltys@eva.mpg.de

Dr. Janet Song
Harvard University
janetsong@fas.harvard.edu

Dr. Ramprakash Srinivasan
Calico Life Sciences
rams@calicolabs.com

Jaya Srivastava
National Institutes of Health
jaya.srivastava@nih.gov

Mr. Stephen Staklinski
Cold Spring Harbor Laboratory
staklins@cshl.edu

Ms. Margaret Starostik
Johns Hopkins University
mstaros1@jhu.edu

Mr. Alexander Starr
Stanford University
astarr97@stanford.edu

Mr. Jiayu Su
Columbia University
js5756@cumc.columbia.edu

Michelle Sun
University of Chicago
michellesun@uchicago.edu

Ms. Yuxuan Sun Sun
Clemson University
yuxuans@clemson.edu

Mengyi Sun
Cold Spring Harbor Laboratory
msun@cshl.edu

Dr. Hillary Sussman
Genome Research, Executive Editor
hsussman@cshl.edu

Maha Syed
Cold Spring Harbor Lab
syed@cshl.edu

Dr. Kar-Tong Tan
National University of Singapore
ktan@nus.edu.sg

Taotao Tan
Baylor College of Medicine
Taotao.Tan@bcm.edu

Fahim Rejanur Tasin
University of Florida
fa.tasin@ufl.edu

Dr. Thais Tavares
Morehouse School of Medicine
ttavares@msm.edu

Dr. Shaolei Teng
Howard University
shaolei.teng@howard.edu

Dr. Jitendra Thakur
Emory University
jthakur@emory.edu

Ms. Polina Tikhonova
PhD student
pmt5304@psu.edu

Dr. Winston Timp
Johns Hopkins University
wtimp@jhu.edu

Edmundo Torres-Gonzalez
University of Minnesota-Twin Cities
edmundo@umn.edu

Dr. Michelle Trenkmann
Springer Nature AG & Co. KG aA
michelle.trenkmann@nature.com

Dr. Lydia Tressel
Cold Spring Harbor Laboratory
tressel@cshl.edu

Mihir Trivedi
University of Washington
mihir28@uw.edu

Devon Truax
ASHG
dtruax@ashg.org

Ms. Vasiliki Tsapalou
European Molecular Biology Laboratory (EMBL)
vasiliki.tsapalou@embl.de

Maria Tsikala Vafea
Icahn School of Medicine at Mount Sinai
maria.tsikalavafea@mssm.edu

Christopher Tuggle
Iowa State University
cktuggle@iastate.edu

Prof. Jenny Tung
Max Planck Institute for Evolutionary Anthropology/ Duke University
jtung@eva.mpg.de

Dr. Yasin Uzun
Penn State College of Medicine
yxu5009@psu.edu

Sarah Vaccaro
Stony Brook University
sarah.vaccaro@stonybrook.edu

Dr. Krishna Veeramah
Stony Brook University
krishna.veeramah@stonybrook.edu

Prof. Benjamin Vernot
University of Vienna
bvernot@gmail.com

Dr. Han Wang
Peking University Health Science Center
3200100310@zju.edu.cn

Ms. Wenjing Wang
National University of Singapore
e1101919@u.nus.edu

Mr. Tongtong Wang
The University of Melbourne
tongtong.wang@petermac.org

Dr. Michelle Ward
University of Texas Medical Branch
miward@utmb.edu

Dr. Doreen Ware
USDA/ CSHL
Ware@CSHL.edu

Dr. Doreen Ware
USDA/ CSHL
Ware@CSHL.edu

Prof. Wesley Warren
University of Missouri
warrenwc@missouri.edu

Dr. Marina Watowich
Vanderbilt University
marina.watowich@vanderbilt.edu

Prof. Xinzhu (April) Wei
Cornell University
aprilwei@cornell.edu

Ms. Sharon Wei
Cold Spring Harbor Labs
weix@cshl.edu

Dr. Josh Weinstock
Emory University
josh.weinstock@emory.edu

Dr. Sherman Weissman
Yale University
sherman.weissman@yale.edu

Dr. Jordan Welker
NYU Langone Health
jordan.welker@nyulangone.org

Dr. Sarah Wheelan
NIH/NHGRI
wheelansj@nih.gov

Dr. Prabhavi Wijesiriwardhana
National Institute of Health
prabhavi.wijesiriwardhana@nih.gov

Mr. Brandon Wilk
University of Alabama Birmingham
bwilk777@uab.edu

Dr. Melissa Wilson
NIH/NHGRI
melissa.wilson3@nih.gov

Dr. Emily Wong
Victor Chang Cardiac Research Insitute
e.wong@victorchang.edu.au

Dr. Peipei Wu
Cold Spring Harbor Laboratory
pwu@cshl.edu

David Wu
University of Pennslyvania
david.wu@pennmedicine.upenn.edu

Ruoxuan Wu
University of Chicago
rxwu@uchicago.edu

Dr. Yan Xia
Yale university
y.xia@yale.edu

Yunxin Xie
Cold Spring Harbor Laboratory
yxie@cshl.edu

Dr. Jiawei Xing
Cold Spring Harbor Laboratory
xing@cshl.edu

Mr. Liaoyi Xu
University of Texas at Austin
xliaoyi@utexas.edu

Dr. Angli Xue
Garvan Institute of Medical Research
a.xue@garvan.org.au

Prof. Kazuhiko Yamamoto
RIKEN Center for Integrative Medical Sciences
kazuhiko.yamamoto@riken.jp

Dr. Chao Yan
New York Genome Center
cyan@nygenome.org

Mr. Zikun Yang
shanghai jiao tong university
ericyangzk@gmail.com

Prof. Xiaoxu Yang
University of Utah
xiaoxu.yang@genetics.utah.edu

Dr. Emma Yee
Cell
eyee@cell.com

Mr. Alex Yenkin
Harvard University
ayenkin@g.harvard.edu

Ms. Ke Yi
University of Georgia
ky96157@uga.edu

Dr. Feyza Yilmaz
The Jackson Laboratory
feyza.yilmaz@jax.org

Kwontae You
Illumina
kyou@illumina.com

Zhezhen Yu
Cold Spring Harbor Laboratory
zhezhen@cshl.edu

Aziz Zafar
Columbia University Irving Medical Center
az2798@cumc.columbia.edu

Dr. Laura Zahn
Cell Genomics/Cell Press
lzahn@cell.com

Arslan Zaidi
University of Minnesota
aazaidi@umn.edu

Dr. Hugo Zeberg
Karolinska Institutet
hugo.zeberg@ki.se

Dr. Lu Zeng
Columbia University
lz2838@cumc.columbia.edu

Dr. Xin Zeng
Cold Spring Harbor Laboratory
zeng@cshl.edu

Dr. Jian Zhou
University of Chicago
jianzhou@uchicago.edu

Dr. Justin Zook
National Institute of Standards and Technology
justin.zook@nist.gov

CODE OF CONDUCT FOR ALL PARTICIPANTS IN CSHL MEETINGS

Cold Spring Harbor Laboratory (CSHL or the 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.

Consistent with the Laboratory's missions, commitments and policies, the purpose of this Code is to set forth expectations for the professional conduct of all individuals participating in the Laboratory's meetings program, both in person and virtually, including organizers, session chairs, invited speakers, presenters, attendees and sponsors. This Code's prohibition against discrimination and harassment is consistent with the Laboratory's internal policies governing conduct by its own faculty, trainees, students and employees.

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 (*see below)*.

Similarly, meeting participants agree to refrain from:

1. Harassment and discrimination, either in person or online, in violation of Laboratory policy based on actual or perceived sex, pregnancy status, gender, gender identity or expression, sexual orientation, race, ethnicity, color, religion, creed, nationality or national origin, immigration or citizenship status, mental or physical disability status, veteran status, military status, marital or partnership status, marital or partnership status, familial status, caregiver status, age, genetic information, status as a victim of domestic violence, sexual violence, or stalking, sexual reproductive health decisions, 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.

BREACHES OR VIOLATIONS OF THE CODE OF CONDUCT

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.

If you have questions or concerns, you can contact the meeting organizers, CSHL staff.

For meetings and courses funded by NIH awards:

Participants may 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.

For meetings and courses funded by NSF awards:

Participants may file a complaint with the NSF. See this page for information on how to file a complaint with the NSF.

Law Enforcement Reporting:

- For on-campus incidents, reports to law enforcement can be made to the Security Department at 516-367-5555 or x5555 from a campus phone.
- For off-campus incidents, report to the local department where the incident occurred.

In an emergency, dial 911.

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 is any unwelcome verbal, visual, written, or physical conduct that occurs with the purpose or effect of creating an intimidating, hostile, degrading, humiliating, or offensive environment or unreasonably interferes with an individual's work performance. Harassment 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 Harassment includes harassment on the basis of sex, sexual orientation, self-identified or perceived sex, gender expression, gender identity, and the status of being transgender. Sexual harassment is not limited to sexual contact, touching, or expressions of a sexually suggestive nature. Sexual harassment includes all forms of gender discrimination including gender role stereotyping and treating employees differently because of their gender. *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

MEETING ALCOHOL POLICY

Consumption of alcoholic beverages is not permitted in CSHL's public areas other than at designated social events (wine and cheese reception, picnic, banquet, etc.), in the Blackford Bar, or under the supervision of a licensed CSHL bartender.

No provision of alcohol by meeting sponsors is permitted unless arranged through CSHL.

Meeting participants consuming alcohol are expected to drink only in moderation at all times during the meeting.

Excessive promotion of a drinking culture at any meeting is not acceptable or tolerated by the Laboratory. No meeting participant should feel pressured or obliged to consume alcohol at any meeting-related event or activity.

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 ***(call for appointment)*** 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**
Pharmacy Value Drugs 391 W. Main Street, Huntington	**631-427-2919**

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
Open 24 hours (with PIN# or CSHL ID)
Staff Hours: 9:00 am – 9:00 pm
Use PIN# (On your registration envelope) to enter Library
See Library staff for photocopier code.
Library room reservations (hourly) available on request between 9:00 am – 9:00 pm

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.

Cold Spring Harbor Laboratory Bookstore & Gift Shop

Main campus, lower level of Grace Auditorium

Store Hours

Did you miss your chance to shop at the CSHL Bookstore and Gift Shop during the conference?

No problem! Visit our Online Bookstore and Gift Shop.

It's a great way to bring home a piece of the experience!

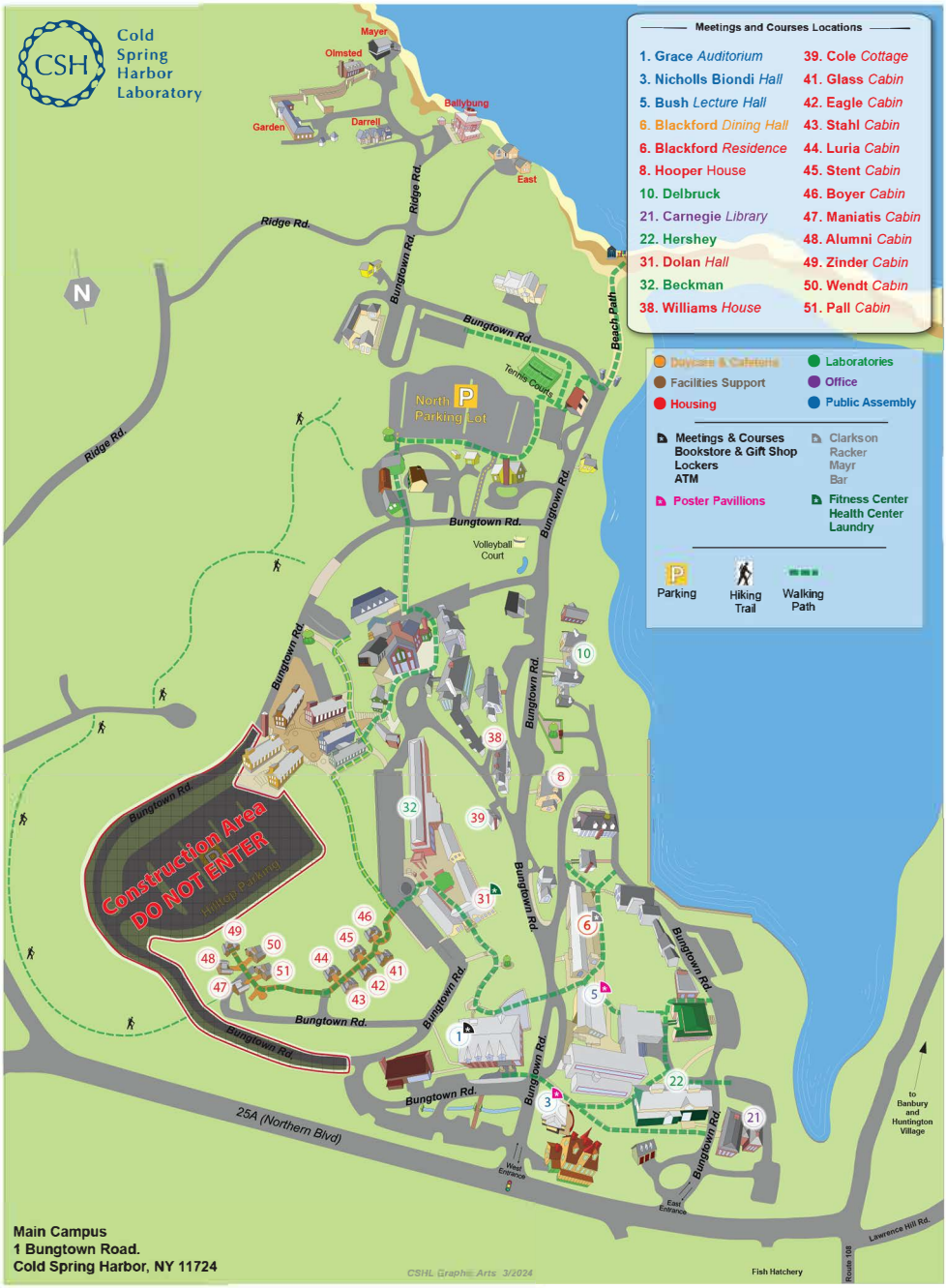
Contact Us
bookstore@cshl.edu
x8837

science books
CSHL-branded merchandise
unique gifts
souvenirs
. . . and more!

Visit our website
cshlvirtualstore.com

CSHL Campus Map

CSHL Map



Meet the Editors

Jennifer DeLeon, Ahmet Denli,
Hillary Sussman



Wednesday

7:00–9:00 PM

and

Friday

1:30–3:30 PM

Nicholls Biondi Hall

CSH PRESS

www.genome.org

How to fold your origami DNA