Abstracts of papers presented
at the 2025 meeting on

BIOLOGY OF GENOMES

May 6–May 10, 2025

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Abstracts of papers presented
at the 2025 meeting on

BIOLOGY OF GENOMES

May 6–May 10, 2025

Arranged by

Tuuli Lappalainen, *New York Genome Center, KTH Royal Institute of Technology & SciLife Lab*
Matthew Meyerson, *Dana-Farber Cancer Institute*
Aaron Quinlan, *University of Utah*
Jenny Tung, *Max Planck Institute for Evolutionary Anthropology & 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; Merck; Oxford Nanopore Technologies; PacBio;** and the **JT Scholarship Fund.**

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

Corporate Benefactors

Regeneron

Corporate Sponsors

Agilent Technologies
Biogen
Calico Labs
New England Biolabs
Novartis Institutes for Biomedical Research

Cover credit: Shuyu He, Max Planck Institute for Evolutionary Anthropology.

BIOLOGY OF GENOMES
Tuesday, May 6– Saturday, May 10, 2025

Tuesday	7:30 pm – 10:30 pm	**1** Population Genomics
Wednesday	9:00 am – 12:00 pm	**2** Cancer Genomics
Wednesday	2:00 pm – 5:00 pm	**3** Functional 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** Evolutionary / Non-Human Genomics
Thursday	1:30 pm – 4:30 pm	**5** Computational and Statistical Genomics
Thursday	5:00 pm – 6:00 pm	ELSI Panel and Discussion
Thursday	7:30 pm – 8:15 pm	Keynote Speaker I
Thursday	8:15 pm – 10:30 pm	**Poster Session II**
Friday	9:00 am – 12:00 pm	**6** Complex Traits and Genomic Medicine
Friday	2:00 pm – 5:00 pm	**Poster Session III**
Friday	5:15 pm – 6:00 pm	Keynote Speaker II
Friday	6:00 pm	*Cocktails and Banquet*
Saturday	9:00 am – 12:00 pm	**7** Emerging Methods and Technologies

Workshop (immediately following morning session)

Oxford Nanopore Technologies, Wednesday, May 7

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.

TUESDAY, May 6—7:30 PM

SESSION 1 POPULATION GENOMICS

Chairpersons: **Simon Gravel,** McGill University, Montreal, Canada
Andrew Kern, University of Oregon, Eugene

WEDNESDAY, May 7—9:00 AM

SESSION 2 CANCER GENOMICS

Chairpersons: **Nada Jabado,** McGill University, Montreal, Canada
Nicholas Navin, MD Anderson Cancer, Houston, Texas

SESSION 3 FUNCTIONAL GENOMICS

Chairpersons: **Nadav Ahituv,** University of California, San Francisco
Soumya Raychaudhuri, Broad Institute of MIT and Harvard, Cambridge, Massachusetts

WEDNESDAY, May 7—5:00 PM

Wine and Cheese Party

WEDNESDAY, May 7—7:30 PM

POSTER SESSION I

See p. xvii for List of Posters

SESSION 4 EVOLUTIONARY / NON-HUMAN GENOMICS

Chairpersons: **Nancy Chen,** University of Rochester, New York
Peter Sudmant, University of California, Berkeley

Dynamics of inbreeding and gene flow in a pedigreed wild population of Florida Scrub-Jays
Faye Romero, Jeremy Summers, James Schmidt, Daniel Seidman, Sahas Barve, John Fitzpatrick, Nancy Chen.
Presenter affiliation: University of Rochester, Rochester, New York. 25

The genetic basis of a unique structural coloration trait in the platyfish, *Xiphophorus evelynae*
Nadia B. Haghani, John J. Baczenas, Theresa R. Gunn, Tristram O. Dodge, Qinliu He, Sashoya Dougan, Paola Fascinetto-Zago, Zihao Ou, Gabe A. Preising, Sophia Haase Cox, Kang Du, Manfred Schartl, Dan Powell, Guan-Zhu Han, Molly Schumer.
Presenter affiliation: Stanford University, Stanford, California; Centro de Investigaciones Cientificas de las Huastecas "Aguazarca", Calnali, Mexico. 26

History repeats itself—Comparative genomics reveals new genes underlying convergent changes in vision, hair, and sperm locomotion
Nathan L. Clark, Maria Chikina, Courtney Charlesworth, Jered Stratton, Dwon Jordana, Amanda Kowalczyk, Emily Kopania.
Presenter affiliation: University of Pittsburgh, Pittsburgh, Pennsylvania. 27

Fast and furious mutation at tandem repeats in a large, four-generation family
Thomas A. Sasani, Michael E. Goldberg, Thomas J. Nicholas, Tom Mokveld, Egor Dolzhenko, Eli Kaufman, David Porubsky, Michael A. Eberle, Evan E. Eichler, Paul Valdmanis, Aaron R. Quinlan, Harriet Dashnow.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 28

THURSDAY, May 8—1:30 PM

SESSION 5 COMPUTATIONAL AND STATISTICAL GENOMICS

Chairpersons: **Stephanie Hicks,** Johns Hopkins University, Baltimore, Maryland
Sara Mostafavi, University of Washington, Seattle

THURSDAY, May 8—5:00 PM

ELSI PANEL and DISCUSSION

Artificial Intelligence and Machine Learning (AI/ML) in Genomics—Navigating the Opportunities and ELSI Challenges

Moderator: **Sheethal Jose,** NHGRI, National Institutes of Health

Panelists: **Alexis Battle,** Johns Hopkins University
Nicole Martinez-Martin, Stanford University
Diane M. Korngiebel, Google and University of Washington School of Medicine

The rapid advancement of AI/ML in genomics presents exciting opportunities and raises new ELSI questions. The session will explore the current landscape of AI/ML in genomics and where it shows the most promise for implementation. Panelists will discuss the importance of establishing robust development and post-deployment evaluation benchmarks and thresholds for AI/ML in genomics. These tools will be discussed in the context of ELSI issues, such as algorithmic bias, data privacy, public acceptance, and responsible development and deployment. We will discuss how to engage in proactive issue spotting and identifying potential issues before they arise in real-world settings.

THURSDAY, May 8—7:30 PM

KEYNOTE SPEAKER

Patricia Wittkopp
University of Michigan, Ann Arbor

"Arrival and survival of the fittest—Genetic variation affecting gene expression in yeast"

THURSDAY, May 8—8:15 PM

POSTER SESSION II

See p. xxx for List of Posters

SESSION 6 COMPLEX TRAITS AND GENOMIC MEDICINE

Chairpersons: **Gemma Carvill,** Northwestern University, Chicago, Illinois
Nicholas Mancuso, University of Southern California, Los Angeles

Viral DNA load is polygenic and confers lymphoma risk
Nolan Kamitaki, Steven A. McCarroll, Po-Ru Loh.
Presenter affiliation: Brigham and Women's Hospital, Boston, Massachusetts; Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Harvard Medical School, Boston, Massachusetts. 45

Widespread effects of medications on gut microbiome composition and function
Ashwin Chetty, Ramanujam Ramaswamy, Nicholas Dylla, Huaiying Lin, Matthew Odenwald, Eric Pamer, Ran Blekhman.
Presenter affiliation: University of Chicago, Chicago, Illinois. 46

FRIDAY, May 9—2:00 PM

POSTER SESSION III

See p. xliii for List of Posters

FRIDAY, May 9—5:15 PM

KEYNOTE SPEAKER

Steven McCarroll
Harvard Medical School
Broad Institute of MIT and Harvard

FRIDAY, May 9—6:00 PM

COCKTAILS and BANQUET

SESSION 7 EMERGING METHODS AND TECHNOLOGIES

Chairperson: **Gilad Evrony,** NYU Grossman School of Medicine, New York
Gloria Sheynkman, University of Virginia School of Medicine, Charlottesville

POSTER SESSION I

POSTER SESSION II

POSTER SESSION III

Functional and epigenetic characterization of African pan-genome contigs—Implications for reference genome bias and human genomic diversity
Rachel Martini, Abdulfatai Tijjani, Kyriaki Founta, Daniel Cha, Sebastian Maurice, Jason White, Melissa Davies, Nyasha Chambwe.
Presenter affiliation: Feinstein Institutes for Medical Research, New York, New York. 266

Dissecting functional elements in giant genomes using the first generation of high-quality salamander assemblies
Nataliya Timoshevskaya, S. Randal Voss, Jeramiah J. Smith.
Presenter affiliation: University of Kentucky, Lexington, Kentucky. 267

Insights into the genetic diversity and adaptation mechanisms of the Andean blueberry to extreme environments using genomic approaches
Maria de Lourdes Torres, Chelsea Specht, Milton Gordillo, Jacob Landis, Martina Albuja.
Presenter affiliation: Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito (USFQ), Quito, Ecuador. 268

Assessing cellular contexts of type 2 diabetes-associated variants at scale
Adelaide Tovar, Amy Etheridge, Romy Kursawe, Kirsten Nishino, Jonathan D. Rosen, Ziwei Chen, Daniel Dicorpo, James Meigs, Alisa Manning, Anshul Kundaje, Kimberly Lorenz, Benjamin F. Voight, Sarah Schoenrock, Ryan Tewhey, Michael Stitzel, Karen Mohlke, Jacob O. Kitzman, Stephen C. Parker.
Presenter affiliation: University of Michigan, Ann Arbor, Michigan. 269

SCiMS—Sex calling in metagenomic sequences
Hanh N. Tran, Kobie J. Kirven, Emily R. Davenport.
Presenter affiliation: Pennsylvania State University, University Park, Pennsylvania. 270

Detecting germline mutations in low-coverage sequence data using pedigrees
Georgia Tsambos, Daniel Seidman, Kelley Harris, Nancy Chen.
Presenter affiliation: University of Washington, Seattle, Washington. 271

Extensive ADAR-mediated RNA editing shapes KRAB-ZFP diversity through dynamic modification of DNA-binding domains
Itamar Twersky, Erez Y. Levanon.
Presenter affiliation: Bar-Ilan University, Ramat Gan, Israel. 272

A novel framework for building cell-specific gene regulatory networks with single-cell multi-omics
Yasin Uzun, Eric Moeller, Karamveer Karamveer, Hannah Valensi.
Presenter affiliation: Penn State College of Medicine, Hershey, Pennsylvania. 273

Non-coding mutations in diffuse large B-cell lymphoma—A cross-species study
Anna D. van der Heiden, Suvi Mäkeläinen, Raphaela Pensch, Sergey V. Kozyrev, Sophie Agger, Cheryl London, Jaime F. Modiano, Karin Forsberg Nilsson, Maja L. Arendt, Kerstin Lindblad-Toh.
Presenter affiliation: Uppsala University, Uppsala, Sweden. 274

IndiGene (GENEtics of INDIviduality)—RNA-seq analysis of selected tissues and different environments in Medaka fish
Christina Vasilopoulou, Tomas Fitzgerald, Ian Brettell, Adrien Leger, Nadeshda Wolf, Natalja Kusminski, Jack Monahan, Carl Barton, Cathrin Herder, Narendar Aadepu, Jakob Gierten, Clara Becker, Omar T Hammouda, Eva Hasel, Colin Lischik, Katharina Lust, Natalia Sokolova, Risa Suzuki, Erika Tsingos, Tinatini Tavhelidse, Thomas Thumberger, Philip Watson, Bettina Welz, Nadia Khouja, Kiyoshi Naruse, Ewan Birney, Joachim Wittbrodt, Felix Loosli.
Presenter affiliation: European Molecular Biology Laboratory, Hinxton, Cambridge, United Kingdom. 275

Extensive structural variation and longevity-associated adaptations in nearctic *Myotis* bats
Juan M. Vazquez, Mary E. Lauterbur, David Bahry, Meaghan Birkemeier, Eric Chen, Petar Pajic, Sarah Kassem, Omer Gokcumen, Michael Singer, Sarah Villa, Saba Mottaghinia, Carine Rey, Sarah Maesen, Michael Buchalski, Lucie Etienne, David Enard, Vincent J. Lynch, Peter Sudmant.
Presenter affiliation: University of California, Berkeley, Berkeley, California. 276

Immune pleiotropy and evolution in the response to *Yersinia pestis*
Tauras Vilgalys, Anne Dumaine, Mari Shiratori, Luis Barreiro.
Presenter affiliation: University of Chicago, Chicago, Illinois. 277

Investigating how poison exons modulate alternative splicing to shape transcriptomes in pluripotency and differentiation
Isha A. Walawalkar, Nathan Leclair, Mattia Brugiolo, Olga Anczukow.
Presenter affiliation: The Jackson Laboratory, Farmington, Connecticut; University of Connecticut Health Center, Farmington, Connecticut. 278

Compressive pangenomics using mutation-annotated networks
Sumit Walia, Harsh Motwani, Kyle Smith, Yu-Hsiang Tseng, Russell Corbett-Detig, Yatish Turakhia.
Presenter affiliation: University of California San Diego, San Diego, California. 279

Estimating recent population split times in non-panmictic populations
Jeff Wall.
Presenter affiliation: Oregon Health and Science University, Beaverton, Oregon. 280

Airqtl dissects cell state-specific causal gene regulatory networks with efficient single-cell eQTL mapping
Lingfei Wang.
Presenter affiliation: University of Massachusetts Chan Medical School, Worcester, Massachusetts. 281

Single-molecule sequence models to decode the regulatory genome
Ruoyu Wang, Junru Jin, Jian Zhou.
Presenter affiliation: University of Texas Southwestern Medical Center, Dallas, Texas. 282

Revisit global expression change in single-cell perturbation data
Shuyue Wang, Han Xu.
Presenter affiliation: MD Anderson Cancer Center, Houston, Texas; MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, Texas. 283

COUTURE—facilitating interpretation from genotype to molecular and functional phenotype in single-cell CRISPR screening
Jun Cao, Xiaoyue Wang.
Presenter affiliation: Institute of Clinical Medicine and Peking Union Medical College Hospital, Beijing, China. 284

Comprehensive functional assessment of *NF1* and *NF2* variants with high-resolution base editing screens
Jiayu Wu, Guangyu Li, Liheng Luo, Chenyu Ma, Shangqi Zhao, Zhuang Du, Xiaoyue Wang.
Presenter affiliation: Institute of Clinical Medicine and Peking Union Medical College Hospital, Beijing, China. 285

Allele specific expression in Alzheimer's disease
Zishan Wang, Delowar Hossain, Varun Subramaniam, Bin Zhang, Minghui Wang, Kuan-lin Huang.
Presenter affiliation: Icahn School of Medicine at Mount Sinai, New York, New York. 286

Industrialization influences biological and molecular mechanisms of aging in immune cells in three non-industrial populations
Marina M. Watowich, Amy Longtin, Julien F. Ayroles, Kenneth Buetow, Hillard Kaplan, Yvonne Lim, Dino Martins, Kee-Seong Ng, Sospeter Njeru, Jonathan Stieglitz, Benjamin Trumble, Vivek V. Venkataraman, Ian J. Wallace, Michael Gurven, Thomas S. Kraft, Alexander G. Bick, Amanda J. Lea.
Presenter affiliation: Vanderbilt University, Nashville, Tennessee. 287

Network-level convergence of rare and common variants underlying complex traits
Sarah N. Wright, Trey Ideker.
Presenter affiliation: University of California, San Diego, La Jolla, California. 288

High-throughput *in silico* screen discovered novel regulators of 3D genome organization
Jiangshan Bai, Qingji Lyu, Jimin Tan, Bailey Tischer, Xinyu Ling, Viraat Goel, Aristotelis Tsirigos, Bradley E. Bernstein, Anders S. Hansen, Bo Xia.
Presenter affiliation: Broad Institute of MIT and Harvard, Cambridge, Massachusetts; Harvard University, Society of Fellows, Massachusetts. 289

Characterization of codon and amino acid frequency variation in the human genome
Zhuorui Xie, Ziyue Gao.
Presenter affiliation: University of Pennsylvania, Philadelphia, Pennsylvania. 290

Defining the landscape of poison exons and their involvement in human diseases
Huilin Xu, Paolo Pigini, Yan Ji, Hannah Lindmeier, Maria Catarina Lima Da Silva, Dadi Gao, Elisabetta Morini.
Presenter affiliation: Massachusetts General Hospital Research Institute, Boston, Massachusetts; Massachusetts General Hospital Research Institute and Harvard Medical School, Boston, Massachusetts; Broad Institute of Harvard and MIT, Cambridge, Massachusetts. 291

< Back

DEEP LEARNING FOR POPULATION GENETICS

Andrew D Kern

University of Oregon, Institute of Ecology and Evolution and Department of Biology, Eugene, OR

As genomic datasets continue to expand, researchers face the challenge of extracting meaningful insights from an overwhelming volume of data. To address this, computational methodologies are evolving to leverage large-scale genomic sequence data for evolutionary genetic inference. In this talk, I will discuss my group's recent work on integrating deep learning approaches to population inference. Two main topics will be highlighted: 1) adapting insights from large language models (LLMs) for casting inference of coalescent times from genomic sequences as a translation problem and 2) using neural posterior estimation (NPE) as a principled way of getting uncertainty estimates from deep learning-based population genetic inference of key parameters such as recombination rates and demographic histories.

GLOBAL PATTERNS OF NATURAL SELECTION INFERRED USING ANCIENT DNA

Laura L Colbran[1], Jonathan Terhorst[2], Iain Mathieson[1]

[1]University of Pennsylvania, Department of Genetics, Philadelphia, PA,
[2]University of Michigan, Department of Statistics, Ann Arbor, MI

Ancient DNA has revolutionized our understanding of human history and, in recent years has begun to illuminate human evolution and natural selection. However, these efforts have largely been limited to Europe, excluding populations in other parts of the world. Many selective pressures have been local to specific populations but others, for example the development of agriculture, may have been more universal. Studying a broad range of global populations can therefore identify both examples of local adaption and more general principles of human adaptation.

In this study, we leveraged new ancient DNA data to test for selection in 7244 individuals from 13 ancient and 19 present-day populations across five regions–Europe, East Asia, South Asia, Africa and the Americas. In each region, we tested for selection using a maximum likelihood approach that models expected allele frequencies based on patterns of admixture. We identify 31 genome-wide significant signals of selection, including both known loci such as *LCT* and *SLC45A2* in Europe, and novel loci such as *MIF* in the Americas. Two loci, *ADH1B* and *FADS1*, were genome-wide significant in Europe and East Asia, and there was a high degree of shared signal across all non-African regions.

We developed a new approach to model time series data in admixed populations and used it to identify fine-scale changes in the strength and targets of selection. We find that the strength of selection on variants associated with agricultural products tended to increase with the intensity of agriculture in both Europe and East Asia. We also find more complex patterns at some loci, for example decreasing selection at *LCT* in the recent past, and long-term balancing selection at the eye color locus *OCA2* in Europe.

We next developed a test for polygenic selection on complex traits by modelling the frequencies of trait-increasing alleles identified in GWAS. We tested for selection jointly across regions, avoiding the confounding effect of population stratification by excluding the European or East Asian GWAS population from the selection test. We find evidence for directional selection on pigmentation and immune traits, and that strong stabilizing selection on female waist-hip ratio was universal across human populations suggesting a fundamental constraint on human morphology.

This study represents the first comprehensive aDNA-based comparison of selective pressures across the world. We highlight the overlap across populations in loci and responses to selection, and demonstrate the utility of aDNA in disentangling the effects of selection from complex demographic histories.

LEVERAGING ARGs FOR POPULATION-LABEL-FREE PRS PREDICTION

Nurdan Kuru[1], Shareef Khalid[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 based on genetic variation. However, their accuracy is limited by biases in genome-wide association studies and broad ancestry labels that overlook human genetic diversity. These limitations particularly affect individuals from underrepresented populations, reducing the clinical utility of PRS in diverse groups.

To address this issue, we introduce a novel approach that models individual genomes using Ancestral Recombination Graphs (ARGs), which capture the true evolutionary history of genetic variants. Instead of relying on predefined population labels, we utilize local genealogies enabling more precise modeling of individual-level genetic relationships.

Our method applies phylogenetic linear mixed models, incorporating ARG-derived variance-covariance matrices to account for genealogical relatedness at individual loci. We model average genetic effects (fixed effects), shared across ancestries, and individual-specific deviations (random effects) which capture differences in genetic backgrounds through local ARGs. Our framework incorporates new individuals by threading them into the ARG and using shared ancestry to compute their PRS predictions.

We evaluated our framework against PRS methods using population-level trees, genome-wide averaged ARGs, non-tree-based approaches, and tools like Lassosum, PROSPER, and DendroPRS. In simulated datasets, our full ARG-based approach improved predictive accuracy by up to 40%, with higher R^2 values and lower prediction loss across different scenarios. Notably, it showed the strongest gains in African and admixed populations, groups traditionally underrepresented in genetic research. Analysis of All of Us data on LDL and height further confirmed its robustness, demonstrating consistent improvements over existing PRS methods.

By incorporating the rich evolutionary history encoded in ARGs, our approach provides a biologically informed alternative to traditional PRS methods. This framework has the potential to enhance genetic risk prediction across all populations, reduce health disparities, and improve the applicability of PRS in diverse ancestry groups.

GENETIC CONTROL OF LOCAL MUTATION RATES

Madison Caballero, Amnon Koren

Roswell Park Comprehensive Cancer Center, Molecular and Cellular Biology, Buffalo, NY

Mutations are the source of evolutionary novelty but also the cause of genetic diseases and cancer. Previous studies have argued that *trans*-mutator loci that increase the genomic mutation rate are unlikely to adaptively evolve in sexually reproducing organisms. However, DNA replication timing – a main effector of mutations – has been shown to vary among humans at thousands of regions along the genome, raising the possibility that human mutation rates vary at local scales. To test this, we analyzed the chromosomal distribution of somatic mutations in 1,662 individuals, controlling for the confounding effects of DNA replication timing on local mutation rates and of *trans*-acting modulators on global somatic mutation rates. We describe substantial inter-individual variation in mutation rates spanning close to one hundred megabase-sized regions of the human genome. By comparing mutation rates to individuals' genotypes, we identified 35 instances in which genetic variants associate with mutation rates in their vicinity. We call these mutation quantitative trait loci (mutQTLs). Although mutQTL were identified in lymphoblastoid cell lines, they also associated with the rate of mutations in chronic lymphocytic leukemia, as well as with the density of nearby germline genetic variants, establishing them as effectors of both somatic and germline mutation rates. Evolutionary analysis of mutQTLs suggests the emergence of *cis*-mutators – novel genetic variants that confer an increased rate of mutation in their chromosomal vicinity. The strongest mutQTL, on chromosome 19, was associated with a hotspot of genetic variation related to a proposed arms race between zinc finger transcription factors and transposable elements. Taken together, these results show that the mutation landscape is subject to ongoing evolution, with mutQTLs providing a portal into the evolution of mutation rate heterogeneity across the genome and across individuals.

GENOMICS IN A LARGE HUMAN GENEALOGY

Simon Gravel[1], Luke Anderson-Trocmé[2], Georgette Femerling-Romero[1], Alejandro Mejia-Garcia[1], Andrii Serdiuk[1]

[1]McGill University, Human genetics, Montreal, Canada, [2]University of Chicago, Human Genetics, Chicago, IL

The population of Quebec, Canada has experienced many founder events. From the 1600s, these founder events have been exquisitely documented by church and civil records, and digitized to form a uniquely complete pedigree, in which about 8500 individuals have contributed an important fraction of the genetic material present in the 8.5 million present-day Quebec residents.

The CARTaGENE cohort recently generated whole-genome data for 30,000 individuals from multiple ancestries in Quebec, 10,000 of which have been linked to the genealogical database managed by the BALSAC project.

I will discuss how ancient and recent migrations have shaped genetic diversity in contemporary Quebec, and highlight opportunities in evolutionary research, medical research, and public health that arise from the combination of large-scale genetic and genealogical data.

Finally, I will discuss the feasibility of reconstructing ancestral genomes within the pedigree going to the 17th century, including recent statistical and algorithmic progress towards that goal.

CHARACTERIZING DE NOVO STRUCTURAL VARIATION IN THE AGING GERMLINE

Stacy Li[1,2], Joseph G Gleeson[3,4], Aaron R Quinlan[5,6], Kenneth I Aston[7,8], Raheleh Rahbari[9], Richard Durbin[10], Peter H Sudmant[1,2]

[1]University of California, Berkeley, Center for Computational Biology, Berkeley, CA, [2]University of California, Berkeley, Department of Integrative Biology, Berkeley, CA, [3]University of California, San Diego, Department of Neurosciences and Pediatrics, San Diego, CA, [4]Rady Children's Hospital, Institute for Genomic Medicine, San Diego, CA, [5]University of Utah, Department of Biomedical Informatics, Salt Lake City, UT, [6]University of Utah, Department of Human Genetics, Salt Lake City, UT, [7]University of Utah School of Medicine, Department of Surgery (Urology), Salt Lake City, UT, [8]University of Utah School of Medicine, Andrology and IVF Laboratory, Salt Lake City, UT, [9]Wellcome Trust Sanger Institute, Cancer, Ageing and Somatic Mutation (CASM), Hinxton, United Kingdom, [10]University of Cambridge, Department of Genetics, Cambridge, United Kingdom

Aging is an emergent phenomenon hallmarked by the deterioration of physiological processes over time. *De novo* germline mutations are directly transmissible to offspring: increased *de novo* mutation frequency in the male germline in aging poses significant risk to reproductive success. *De novo* structural variants (dnSVs) affect large genomic regions and are known contributors to congenital developmental disorder. Notably, autism spectrum disorder (ASD) is associated with both an elevated dnSV burden and advanced paternal age at conception. Advances in highly accurate single-molecule long-read sequencing now enable direct characterization of dnSVs without relying on proxy measures (i.e. read depth).

To address this, we applied PacBio HiFi long-read sequencing to identify dnSVs in bulk sperm samples from nineteen donors aged 27-62, including fathers of children with ASD and TSC, a genetic disorder driven primarily by *de novo* mutations. We created highly contiguous phased assemblies (average N50 = 70Mb) for each donor and used personal genome alignment to identify clonal (shared amongst sperm descended from a common progenitor) and unique (private to <4 gametes generated during meiosis) variants. Our self-assembly approach significantly reduced false calls compared to conventional reference alignment, enabling detection of both clonal dnSVs and unique meiotic events.

To validate our approach, we characterized the frequency and distribution of *de novo* retrotransposition events captured within single reads. We identified multiple high-confidence retrotransposition events, primarily from AluYb8, AluYa5, and L1HS elements, with frequencies ranging from 2.1 to 10.2 events per 100 cells. We observed a significant increase in event frequencies in older donors, with a stronger age association (R^2=0.608, p = 0.003) in samples from healthy individuals. Preliminary results suggest potentially elevated mutation frequencies in samples from fathers of children with ASD and TSC. Notably, we identified several complex clonal dnSVs composed of sequential *de novo* insertions and duplications, potentially mediated by non-allelic homologous recombination. These findings demonstrate the utility of long-read sequencing and multi-modal analysis for comprehensive dnSV detection in the aging germline, with implications for understanding both genome evolution and genetic disease risk transmission.

FROM NORTH ASIA TO SOUTH AMERICA: TRACING THE LONGEST HUMAN MIGRATION THROUGH GENOMIC SEQUENCING

Elena Gusareva[1,2], Amit G Ghosh[1,2], Stephan C Schuster[2], Vadim A Stepanov[3], Hie Lim Kim[1,2]

[1]Nanyang Technological University, Asian School of the Environment, Singapore, Singapore, [2]Nanyang Technological University, SCELSE, Singapore, Singapore, [3]Tomsk National Medical Research Center, Research Institute of Medical Genetics, Tomsk, Russia

We analyzed genome sequencing datasets from 1,537 individuals from 139 ethnic groups to reveal the genetic characteristics of understudied populations in North Asia and the Americas. Our analysis demonstrated that the nomadic hunter-gatherer West Siberians, represented by the Kets and Nenets, contributed to the genetic ancestry of most Siberian populations. This is supported by our estimation that West Siberians were the largest population in North Eurasia 10,000-13,900 years ago. We also found that West Beringians, including the Koryaks, Inuit, and Luoravetlans, exhibit genetic adaptation to the Arctic climate. In South America, our analysis showed that the earliest inhabitants of South America split into four ancestral groups – Amazonians, Andeans, Chaco Amerindians, and Patagonians – ~13,900 years ago. This population structure corresponds closely with biogeographic boundaries in South America, suggesting the impact of environmental factors on population history. Their longest migration and spatial isolation due to the vastness of the South American continent led to population decline and reduced genetic diversity. The Patagonian ethnicities, including the Yagan, Kawésqar, and Chaco Amerindians, show the smallest effective population size in our dataset. This, in turn, has resulted in reduced immunogenic diversity, likely increasing their vulnerability to past and current pathogens. These findings highlight how population history and environmental pressures shaped the genetic architecture of human populations across North Asia and South America.

DISENTANGLING GENETIC AND PHENOTYPIC RESPONSES TO SELECTION IN THE BODY MASS OF WILD KALAHARI MEERKATS

Alexander E Downie[1], Elizabeth Mittell[2], Kasha Strickland[2], Tim Clutton-Brock[3,4], Marta Manser[4,5], Loeske Kruuk[2], Jenny Tung[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Primate Behavior and Evolution, Leipzig, Germany, [2]University of Edinburgh, Institute of Ecology and Evolution, Edinburgh, United Kingdom, [3]University of Cambridge, Department of Zoology, Cambridge, United Kingdom, [4]Kalahari Research Center, Van Zylsrus, South Africa, [5]University of Zurich, Department of Evolutionary Biology and Environmental Studies, Zurich, Switzerland

Trait evolution in response to natural selection is the engine that drives adaptation. However, classical methods for predicting the phenotypic response to directional selection frequently fail in natural populations. These failures could be explained by genetic responses to selection that are masked by environmental change, error in estimating selection coefficients or trait heritability, or other challenges (e.g., indirect genetic effects, trait genetic covariance) that confound simple predictions. Disentangling these explanations is difficult, in part because complementary data on individual-level fitness, trait variation, and genome-wide genotype data remain rare for natural populations.

To confront this challenge, we draw on long-term records and genome-wide resequencing data for more than 3,000 individually recognized meerkats studied since 1993 in the Kalahari Desert of South Africa. We focus specifically on body mass, an intensively phenotyped trait (mean 221 repeated measures per individual) that strongly predicts lifetime reproductive success and yet exhibits evidence of decline across the 30 years of the study. We generated a reference panel of 100 deeply-sequenced individuals and imputed genotype data at over 400k single nucleotide variants for 2,950 individuals from the extended pedigree that were sequenced to low coverage. Together, these data reveal an h^2 for body mass of 0.11 and a very strong standardized selection differential of 1.46, favoring heavier animals. We then performed a genome-wide association study to produce effect estimates at each SNV for body mass. We combined said estimates with gene-dropping simulations through the 11-generation pedigree to test whether alleles associated with heavier body mass show evidence of the directional selection predicted by the selection differential but in opposition to the phenotypic trend. Together, our findings highlight the complexity of predicting responses to selection in natural populations based on genetic or phenotypic data alone, including the value added by breakthroughs in sequencing capacity and imputation to map traits and describe allelic dynamics in finer detail.

H3K27me3 SPREADING ORGANIZES CANONICAL PRC1 CHROMATIN ARCHITECTURE TO REGULATE DEVELOPMENTAL PROGRAMS

Nada Jabado[1], Brian Krug[1], Chao Lu[2]

[1]McGill University, McGill University Health Center Research Institute, Pediatrics and Human Genetics, Montreal, Canada, [2]McGill University, Human Genetics, Montreal, Canada, [3]Columbia University, Genetics & Development CUMC, New York, NY

Polycomb Repressive Complex 2 (PRC2)-mediated histone H3K27 tri-methylation (H3K27me3) recruits canonical PRC1 (cPRC1) to maintain heterochromatin. In early development, polycomb-regulated genes are tethered by long-range 3D interactions, many of which are lost during lineage differentiation. We show that polycomb-anchored looping is controlled by H3K27me3 spreading and regulates target gene silencing and cell fate specification. Using glioma-derived H3 Lys-27-Met (H3K27M) mutations as tools to restrict H3K27me3 deposition, we show that H3K27me3 confinement concentrates the chromatin pool of cPRC1, resulting in heightened 3D interactions mirroring chromatin architecture of pluripotency, and heightened transcriptional repression that maintains cells in progenitor states to facilitate tumor development. Conversely, H3K27me3 spread in pluripotent stem cells, dilutes local cPRC1 chromatin concentration which weakens polycomb contact frequencies. These results identify the regulatory principles and disease implications of polycomb looping and nominate histone modification-guided distribution of reader complexes as an important mechanism for nuclear compartment organization.

AN EGFR HOTSPOT MUTATION INTERACTS WITH RBM10 TO INFLUENCE LUNG CANCER RISK IN EAST ASIANS

Anders B Dohlman[1,2,3], Ethan Shurberg[1,2,3], Meagan Montesion[4], Smruthy Sivakumar[4], Owen Hirschi[1,2,3], Alaina Shumate[1,2,3], Garrett Frampton[4], Alexander Gusev[1,3], Matthew Meyerson[1,2,3]

[1]The Dana-Farber Cancer Institute, Medical Oncology, Boston, MA, [2]The Broad Institute, Cancer Program, Cambridge, MA, [3]Harvard Medical School, Genetics, Boston, MA, [4]Foundation Medicine Inc., Boston, MA

Lung cancer remains the leading cause of cancer-related death worldwide, killing more than 1.8 million people each year. A key unresolved question about lung cancer is the relationship between patient ancestry, sex, and somatic mutations in epidermal growth factor receptor (EGFR). In lung adenocarcinoma, EGFR mutations are significantly more common patients of East Asian descent (~50%) than in patients of European or African descent (~10%), as well as in women and non-smokers.

To explore the underlying factors contributing to these disparities, we performed an analysis of lung cancer genomes from diverse ancestries using data from Foundation Medicine (n = 64,052) and GENIE (n = 21,215). We found that the frequency of in-frame deletions in Exon 19 and point mutations at L858R – together accounting for 90% of EGFR mutations – varied significantly and independently by patient ancestry and sex. Specifically, patients of African and South Asian ancestry with EGFR mutations were twice as likely to harbor Exon 19 deletions compared to those of European ancestry. In contrast, L858R mutations were independently enriched in patients of East Asian ancestry, in women, and in patients with cooccurring RBM10 loss-of-function mutations.

RBM10, an X-linked splicing factor, exhibited a higher mutant allele frequency was more likely to cooccur with EGFR L858R mutations in East Asians than in other groups, even after controlling for patient age and sex, implicating this gene in observed disparities in lung cancer mutations. These findings strongly suggest the presence of a sex-modulated germline risk factor influencing EGFR mutation risk in lung across ancestries, highlighting a novel interaction between RBM10 and EGFR mutations.

DISCRETE PHASES OF GENOME EVOLUTION UNDERLIE SARCOMAGENESIS

Rodrigo Gularte Mérida[1], Timour Baslan[2], Corey Weistuch[3], Evan Seffar[4], Sienna Linden[1], Nicole Blekhter[1], Kalyani Chadalavada[1], Cristina R Antonescu[5], Narasimhan Agaram[5], Tomoyo Okada[1], Nicholas D Socci[6], Samuel Singer[1]

[1]Memorial Sloan Kettering Cancer Center, Department of Surgery, New York, NY, [2]University of Pennsylvania, Department of Biomedical Sciences, Philadelphia, PA, [3]Memorial Sloan Kettering Cancer Center, Department of Physics, New York, NY, [4]Memorial Sloan Kettering Cancer Center, Computational Oncology, New York, NY, [5]Memorial Sloan Kettering Cancer Center, Department of Pathology, New York, NY, [6]Memorial Sloan Kettering Cancer Center, Bioinformatics Core, New York, NY

Understanding how cancer genomes evolve is important in devising diagnostic and therapeutic approaches for cancer patients. To date, most studies have relied on the analysis of deep sequencing data of single-tumor biopsies or shallow-depth sequencing of multi-region samples acquired across a tumor mass. It is not known what novel insights can be gleaned from both deep and multi-region profiling of human cancer genomes. Here, multi-region, single-cell DNA copy-number profiling applied to the most commonly diagnosed adult soft tissue sarcoma; Liposarcoma, a heterogenous disease driven by complex 12q amplifications and copy number alterations (CNA), reveals novel insights into the forces driving disease evolution. Clonal decomposition via consensus clustering of single-cell copy number identifies a novel, low-frequency subpopulation of cells, termed Earliest Detectable Clones (EDC), that harbor low-level 12q14 amplification and found at early/well-differentiated (WDLS) as well as late/de-differentiated (DDLS) stages of the disease, suggestive of persistence of liposarcoma initiating cells. Structural variant analysis and pseudo-temporal ordering of 12q amplifications in single cells reveals progressive and ordered acquisition of amplified genes with MDM2 acquired first, CDK4 second, followed by other 12q amplified genes, including known driver (e.g. HMGA2) and novel (e.g. CDK13) genes. In addition, we find that further 12q amplicon evolution mediates the acquisition of progression-associated amplifications on other chromosomes, such as JUN (1p32) and TCF21 (6q23-25), and that their acquisition occurs at early stages of liposarcoma genome evolution, indictive of early commitment to dedifferentiated disease. After an amplification-centric phase, dedifferentiated clones selectively acquire recurrent, broad deletions at 11q, 13q, and 19q that are timed relatively late during DDLS progression. We find these late, prognostically relevant alterations to be acquired independently in multiple clones within the tumor mass (i.e. converged upon), and present as localized and geographically restricted clones in the tumor mass. Finally, we link 11q deletions, encompassing ATM, to rapid proliferative bursts that result in regionally localized genomic instability. Together, the results revise the temporal and spatial genetic architecture that underlies WD/DDLS initiation and progression and suggest management strategies to target this heterogeneous disease.

DECRYPTING THE COLON: LEVERAGING A TRIAD OF TECHNIQUES TO INVESTIGATE CRYPT-SPECIFIC SOMATIC MOSAICISM

Laurel Hiatt, Hannah Happ, Aaron Quinlan

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

Somatic mosaicism results from accumulated mutations in non-germline cells throughout an organism's lifetime. Somatic mutations play a central role in pathogenesis, from developmental syndromes to cancer, and there is growing consensus that somatic mosaicism in healthy tissue influences tissue phenotype and disease predisposition. The colon is a model organ for studying mosaicism: ample tissue can be sourced via cadaver donation and routine colonoscopy, and the functional unit of the colon—the colon crypt, ~2000-cell epithelial sheets—is highly clonal due to its contained microstructure and shared ancestor stem cell. Significant associations between somatic mosaicism and colorectal pathologies motivates research in colon crypts across disease states. Nevertheless, essential clinical questions remain, such as the origin of regional variability in disease predisposition heretofore unexplained by current technologies.

We leverage three technologies still untested in the field of somatic mosaicism to interrogate regional somatic mosaicism across pre-malignant colon. We use the MMI CellCut system to efficiently capture complete, individual crypts, Watchmaker Genomics library prep kits to create libraries from these crypts, and Element Biosciences sequencing to generate exceptionally high-quality data. Each of these techniques provides significant and cumulative advantages in the generation of data from low tissue input. The MMI CellCut system is a laser microdissection tool optimized for high quality and quantity DNA recovery given its highly precise laser capabilities and capacity for informative metadata. Watchmaker Genomics library preparation uses optimized enzymatic efficiencies to mitigate DNA loss and minimize artifacts that can complicate downstream somatic mosaicism analyses. In conjunction with Element sequencing, these methodologies coalesce into a workflow with numerous potential applications in the evolving field of somatic mosaicism research.

To illustrate the potential of this workflow, we present ongoing progress evaluating region-based somatic mosaicism in individual colon crypts, motivated by regional disease presentations of colorectal pathologies. By extracting mutational signatures indicative of mutational etiology, we evaluate whether certain mutagenic processes may be enriched in specific parts of the organ. Investigating these processes can provide insight into regional pathologies and their genetic etiologies.

WHEN IS A CANCER REALLY A CANCER? ANEUPLOIDY IN NORMAL BREAST TISSUES

Nicholas Navin

MD Anderson Cancer Center, Systems Biology, Houston, TX

Aneuploid epithelial cells are common in breast cancer, however their presence in normal breast tissues is not well understood. To address this question, we applied single cell DNA sequencing to profile copy number alterations (CNAs) in 83,206 epithelial cells from breast tissues of 49 healthy women and single cell DNA&ATAC co-assays to 19 women. Our data shows that all women harbored rare aneuploid epithelial cells (median 3.19%) that increased with age. Many aneuploid epithelial cells (median 82.22%) in normal breast tissues underwent clonal expansions and harbored CNAs reminiscent of invasive breast cancers (gains of 1q, losses of 10q, 16q and 22q). Co-assay profiling showed that the aneuploid cells were mainly associated with the two luminal epithelial lineages, while spatial mapping showed that they localized in ductal and lobular structures with normal histopathology. Collectively, these data show that even healthy women have clonal expansions of rare aneuploid epithelial cells in their breast tissues.

A COMPUTATIONAL MODELING FRAMEWORK FOR SINGLE-CELL GENE EXPRESSION EVOLUTION LEVERAGING LINEAGE PHYLOGENY OF METASTATIC CANCER

Jiawei Xing[1], Stephen Staklinski[1], Nurdan Kuru[1], Dawid Nowak[2], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Weill Cornell Medicine, Meyer Cancer Center, New York, NY

Metastasis is the leading cause of cancer-related mortality, yet the molecular mechanisms driving metastasis across different tissues remain poorly understood. Traditional detections of metastatic gene expression rely on differential expression analyses of scRNA-seq data by grouping cells from primary and metastatic sites into clones. However, these methods overlook cancer lineage phylogenies, which encode crucial information on tumor heterogeneity and metastatic history. Recent advances have adapted computational models from species evolution to model gene expression along cancer cell lineages, enabling hypothesis testing of lineage-specific adaptations. However, the sparsity of single-cell sequencing data limits the power of these models in detecting differential expression targets. To address these challenges, we present a modeling framework that integrates the Ornstein-Uhlenbeck process with Poisson-distributed read counts using mean-field variational inference and variational autoencoders. Our method jointly models genes with sparse read counts across gene sets, enabling the detection of gene pathways undergoing expression shifts across metastatic sites. Additionally, we integrate our approach with BEAM, our Bayesian framework for phylogeny inference, which accounts for uncertainties in the lineage-tracing data and collectively generates tree samples with corresponding metastatic events. We validate our framework using synthetic sequencing data from cancer cell simulations and present accurate predictions on differentially expressed genes during metastasis, which outperforms the current model from species evolution and the standard differential expression analysis. Applications to metastatic lung and prostate cancer datasets identify both well-established and novel gene targets as potential positive or negative regulators of tissue-specific metastasis. Our approach offers a powerful modeling framework for integrating lineage information into single-cell transcriptomics, offering novel biological insights into gene expression evolution during cancer metastasis. Looking ahead, we aim to extend this model by integrating lineage-tracing with spatial transcriptomics and multi-omics data, moving towards constructing a comprehensive atlas of metastatic cancer evolution in both spatial and temporal scales.

A TEMPORAL ATLAS OF REGULATORY ACTIVITY IN THE HUMAN GENOME DURING CELL FATE TRANSITIONS

Beatrice Borsari*[1], Silvia González-López*[1,2], Amaya Abad[1], Vasilis F Ntasis[1], Cecilia C Klein[1], Ramil Nurtdinov[1], Diego Garrido-Martín[1], Carme Arnan[1], Alexandre Esteban[1], Emilio Palumbo[1], Marina Ruiz-Romero[1], Raül G Veiga[1], Maria Sanz[1], Bruna R Correa[1], Rory Johnson[1], Sílvia Pérez-Lluch[1], Roderic Guigó[1,2]
[1]Center for Genomic Regulation, Computational Biology and Health Genomics, Barcelona, Catalonia, Spain, [2]Universitat Pompeu Fabra, MELIS, Barcelona, Catalonia, Spain

* Equal contribution

Eukaryotic genomes are rich in regulatory regions that orchestrate gene expression, shaping cell identity and responses to stimuli. Histone post-translational modifications (hPTMs) play a key role in fine-tuning these regulatory elements and, consequently, gene expression. Among the most comprehensive efforts to compile these regions is the ENCODE Consortium's registry of candidate *cis*-regulatory elements (cCREs). However, this catalog mostly characterizes regulatory activity in static tissues and cell types, lacking the temporal resolution needed to capture changes in genome regulation over time. While the registry has expanded significantly since its initial release, the underlying logic of cCRE function remains incompletely understood, in part due to the limited availability of time-resolved multiomic datasets to investigate their role in development and differentiation.
As part of the ENCODE Consortium, we have addressed this question by generating an unprecedented, highly temporally-resolved multiomic dataset, capturing both epigenetic and transcriptional profiles during the transdifferentiation of human B-cell precursors into macrophages. This dataset is complemented by publicly available ATAC-seq and Hi-C data. Our results show that hPTMs are deposited in a cumulative fashion, following a precise order that is conserved across most cCREs: first H3K4me1/2, followed by H3K27ac, H3K9ac, and finally H3K4me3. We next analyzed the major epigenetic trajectories of cCREs over time, and found that the most dynamic changes involve cell type-specific distal regulatory elements, with myeloid and B-cell-specific enhancers showing progressive gain and loss of histone marking, respectively.
To delve into the functional implications of this, we integrated our transcriptomic data. We found that hPTMs are more tightly linked with *de novo* gene activation rather than upregulation of already expressed genes. Moreover, while some hPTMs anticipate gene expression, others like H3K4me3 are delayed, and so is marking at associated enhancers, indicating that the latter could be a consequence of epigenetic marking at promoters.
All in all, our densely-spaced time course transdifferentiation system has allowed us to uncover, at unprecedented resolution, how chromatin modifications are regulated along dynamic processes.

COMPARATIVE ANALYSIS OF NON-CODING CONSTRAINT MUTATIONS IN CANINE AND HUMAN OSTEOSARCOMA REVEALS A SHARED UNDERLYING DISEASE NETWORK

Raphaela Pensch[1,2], Sophie Agger[1,2,3], Suvi Mäkeläinen[1,2,4], Sergey Kozyrev[1,2], Sharadha Sakthikumar[1,2,5], Anna D van der Heiden[1,2], Ananya Roy[2,4], Katarina Tengvall[1,2], Lauren E Burt[6], Jaime F Modiano[6], Karin Forsberg-Nilsson[2,4], Maja L Arendt[1,2,3], Kerstin Lindblad-Toh[1,2,5]

[1]Uppsala University, Department of Medical Biochemistry and Microbiology, Uppsala, Sweden, [2]Uppsala University, SciLifeLab, Uppsala, Sweden, [3]University of Copenhagen, Department of Veterinary Clinical Sciences, Copenhagen, Denmark, [4]Uppsala University, Department of Immunology, Genetics and Pathology, Uppsala, Sweden, [5]Broad Institute, Department of Vertebrate Genomics, Cambridge, MA, [6]University of Minnesota, Department of Veterinary Clinical Sciences, St. Paul, MN

Canine osteosarcoma (OSA) is a common and aggressive bone cancer that serves as a valuable model for the rare human cancer. Besides protein-coding mutations, which have been widely investigated, non-coding mutations play a crucial role in driving cancer. However, their contribution to OSA tumorigenesis and their translational relevance remain unexplored. In this comparative study, we leveraged tumor/normal whole-genome sequencing data of 116 canine and 38 human OSA patients to study the role of non-coding mutations in OSA. Utilizing the evolutionary constraint scores from the Zoonomia project, candidate driver genes with an enrichment of likely functional non-coding constraint mutations (NCCMs) were identified. Eight candidates, including the oncogenes *SOX2* and *BCL11A*, were shared across species and their NCCMs were proposed as novel non-coding drivers. To investigate if the dog and human OSA gene sets shared underlying biological mechanisms, we applied gene interaction network propagation and network colocalization analysis. This revealed a colocalized network that was conserved between both species and significantly associated with mammalian phenotype ontologies related to tumor-free survival time, tumor incidence and abnormal osteoblast differentiation. Furthermore, genes transcriptionally regulated by the methyl-CpG binding protein MECP2 were overrepresented in the colocalized network and binding motifs of the MECP2 co-repressor complex were disrupted by NCCMs in multiple known target genes in dogs and humans. These findings suggest transcriptional regulation by MECP2 as a novel cross-species driver pathway and indicate that NCCMs are likely involved in its dysregulation in OSA. Overall, our results demonstrate the power of evolutionary constraint in identifying novel candidate drivers that might be used to develop better diagnostic, prognostic and treatment strategies in the future. Moreover, our study emphasizes the significance of canine OSA as a suitable model for studying the human disease.

FUNCTIONALLY DEFINING THE GENE REGULATORY EFFECTS OF COMPLEX AUTOIMMUNE DISEASE ALLELES

Soumya Raychaudhuri[1,2,3]

[1]Brigham and Women's Hospital, Divisions of Rheumatology and Genetics, Boston, MA, [2]Harvard Medical School, Medicine, Biomedical Informatics, Boston, MA, [3]Broad Institute, Medical and Population Genetics, Cambridge, MA

Rheumatoid arthritis is a canonical autoimmune disease affecting up to 1% of the population. It is associated with autoantibodies and causes chronic inflammatory arthritis, leading to pain, disability, and a reduction in lifespan. While ~150 risk alleles have been identified in addition to those within the HLA, disease mechanisms continue to be elusive. Here, we present data on using single-cell data to define key cell states in autoimmune diseases and regulatory mechanisms within rheumatoid arthritis and other immunological diseases. We investigate how single-cell data can be used to define gene regulatory mechanisms that risk alleles act on to cause autoimmune disease. We demonstrate how we can construct cell-state-specific enhancer-gene maps from multimodal data with RNA and ATAC-seq data capturing transcription and open chromatin jointly. Then, we demonstrate how we can use CRISPR gene editing to confirm the precise functional effect of these alleles in cell lines and primary cells. Our approach is a powerful single-cell assay, MINECRAFT-seq, that sequences DNA to capture edited cells alongside surface marker tags and whole-genome RNA. This approach enables us to define the precise cis and trans molecular effects of disease alleles. With these approaches, we take steps to advance the understanding of the genetic mechanisms of autoimmunity and start building the essential gene regulatory network underlying autoimmune conditions.

FOS BINDING SITES ARE A HUB FOR THE EVOLUTION OF ACTIVITY-DEPENDENT GENE REGULATORY PROGRAMS IN HUMAN NEURONS

Ava C Carter[1], Janet H Song[2,3,4,5], Gabriel T Koreman[1,7], Jillian E Petrocelli[1], Josephine E Robb[1], Evan Buchinsky[2,3,4,5], Sara K Trowbridge[1,8], David M Kingsley[6,9], Christopher A Walsh[2,3,4,6], Michael E Greenberg[1]

[1]Harvard Medical School, Neurobiology, Boston, MA, [2]Boston Childrens Hospital, Division of Genetics and Genomics, Boston, MA, [3]Harvard Medical School, Pediatrics, Boston, MA, [4]Harvard Medical School, Neurology, Boston, MA, [5]Harvard Medical School, Allen Discovery Center for Human Brain Evolution, Boston, MA, [6]Howard Hughes Medical Institute, NA, Chevy Chase, MD, [7]Harvard Medical School, Program in Neuroscience, Boston, MA, [8]Boston Childrens Hospital, Department of Neurology, Boston, MA, [9]Stanford University, Department of Developmental Biology, Stanford, CA

Humans have evolved dramatic modifications to brain development, maturation, and plasticity that likely underlie our unique cognitive abilities. After birth, sensory inputs to neurons trigger the induction of activity-dependent gene expression (ADGE) that mediates many aspects of neuronal maturation, a process that is uniquely and dramatically protracted humans. It is unknown whether evolutionary changes to ADGE programs underlie human-specific aspects of brain development. To identify human-specific features of the ADGE program, we characterized this gene program in neurons derived from human-chimpanzee tetraploid pluripotent stem cells, where alleles from the two species share the same nuclear environment and can be stimulated in a synchronized manner. This system allows us to distinguish cis from trans regulatory effects on gene expression and ultimately identify regulatory regions where sequence changes drive gene expression differences. We identified 235 activity-dependent genes that are differentially expressed in cis between humans and chimpanzees. We further found that the nearby activity-dependent open chromatin regions that may regulate these genes are highly enriched for AP-1 motifs and are bound by the activity-dependent transcription factor FOS. A quantitative assessment of these FOS-binding sites revealed that over half of them display biased FOS binding to the human or chimpanzee alleles and are enriched for single nucleotide variants between humans and chimpanzees that would be predicted to eliminate FOS binding. Targeting the FOS-bound enhancers for a subset of human-biased AD genes significantly reduces ADGE and changes neuronal firing dynamics as measured on multielectrode arrays. Taken together, our findings indicate that FOS-bound AP-1 enhancers are sites of frequent evolution in humans that lead to human-specific features of ADGE and may contribute to the unusually protracted and complex process of postnatal brain development.

UNCOVERING THE ROLE OF REGULATORY VARIANTS IN HUMAN EVOLUTION

David Gokhman[1], Ryder Easterlin[2], Nadav Mishol[1], Yizhi Yan[3], Katharina Lange[1], Simon Fishilevich[1], Nadav Ahituv[2], Fumitaka Inoue[3]

[1]Weizmann Institute of Science, Molecular Genetics, Rehovot, Israel, [2]University of California, San Francisco, Institute for Human Genetics, San Francisco, CA, [3]Kyoto University, ASHBi, Kyoto, Japan

Changes in gene regulation are key drivers of human evolution. However, which regulatory changes shaped human adaptations, and especially how, remains largely unknown. Here, we employed massively parallel reporter assays in skeletal, neural, and skin cells to uncover the functional role of all of the 71,443 variants distinguishing Neanderthals and Denisovans from modern humans, as well as the 541,851 variants distinguishing all human lineages from other great apes. This extensive catalog allowed us to discover thousands of single-nucleotide variants that altered human gene expression levels. To identify the genes affected by these expression-altering variants, we generated human-chimp and human-gorilla hybrid cells, a powerful system to detect cis-regulatory expression differences driven by nearby variants. Synergizing these approaches, we found three systems that likely experienced unique selective pressures during human evolution: the face, vocal tract, and cerebellum. Interestingly, we detected several examples of convergent evolution between modern and archaic humans. For example, both lineages completely silenced the activity of an enhancer of *KDM8*, a gene involved in tumor progression, but Neanderthals and Denisovans achieved this through motif disruption, while modern humans accomplished this through hypermethylation. Finally, we focused on two regions, and using CRISPR/Cas9 and mouse models, we uncovered their central role in shaping human-specific traits. These include a variant that doubled the expression of *EVC*, a regulator of facial development, likely resulting in the reshaping of our face, and a variant that silenced *IRF4*, a regulator of skin pigmentation. Overall, this work helps uncover key evolutionary changes by synergizing two powerful systems – reporter assays and human-ape hybrids. Together, this allowed us to generate the first comprehensive catalog of functionally important noncoding variants in deep and recent human evolution, and to shed light on the regulatory changes underlying pivotal human adaptations.

CHARACTERIZATION OF CELL TYPE-SPECIFIC ISOFORM EXPRESSION IN THE ADULT HUMAN CORTEX USING LONG-READ RNA SEQUENCING

Yoav Hadas[1,2,3,4], Xiao Lin[1,2,3,4], Emma Monte[6], Tao Wang[6], Maya Fridrikh[2], Soumya Kundu[6], Robert Sebra[2,4], Harm van Bakel[2,4,7,8], Michael Snyder[6], Joachim Hallmayer[6], Alexander Urban[6], Dalila Pinto[1,2,3,4]

[1]Icahn School of Medicine at Mount Sinai, Department of Psychiatry, New York, NY, [2]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomics Sciences, New York, NY, [3]Icahn School of Medicine at Mount Sinai, The Mindich Child Health and Development Institute, New York, NY, [4]Icahn School of Medicine at Mount Sinai, Icahn Genomics Institute, New York, NY, [5]Icahn School of Medicine at Mount Sinai, Friedman Brain Institute, New York, NY, [6]Stanford University, School of Medicine, Palo Alto, CA, [7]Icahn School of Medicine at Mount Sinai, Department of Microbiology, New York, NY, [8]Icahn School of Medicine at Mount Sinai, Department of Artificial Intelligence and Human Health, New York, NY

Our previous work has shown that alterations in gene and isoform expression play a key role in the etiology of neuropsychiatric disorders. To obtain an accurate transcriptome reference of the human brain, we have used long-read RNA sequencing to construct a high-resolution map (IsoHuB) of full-length mRNA in the human prefrontal cortex (PFC), revealing thousands of novel isoforms encoding novel proteins. Here, we expanded on this work by analyzing isoform expression at single-nucleus resolution.

We isolated single nuclei (sn) from the postmortem PFC of 32 neurotypical donors and generated short-read (10x snRNA-seq, 10x Multiome, and snSMART-seq) and long-read sequencing data (snIso-Seq and MAS-Iso-Seq). Additionally, we incorporated 11 publicly available snIso-Seq and snSMART-seq cortical datasets. The assembled datasets were aligned to our isoHuB long-read reference supplemented with GENCODE 41 and used to quantify gene abundance for cell type assignment as well as for isoform characterization and quantification. A total of 210,000 nuclei across 43 donors passed QC, and at least 27 cortical cell subtypes were identified across all data types. More than 160,000 full-length isoforms were characterized using snIso-Seq and MAS-Iso-Seq data, and more than 300,000 isoforms were quantified using snSMART-seq data. Over 1,000 novel isoforms belong to previously unannotated loci. Among the newly detected loci, 38 exhibited cell type specificity.

We further identified widespread differences in isoform abundance and usage between PFC cell types, including isoforms with combinations of distant splice sites that cannot be resolved using short-read RNA-seq data alone. To explore the biological functions of isoforms across different cell types, we applied multiple algorithms to predict protein domains and localization signals of the novel isoforms. We identified thousands of protein isoforms with novel functions, many of which are specific to neuronal or glial cell types.

Overall, our results demonstrate that combining short- and long-read RNA sequencing at single-nucleus resolution is an efficient approach for characterizing cell type-specific RNA isoform usage.

FUNCTIONAL CHARACTERIZATION OF GENE REGULATORY ELEMENTS

Nadav Ahituv[1,2]

[1]UCSF, Bioengineering and Therapeutic Sciences, San Francisco, CA,
[2]UCSF, Institute for Human Genetics, San Francisco, CA

Nucleotide variation in gene regulatory elements is a major determinant of phenotypes including morphological diversity between species, human variation and human disease. Despite continual progress in the cataloging of these elements, little is known about the code and grammatical rules that govern their function. Deciphering the code and their grammatical rules will enable high-resolution mapping of regulatory elements, accurate interpretation of nucleotide variation within them and the design of sequences that can deliver molecules for therapeutic purposes. To this end, we are using massively parallel reporter assays (MPRAs) to simultaneously test the activity of thousands of gene regulatory elements in parallel. By designing MPRAs to learn regulatory grammar or to carry out saturation mutagenesis of nucleotide changes in disease causing gene regulatory elements, we are increasing our understanding of the phenotypic consequences of gene regulatory mutations.

HIGH-THROUGHPUT TARGET DISCOVERY FOR NON-CODING AUTOIMMUNE GWAS LOCI IN PRIMARY HUMAN IMMUNE CELLS.

Viacheslav A Kriachkov[1], Davide Vespasiani[1], Jeralyn Ching Wen Hui[1], Vanessa Bryant[2], Liam Gubbels[3], Melanie Neeland[3], Shivanthan Shanthikumar[3], Hamish W King[1]

[1]Walter and Eliza Hall Institute, Genetics and Gene Regulation, Melbourne, Australia, [2]Walter and Eliza Hall Institute, Immunology, Melbourne, Australia, [3]Murdoch Children's Research Institute, Melbourne, Australia

Genome-wide association studies have discovered thousands of genetic variants linked with autoimmune disease, and yet the underlying molecular pathways have remained elusive. A key challenge is that >90% of identified GWAS hits are in non-coding genomic regions that makes it difficult to predict their contribution to disease. Here, we have experimentally tested the regulatory function of >750 autoimmune risk loci in human B cells - a highly relevant cell type given that auto-antibody production by self-reactive B cells is common to many autoimmune diseases.

Analysis of fine-mapped genetic variants from over 30 different autoimmune traits (including common conditions such as lupus, Crohn's disease, and multiple sclerosis) revealed that 146/818 (17.8%) of autoimmune risk variants in B cell open chromatin are associated with multiple disease traits, suggesting shared molecular pathways underlying diverse autoimmune diseases. To test the functional relevance of these non-coding risk loci, and identify their potential target genes, we coupled a CRISPR activation (CRISPRa) screen with single-cell RNA-seq and surface protein detection in primary human B cell cultures. To our knowledge, this is the first single-cell CRISPRa screen performed in primary human B cells. We tested 762 autoimmune risk loci (encompassing 780 SNPs) and identified significant 572 locus-target changes in cis (<500kb) for 385 (50%) unique risk loci, at a mean distance of 60.8kb. While the nearest gene was targeted 83.9% of the time, nearly half (47%) of autoimmune risk loci had multiple gene targets and we uncover many examples of complex regulatory landscapes including gene skipping and co-regulation of multiple promoters. Analysis of surface protein levels confirmed concordant changes in surface protein and gene expression changes for 29 autoimmune risk loci, including for chemokine receptors known to mediate B cell migration and survival. Our CRISPRa strategy allowed identification of lowly expressed gene targets for autoimmune loci, including cytokines and transcription factors. Finally, we quantify the downstream *trans*-regulatory networks of *cis*-target transcription factors by examining differential gene and enhancer-dependent RNA transcription in our single-cell transcriptomic dataset. Combined with massive parallel reporter assays, phenotypic screens, and prime editing, our study has discovered the gene regulatory targets and consequences of non-coding autoimmune genetic variation in human B cells and represents a major advance in our understanding of the genetic networks that may drive autoimmune disease.

SYSTEMATIC ANALYSIS OF THE IMPACT OF SHORT TANDEM REPEATS ON GENE EXPRESSION

Xuan Zhang[1], Lingzhi Zhang[1], Susan Benton[1], Ellice Wang[1], Eric Mendenhall[2], Alon Goren[1], Melissa Gymrek[1]

[1]University of California, San Diego, Department of Medicine, La Jolla, CA, [2]HudsonAlpha Institute for Biotechnology, Biological Sciences, Huntsville, AL

Short tandem repeats (STRs) represent one of the most dynamic elements of the human genome. These repetitive sequences, ranging from 1-6 bp, exhibit mutation rates orders of magnitude higher than single nucleotide changes. Recent studies have revealed widespread associations between STR variation and complex traits, including gene expression and splicing. However, how STRs mechanistically influence gene regulation has remained largely unexplored due to technical challenges in synthesizing, manipulating, sequencing, and measuring the effects of these repetitive sequences.

We optimized a massively parallel reporter assay (MPRA) to characterize over 33,000 promoter-proximal STR loci by designing a library with 3-4 variants per locus. Each element is flanked by 55-76 bp of surrounding genomic context, tagged with random barcodes, and cloned upstream of a reporter gene. We employed complementary sequencing technologies to capture a comprehensive spectrum of STR variants, including challenging homopolymers longer than 25 bp.

Of the 19,818 STRs that passed filtering, we identified 1,366 loci with significant associations between repeat copy number and gene expression. These loci showed a bias toward positive effect sizes, with longer repeats linked to higher expression. Notably, GC-rich repeats consistently increased expression levels. We selected the top 300 loci and designed a second array for detailed perturbation studies, manipulating repeat unit, orientation, and length, with 200-300 perturbations per locus. Our results revealed that the repeat unit sequence is the primary driver of expression differences, while strand orientation and flanking sequence context had weaker effects. The high resolution of this perturbation-MPRA enabled us to detect both non-linear effects (e.g., AAAC/GTTT) and linear effects (e.g., AGCG), which emerge beyond a certain copy number threshold. Finally, we observed crosstalk between the GC content of the repeat unit and the surrounding flanking regions, particularly affecting AT-rich repeat units.

We demonstrate that STRs can directly modulate gene expression depending on sequence composition, repeat unit size, and genomic context. Through systematic dissection of numerous regulatory STRs, we reveal unexpected complexity in how these elements influence transcription, including non-linear effects and sequence-specific thresholds. Our results establish STRs as a source of regulatory variation and provide a framework for understanding how this dynamic form of genetic variation shapes gene expression.

GENERAL PRINCIPLES AND CELL TYPE-SPECIFICITY OF THE HUMAN RNA-DNA INTERACTOME

Alice Lambolez[1], The FANTOM6 Consortium[1], Hazuki Takahashi[1], Piero Carninci[1,2]

[1]RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan, [2]Human Technopole, Research Center for Genomics, Milan, Italy

Ever since its foundation in 2000, the FANTOM project has aimed to decode the mammalian genes and their regulation. 20 years ago, the consortium unveiled that a large portion of the transcriptome does not necessarily code for proteins. In its on-going 6th edition, FANTOM focuses on the tens of thousands of RNAs that interact with and could regulate the chromatin, and for which the functions and mechanisms remain mostly poorly understood. To this end, we applied and intersected a wide array of whole genome sequencing technologies, such as CAGE, CUT&Tag, ATAC-seq, Hi-C and RADICL-seq, in 16 different human cell types, including differentiating iPS cells and activated immune cells. This massive endeavor allowed us prepare the first global RNA-DNA interaction atlas in a cell type-specific context.

Network analyses on RADICL-seq data revealed that this interactome is highly organized and evolves during cell differentiation. RNAs that bind to chromatin do not merely recapitulate the transcriptional landscape and primarily derive from introns and exons of protein- and lncRNA-coding loci. Moreover, we show that the interactome is highly cell type-specific, both in terms of pool of DNA-binding RNAs and of targeted regions.
These targeted loci can be located either close to the RNA source or as far as on a different chromosome. The proportion of short- *versus* long-distance interactions varies in function of the cell type, the coding nature of the RNA, or whether it contains repeat elements. We also observed a clear link between the 3D conformation of the chromatin and the localization of RNA-DNA contacts, which tend to occur within the same TAD. Furthermore, RNAs preferentially bind to open chromatin regions and their interaction landscape appears to be modulated by changes in chromatin accessibility between cell types. Notably, different types of chromatin sub-compartments, which are associated with distinct epigenetic landscapes, show different levels of RNA-DNA interactions.

DNA-binding RNAs significantly overlap those involved in non-self RNA-RNA interactions and are enriched in elements related to RNA-binding proteins, implying that RNA-DNA interactions contribute to the formation of complex structures on the chromatin. In particular, we detected numerous cell-type specific RNA-aggregation hotspots which are favored by snoRNAs, enriched in active regions, and likely correspond to RNA-processing platforms.
Finally, regulatory elements targeted by RNAs tend to show strong changes in activity when they are differentially interacted, suggesting that RNA-DNA interactions function in part as regulators of the target's expression. Both RNAs and targeted elements being significantly associated with cell type-relevant and/or disease-related traits, our results highlight the important role played by these interactions in the proper functioning of the cell.

DYNAMICS OF INBREEDING AND GENE FLOW IN A PEDIGREED WILD POPULATION OF FLORIDA SCRUB-JAYS

Faye Romero[1], Jeremy Summers[1], James Schmidt[1], Daniel Seidman[1], Sahas Barve[2], John Fitzpatrick[3], Nancy Chen[1]

[1]University of Rochester, Biology, Rochester, NY, [2]Archbold Biological Station, Avian Ecology, Venus, FL, [3]Cornell University, Ecology and Evolutionary Biology, Ithaca, NY

Inbreeding depression, or the reduced fitness of offspring of related parents, is a common phenomenon that can cause the decline and eventual extinction of natural populations. A potentially powerful approach for mitigating the effects of drift and inbreeding is by introducing genetic variation via gene flow. However, gene flow can also introduce more deleterious variation or lead to outbreeding depression, and genetic or fitness effects may change across generations. As a result, assisted translocations as a management strategy remains controversial. Our understanding of the long-term effects of gene flow is limited by our inability to directly measure the reproductive success of immigrants over time, except in a few special study systems. Here, we quantify the multigenerational impacts of gene flow in the endangered Florida Scrub-Jay (*Aphelocoma coerulescens*). A population of Florida Scrub-Jays at Archbold Biological Station has been studied since 1969, resulting in complete life histories for thousands of individuals on a 16-generation pedigree. We have identified all immigrant individuals into the study population since 1990, providing a rare opportunity to precisely measure the impact of immigration on levels of genetic variation and fitness over time. Previous work in our study population showed that decreased immigration from smaller peripheral populations resulted in increased inbreeding and reduced fitness via inbreeding depression, with evidence of fitness benefits of immigrant ancestry in females but outbreeding depression in males. To further elucidate the effects of gene flow in our population, we performed whole genome sequencing of >500 individuals and characterized levels of inbreeding and deleterious variation across the genome for recent immigrants, residents, and different generations of immigrant descendants. We found that immigrant birds were more inbred than resident birds, exhibiting more abundant and longer homozygous regions, with varying levels of deleterious variation. We test how levels of genetic diversity and fitness change in admixed lineages by comparing homozygous regions of the genome and levels of deleterious variation with individual fitness in different generations of immigrant descendants. This study provides a detailed look at the importance of immigrants to population genetic diversity over time. Our results have important implications for the potential success of genetic rescue or other assisted translocation strategies in conservation management.

THE GENETIC BASIS OF A UNIQUE STRUCTURAL COLORATION TRAIT IN THE PLATYFISH, *XIPHOPHORUS EVELYNAE*

Nadia B Haghani[1,2], John J Baczenas[1], Theresa R Gunn[1,2], Tristram O Dodge[1,2], Qinliu He[3], Sashoya Dougan[1], Paola Fascinetto-Zago[1,2], Zihao Ou[5], Gabe A Preising[1,2], Sophia Haase Cox[1], Kang Du[6], Manfred Schartl[6], Dan Powell[1,7], Guan-Zhu Han[3], Molly Schumer[1,2,4]

[1]Stanford University, Department of Biology, Stanford, CA, [2]Centro de Investigaciones Cientificas de las Huastecas "Aguazarca", Calnali, Mexico, [3]Nanjing Normal University, College of Life Sciences, Nanjing, China, [4]Howard Hughes Medical Institute, Stanford, CA, [5]Stanford University, Materials Science and Engineering, Stanford, CA, [6]Texas State University, Xiphophorus Genetic Stock Center, San Marcos, TX, [7]Louisiana State University, Biological Sciences, Baton Rouge, LA

The diversity of pigmentation patterns and colors across the tree of life is extraordinary, reflecting the essential process of adaptation to the environment. While many organisms use pigment molecules to produce these colors, some manipulate light in unique ways to generate complex visual cues, known as structural coloration. Structural colors have convergently evolved across many lineages yet remain understudied at the genetic and evolutionary levels. Swordtail fish of the genus *Xiphophorus* harbor a diverse array of heritable structural coloration traits and are becoming a tractable system to study the genetic basis of adaptation. We discovered a natural population of *Xiphophorus evelynae* that is polymorphic for a unique reflective pattern in their scales, causing these animals to sparkle as they swim through the water. This sparkle trait presents an exciting opportunity to investigate the genetic mechanisms that underlie trait evolution. To first identify genetic loci associated with sparkle, we performed a genome-wide association study (GWAS), which revealed a significant peak on chromosome 15 that spans a 72-gene region. To narrow candidate genes likely contributing to sparkle, we measured transcriptional profiles of the epidermis and found 2 differentially expressed genes that fall within the GWAS peak. One of these genes, anaplastic lymphoma kinase ligand 2 (alkal2), is known to activate a receptor tyrosine kinase and drive the differentiation of pigment progenitor cells into reflective iridophore cells. We demonstrated that the sparkle trait is disrupted upon treatment with a small-molecule inhibitor of the alkal2 receptor. Operating under the assumption that cis-regulatory elements drive the overexpression of alkal2 in sparkle individuals, we assembled multiple X. evelynae PacBio HiFi genomes to investigate genetic differences in the surrounding region. We discovered that an 18kb transposable element (TE) sequence of an active Endogenous Retrovirus (ERV) family is perfectly associated with the sparkle trait. To assess whether this insertion changes chromatin accessibility, we conducted ATAC-sequencing and identified regions in the ERV insertion with high chromatin accessibility, suggesting that this sequence acts as an enhancer. ERVs are often co-opted as cis-regulatory elements that contribute to regulatory variation. However, active retroviruses are typically associated with disease, and to our knowledge, have not been previously identified as regulators of adaptive traits. In this work, we investigate the genetic mechanism that underlies the evolution of an understudied, adaptive pigmentation trait.

HISTORY REPEATS ITSELF: COMPARATIVE GENOMICS REVEALS NEW GENES UNDERLYING CONVERGENT CHANGES IN VISION, HAIR, AND SPERM LOCOMOTION

Nathan L Clark[1], Maria Chikina[2], Courtney Charlesworth[1], Jered Stratton[1], Dwon Jordana[1], Amanda Kowalczyk[2], Emily Kopania[1]

[1]University of Pittsburgh, Biological Sciences, Pittsburgh, PA, [2]University of Pittsburgh, Computational and Systems Biology, Pittsburgh, PA

A combination of comparative genomics and phylogenetics is increasingly being used to assign genotype to phenotype through an approach called PhyloG2P. PhyloG2P examines species sharing a convergently evolved phenotype to identify related genes that experienced parallel changes. Our PhyloG2P methods search for positive selection, shifts in evolutionary rate, and pseudogenization events associated specifically with the branches over which a convergent phenotype evolved. Our studies have identified new genes and regulatory regions controlling traits important to fitness and health. In blind mammals, PhyloG2P readily identified ocular genes since many undergo a rate acceleration due to relaxed functional constraint. We take this approach further in a new innovation to assign quantitative trait scores to 420 mammalian species reflecting their visual ability, using solely their genome sequences. Visual trait scores provide an objective ranking of species ranging from poorly sighted species (bats, moles, armadillo) to species with high visual acuity (primates, felids). This continuous trait then serves as an even more powerful method to identify previously unknown ocular genes. We present functional characterization of 5 new ocular genes identified thus. These include a developmental caspase, lens-specific proteins, and neurological proteins contributing to visual perception, as measured in visual acuity tests in mutant embryonic zebrafish. Further application of trait scores to “bald” mammal species – those with reduced body hair (elephants, mole-rats, humans) - identified new microRNAs involved in hair follicle formation and maintenance that were experimentally validated in mice. Finally, our PhyloG2P studies identified new sperm proteins by studying sperm competition in rodents and primates. One exciting candidate is a newly evolved protein that originated in placental mammals and has been undergoing rapid adaptive evolution ever since. It is localized to the sperm midpiece, is only expressed in testis, and itself is a fragment of an ancient centrosomal protein. Given the involvement of centrosomal proteins in the sperm flagellum, we propose it modifies sperm locomotion and hence competition between conspecific males. A knock-out mouse will soon reveal its functional role. In general, PhyloG2P methods are already providing new high level functional assignments proteins, microRNAs, and regulatory regions, thereby accelerating the rate of biological discovery.

FAST AND FURIOUS MUTATION AT TANDEM REPEATS IN A LARGE, FOUR-GENERATION FAMILY

Thomas A Sasani[1], Michael E Goldberg[1], Thomas J Nicholas[1], Tom Mokveld[2], Egor Dolzhenko[2], Eli Kaufman[4], David Porubsky[3], Michael A Eberle[2], Evan E Eichler[3], Paul Valdmanis[4], Aaron R Quinlan[1], Harriet Dashnow[5]

[1]Univ. of Utah, Dept. of Human Genetics, Salt Lake City, UT, [2]PacBio, PacBio, Menlo Park, CA, [3]Univ. of Washington, Dept. of Genome Sciences, Seattle, WA, [4]Univ. of Washington, Division of Medical Genetics, Dept. of Medicine, Seattle, WA, [5]Univ. of Colorado, Dept. of Biomedical Informatics, Aurora, CO

Every generation, germ cells deploy a complex network of proteins to create near-perfect copies of our genomes. Infrequently, *de novo* mutations occur and provide the substrate for both evolution and genetic disease.
By sequencing thousands of families, we've carefully estimated the rate at which *de novo* single-nucleotide mutations occur in the human germline. But what about other types of mutation, such as expansions and contractions at tandem repeats (TRs)? These repetitive stretches of nucleotides — like a sequence of fifty CAG motifs, or a series of one hundred ATTTTTs — comprise almost 8% of the human genome and pose a challenge for DNA replication machinery. Polymerases may stall and stutter as they process over TR sequences, accidentally incorporating extra motifs (an "expansion") or deleting existing motifs (a "contraction") along the way. Since TRs are highly repetitive and can measure thousands of base pairs, they are also difficult to interrogate using short sequencing reads. To overcome these challenges, we sequenced a 28-member, four-generation CEPH/Utah family with the PacBio long-read platform, genotyped nearly 8 million TR loci, and performed a detailed analysis of TR mutagenesis in the human genome.
The average mutation rate at TR loci was 4.7×10^{-6} per locus per generation, orders of magnitude higher than at single-nucleotide sites. Not all TRs are created equal, however. Thirty-two loci were hyper-mutable, experiencing up to a dozen expansions or contractions across 44 meioses in the pedigree. Why were these loci hotspots for mutagenesis, while millions of others remained unchanged from generation to generation? We found that mutation rates were conditional on the motif composition of TR loci. At one locus, for example, we observed ten haplotypes with recurrent expansions and contractions of a particular 19 base pair motif. Haplotypes with a subtly different version of that motif — comprising just two additional nucleotides — remained relatively stable throughout the pedigree, and were much less polymorphic in an unrelated collection of 47 long-read human genomes. Our results reveal some of the complex factors that govern mutability at tandem repeats; a difference of just a few nucleotides might awaken a previously dormant motif, potentiating recurrent mutations in future generations.

THE EVOLUTION OF STRUCTURAL AND SINGLE NUCLEOTIDE MUTATION ACROSS HAPLOTYPE-RESOLVED VERTEBRATE GENOME ASSEMBLIES

Nicolas R Lou[1], Daven Lim[1], Minoli Daigavane[1], Nilah M Ioannidis[2], Peter H Sudmant[1]

[1]University of California Berkeley, Integrative Biology, Berkeley, CA, [2]University of California Berkeley, EECS, Berkeley, CA

Structural variants (SVs) contribute disproportionately to genetic variation, adaptation, and disease. However, due to challenges in characterizing SVs, their mutational properties, distribution, and diversity has remained understudied compared to single nucleotide variants (SNVs), particularly in non-model organisms. Here, taking advantage of recently generated haplotype resolved diploid genome assemblies across ~500 vertebrate species, we survey SV diversity across the vertebrate tree of life and study genomic features that drive this variation using machine learning methods. By generating SV length profiles and performing de-novo repeat annotations on SVs, we characterize lineage-specific expansions of transposable element activities across different taxonomic groups (e.g. Alu and L1 elements in primates, LTR elements in birds). While the levels of diversity in SVs and SNVs are generally positively correlated, we find that fish and amphibians tend to have more SVs than mammals, reptiles, and birds given the same number of SNVs. These results suggest that mutational mechanisms driving SV formation are fundamentally distinct in amniotes. In addition, species that are endangered or threatened tend to have lower levels of diversity in both SVs and SNVs. We train machine learning models based on sequence features to predict SV events across species and identify genomic features that drive structural variation. Using an ensemble of random forests and convolutional neural networks, SV occurrences can be consistently predicted (AUC >0.8) across species. The predictive ability of these models demonstrates that structural variation is often driven by specific sequence features. Features that contribute significantly to the model predictions include base composition, neighboring genetic variants, genes, and repetitive regions. These features are similar across species, suggesting that mechanisms driving SVs are broadly conserved across vertebrate genomes. In addition, we observe kmers of varying lengths that are significantly enriched near SV breakpoints, such as specific dinucleotide repeats and G-quadruplexes, which are also recognized by the convolutional neural network as strong predictors of SVs. These kmers have been shown to form unstable DNA structures and can thus contribute to chromosomal instability and double-stranded breaks. Together our research suggests that while the overall distribution of SVs has dramatically shifted over the last 500 million years, many of the underlying mutational mechanisms and selective drivers of SV diversity have remained conserved.

UNDERSTANDING MUTATIONAL PROCESSES FROM *ARABIDOPSIS* PANGENOME GRAPHS

Zhigui Bao[1], Fernando A Rabanal[1], Andrea Guarracino[2], Sebastian Vorbrugg[1], Wenfei Xian[1], Erik Garrison[2], Detlef Weigel[1]

[1]Max Planck Institute for Biology Tübingen, Molecular Biology, Tübingen, Germany, [2]University of Tennessee Health Science Center, Genetics, Genomics and Informatics, Memphis, TN

The precise enumeration of genetic variants distinguishing a collection of individuals is fundamental to understanding evolution. However, even aligning two genomes remains challenging because fixed parameters are applied across chromosomes despite varying evolutionary divergence, making population-level comparisons even more complex. While pangenome graphs theoretically offer an ideal solution for all-against-all whole genome comparison, their computational intensity prevents scaling to large populations.

To address these limitations, we curated over 600 long-read assemblies of *Arabidopsis thaliana* genomes, comparing variant detection by conventional short- or long-read mapping and multiple whole-genome alignment approaches. While all alignment-based methods effectively captured small variants, including those in regions with spurious read mapping and low mappability, there was considerable disagreement with structural variants (SVs, >50 bp), with only two-thirds consistently identified, even in single-genome comparisons. Population-level variation was particularly high in highly divergent regions characterized by clustered SVs, but current clustering methods could not adequately identify allelic groups. This in turn makes it difficult to decipher the underlying mutational mechanisms.

To enable efficient large-scale genomic comparisons, we developed GRASP (Graph Reconstruction Anchored by Subgraph Partitioning), a computational framework that rapidly characterizes structural diversity at orthologous loci across populations. Using GRASP, we uncovered extensive long-range linkage disequilibrium between functional alleles. This comprehensive and unbiased cataloging of genetic variation has the potential to revolutionize the understanding of the genotype-phenotype map and the prediction of favorable allelic combinations that can fuel adaptation to a rapidly changing environment. While developed for A. thaliana, GRASP should be broadly applicable to any species with multiple high-quality genomes.

THE MUC19 GENE IN DENISOVANS, NEANDERTHALS, AND MODERN HUMANS: AN EVOLUTIONARY HISTORY OF RECURRENT INTROGRESSION AND NATURAL SELECTION

Fernando Villanea*[2], David Peede*[1,9], Eli Kaufman[3], Valeria Añorve-Garibay[1,9], Elizabeth Chevy[1,9], Viridiana Villa-Islas[4], Kelsey Witt[5], Roberta Zeloni[6], Davide Marnetto[6], Priya Moorjani[7], Flora Jay[8], Paul Valdmanis[3], Maria Avila-Arcos[4], Emilia Huerta-Sanchez[1,9]

[1]Brown University, Ecology Evolution and Organismal Biology, Providence, RI, [2]University of Colorado Boulder, Anthropology, Boulder, CO, [3]University of Washington, Division of Medical Genetics, Department of Medicine, Seattle, WA, [4]Universidad Nacional Autónoma de Mexico, International Laboratory for Human Genome Research, Queretaro, Mexico, [5]Clemson University, Center for Human Genetics and Department of Genetics and Biochemistry, Clemson, SC, [6]University of Turin, Neurosciences "Rita Levi Montalcini", Turin, Italy, [7]University of California, Berkeley, Molecular and Cell Biology, Berkeley, CA, [8]Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France, [9]Brown University, Center for Computational Molecular Biology, Providence, RI

The study of archaic introgression has already illuminated candidate genomic regions that affect the health and overall fitness of some populations. In this talk, I will focus on identifying regions where archaic ancestry could have been a useful source of standing genetic variation as the early Indigenous American populations adapted to new environments. We pinpoint a gene MUC19, for which modern humans carry an archaic haplotype, which harbors a high density of Denisovan-specific variants, including nine of which are missense variants. We find the Denisovan-specific variants for the archaic MUC19 haplotype have risen to high frequencies in admixed Latin American individuals among global populations, and at highest frequency in 23 ancient Indigenous American individuals. Additionally, we find that the archaic MUC19 haplotype carries a higher copy number of a 30 base pair variable number tandem repeats (VNTR), and that copy number variation of this repeat is at high frequency in admixed Latin American populations and is associated with the number of introgressed haplotypes within an individual at MUC19. Finally, we find that some Neanderthals carry the Denisovan-like MUC19 haplotype, and that it was likely introgressed into human populations through Neanderthal introgression rather than Denisovan introgression. This study provides one of the first examples of positive selection acting on VNTRs on an archaic genetic background. Our results show that investigating patterns of archaic ancestry in populations from the Americas can identify exciting candidate loci that can expand our understanding of adaptation from archaic standing variation.

UNCOVERING GENE REGULATORY DIFFERENCES BETWEEN HUMAN AND CHIMPANZEE NEURAL PROGENITORS

Janet H Song[1,2,3,4], Ava C Carter[1,5], Evan M Bushinsky[1,2,3,4], Samantha G Beck[1,2,3,4], Jillian E Petrocelli[1,5], Gabriel T Koreman[1,5], Juliana Babu[1,2,3,4], David M Kingsley[4,6], Michael E Greenberg[1,5], Christopher A Walsh[1,2,3,4]

[1]Allen Discovery Center, Human Brain Evolution, Boston, MA, [2]Boston Children's Hospital, Genetics and Genomics, Boston, MA, [3]Harvard Medical School, Pediatrics and Neurology, Boston, MA, [4]Howard Hughes Medical Institute, Chevy Chase, MD, [5]Harvard Medical School, Neurobiology, Boston, MA, [6]Stanford University, Developmental Biology, Stanford, CA

Although comparisons of human and non-human primate brains have identified thousands of molecular differences, it has been difficult to identify the human-specific sequence variants that underlie the dramatic modifications to brain size, connectivity, and function found in humans. One hurdle is that current comparative approaches cannot distinguish *cis*-regulated genes, which change in expression due to nearby sequence variants on the same DNA molecule, from *trans*-regulated genes, which change in expression due to changes in diffusible factors in the cellular environment (like the levels of *cis*-regulated transcription factors (TFs)). To distinguish *cis* from *trans* changes, we generated human-chimpanzee tetraploid stem cell lines as a genetic model where the human and chimpanzee genomes are in the same cellular environment and only *cis*-regulated changes are observed. We have now used this system to profile *cis*- and *trans*-regulated genes and open chromatin regions in neural progenitor cells (NPCs), in order to identify genetic changes that underlie the expansion in size and neuron number in the human brain. Genes that are more highly expressed in humans are enriched for processes related to the cell cycle, consistent with increased neurogenesis in humans. We identify *cis*-regulated TFs, including *FOSL2* and *MAZ*, whose motifs are enriched at *trans*-regulated open chromatin peaks, suggesting that these TFs may be major drivers of epigenomic and transcriptomic rewiring between human and chimpanzee NPCs. To identify human-specific variants that underlie *cis*-regulated gene expression changes, we linked *cis*-regulated open chromatin peaks that contain derived sequence changes in humans to nearby *cis*-regulated genes. A CRISPR inhibition screen of 106 *cis*-regulated peaks identified species-specific enhancers, including one near *TNIK*. Further characterization of *cis*-regulated TFs and non-coding regions in NPCs, along with the application of this model to additional cell types and paradigms, will advance our understanding of how human-specific sequence changes contribute to increased brain size, as well as other phenotypes that have arisen in the human lineage.

SPATIALLY-AWARE QUALITY CONTROL FOR SPATIAL TRANSCRIPTOMICS

Michael Totty[1], Stephanie C Hicks[1,2], Boyi Guo[1]

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

Quality control (QC) is a crucial step to ensure the reliability of data obtained from RNA sequencing experiments, including spatially-resolved transcriptomics (SRT). Existing QC approaches for SRT that have been adopted from single-cell/nucleus RNA-sequencing (sc/snRNA-seq) methods are confounded by spatial biology and are inappropriate for SRT data. Here, we introduce SpotSweeper, spatially-aware QC methods that leverages local neighborhoods to correct for biological confounds in order to identify both local outliers and regional artifacts in SRT. Using SpotSweeper on publicly available data, we identified a consistent set of Visium barcodes/spots as systematically low quality and demonstrate that SpotSweeper accurately identifies two distinct types of regional artifacts. SpotSweeper represents a significant advancement in spatially-resolved transcriptomics quality control, providing a robust, generalizable framework to ensure data reliability across diverse experimental conditions and technologies.

DECODING SEQUENCE DETERMINANTS OF GENE EXPRESSION IN DIVERSE CELLULAR AND DISEASE STATES

Avantika Lal[1], Alexander Karollus[1,2], Laura Gunsalus[1], David Garfield[1], Surag Nair[1], Alex M Tseng[1], M G Gordon[3], John Blischak[4], Bryce van de Geijn[4], Tushar Bhangale[4], Jenna L Collier[1], Nathaniel Diamant[1], Tommaso Biancalani[1], Hector Corrada Bravo[1], Gabriele Scalia[1], Gokcen Eraslan[1]

[1]Genentech, gRED Computational Sciences, South San Francisco, CA, [2]Technical University of Munich, School of Computation, Information and Technology, Munich, Germany, [3]Genentech, Cellular and Tissue Genomics, South San Francisco, CA, [4]Genentech, Human Genetics, South San Francisco, CA

Sequence-to-function deep learning models have transformed our understanding of gene regulation. However, existing models are largely trained on bulk data from healthy tissues, and so cannot model gene expression in specific cell types or disease contexts. In contrast, single-cell RNA sequencing (scRNA-seq) datasets profile expression in diverse cell types, states, and diseases, yet revealing regulatory mechanisms from such data is challenging.

We present Decima, a sequence-to-function model trained on atlas-scale scRNA-seq data. Decima is initially trained to predict genome-wide coverage in bulk sequencing assays, similarly to the Borzoi model. Unlike existing models, it is then fine-tuned on pseudobulked scRNA-seq data from over 22 million cells. Given the 500 kb sequence surrounding an unseen gene, Decima can predict its expression in hundreds of cell types, across diverse cellular states, tissues, and diseases.

Decima identifies regulatory elements, including cell type-specific distal enhancers, and binding sites for transcription factors (TFs) that define cell type identities. It can predict subtle changes in expression between the same cell type across different tissues, states, or diseases, revealing the underlying regulatory grammar. For example, it identified TF motifs associated with fibroblasts in specific tissues and highlighted TF interactions driving fibroblast-specific disease responses.

Decima directly predicts non-coding variant effects at cell type resolution. On diverse single-cell eQTL and GWAS datasets, it achieves state-of-the-art performance in identifying expression-altering variants, predicting the magnitude and direction of their effect, and identifying the affected cell types. Moreover, its attributions yield mechanistic hypotheses to explain variant effects.

Finally, Decima can design regulatory elements that drive cell type specific expression. Using Decima, we designed fibroblast-specific promoters, and even promoters with elevated activity in inflammatory fibroblasts. This may enable novel gene therapies that are activated specifically in diseased cells. Altogether, Decima combines the utility of sequence models with the richness of single-cell datasets, unlocking new applications in biology.

IDENTIFICATION OF RULES UNDERLYING HOW INDIVIDUAL CELL TYPES GIVE RISE TO CONVERGENT PHENOTYPES IN THE BRAIN

Hongru Hu[1,2], Gerald Quon[1,2]

[1]University of California-Davis, Molecular and Cellular Biology, Davis, CA, [2]University of California-Davis, Genome Center, Davis, CA

Tissues, organs and other complex biological samples are made up of collections of diverse cell types, that interact according to some underlying biological "ruleset" to produce convergent sample-level phenotypes. A major unaddressed challenge in single cell genomics is to predict these rulesets from scRNA-seq data. We have developed a deep learning framework that takes as input a collection of cells from a single complex sample, passes the collection of single cell measurements through a series of "sensors" trained to identify interactions between cells, and makes sample-level predictions of phenotype. Our strategy thus bridges the gap between cellular heterogeneity and sample-level traits and helps uncover the biological mechanisms underlying phenotypic diversity.

We have applied our framework to single cell atlases of the human neural organoid (HNOCA) and developing human brain (HDBCA) to uncover rules underlying developmental stage and neurological disease status. It accurately predicted the developmental order of 282 HNOCA samples (validation correlation = 0.96), and identified two key contributing cell types towards defining organoid age, ventral and Dorsal Telencephalic neurons. Our results broadly agree with the established knowledge that these cell populations primarily generate GABAergic interneurons and glutamatergic neurons, respectively. To test generalizability, we applied the framework trained on HNOCA to HDBCA. The model successfully predicted unseen fetal brain development stages (correlation = 0.87), showing that the model captured shared developmental patterns between lab-grown organoids and natural brain tissue. Unlike the organoid samples, in the natural brain tissues, the radial glia cell type was predominantly identified as driving prediction of organoid age. These results also agree with the role of radial glia as the primary progenitors during neurodevelopment. Visualization of the summary statistics our framework computes over each sample from both HNOCA and HDBCA revealed that organoid developmental age follows a parallel trajectory similar to human brain development, which validates organoids as effective models of human brain growth.

Beyond development, we also trained the framework on human neural organoids datasets that model 10 different major neurological disorders, where we effectively distinguished these conditions (validation F1 score = 0.61), and identified influential cell types for disease classification. These findings suggest that different disorders impact specific cell populations in unique ways, offering insights into disease-specific cellular mechanisms.

PREDICTING DELETERIOUS PROMOTER MUTATIONS WITH DEEP LEARNING

Kishore Jaganathan[1], Nicole Ersaro[1], Gherman Novakovsky[1], Yuchuan Wang[1], Evin Padhi[2], Ziming Weng[2], Jeremy Schwartzentruber[1], Petko Fiziev[1], Irfahan Kassam[1], Ashley Lim[1], Grace Png[1], Jacob Ulirsch[1], Anshul Kundaje[1,3], Anne O'Donnell-Luria[4], Stephan Sanders[5], Heidi Rehm[4], Stephen Montgomery[2], Kyle Farh[1]

[1]Illumina Inc, Artificial Intelligence Department, Foster City, CA, [2]Stanford University, Department of Pathology, Stanford, CA, [3]Stanford University, Department of Genetics, Stanford, CA, [4]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [5]University of California San Francisco, Department of Psychiatry and Behavioral Sciences, San Francisco, CA

Only a minority of patients with rare genetic disorders are currently diagnosed by exome sequencing, suggesting that additional unrecognized pathogenic variants may reside in non-coding regions. We introduce PromoterAI, a deep neural network that accurately identifies non-coding promoter variants which dysregulate gene expression. PromoterAI is initially trained to predict histone modifications, DNA accessibility, transcription factor binding, and strand-specific CAGE signals at base-pair resolution around transcription start sites. The model is subsequently fine-tuned on a curated set of rare promoter variants associated with outlier gene expression in the GTEx cohort, leveraging novel cis- and trans-factor corrections to increase outlier size and quality.

This fine-tuning process significantly improves performance across extensive validation benchmarks, generalizing to both unseen genes and independent datasets. The fine-tuned model also shows improved alignment with evidence from sequence conservation and is less prone to mispredicting the direction of effect for regulatory motifs. We show that promoter variants with predicted expression-altering consequence experience strong negative selection in human populations, and produce outlier expression at both RNA and protein levels in thousands of individuals. We observe that clinically relevant genes in rare disease patients are significantly enriched for such variants, and validate their functional impact through MPRA experiments. Our estimates suggest that promoter variation accounts for nearly 6% of the genetic burden associated with rare diseases, underscoring the importance of systematically investigating non-coding promoter regions in clinical genomics.

SIMBA+: DISSECTING GENETIC VARIANT FUNCTION THROUGH SINGLE-CELL MULTIOMICS INTEGRATION

Jayoung Ryu[1], Junxi Feng[1], David Barzideh[1], Elizabeth Dorons[2], Karthik Guruvayurappan[3], Anatori E Prieto[4], Zixuan E Zhang[5], Kushal Dey[3], Steven Gazal[5], Martin J Zhang[4], Luca Pinello[1]

[1]MGH/Harvard Medical School/BROAD, Pathology, Boston, MA, [2]Harvard T.H. Chan School of Public Health, Epidemiology, Boston, MA, [3]MSK, Cancer Center, NY, NY, [4]CMU, Biological Sciences, Pittsburgh, PA, [5]USC, Population and Public Health Sciences, Los Angeles, CA

Genome-wide association studies (GWAS) have identified thousands of disease-associated loci, yet translating these findings into mechanistic insights remains challenging due to difficulties in identifying causal variants, their target genes, and the specific cellular contexts in which they operate. Here we present SIMBA+, a computational framework that leverages single-cell multiomics to characterize variant function by jointly analyzing chromatin accessibility and gene expression data through a novel probabilistic graph based modeling.

SIMBA+ constructs a probabilistic graph integrating GWAS signals with regulatory relationships captured in single-cell multiomics data. Through metapath analysis on this graph, our method identifies: 1) likely causal variants and their target genes, outperforming existing correlation and regression-based approaches, particularly for distal regulatory interactions; 2) specific cellular contexts where variants regulate their targets; and 3) cell-state-specific heritability through a novel factor analysis formulation that decomposes both RNA and ATAC signals.

We validated SIMBA+'s predictions using several orthogonal datasets including tissue and cell type specific eQTLs from GTEx, OneK1K, and CRISPR-based enhancer-gene validations. SIMBA+ successfully nominated validated target genes and their relevant cell states for several phenotypes. Our systematic benchmark demonstrated superior performance in linking variants to their target genes, with particular strength in capturing distal regulatory relationships that are typically missed by distance-based approaches.

Importantly, we show how to formulate single-cell level heritability estimation through multiomics decomposition, applying SIMBA+ to 75 GWAS traits across three single-cell multiome datasets. This unprecedented granular analysis revealed disease-relevant cell states and captured additional heritability beyond pseudobulk approaches. The method provides interpretable decomposition of heritability that connects to gene sets, marker genes, and regulatory elements, enabling mechanistic understanding of genetic associations.

By providing a unified framework for interpreting genetic variants through the lens of single-cell regulatory mechanisms, SIMBA+ advances our ability to translate GWAS findings into biological insights and therapeutic hypotheses.

NEURAL NETWORK MODELS PREDICT PROTEIN LEVELS FROM SEQUENCES ACROSS INDIVIDUALS AND GENES

Eduarda Vaz[1], Alexis Battle[2,3]

[1] Johns Hopkins University, Department of Chemical and Biomolecular Engineering, Baltimore, MD, [2] Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [3] Johns Hopkins University, Department of Computer Science, Baltimore, MD

Machine learning models that predict the molecular consequences of sequence variation across individuals promise to help identify causal disease variants, predict the impact of rare variation, and help characterize regulatory mechanisms. State-of-the-art sequence-based deep learning models capture regulatory motifs and long-range dependencies, enabling context-dependent predictions of gene expression. However, existing models have struggled to predict individual gene expression from personal genomes, indicating they have not fully captured the effects of genetic variation. Here, we explore fine tuning models for predicting protein expression across both individuals and genes, seeking to better capture the impact of sequence variation on gene regulation through improved training schemes and data. We leverage a large cohort of paired whole-genome sequencing (WGS) and proteomics data from the UK Biobank Pharma Proteomics Project (UKB-PPP), encompassing measurements from 43,000 individuals across a panel of approximately 3,000 circulating blood proteins. For each individual and gene, we input DNA sequence centered at the transcription start site (TSS), incorporating all single-nucleotide variants (SNVs) to train single-gene models. Our fine-tuned model outperforms the pre-trained models, generalizing well to unseen individuals, and with Pearson correlation coefficient of up to 0.96 across individuals for the strongest genes. Moreover, the fine-tuned model effectively prioritizes functional variants that induce significant changes in transcription factor binding affinity, outperforming linear models in identifying regulatory variants with potential biological impact. Finally, we explore a multi-gene model that leverages both context-dependent regulatory elements across genomic regions and personal genetic effects. This approach aims to generalize effectively to genetic variation even for previously unseen genes, a significantly more challenging task where existing models have underperformed. We demonstrate that sequence-to-function models can be fine tuned to predict protein levels across individuals and identify genetic variants impacting protein levels. Furthermore, larger cohorts and improved training schemes push the boundaries of sequence-to-function model performance across individuals, bringing us closer to the goals of personal genomics.

LONG-READ TRANSCRIPTOMICS OF DIFFERENTIATING NEURONS IDENTIFIES CELL TYPE SPECIFIC SPLICE ISOFORMS WITH FUNCTIONALLY DISTINCT REGULATORY ELEMENTS AND ENCODED PEPTIDES.

Pieter Spealman[1,2], Yu-Han Hsu[1,2], Greta Pintacuda[1,2], Akanksha Khorgade[3], Asa Shin[3], Houlin Yu[3], Aziz M Al'Khafaji[3], Kasper Lage[1,2,4]

[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]Broad Institute of MIT and Harvard, Cambridge, MA, [4]Copenhagen University Hospital, Institute of Biological Psychiatry, Mental Health Services, Copenhagen, Denmark

Alternative transcript isoforms varying in transcription start, transcription termination, or splicing sites have been shown to vary over the course of human neuron cellular differentiation. These isoforms may gain or lose *cis*-acting regulatory elements or generate alternative protein isoforms. Because of these broad properties, identifying the functional consequences of these isoforms, and the potential role they play in neuropathology, has been historically difficult.

To better understand the functional consequences of alternative isoforms we differentiate iPSC (day 0) into NPC (day 4) and Excitatory Neurons (ExN, day 31) using a protocol that combines NGN2 induction with small molecule patterning. We selected excitatory neurons for their proposed role in schizophrenia. We then performed long-read transcriptome sequencing (MAS-ISO-Seq) on each timepoint. This was highly replicated with a minimum 10 replicates per time point, enabling us to identify 26K alternative isoforms with very high confidence (each isoform required identification in at least 6 replicates). Of these, 6K are the majority isoform in only one cell type with 87 of those being GWAS genes previously identified as being significantly associated with schizophrenia.

Of the high confidence alternative isoforms, we find the most common vary in exons that contain non-coding regions (predominantly the 5'UTR), which may potentially give rise to upstream open reading frames (uORFs) or N-terminal extensions. Using previously published RiboSeq data we identify numerous potential changes in translation; such as a LMNA isoform with a uORF containing exon in NPC and ExN but absent in iPSC.

Finally, we find that highly cell specific isoforms (greater than 80% present in only one cell type) are significantly enriched in schizophrenia associated SNVs identified by GWAS studies, (avg. 2.3-fold higher than random hypergeometric test, avg. p-value = 0.02). Several notable examples are IRF3 (rs2304204) that alters a uORF codon from (Q > L); ZNF823 (rs72986630) that alters a uORF codon (A > C or P); and ENOX1 with an intron variant (rs145463728) located near an exon encoding a very highly translated double ATG start codon uORF.

EFFICIENT COUNT-BASED MODELS IMPROVE POWER AND ROBUSTNESS FOR LARGE-SCALE SINGLE-CELL eQTL MAPPING

Zixuan Zhang[1], Artem Kim[1], Noah Suboc[1], Steven Gazal[1,2,3], Nicholas Mancuso[1,2,3]

[1]Keck School of Medicine, University of Southern California, Dept of Population and Public Health Sciences, Los Angeles, CA, [2]University of Southern California, Dept of Quantitative and Computational Biology, Los Angeles, CA, [3]Keck School of Medicine, University of Southern California, Norris Comprehensive Cancer Center, Los Angeles, CA

Population-scale single-cell transcriptomic technologies (scRNA-seq) enable characterizing variant effects on gene regulation at the cellular level (e.g., single-cell eQTLs; sc-eQTLs). However, existing sc-eQTL mapping approaches are either not designed for analyzing sparse counts in scRNA-seq data or can become intractable in large datasets. Here, we propose jaxQTL, a flexible and efficient sc-eQTL mapping framework using highly efficient count-based models given pseudo-bulk data. Using extensive simulations, we demonstrated that jaxQTL with a negative binomial model outperformed other models in identifying sc-eQTLs, while maintaining a calibrated type I error. We applied jaxQTL across 14 cell types of OneK1K scRNA-seq data (N=982), and identified 11-16% more eGenes compared with existing approaches, primarily driven by jaxQTL ability to identify lowly expressed eGenes. We observed that fine-mapped sc-eQTLs were further from transcription starting site (TSS) than fine-mapped eQTLs identified in all cells (bulk-eQTLs; P=1x10-4) and more enriched in cell-type-specific enhancers (P=3x10-10), suggesting that sc-eQTLs improve our ability to identify distal eQTLs that are missed in bulk tissues. Overall, the genetic effect of fine-mapped sc-eQTLs were largely shared across cell types, with cell-type-specificity increasing with distance to TSS. Lastly, we observed that sc-eQTLs explain more SNP-heritability (h2) than bulk-eQTLs (9.90 ± 0.88% vs. 6.10 ± 0.76% when meta-analyzed across 16 blood and immune-related traits), improving but not closing the missing link between GWAS and eQTLs. As an example, we highlight that sc-eQTLs in T cells (unlike bulk-eQTLs) can successfully nominate IL6ST as a candidate gene for rheumatoid arthritis. Overall, jaxQTL provides an efficient and powerful approach using count-based models to identify missing disease-associated eQTLs.

CRISPRi PERTURBATION SCREENS AND eQTLs PROVIDE COMPLEMENTARY AND DISTINCT INSIGHTS INTO GWAS TARGET GENES

Samuel Ghatan[1], Winona Oliveros[1], Jasper Panten[2], Neville E Sanjana[1], John Morris[3], Tuuli Lappalainen[1,2]

[1]New York Genome Center, New York, NY, [2]Scilifelab & KTH Royal Institute of Technology, Stockholm, Sweden, [3]University of Toronto, Toronto, Canada

Most complex trait-associated variants reside in non-coding regions, and identifying their target genes in cis has been a key challenge for the field. The traditional eQTL mapping approaches is not accompanied by pooled CRISPRi screens with single-cell transcriptomics. In this study, we systematically quantified and characterized the similarities and differences between these methods. We analyzed GWAS loci from 44 blood trait studies from the UK Biobank and Blood Cell Consortium, colocalizing them with eQTLs from 50 blood-related studies and overlapping them with enhancer CRISPRi perturbation data from 12 studies. We obtained 873 CREs that intersected a CRISPRi gRNA target, of which 144 CREs had a CRISPR target gene (cGene; FDR < 0.10) and 511 had a colocalized eGene (PP.H4 > 0.5). In 94 CREs both an eGene and cGene were found, and gene targets intersected at 64 CRE-gene target pairs. Notably, the eGene yield was highest for bulk cell type eQTLs, with single-cell eQTLs providing few hits and overlaps with cGenes due to low power of these data. Altogether, while the eQTL and CRISPRi approaches had significant target gene overlap, the differences are substantial. We next characterized the reasons for this. GWAS eGenes were significantly more likely to be distal to GWAS-CREs than cGenes, with 60% of eGenes and 23% of cGenes being beyond the 2nd closest gene (7.7x10-28); as many as 51% of cGenes were within 10kb of the CRE. Most cGene-linked CREs (83%) but only 39% of eQTLs were associated with a single gene (P = 1.1x10-20), demonstrating the higher sensitivity of eQTL mapping, further dissected by power simulations. A higher proportion of cGenes (29%) exhibited predicted loss-of-function intolerant (pLI) scores > 0.9 compared to eGenes (22%) (P = 0.05), and had a higher number of enhancers (P = 2x10-04) suggesting that CRISPRi analysis is better at identifying constrained and tightly regulated target genes. Finally, we compiled a list of 290 gold-standard blood trait genes in our 873 loci based on rare coding variant and Mendelian associations. While cGenes had a higher proportion of gold standard genes (17% vs. 8%, P = 5.2x10-03), eQTL studies captured more genes overall (51 vs. 22) although also more often finding additional genes in these loci. Taken together, these findings illustrate how eQTL mapping and CRISPRi screens each capture unique dimensions of the regulatory landscape, offering complementary approaches to gene mapping. While eQTL analyses often link GWAS variants to multiple genes, highlighting complex regulatory activity, CRISPRi experiments typically pinpoint direct, proximal targets that, although more strongly supported, may be less novel or surprising as gene targets

MODELING THE EVOLUTION OF GENE REGULATORY COMPLEXITY AND ITS ROLE IN COMPLEX DISEASE

Carl G de Boer[1,2], Madison Chapel[1]

[1]University of British Columbia, Bioinformatics, Vancouver, Canada,
[2]University of British Columbia, Biomedical Engineering, Vancouver, Canada

Each eukaryotic gene is typically under complex regulation, under the control of many transcription factors (TFs), resulting in a highly inter-connected GRN. This high degree of interconnectivity means that many genetic variants across the whole genome can influence complex traits via changes that propagate through the GRN, eventually perturbing core disease genes (ie. omnigenic model). We explore the evolution of complex traits using a combination of modeling and machine learning. Using a deep learning oracle that predicts yeast promoter activity from sequence, we show that regulatory complexity is generally not selected for. Instead, selection appears to act on the expression level, without regard for the regulatory mechanisms that generated it. In contrast to eukaryotic GRNs, prokaryotic GRNs are relatively simple, with few TFs regulating each gene. Since prokaryotes primarily reproduce asexually and eukaryotes primarily sexually, we hypothesized that recombination could explain why eukaryotes evolved more complex GRNs. However, using simulations, we show that recombination does not result in more complex GRNs, but does result in GRNs that are more robust to genetic variation. We next asked how genetic variation combines to impact complex phenotypes. Human genetic studies have shown how variants tend to combine linearly on the log odds ratio of disease. However, due to the non-linear relationship between the log odds ratio and disease prevalence, risk variants have more severe phenotypic consequences in high-risk polygenic backgrounds. Meanwhile, selection has limited power to eliminate risk variants because they have minimal impact in low polygenic risk backgrounds. Increasing polygenicity strengthens selection because more individuals achieve genetic and phenotypic extremes, shrinking effect sizes. Our studies reveal the importance of considering polygenic background when considering variant effects in cell-based studies, and in modeling evolution, and find little evidence that regulatory complexity was directly selected for at the origin of eukaryotes.

NOVEL VARIANT DISCOVERY AND IMPLICATIONS FOR INTERPRETATION FROM LONG-READ SEQUENCING ON 1,027 AFRICAN AMERICANS IN *ALL OF US*

Qiuhui Li

Johns Hopkins University, Department of Computer Science, Baltimore, MD

The *All of Us* Research Program (AoU) is building a comprehensive national biobank comprised of one million participants with matched genomic, trait, and health data. This effort aims to facilitate genotype-to-phenotype associations across various human traits and diseases, representing a major commitment to the future of genomic medicine in the U.S. Short-read genomic sequencing (SRS) data are currently available for the first >400,000 enrolled individuals. However, several studies have shown long-read genome sequencing (LRS) provides far more comprehensive and accurate variant ascertainment than SRS, especially for certain genic regions, structural variants (SVs), and repetitive genomic regions.

Addressing this need, we report on the phase I long-read sequencing of the AoU cohort which spans 1,027 individuals self-identifying as African American or Black. For this, we use a mid-pass (~8x coverage) LRS strategy that substantially increases the structural variant discovery rate compared to SRS with more manageable costs. Using these data, we built a new comprehensive variant resource, including SNVs, indels, SVs, repeat expansions, Cyp2d6 haplotypes, HLA alleles, and more. This resource includes nearly 1M structural variants, including >190,000 previously undetected, many with potential medical implications. Using CADD-SV, we assessed the deleteriousness of identified SVs and found 426 variants classified as likely pathogenic, 68 of which were absent from the AoU SRS and 1000 Genomes project LRS SV callsets. Notably, LRS-based tandem repeat analysis revealed multiple mutations in FMR1 and other clinically relevant regions, including repeat motif changes not observed in prior SRS data. Additionally, we detected repeat recombination events affecting BRCA2, TP53, GSTM1, and ALS2 within the cohort. We then genotyped these variants in 731 samples from the 1000 Genomes Project with matched RNA sequencing data to assess their impact on gene expression. This revealed the identification of 906 new SV-eQTLs including for IL2RB, implicated in immune system disorders, and PIK3R1, a known tumor suppressor. Moreover, we identified thousands of AoU SVs in strong linkage disequilibrium with significant variants from NHGRI-EBI GWAS catalog, indicating their associations with diverse traits and diseases, including gastrointestinal disorders (e.g., gastric cancer and peptic ulcers) and cardiovascular conditions (e.g., abdominal aortic aneurysm, coronary artery disease and cardiovascular disease). This LRS dataset, along with the phase II long-read dataset of >10,000 samples available later this year, serves as a valuable resource for deciphering the complex genetic architecture of the human genome and advancing the development of precision medicine.

SINGLE-CELL eQTL MAPPING OF IMMUNE RESPONSE REGULATION IN SYSTEMIC LUPUS ERYTHEMATOSUS PATIENTS

Haerin Jang[1], Catherine Sutherland[1], Niek de Klein[1,2], Tarran Rupall[1,2], Bess Chau[1,2], Wanseon Lee[1], Norzawani Buang[3], Magdalena West[4], Emily Holzinger[5], Sarah Middleton[6], Virginia Savova[7], Matthew C Pickering[3], Marina Botto[3], Carla Jones[1], Timothy Vyse[4], James Peters[3], Gosia Trynka[1,2], Emma Davenport[1]

[1]Wellcome Sanger Institute, Human Genetics, Hinxton, United Kingdom, [2]Open Targets, Hinxton, United Kingdom, [3]Imperial College London, London, United Kingdom, [4]King's College London, London, United Kingdom, [5]Bristol Myers Squibb, Cambridge, MA, [6]GSK, Genomic Sciences, Collegeville, PA, [7]Sanofi, Cambridge, MA

Background: Systemic lupus erythematosus (SLE) is a complex, multisystem autoimmune disease driven by cell-type-specific regulatory networks. Single-cell expression quantitative trait locus (sc-eQTL) mapping provides a powerful approach to uncover how genetic variation influences immune cell function and transcriptional responses in SLE.
Methods: SLE patients with African or Afro-Caribbean, European, and South-Asian ancestry were recruited from two sites in London. We generated whole genome sequencing data and single cell gene expression (RNA-seq), surface protein (CITE-seq) and immune receptor (VDJ-seq) data from peripheral blood mononuclear cells. We mapped sc-eQTLs using tensorQTL with mean pseudobulked gene expression.
Results: We sequenced 663,433 single cells from 287 SLE patients and 11 healthy controls. sc-eQTLs were mapped in five immune cell types, identifying 4,090, 2,806, 1,044, 1,407, and 745 eGenes in CD4+ T cells, CD8+ T cells, B cells, NK cells, and monocytes, respectively. The number of eGenes detected significantly correlated with the mean number of cells per donor for each cell type (Pearson p-val= 0.0014). Among eGenes detected, 2,626 genes (50%) were found in more than one cell type and included eGenes that are already known to colocalize with SLE-associated loci, such as *ORMDL3* in B cells (q-val=9.4e-14, beta=-0.73) and CD8+ T cells (q-val=5.1e-26, beta=-0.76). A total of 2,647 eGenes were unique to a single cell type, many of which have cell-type-specific immune functions, such as *FCRL5* in B cells (q-val= 6.7e-13, beta=0.69) and *C3AR1* in monocytes (q-val=7.8e-36, beta=0.84). We intend to colocalize our results with SLE GWAS loci to gain additional insight into how regulation of gene expression varies in specific cell types in SLE. In parallel, we are assessing how to increase power for eQTL detection by modelling discreet single-cell read counts with saigeQTL.
Conclusions: We generated a multi-omics single-cell dataset from a diverse cohort of patients to map sc-eQTLs and uncover the genetic regulation of cell-type-specific gene expression and regulatory networks underlying SLE.

VIRAL DNA LOAD IS POLYGENIC AND CONFERS LYMPHOMA RISK

Nolan Kamitaki[1,2,3], Steven A McCarroll[2,3,4], Po-Ru Loh[1,2]

[1]Brigham and Women's Hospital, Division of Genetics, Department of Medicine, Boston, MA, [2]Broad Institute of MIT and Harvard, Program in Medical and Population Genetics, Cambridge, MA, [3]Harvard Medical School, Department of Genetics, Boston, MA, [4]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA

Many viruses have adapted to persist in infected humans for life. The amount of viral genetic material present, or viral load, can vary significantly between infected individuals and is heritable. However, little is known about the specific human genetic variants that influence viral load, or about the causal relationship between elevated viral load and diseases. Here, we analyzed viral load of 31 DNA viruses in human blood whole-genome sequencing (WGS) data (n=490,401 UK Biobank participants) and saliva WGS data (n=12,519 SPARK participants). Viral DNA was infrequently observed in blood WGS data: although >90% of adults are infected with Epstein-Barr virus (EBV), only 15% of blood samples contained EBV DNA detectable from WGS, suggesting higher viral load in these samples. Genome-wide association analysis of viral DNA load identified human genetic influences on load of six viruses (EBV, human herpesvirus (HHV) 7, HHV-6B, and three anelloviruses). For all six viruses, the strongest host genetic effects on viral load localized to the human leukocyte antigen (HLA) complex, similar to previous results on HIV and hepatitis B. Each virus appeared to be influenced by different *HLA* alleles: EBV viral DNA load associated most strongly with the class II *HLA-DRB1*04:04* allele (p=1.7x10^{-560}), whereas HHV-7 load associated with class I residues such as *HLA-B* R180D (p=1.2x10^{-538}). Beyond HLA, genetic variation at 39 other loci throughout the genome associated with EBV DNA load. These included genes encoding proteins that process peptides for antigen presentation, such as *ERAP1*/*ERAP2* (p=6.8x10^{-63}) and *CPVL* (p=6.0x10^{-10}), which exhibited an allelic series including missense and loss-of-function variants.

These polygenic influences on EBV DNA load provided dozens of genetic instruments that enabled assessing the causality of associations with diseases. Mendelian randomization (MR) indicated that EBV DNA load is unlikely to causally affect risk for autoimmune conditions, despite being elevated with systemic lupus erythematosus (p=2.2x10^{-10}), rheumatoid arthritis (p=2.1x10^{-31}), and multiple sclerosis (p=0.018). In contrast, MR found a strong causal effect of detectable EBV DNA on risk of Hodgkin lymphoma (OR=4.30, [2.11-6.94], p=1.2x10^{-3}) and a weaker effect on non-Hodgkin lymphoma (OR=2.31, [0.98-3.93], p=2.8x10^{-3}). This suggests that higher chronic EBV viral load increases lymphoma risk, whereas associations of EBV load with autoimmune conditions may reflect immune response.

WIDESPREAD EFFECTS OF MEDICATIONS ON GUT MICROBIOME COMPOSITION AND FUNCTION

Ashwin Chetty[1], Ramanujam Ramaswamy[2], Nicholas Dylla[2], Huaiying Lin[2], Matthew Odenwald[3], Eric Pamer[2,4], Ran Blekhman[5]

[1]University of Chicago, Committee on Genetics, Genomics and Systems Biology, Chicago, IL, [2]University of Chicago, Duchossois Family Institute, Chicago, IL, [3]University of Chicago, Department of Medicine, Section of Gastroenterology, Hepatology, and Nutrition, Chicago, IL, [4]University of Chicago, Department of Medicine, Section of Infectious Diseases and Global Health, Chicago, IL, [5]University of Chicago, Section of Genetic Medicine, Department of Medicine, Chicago, IL

Medications exert strong effects on the gut microbiome, ultimately impacting host physiology and affecting health outcomes. Previous work has characterized the effects of specific medication classes on microbial taxa in vitro. However, there is little research examining the effects of medications on the microbiome in vivo, especially in clinical settings and across multiple time points. In this study, we leverage longitudinal stool metagenomics, metabolomics, and electronic health records from 3,469 samples representing 1,122 patients at the University of Chicago Medical Center between 2020 and 2023. We employ a model that compares patient microbiome composition before and after starting a medication while controlling for the effects of medical procedures, diagnoses, demographic variables, and other drugs taken simultaneously. In total, we identify 36,637 associations relating 138 of the most commonly-given medications, such as oxycodone and pantoprazole, with 106 microbial genera, 204 microbial species, 627 microbial pathways, and 124 gut metabolites. We find, for example, that oral prednisone is associated with decreases in Parabacteroides and Enterobacter abundance and increases in microbial ergosterol and long-chain fatty acid biosynthesis pathways. Furthermore, the selective serotonin reuptake inhibitor sertraline is associated with increases in Alistipes abundance, increases in microbial pathways related to dopamine degradation, and increases in fecal concentration of the related metabolite kynurenic acid. Considering interaction effects between medications, microbial genera, and lab tests, we identify 27,321 significant interactions, including evidence that Enterococcus interacts with heparin in vivo. Overall, our results provide evidence of medication-microbiome interactions spanning a variety of medication classes and identify potential microbial taxa and pathways that may mediate these interactions.

NEW FRONTIERS IN PROTEOMICS: TECHNOLOGIES TO ILLUMINATE PROTEOFORMS IN COMPLEX TRAITS AND DISEASE

Gloria Sheynkman

University of Virginia, Molecular Physiology and Biological Physics, Charlottesville, VA

Genome-wide association studies (GWAS) and quantitative trait loci (QTL) analyses (sQTLs, phosphoQTLs, apaQTLs, etc.) have pinpointed numerous genomic variants that modulate complex traits and disease. Yet many of these variants lie in noncoding regions, making unclear their impact on protein products. Increasing evidence indicates that the ultimate phenotypic effect often depends on distinct "proteoforms"—molecularly diverse protein forms arising from a single gene via mechanisms such as alternative splicing and post-translational modification. However, there remains a critical gap: precisely how different variations co-occur in the final protein product, and how they contribute to pathogenesis, is poorly understood.

Recent technological convergences in transcriptomics and proteomics offer new pathways to address these questions. Long-read RNA sequencing can define full-length transcript isoforms, including novel splice events, thereby refining the protein reference for proteomic searches. Mass spectrometry-based proteogenomics then leverages these sample-matched isoform sequences to detect low-abundance or cryptic peptides that otherwise elude discovery. Beyond these approaches, an exciting frontier in "next-generation proteomics" has emerged—single-molecule peptide sequencing platforms. These new technologies augment traditional MS by providing direct, residue-by-residue readouts that can resolve subtle variation (e.g., single amino acid substitutions, phosphorylations) in proteoforms.

In this presentation, I will highlight emerging and cutting-edge technologies and workflows that bridge the gap between genetic variation and proteomic complexity. We will explore how they illuminate proteoforms in large-scale population studies and disease contexts, and discuss the future of isoform-level proteomic analyses for understanding gene function, informing biomarker discovery, and ultimately advancing precision medicine.

COMPREHENSIVE DIPLOID CHROMATIN MAP OF A SINGLE CELL USING DEAMINASE-ASSISTED FIBER-SEQ (DAF-SEQ)

Elliott G Swanson[1], Yizi Mao[2], Benjamin J Mallory[1], Mitchell R Vollger[2], Jane Ranchalis[2], Stephanie C Bohaczuk[2], Nancy L Parmalee[3], James T Bennett[3,4], Andrew B Stergachis[1,2,5]

[1]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [2]University of Washington School of Medicine, Division of Medical Genetics, Seattle, WA, [3]Seattle Children's Research Institute, Center for Developmental Biology and Regenerative Medicine, Seattle, WA, [4]University of Washington, Department of Pediatrics, Seattle, WA, [5]Brotman Baty Institute for Precision Medicine, Mutational Scanning, Seattle, WA

Gene regulation is orchestrated within single cells by millions of proteins co-occupying individual chromosome-length chromatin fibers. Although numerous methods exist for resolving protein occupancy along individual chromatin fibers or within single cells, our current understanding of the chromatin architecture of a single cell is fragmented and sparse. Specifically, existing maps of the chromatin epigenome of single cells resolve at most 0.1% of the entire genome of each cell, limiting our understanding of how the chromatin epigenome of a single cell is truly regulated. To overcome these limitations, we developed Deaminase-Assisted single-molecule chromatin Fiber sequencing (DAF-seq), which leverages a non-specific double-stranded DNA cytidine deaminase to efficiently stencil protein occupancy along DNA molecules via selective deamination of accessible cytidines, which are preserved via C-to-T transitions upon DNA amplification. We demonstrate that DAF-seq permits single-molecule footprinting at near single-nucleotide resolution, enabling the precise delineation of proteins that cooperatively occupy chromatin fibers. Furthermore, DAF-seq enables the synchronous identification of single-molecule chromatin and genetic architectures – resolving the functional impact of somatic variants, as well as transitional chromatin epi-states. Finally, we demonstrate that combining DAF-seq with single-cell Primary Template-directed Amplification (PTA) whole-genome amplification enables the accurate reconstruction of the diploid genome and epigenome from a single cell, observing chromosome-scale protein co-occupancy across 96% of a single cell's mappable genome, with at least 74% of the genome and chromatin epigenome from a single cell accurately haplotype phased. Furthermore, the unique deamination profiles of each read enable us to assemble 5,144 ultra-long (>100 kb in length) consensus reads from a single cell, with a single-cell N50 of 33.2 kb. Using single-cell DAF-seq data across 12 individual cells, we demonstrate that a cell's accessible regulatory landscape can diverge by as much as 63% while still retaining the cell's identity. Overall, DAF-seq enables the comprehensive characterization of protein occupancy across entire chromosomes with single-nucleotide, single-molecule, single-haplotype, and single-cell precision.

A NOVEL FRAMEWORK FOR MULTIPLEX MEASUREMENTS OF THE ABUNDANCE AND INTERACTION OF PROTEINS

Tianyao Xu[1], Jingyao Wang[1], Yoonju Shin[1], Yuwei Cao[1], Lingzhi Zhang[1], Eduardo Modolo[1], Tamar Dishon[1], Eric Mendenhall[2], Sven Heinz[3], Christopher Benner[3], Alon Goren[1]

[1]UC San Diego, Department of Medicine, Division of Genomics & Precision Medicine, La Jolla, CA, [2]HudsonAlpha, Institute for Biotechnology, Huntsville, AL, [3]UC San Diego, Department of Medicine, Division of Endocrinology & Metabolism, La Jolla, CA

Protein-protein interactions (PPIs) and protein abundance are integral for the majority of cellular processes such as chromatin structure and function, signal transduction and growth regulation. Yet, current methods to study PPIs cannot fully capture this complexity in an unbiased manner. Thus, most studies only focus on a handful of proteins and do not address the multi-factorial manner in which complexes affect the physiology of normal or disease conditions.

To overcome this challenge, we developed Prod-seq: high-throughput identification and characterization of multiple PPIs along with quantifying the proteins surveyed. Prod-seq uses antibodies conjugated to barcoded oligonucleotides (Ab-oligos) with UMIs. Prod-seq builds on "DNA-caliper", a DNA-oligo with two 3' end arms used for hybridization to proximal Ab-oligos. Following the hybridization, both DNA-caliper arm strands are extended to covalently capture proximal Ab-oligo barcodes and UMIs on a single molecule. The extended DNA-calipers are then converted into an Illumina library and sequencing allows identifying and counting (UMIs) proximal proteins and quantifying the proteins targeted.

We established Prod-seq using an array of recombinant complexes to rigorously benchmark our method both for protein abundance quantifications and PPI detection. The recombinant proteins allowed us to improve the signal-to-noise ratio and perform optimizations.

Here, we focused on the polycomb group (PcG) complex and the diffuse intrinsic pontine glioma (DIPG) H3K27M mutation, and compared PPIs and protein abundances for a set of PcG members and relevant histone modifications in multiple cellular systems. These include a cell line with a knockout of EZH2, the H3K27me3 writer component of the PcG complex (HEK293T EZH2-KO and EZH2-WT cells) as well as human induced pluripotent stem cells (hiPSCs) expressing H3K27M or treated with the EZH2-specific inhibitor Tazemetostat. We demonstrate the ability of our tool to reproducibly quantify protein abundances and detect and quantify PPIs and changes in the stoichiometry of PcG complex component members and the target histone modifications when EZH2 is perturbed. Together, we present Prod-seq and apply it to study changes in the PcG complex and the H3K27M mutation. Our tools are built for easy dissemination, including features such as low cost and minimal operational requirements.

POISEN: A BIOINFORMATICS PIPELINE TO IDENTIFY POISON EXONS IN LONG-READ TRANSCRIPTOMES

Mia S Broad, Jung Hong, Kay-Marie Lamar, Jeffrey D Calhoun, Gemma L Carvill

Northwestern University, Neurology, Chicago, IL

Alternative poison exon (PE) splicing is a regulatory mechanism that tightly controls protein abundance with cell-type specificity. When included in an mRNA transcript, PEs introduce a premature termination codon, triggering nonsense-mediated mRNA decay (NMD). Previously, we identified rare pathogenic variants near PE splice sites and adjacent intronic regions in *SCN1A* in patients with Dravet syndrome. These variants cause aberrant PE splicing in iPSC-derived neurons, leading to low protein abundance consistent with *SCN1A* haploinsufficiency. Pathogenic variants near PE splice sites in *SYNGAP1* and *FLNA* have also been linked to neurodevelopmental disorders (NDDs), including developmental epileptic encephalopathy, autism spectrum disorder, and periventricular nodular heterotopia. Despite their significance, PEs remain understudied due to several challenges. PE-containing isoforms are rapidly degraded by NMD, precluding the study of their natural biology and pathogenic disruptions. Likewise, short-read RNA sequencing cannot precisely determine splice junction order due to read length limitations, hindering computational resolution of PE locations. To address this, we developed POISEN (**P**oison ex**O**n d**IS**covery for long-r**E**ad tra**N**scriptomes), a bioinformatics pipeline to identify PEs in long-read transcriptomes. POISEN integrates existing long-read annotation tools with customized approaches to locate PEs in NMD-predicted mRNA transcripts. We sequenced day 20 and day 60 cerebral organoids (COs) using PacBio Iso-Seq to investigate PEs in a model of neurodevelopment. Our findings demonstrate that POISEN annotated 201,079 putative PE-containing isoforms in the CO transcriptomes, with ~37% of the PE-containing genes representing novel discoveries. We validated a subset of these PEs by inhibiting NMD in HAP1 and HepG2 primary cell lines, followed by Nanopore ONT long-read sequencing. This analysis revealed a significant increase in the PSI of known PEs compared to control ($p.adj \leq 0.001$). Additionally, these PEs were significantly enriched in NDD-associated genes (n=1,203; p-value < 2.2e-16), with a ~2.3-fold enrichment. We will curate the PEs identified by POISEN in an online repository and Shiny app for the scientific community to facilitate further research into the roles of PEs in neurodevelopment and Mendelian disorders. Additionally, antisense oligonucleotides have been shown to effectively target PE splicing to rescue haploinsufficiency, with clinical trials underway for *SCN1A*-related epilepsy. Therefore, our PE repository will serve as a promising list of potential therapeutic targets for treating NDDs and present new options for patient care.

ADVANCING THE FIDELITY AND SPEED OF SOMATIC MUTATION DETECTION

Gilad Evrony

NYU Grossman School of Medicine, Center for Human Genetics and Genomics, New York, NY

Most somatic mutations are present at low levels in tissues, necessitating higher fidelity than standard DNA sequencing can achieve. Additionally, the speed at which current methods can detect and quantify somatic mutations precludes their use in clinical settings that require real-time decision making, such as cancer surgeries. Overcoming these two key limitations in profiling somatic mutations--fidelity and speed--can help advance our understanding of somatic mutations and can enable new clinical advances. Here, we present two approaches to addresss each of these limitations in new ways. First, we present a single-molecule sequencing approach that achieves single-molecule fidelity for both mutations as well as single-strand DNA mismatches and damage. This reveals single-strand precursors of mutations in cancer-predisposition syndromes, in sperm, and in the mitochondrial genome, as well as DNA damage stemming from APOBEC and standard handling of DNA in the laboratory . The second technology we introduce, ultra-rapid droplet digital PCR, achieves highly sensitive and specific quantification of cancer hotspot mutations from tissue to result in only 15 minutes, which we have implemented intraoperatively in brain tumor surgeries. These two approaches improve on the state-of-the-art in terms of fidelity and speed of profiling somatic mutations, opening new avenues for studying somatic mutations and utilizing them clinically.

WIDESPREAD VARIATION IN MOLECULAR INTERACTIONS AND REGULATORY PROPERTIES AMONG TRANSCRIPTION FACTOR ISOFORMS

Luke Lambourne*[1,2,3], Kaia Mattioli*,***[4], Clarissa Santoso*[5,6], Gloria Sheynkman**[1,2,3], Sachi Inukai**[4], Babita Kaundal**[7], Anna Berenson[5], Kerstin Spirohn-Fitzgerald[1,2,3], Tong Hao[1,2,3], Adam Frankish[8], Josh A Riback[9], Nathan Salomonis[10,11], Michael A Calderwood[1,2,3], David E Hill[1,2,3], Nidhi Sahni***[7], Marc Vidal***[1,2,3], Martha L Bulyk***[1,4,12], Juan I Fuxman Bass***[1,5,6]

[1]Dana Farber Cancer Institute, Center for Cancer Systems Biology (CCSB), Boston, MA, [2]Blavatnik Institute, Harvard Medical School, Department of Genetics, Boston, MA, [3]Dana Farber Cancer Institute, Department of Cancer Biology, Boston, MA, [4]Brigham and Women's Hospital and Harvard Medical School, Division of Genetics, Department of Medicine, Boston, MA, [5]Boston University, Department of Biology, Boston, MA, [6]Boston University, Bioinformatics Program, Boston, MA, [7]The University of Texas MD Anderson Cancer Center, Department of Epigenetics and Molecular Carcinogenesis, Houston, TX, [8]European Bioinformatics Institute, Wellcome Genome Campus, European Molecular Biology Laboratory, Hinxton, Cambridge, United Kingdom, [9]Baylor College of Medicine, Department of Molecular and Cellular Biology, Houston, TX, [10]University of Cincinnati College of Medicine, Department of Pediatrics, Cincinnati, OH, [11]Cincinnati Children's Hospital Medical Center, Division of Biomedical Informatics, Cincinnati, OH, [12]Brigham and Women's Hospital and Harvard Medical School, Department of Pathology, Boston, MA

Most human transcription factors (TFs) genes encode multiple protein isoforms differing in DNA binding domains, effector domains or other protein regions. The global extent to which this results in functional differences between isoforms remains unknown. Here, we systematically compared 693 isoforms of 246 TF genes, assessing DNA binding, protein binding, transcriptional activation, subcellular localization, and condensate formation. Relative to reference isoforms, two-thirds of alternative TF isoforms exhibit differences in one or more molecular activities, which often could not be predicted from sequence. We observed two primary categories of alternative TF isoforms: “rewirers” and “negative regulators”, both of which were associated with differentiation and cancer. Our results support a model wherein the relative expression levels of, and interactions involving, TF isoforms add an understudied layer of complexity to gene regulatory networks, demonstrating the importance of isoform-aware characterization of TF functions and providing a rich resource for further studies.

*These authors contributed equally
**These authors contributed equally
***Corresponding authors

NEW ASSEMBLY-BASED METHODS FOR DETECTING LARGE COMPLEX STRUCTURAL REARRANGEMENTS IN HUMAN GENOMES

Peter A Audano, Christine R Beck

The Jackson Laboratory, Genomic Medicine, Farmington, CT

Structural variation is a major contributor to human diversity, adaptation, and disease. Simple structural variant (SV) types include deletions, insertions, duplications, inversions, and translocations, and SVs account for most of the variable bases between genomes. Complex structural variants (CSVs) that consist of one or more simple events *in cis* appear more frequently in diseases and cancers where DNA repair, apoptosis, and cell cycle checkpoints are compromised, although CSVs also comprise germline variation in healthy individuals contributing to human diversity and evolution. CSVs are often characterized by homologous sequences at breakpoints, and while large repeats mediate large CSVs, smaller tracts of homology mediate CSVs in non-repetitive loci. Long-read assemblies have increased the size of detectable SVs and have expanded variant detection into more complex regions of the genome, and yet methods for identifying CSVs from assemblies are limited. Here, we have developed a new assembly-based approach for detecting rearrangements by tracing through reference and assembly sequences in order to increase sensitivity in large segmental duplications, expand the size of detectable CSVs, and refine CSV breakpoints. We have integrated this method into our publicly available variant caller, PAV. When we compare 130 phased haplotypes from 65 individuals assembled by the Human Genome Structural Variation Consortium (HGSVC) to the T2T-CHM13v2.0 reference sequence, we identify 148 CSVs per genome. Furthermore, of the 2,410 distinct CSVs identified across these genomes, we find 117 unique complex structures. CSVs in highly repetitive regions can now be detected, including several distinct complex events in repetitive NBPF genes that were previously missed by short-read, long-read, assembly-based, and optical mapping approaches. Compared to SVision-pro, we find between 16 and 57 CSV per haplotype called by both tools, 36 to 83 CSVs per haplotype are called by only by PAV (2.3 to 68 kbp median size), and 724 to 1,587 CSVs per haplotype are called only by SVision-pro (538 to 706 bp median size). This highlights that large CSVs within genomically complex regions are now regularly detected by PAV, including the NBPF CSVs which are missed by other callers including SVision-pro. We applied PAV to six near-T2T nonhuman primate (NHP) from the Primate T2T Consortium. Using the chimpanzee genome, we were able to determine the ancestral state for 1,496 of 2,410 unique human CSVs (62%). Of these, 56 CSVs were also found in the chimpanzee genome with 23% and 60% allele frequencies in human and nonhuman primates suggesting that these are likely common CSVs that became reference alleles by chance. With thousands of phased assemblies now in production, the CSV CSV method now embedded in PAV represents a critical step toward understanding the impacts of complex rearrangements on the genome, gene expression, evolution, and disease.

MIC-DROP-SEQ: SCALABLE GENETIC SCREENING OF VERTEBRATE DEVELOPMENT WITH CELLULAR RESOLUTION

Clayton M Carey[1], Saba Parvez[2,3], Zachary J Brandt[2], Randall T Peterson[2], James A Gagnon[1]

[1]University of Utah, School of Biological Sciences, Salt Lake City, UT, [2]University of Utah, Department of Pharmacology and Toxicology, Salt Lake City, UT, [3]Northwestern University, Department of Cell & Developmental Biology, Chicago, IL

Understanding how genetic variation shapes animal traits and disease requires a comprehensive mapping of the molecular links between genotype and phenotype. Comparative cellular atlases generated using single-cell RNA sequencing (scRNA-seq) enable the mapping of these connections with high resolution, but scaling this approach in large genetic screens with whole animals remains challenging. To address this, we developed a method combining Multiplexed Intermixed CRISPR Droplets with scRNA-seq (MIC-Drop-seq), which couples high-throughput CRISPR gene disruption in zebrafish embryos with molecular phenotyping via multiplexed scRNA-seq. Genotype is determined at the cellular level through direct gRNA capture and sequencing, allowing for the demultiplexing of mutants from large pools of heterogeneous cells. This approach enables phenotypic comparisons across dozens of mutants in parallel, within each embryonic cell type. In a single MIC-Drop-seq experiment, we intermixed 50 droplet types, each targeting a transcriptional regulator active in early development, and injected them randomly into 1,000 zebrafish embryos. Pooled scRNA-seq was then performed at 24 hours post-fertilization. Tissue-specific gene expression and cell abundance analysis of demultiplexed mutant cells recapitulated known mutant phenotypes while also revealing novel roles for several targeted transcription factors in brain and mesoderm development, which were validated using orthogonal methods. Cross-referencing phenotypic changes with target gene expression patterns uncovered pervasive yet factor-specific cell-extrinsic effects, demonstrating the method's ability to detect unanticipated developmental phenotypes. Thus, MIC-Drop-seq provides a powerful and scalable platform for in vivo mapping of genotype-phenotype connections at single-cell resolution.

EXPRESSION AND ISOLATION OF THE MAP3 PHEROMONE RECEPTOR FROM YEAST FOR STRUCTURAL AND GENETIC STUDIES

Bethlehem D Abebe, Steven Z Chou

University of Connecticut Health Center, Molecular Biology & Biophysics, Farmington, CT

Cell signaling pathways mediated by pheromone receptors play a pivotal role in cellular communication and reproductive processes across various organisms. In fission yeast, *Schizosaccharomyces pombe*, there are two pheromones, P-factor and M-factor which bind to the Mam2 and Map3 pheromone receptor respectively, regulating mating processes. These receptor's functions may extend to broader implications in cellular physiology. Fungi, including pathogenic species, utilize pheromones for communication and mating. *Candida albicans* utilizes pheromone receptors, Ste2 (a-factor) and Ste3 (α-factor), in a similar mechanism to S. cerevisiae to initiate meiosis. A deeper understanding of the mating pheromone receptors in fission yeast will shed light on the mechanisms and regulation of key processes in mating, meiosis, and pathogenesis in fungal pathogens such as *C. albicans*. Despite its significance, the intricate molecular details of the Mam2 and Map3 pheromone receptors have yet to be uncovered. This research aims to bridge this gap by elucidating the structural organization, ligand-binding specificity, and downstream signaling events of the Map3 pheromone receptor in *S. pombe*. Using a PCR-based gene targeting technique in fission yeast we express our gene of interest in fission yeast and overexpress in insect cell lines fused with a Green Fluorescent Protein (GFP). The isolated protein will be used to determine the structure of the Map3 signaling complex using single-particle cryo-electron microscopy (cryo-EM). Structural analysis of isolated Map3 protein will provide insights into active sites and binding pockets for potential antifungal drug targeting. This structural understanding can then inform drug design and development efforts to combat diseases caused by fungal pathogens. Complementing these methods, yeast genetic approaches, such as mutagenesis, will be employed to investigate changes in structural, functional, and localization of the pheromone receptor Map3 and their implications on signaling pathways. Ultimately, this knowledge holds great promise for impacting human health and disease through advancements in drug discovery and medicine.

TFXCAN REVEALS TRANSCRIPTIONAL PROGRAMS DRIVING COMPLEX TRAITS AND DISEASES.

Temidayo Adeluwa[1], Sarah Sumner[2], Saideep Gona[1], Festus Nyasimi[3], Sofia Salazar[2], Sylvan Baca[4], Matthew Freedman[4], Alexander Gusev[4], Boxiang Liu[5], Ravi Madduri[6], Guimin Gao[7], Tiffany Amariuta[8], Hae Kyung Im[2]

[1]The University of Chicago, Genetics, Genomics, and Systems Biology, Chicago, IL, [2]The University of Chicago, Department of Medicine, Section of Genetic Medicine, Chicago, IL, [3]The University of Chicago, Department of Human Genetics, Chicago, IL, [4]Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, [5]National University of Singapore, Department of Pharmacy and Pharmaceutical Sciences, Faculty of Science, Singapore, Singapore, [6]Argonne National Laboratory, Data Science and Learning Division, Lemont, IL, [7]The University of Chicago, Department of Public Health Sciences, Biostatistics Laboratory, Chicago, IL, [8]University of California San Diego, Halıcıoğlu Data Science Institute, La Jolla, CA

Genome-wide association studies (GWAS) have identified thousands of trait-associated loci but have failed to assign potential functional mechanisms mediated by these loci. One possibility is that a single nucleotide polymorphism (SNP) disrupts transcription factor (TF) binding, leading to changes in a molecular process or phenotype. Presently, there are no methods to identify TFs whose genetically disrupted binding may contribute to the development of a phenotype.

For the first time, we introduce a framework, TFXcan, to identify TFs involved in a phenotype analogous to transcriptome-wide association studies (TWAS). TFXcan nominates TF/tissue pairs by testing the correlation of the genetically predicted component of TF/tissue binding at a GWAS locus with the phenotype. We applied TFXcan to prostate cancer (PrCa), Type 2 Diabetes (T2D) and metabolic traits by testing the association of 692 TF/tissue pairs (202 TFs and 54 tissues).

Our results suggest that the binding activity of multiple TFs can be associated with a phenotype, consistent with models of collaborative binding. Further analysis of the results reveal transcriptional programs (combinations of TF/tissue pairs), suggesting biological pathways or subtypes of the diseases. For example, TFXcan applied to PrCa reveals a program driven primarily by the androgen receptor (AR) and Forkhead Box A1 (FOXA1) TFs, and a distinct program driven by ERG, a TF that mediates a fusion event with *TMPRSS2* in large proportions of PrCa cases. TFXcan applied to T2D identifies a liver-specific program and various programs driven by TFs involved in beta cell development including SMAD2 and SMAD3.

TFXcan is the first framework capable of the systematic identification of TFs and TF complexes driving diseases. It is user-friendly, requiring only GWAS summary statistics and publicly available reference data (e.g., reference genome and reference population genotype data) as input. Importantly, TFXcan allows us to investigate genetically mediated TF binding as GWAS loci functional mechanisms, and dissect the regulatory programs that drive diseases.

BIOBANK-SCALE MULTI-OMIC MODELLING OF CIRCADIAN DISRUPTION

Clara Albinana[1,2], Naomi Wray[1,3]

[1]Big Data Institute, University of Oxford, Oxford, United Kingdom, [2]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark, [3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

Background: Several common diseases have been associated with imbalances of the circadian rhythm, a 24h-cycle molecular system that maintains body homeostasis. The state-of-the-art methods to infer circadian disruption require longitudinal data in time-consuming and expensive studies, limiting the sample size and therefore the extrapolation of results to the general population. In this study, we introduce a method to infer circadian disruption from large multi-modal Biobank data and explore its association to disease risk.

Material and Methods: Using the time of blood extraction as a proxy for time of day, we leverage 4 omics datasets from the UK Biobank (proteomic, metabolomic, biochemistry and haematological assays), with up to 4,000 variables. We trained LASSO models to predict time of day as "omics time". We define their difference as a measure of individual circadian disruption and explore its effect in the risk of common diseases in the UK Biobank.

Results: Single-variant associations reveal a widespread effect of blood sample timing in the variation of the studied blood biomarkers, with time of day explaining up to 10% of the variation of the studied biomarkers. LASSO models showed up to 70% R2 in time of day, with the proteomics subset dominating the signal. From the 20 most common diseases in the cohort, our measurement of circadian disruption is positively associated with depressive episodes and type-2 diabetes.

Conclusion: In this study, we showcase the widespread effect of blood sample timing in blood biomarker variation and use it as an opportunity to train circadian disruption models, validating hypotheses on the risk of metabolic and psychiatric traits.

INTEGRATED ANALYSIS OF LIVER CELL-TYPE QTL WITH BULK TISSUE REVEALS MECHANISMS OF COMPLEX TRAITS

Abdalla A Alkhawaja*[1], Kevin W Currin*[1], Hannah J Perrin[1], Swarooparani Vadlamudi[1], Amy S Etheridge[1,2], Gabrielle H Cannon[3], Carlton W Anderson[3], Anne H Moxley[1], Erin G Schuetz[4], Federico Innocenti[2], Terrence S Furey[1,5], Karen L Mohlke[1]

[1]Department of Genetics, University of North Carolina, Chapel Hill, NC, [2]Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, [3]Advanced Analytics Core, University of North Carolina, Chapel Hill, NC, [4]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN, [5]Department of Biology, University of North Carolina, Chapel Hill, NC

*Authors contributed equally

The variants, genes, and cell types responsible for disease risk at GWAS signals remain largely unknown. While bulk tissue studies have used variant associations with gene expression and chromatin accessibility (e/caQTL) to implicate genes and regulatory variants, they lack cell-type resolution and may miss effects in less abundant cell populations. We jointly profiled single-nucleus RNA and ATAC-seq on 40 human livers and identified 68,393 nuclei across 7 major cell types, predominantly hepatocytes (68%). We identified 306,706 chromatin accessible regions (peaks), including 17,147 marker peaks enriched for cell-type-specific processes such as angiogenesis in endothelial cells. Comparison to bulk liver ATAC-seq peaks revealed 70,884 peaks only detected in cell types, 73% of which were not detected in hepatocytes. Enrichment of GWAS variants in cell-type peaks linked liver enzyme traits to hepatocytes, blood pressure to mesenchymal cells, and urea to cholangiocytes. We then identified 67 eQTLs and 1,885 caQTLs, including 12 eQTLs and 5 caQTLs not detected in hepatocytes. Colocalization of cell-type caQTLs with complex trait GWAS identified 205 hepatocyte caQTLs shared with 363 GWAS signals, primarily for liver enzymes and cholesterol traits. We next integrated GWAS, eQTL, caQTL, and cell-type data. For instance, at a GWAS signal for blood cell traits and a bulk eQTL for *ITGAD*, we observed an eQTL in Kupffer cells for which the lead is within a peak detected only in Kupffer cells. To further elucidate cell-type effects, we predicted cell types for 35,361 bulk liver caQTLs and mapped 3,112 to a single non-hepatocyte cell type. For example, a bulk caQTL colocalized with a GWAS signal for HDL-cholesterol only overlapped a peak in mesenchymal cells that is proximal to the transcription start site of *NRP1*, involved in the activation of a sub-type of mesenchymal cells. Overall, this study provides a high-resolution map of chromatin accessible regions in liver cell types, shows that single-nucleus profiling captures peaks missed in bulk, and offers insights into genetic mechanisms underlying cell-type gene regulation.

TADs AND LOOPS ARE IMPOSSIBLE OBJECTS

Luay Almassalha[1,2], Marcelo Carignano[2], Ruyi Gong[2], Wing Shun Li[2], Lucas Carter[2], Kyle MacQuarrie[2], Igal Szleifer[2], Vadim Backman[2]

[1]Northwestern Memorial Hospital, Gastroenterology and Hepatology, Chicago, IL, [2]Northwestern University, Biomedical Engineering, Evanston, IL, [3]Northwestern University, Biomedical Engineering, Evanston, IL, [4]Northwestern University, Biomedical Engineering, Evanston, IL

Long-range interactions across the human genome play a crucial role in regulating gene expression, coordinating differentiation, and guiding cellular responses to signaling. While chromatin architecture is often described through topologically associating domains (TADs) and loop extrusion models, these connectivity features may not necessarily define space-filling volumes. We show that these represent "impossible objects", that require consideration of the spatial reasoning that produce the structures. We propose that chromatin organization in individual cells follows geometric constraints to produce packing domains, nanoscale physical structures that act as the reaction volumes regulating transcription. Packing domains act as geometric manifolds that coordinate the positioning of nucleosome remodeling enzymes, transcription factors, and polymerases by physical principles. We then describe how loops serve not as a physical scaffold but as a mechanism for dynamically allocating segments. By producing an allocation, loops instead act to define which segment is intended to work as a coherent unit. The type of transcriptional output produced requires the integration of additional information about the cell state. These state variables (nuclear volume, redox potential, and ionic conditions) guide nucleosome remodeling complexes to generate the reaction volumes that produce RNA. By integrating geometric principles with connectivity, we conclude by highlight a novel framework for understanding genome structure-function.

GENOMIC FLEXIBILITY THROUGH EXTRACHROMOSOMAL AMPLICONS: A *LEISHMANIA* SURVIVAL STRATEGY

Atia Amin[1], Ana Victoria Ibarra-Meneses[2], Christopher Fernandez-Prada[2,4], Mathieu Blanchette[3], David Langlais[1,4]

[1]Dahdaleh Institute for Genomic Medicine, Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada, [2]Department of Pathology and Microbiology, Université de Montréal, Saint-Hyacinthe, QC J2S 2M2, Canada, [3]School of Computer Science, McGill University, Montreal, QC H3A 0E9, Canada, [4]Department of Microbiology and Immunology, McGill Research Centre on Complex Traits, Montreal, QC, Canada

Leishmania is a eukaryotic parasite that causes leishmaniasis, a disease affecting ~350 million people worldwide. Unlike most eukaryotes that regulate gene expression through transcriptional regulation, *Leishmania* mainly modulates its gene expression by changing the number of gene copies. One way it achieves this is by amplifying extrachromosomal DNA amplicons, especially under drug stress. Although hypothesized, no systematic study had confirmed that *Leishmania* generates diverse amplicons through recombining its numerous repeated sequences. Furthermore, the recombination mechanisms that led to the formation of these amplicons remains poorly understood.

In our study, we exposed different *Leishmania* species to gradually increasing drug stress and examined their extrachromosomal amplicon diversity. We found that drug-resistant *Leishmania infantum*—the species responsible for the most severe form of the disease—produces a diverse mix of both linear and circular amplicons carrying resistance genes. While both forms can coexist, increasing drug pressure appears to favor circular amplicons that contain at least two copies of the resistance gene. Using Nanopore long-read sequencing, we delineate the recombination events leading to these amplifications. Our data support a model in which gene duplication begins with an inter-chromatid homologous recombination, which creates an initial intra-chromosomal duplication. This is then followed by a second homologous recombination within the duplicated region, further increasing gene copy number. Interestingly, the two other species, *Leishmania major* and *Leishmania braziliensis*—which cause milder forms of the disease—showed less amplicon diversity under the same drug pressure conditions. This suggests that different *Leishmania* species may have distinct adaptive strategies in response to drug stress.

Using *Leishmania* parasites as models, our findings provide new insights into the recombination-driven plasticity in eukaryotes. We hope that our study will pave the way for further research into extrachromosomal amplicon dynamics and their role in the evolution of parasite genomes, while also highlighting potential new targets for diagnostic and therapeutic strategies to combat drug resistance in *Leishmania*.

OVERCOMING CHALLENGES IN TROPICAL ARCHAEOGENOMICS: A CASE IN EARLY COLONIAL INTERACTIONS AND INDIGENOUS ENSLAVEMENT IN THE SPANISH COLONIZED CARIBBEAN

Beatriz Amorim[1], Lourdes Perez Iglesias[2], Roberto Valcarcel[3], Jason Laffoon[4], Yadira Chinique[5], Miren Iraeta Orbegozo[6], Marcela Sandoval Velasco[6], Jazmin Ramos Madrigal[6], Hannes Schroeder[4,6], Kathrin Nägele[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Archaeogenetics, Leipzig, Germany, [2]Centro de Investigaciones y Servicios Ambientales y Tecnológicos (CISAT), Departamento Centro Oriental de Arqueología, Holguín, Cuba, [3]Nova Southeastern University, Department of Humanities, Fort Lauderdale, FL, [4]Leiden University, Faculty of Archaeology, Leiden, Netherlands, [5]University of Winnipeg, Department of Anthropology, Winnipeg, Canada, [6]Globe Institute, University of Copenhagen, Center for Evolutionary Hologenomics, Copenhagen, Denmark

Our project investigates the El Chorro de Maíta site in Cuba, one of the largest indigenous Caribbean cemeteries dating to the peri-colonial era. This period was characterized by European colonization, which was responsible for the displacement of indigenous populations through systems such as the encomienda. Concurrently, the Caribbean became a significant trading hub among the Trans-Atlantic slave routes.

Excavations at El Chorro, ongoing since the 1980s, have yielded over 150 sets of human remains. Strontium isotope analysis has identified individuals of local Cuban origin, as well as migrants from the Yucatán Peninsula in Mexico. This aligns with historical evidence of forced relocation of indigenous groups within the Caribbean and Central American regions due to the encomienda system. The site has yielded remains of an individual of European ancestry and associated material culture, as well as an individual of “African” descent. The diversity observed at this site exemplifies the complex cultural and genetic exchanges that occurred during the European colonial period in the Americas.

Of the teeth recovered from the site, we successfully obtained the genomes of 32 individuals and conducted population genomics analysis. While our results corroborated the ancestral assignments based on non-genetic evidence, this work presented significant challenges. The tropical environment of the Caribbean, with its high temperatures and humidity, poses difficulties for DNA preservation. Additionally, the lack of representative modern reference groups from the Caribbean and Central and South American mainland complicates our analysis, particularly for imputation methods and ancestry inference. These challenges are amplified when working with small-scale sites that yield limited sample sizes and varying genome quality.

We are working towards overcoming these challenges to obtain high-quality genomic data that we can integrate with archaeological and historical evidence to illuminate the complex demographic dynamics during early colonisation in the Caribbean. By analysing genetic origins alongside mortuary practices, we aim to reconstruct both the social structure at El Chorro and the individual life histories of those interred there.

mtDNA SIGNATURES OF SEX-BIASED ADMIXTURE IN EUROPEAN-AFRICANS FROM MID-SOUTH US

E K Amos-Abanyie[1], S Buonaiuto[1], F Marsico[1], N R Migliore[3], A Tommasi[3], N Boga[1], A Mohammed[2], L K Chinthala[2], T H Finkel[4], R L Davis[2], C W Brown[4], R W Williams[1], D Ashbrook[1], A Achilli[3], V Colonna[1]

[1]UTHSC, Genetics Genomics and Informatics, Memphis, TN, [2]UTHSC, Center for Biomedical Informatics, Memphis, TN, [3]University of Pavia, DBB, Pavia, Italy, [4]UTHSC, Dept of Pediatrics, Memphis, TN

The history of admixture between human populations often shows sex-biased patterns, where source populations contribute unequal proportions of males and females. Understanding these patterns is crucial for reconstructing demographic histories and studying the genetic consequences of historical events. We analyzed mitochondrial DNA (mtDNA) variation in the Biorepository and Integrative Genomics (BIG) Initiative, a diverse pediatric cohort mostly from Memphis, TN, that includes over 13,000 sequenced genomes, with 50% of participants having non-European ancestry. BIG represents one of the largest pediatric cohorts with substantial admixed representation, including 20% African ancestry and 30% showing admixture patterns, primarily European-African and European-American.

Our analysis identified all major worldwide mtDNA haplogroups in BIG except those typically restricted to Oceania and Southeast Asia. As expected, individuals with non-European ancestry showed greater divergence from the reference sequence, with African and European-African admixed individuals exhibiting the highest number of variable sites. The African-associated L haplogroups showed the largest variance in variable sites, confirming previous evidence of higher genetic diversity in African populations from Memphis, from both independent cohorts and nuclear DNA analyses of BIG.

Focusing on European-African admixed individuals (N=1,887), we found a striking predominance of African maternal lineages, with haplogroup L accounting for 77.85% of lineages, while European/Western Eurasian haplogroups H (9.49%) and T, U, and J (collectively 6.51%) were present at lower frequencies. This marked asymmetry in maternal ancestry suggests strong sex-biased admixture, consistent with historical patterns of predominantly European male and African female contributions during the formation of admixed populations in contexts of colonialism and forced migration.

These findings have important implications for medical genetics, particularly for understanding mitochondrial disease risk in admixed populations, where patterns may more closely follow those of African populations despite substantial European genetic ancestry. Our results demonstrate the value of analyzing mtDNA in diverse cohorts and highlight the importance of considering sex-biased admixture patterns in genetic studies of human populations.

VARIANT POSITION MODULATES FUNCTIONAL IMPACT IN MASSIVELY PARALLEL REPORTER ASSAYS

Sambina Islam Aninta[1], Ryan Tewhey[2], Carl G de Boer[1]

[1]University of British Columbia, School of Biomedical Engineering, Vancouver, Canada, [2]The Jackson Laboratory, Bar Harbor, Maine, ME

Genome-wide association studies (GWAS) have identified hundreds of thousands of disease-associated loci; however, the causal variants for the vast majority of these trait-associated loci are yet to be identified and functionally validated. To establish a mechanistic link between disease-causing variants in *cis*-regulatory elements (CREs) and their effects on gene regulatory pathways, massively parallel reporter assays (MPRAs) are performed (Tewhey et al., 2016). The measured effects of alternate and reference alleles in these experiments are used to prioritize causal loci (Mouri et al., 2022).

In a typical MPRA reporter, the variant of interest is positioned at the center before being introduced into the target cell, either via a lentiviral vector or episomal expression (Tewhey et al., 2016). Since the cis-regulatory region is expressed outside its native genomic context in MPRA, accurately capturing its regulatory effects can be challenging. Positioning the variant at the center is believed to help mitigate positional biases and minimize confounding influences from surrounding sequences, ultimately providing a more reliable measurement of its impact on gene expression.

However, it remains unclear whether the center is the optimal position for the variant within the reporter construct and how variant placement influences the prioritization of causal variants. In this study, we explored this question both experimentally and through predictive modelling, analyzing how the variant's effect changes depending on its position within the reporter.

Our results, from both experimental and predictive analyses, reveal that the effect of a single nucleotide mutation (SNP) on expression varies significantly based on its position within the reporter. When the variants are placed further apart, the correlation between their SNP effects decreases. Additionally, variants positioned closer to the promoter in the reporter vector exhibit stronger effects.

Our findings highlight the importance of considering variant positioning in MPRA experiments, emphasizing the need for further studies to determine the optimal approach for testing variant effects.

SWITCH-LIKE GENE EXPRESSION MODULATES DISEASE RISK

Alber Aqil[1], Yanyan Li[2], Saiful Islam[2], Madison Russel[3], Theodora Kallak[4], Marie Saitou[5], Omer Gokcumen[1], Naoki Masuda[2,3]

[1]State University of New York at Buffalo, Biological Sciences, Buffalo, NY, [2]State University of New York at Buffalo, Mathematics, Buffalo, NY, [3]State University of New York at Buffalo, Institute for Artificial Intelligence and Data Science, Buffalo, NY, [4]Uppsala University, Department of Women's and Children's Health, Uppsala, Sweden, [5]Norwegian University of Life Sciences, Faculty of Biosciences, Aas, Norway

A fundamental challenge in biomedicine is understanding the mechanisms predisposing individuals to disease. While previous research has suggested that switch-like gene expression is crucial in driving biological variation and disease susceptibility, a systematic analysis across multiple tissues is still lacking. By analyzing transcriptomes from 943 individuals across 27 tissues, we identified 1,013 switch-like genes. We found that only 31 (3.1%) of these genes exhibit switch-like behavior across all tissues. These universally switch-like genes appear to be genetically driven, with large exonic genomic structural variants explaining five (~18%) of them. The remaining switch-like genes exhibit tissue-specific expression patterns. Notably, tissue-specific switch-like genes tend to be switched on or off in unison within individuals, likely under the influence of tissue-specific master regulators, including hormonal signals. Among our most significant findings, we identified concordantly switched-off genes in the vagina that are linked to vaginal atrophy (44-fold, $p<10\text{-}4$). Experimental analysis of vaginal tissues revealed that low systemic levels of estrogen lead to a significant reduction in both the epithelial thickness and the expression of the switch-like gene *ALOX12*. We propose a model wherein the switching off of driver genes in basal and parabasal epithelium suppresses cell proliferation therein, leading to epithelial thinning and, therefore, vaginal atrophy. Our findings underscore the significant biomedical implications of switch-like gene expression and lay the groundwork for potential diagnostic and therapeutic applications.

DISTINCT MONOALLELIC EXPRESSION SIGNATURES CHARACTERIZE UNCLASSIFIED BREAST TUMORS AND ASSOCIATE WITH PATIENT GENETIC BACKGROUNDS

Mona Arabzadeh, Amartya Singh, Hossein Khiabanian

Rutgers Biomedical Health Sciences, Rutgers Cancer Institute, New Brunswick, NJ

In diploid cells, allelic imbalance occurs when gene alleles are expressed at different levels due to genetic variations, regulatory rearrangements, copy-number changes, or epigenetic inactivation. In normal conditions, it arises through both non-random mechanisms, such as imprinting, and random processes like X-chromosome inactivation. However, in cancer, allelic imbalance becomes particularly significant, providing a selective advantage by impairing tumor suppressors or promoting the expression of altered alleles. While public datasets like The Cancer Genome Atlas (TCGA) have been widely utilized, allelic imbalance remains an underexplored aspect, partly due to the need for informative SNPs to capture these events. Unraveling such complexities also requires considering population heterogeneity, emphasizing the need for large, diverse cohorts to gain deeper insights.

To investigate the allelic imbalance landscape in tumor datasets, we developed Interval-Based Allelic Imbalance Detection (IBAid), a quantitative framework that uses interval arithmetic to distinguish monoallelic from biallelic expression while normalizing for copy number and tumor purity. IBAid computes confidence intervals based on depth measurement uncertainty to ensure robust allelic ratio estimation. We applied this approach to TCGA's BRCA dataset to uncover novel insights into allelic imbalance patterns across molecular subtypes. We assessed per-sample allelic imbalance gene enrichment and identified recurrently imbalanced genes across samples. Unsupervised sample-level analysis revealed the highest number of imbalanced genes in the Basal subtype and the lowest in Lum A, aligning with the supervised findings on the molecular subtypes using PAM50 classification. Gene-level enrichment analysis further identified subtype-specific genes exhibiting high allelic imbalance, implicating potential functional roles in tumor progression.

Through unsupervised gene enrichment analysis across all samples we identified two key gene groups: one exhibiting monoallelic expression across all breast cancer subtypes and another within a subset of BRCA samples that had not been previously classified into any established subtype—a significant novel finding. Notably, this unclassified subgroup was enriched in Black/African American patients, with specific SNPs marking monoallelic expression. Clinically, these tumors exhibit poor survival outcomes, comparable to the aggressive Basal subtype, suggesting a potential link between allelic imbalance and disease severity. This finding underscores the need to explore the genetic and epigenetic mechanisms driving allelic imbalance and their therapeutic implications, as these patterns may serve as novel biomarkers for prognosis and targeted treatment strategies.

ADVANCING scRNA-SEQ ANALYSIS: A NEW PARADIGM FOR GRAPH-PARTITIONING AND CLUSTER REFINEMENT ENABLES IDENTIFICATION OF NOVEL MALIGNANT CELL SIGNATURES LINKED WITH RESISTANCE TO IMMUNE CHECKPOINT BLOCKADE TREATMENTS

Mona Arabzadeh, Amartya Singh

Rutgers Biomedical Health Sciences, Rutgers Cancer Institute, New Brunswick, NJ

Single-cell technologies have been introduced to study cellular heterogeneity, providing deeper insights into individual cell states. Over time, additional omics layers have been integrated, enhancing their resolution and scope. However, a continuous need remains to refine analytical methods, ensuring accurate data interpretation. This process involves selecting variable features, transforming count data, applying dimensionality reduction and graph partitioning, and ultimately assigning precise cell labels to the resulting clusters.
We introduce a comprehensive single-cell analysis pipeline that begins with the 1- critical steps of feature selection and normalization, 2- this is followed by a novel deterministic kNN-based community detection method and 3- a robust cluster refinement approach. Our community detection method addresses two key limitations in traditional graph-partitioning methods used for scRNA-seq analysis. First, it overcomes the constraint of identifying mutually exclusive and exhaustive clusters, a limitation that can obscure transcriptional differences between cells within similar states. Second, it tackles the inherent stochasticity of traditional methods, which causes clustering results to vary each time they are applied to the same kNN graph, impacting the identification of marker gene sets. To further ensure the reliability of the clusters, our robust cluster refinement approach evaluates the stability of the identified communities and refines cell labeling based on the actual expression profiles of the marker genes for each cluster. This approach leads to more consistent and biologically meaningful cluster assignments. All methods provided in a single scRNA-seq analysis R package, named Piccolo, compatible with other single-cell analysis methods and objects.
We demonstrate significant improvements by benchmarking our framework across biological truth-known, as well as simulated counts datasets. In particular, we highlight how our proposed approaches enable us to identify small but robust cell groups, where the conventional pipeline would fail to reliably identify such cell groups. We further applied our scRNA-seq framework to scRNA-seq datasets obtained from studies conducted to uncover and examine cell-specific responses to immune checkpoint blockade (ICB) treatment. In particular, we examined pre- and post-ICB treatment samples from 8 distinct cancer subtypes associated with 4 tissues of origin to identify novel malignant cell gene expression signatures as well as aberrant gene expression signatures exhibited by tumor-associated stromal cells that were missed in the original analyses by the authors of these studies. The identified signatures play a vital role in limiting the efficacy of ICB treatments and thus development of therapies that target them would lead to significant improvements in treatment outcomes for patients.

A NEW BAYESIAN METHOD TO PERFORM DEMOGRAPHIC INFERENCE FROM GENOMIC DATA

Tommaso Stentella[1], 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 Tech, Centre for Computational Biology, Paris, France, [3]Weizmann Institute of Science, Department of Complex Systems, Rehovot, Israel

Segregating sites and their statistical distribution along genomes contain much information about the degree of divergence or relatedness of homologous sequences in a population. The pattern of these sites is dictated by mutation and population dynamics together with recombination. In fact, recombination breaks down genomes into smaller regions which share the same coalescent tree, i.e. are Identical by Descent (IBD). This ensemble of trees then gives rise to what is referred to as the Ancestral Recombination Graph (ARG). The Sequentially Markov Coalescent approximation to the ARG and Hidden Markov Models (e.g. PSMC) are commonly used to infer population history jointly with IBD segments. Here we circumvent the inference of recombination events and introduce a novel methodology, focusing on the demographic history. We argue that it is sufficient to consider only the distribution of distances between segregating sites. This quantity is directly observable and is inherently a marginal distribution over both the ancestral recombination and coalescence events. Using our theoretical framework, we derive an analytical formula for the distribution of distances between segregating sites conditional on an arbitrary panmictic demography.

EXPLORING THE GENOMIC UNDERPINNINGS OF EVOLUTIONARY MISMATCH

Audrey M Arner[1], Jonathan Lifferth[2], Tan Bee Ting A/P Tan Boon Huat[3], Kar Lye Tam[3], Yvonne A Lim[3], Kee-Seong Ng[4], Vivek V Venkataraman[5], Ian J Wallace[6], Thomas S Kraft[7], Amanda J Lea[1]

[1]Vanderbilt University, Department of Biological Sciences, Nashville, TN, [2]Vanderbilt University, Department of Human Genetics, Nashville, TN, [3]Universiti Malaya, Department of Parasitology, Kuala Lumpur, Malaysia, [4]Universiti Malaya, Department of Medicine, Kuala Lumpur, Malaysia, [5]University of Calgary, Department of Anthropology and Archaeology, Calgary, Canada, [6]University of New Mexico, Department of Anthropology, Albuquerque, NM, [7]University of Utah, Department of Anthropology, Salt Lake City, UT

Phenotypes that evolve in a given environment may become disadvantageous if the environment rapidly changes, potentially leading to "evolutionary mismatch" and decreased fitness in the new environment. Recent work has provided clear predictions for mismatch at the genetic level: mismatched loci should exhibit evidence of past selection and function differently in ancestral versus novel environments. To test these predictions, we collected lifestyle questionnaires, white blood cell gene expression (n=686), and whole genome sequencing data (n=353) with the Orang Asli, the Indigenous peoples of Peninsular Malaysia who currently live along a gradient from subsistence-based to more urban and "mismatched" lifestyles. We first found that lifestyle is associated with coordinated changes in gene expression (>1,000 differentially expressed genes; 5% FDR), especially for immune and metabolism-related pathways (GSEA and GO analyses). Using our WGS data, we called 2,497,983 SNPs in this previously uncharacterized population and developed pipelines for effective imputation. The Orang Asli are typically divided into 19 distinct ethnolinguistic groups -- therefore, we also characterized population structure, finding that Orang Asli ethnolinguistic groups cluster most closely with each other than with other Asian populations. Next, we applied the Population Branch Statistic (PBS) and integrated haplotype score (iHS) to identify candidates of past positive selection. We are currently following up to test for overlaps between these putatively selected regions and the differentially expressed genes to understand how past adaptation interacts with current lifestyle change, as well as identifying cis-eQTL that contribute to the environmentally dependent gene expression. This research provides insight into the consequences of rapid shifts toward more urban, industrialized environments in an underrepresented population, with implications for understanding these effects globally.

DYSREGULATION OF CELLULAR SIGNALING NETWORKS UPON RAPID ENVIRONMENT SHIFT

Thomas K Atkins[1], Kristina M Garske[1], Charles M Mwai[1], Julie Peng[1], Matt Chao[1], Emma Gerlinger[1], John Kahumbu[2], Boniface Mukoma[3], Echwa John[3], Patricia Kinyua[3], Anjelina Lopurudoi[3], Nicholas Mutai[3], Dino Martins[4], Amanda Lea[5], Julien F Ayroles[1]

[1]Princeton University, Lewis-Sigler Institute, Princeton, NJ, [2]Harvard University, Department of Organismal and Evolutionary Biology, Cambridge, MA, [3]Turkana Health and Genomics Project, Nairobi, Kenya, [4]Turkana Basin Institute, Nairobi, Kenya, [5]Vanderbilt University, Department of Biological Sciences, Nashville, TN

Stressful environmental perturbations can disrupt the coordinated expression of genes within regulatory networks, leading to a systemic loss of correlation between gene expression levels—a phenomenon known as decoherence. We hypothesize that recent rapid shifts from ancestral human environments characterized by high pathogen load to urbanized hygienic environments leads to the same loss of correlation. To study this question, we work with the Turkana, a community in Northern Kenya that spans an extreme lifestyle gradient that replicates the human environmental transition from rural pastoralism to urban living. The presence of both lifestyles within a single ethnic group provides an unprecedented opportunity to study the immune system across the evolutionary mismatch spectrum, allowing us to uncover gene expression networks that developed in a high pathogen load environment, but have become dysregulated in a hygienic environment.

To characterize these networks, we generated scRNA sequencing data of PBMCs for 250 Turkana individuals, obtaining a dataset of 1.3 million cells. By testing for changes in correlation between expression of genes across cell types, we infer regulatory networks that change across environments. Using the CILP framework, we identify 588 pairs of genes that exhibit a significant change (FDR<0.05) in correlation between urban and pastoralist populations, 62% of which display greater correlation in pastoralist individuals. Of these gene pairs, 14% exhibit a loss of correlation within cell type, while the remaining 86% display loss of correlation across two different cell types. From these results we can begin to map how the evolutionary recent reduction in human pathogen load has reshaped our immune systems.

COMPREHENSIVE PHYLOGENOMIC ANALYSIS OF MYCOBACTERIUM TUBERCULOSIS IN ETHIOPIA

Betselot Z Ayano[1,2], Alemayehu Godana[2], Helen Nigussie[3]

[1]Ethiopian Public Health Institute, Virology and Genomics, Addis Ababa, Ethiopia, [2]Addis Ababa University, Institute of Biotechnology, Addis Ababa, Ethiopia, [3]Addis Ababa University, Microbial, Cellular and Molecular Biology, Addis Ababa, Ethiopia

Background: *Tuberculosis* (TB) remains a major health challenge in Ethiopia. This study applies bioinformatics to analyze *Mycobacterium tuberculosis* (MTB) sequences, focusing on lineage (genetically distinct MTB strain) diversity and resistance-associated mutations.

Methods: We analyzed 580 Ethiopian MTB sequences (2008–2024) using TB-Profiler for lineage classification and resistance prediction, IQ-TREE for phylogenetics, and Python for statistical analysis.

Results: Genome coverage exceeded 99% at 30× depth. Drug-sensitive cases accounted for 70.34%, while multidrug-resistant TB (MDR-TB; 15.52%), pre-extensively drug-resistant TB (pre-XDR-TB; 5.69%), and extensively drug-resistant TB (XDR-TB; 0.34%) comprised the rest. Among drug-resistant TB cases, MDR-TB was prevalent in patients, while isoniazid-resistant TB (HR-TB) dominated in returnees from Saudi Arabia and refugees, with no MDR-TB in refugees. Isoniazid (25.0%) and ethionamide (14.31%) showed the highest resistance among first- and second-line drugs. Two cases resisted Bedaquiline, Clofazimine, Delamanid, and Pretomanid; none resisted linezolid. MDR-TB was linked to katG mutations, while inhA-katG co-mutations were associated with pre-XDR-TB. Cross-resistance was identified in inhA, rrs, fbiC, mmpR5, and gyrA.
A significant association observed between lineages and drug resistance. Lineage 4 (L4) was predominant (53.0%), led by sub-lineage 4.2.2.2 (47.65%). MDR-TB was highest in L4 (23.1%) and lineage 3 (L3) (13.8%), while lineage 7 (L7) remained largely drug-sensitive (98.0%). L4 was prevalent in patients and returnees, L3 in refugees, and L7 (Ethiopian lineage) appeared in a single returnee. The first documented lineage 9 (L9) case in Ethiopia was identified in a refugee. Mixed-species cases (*M. bovis–MTB, M. orygis–MTB*) were observed, with mixed MTB lineages being significantly drug-sensitive. Phylogenetic analysis confirmed L4 and L3 dominance and clustering of MDR-TB and Pre-XDR-TB, suggesting active transmission.

Conclusion: L4's MDR-TB dominance highlights its clinical importance. L7's drug sensitivity suggests unique characteristics. The first Ethiopian L9 case indicates migration-driven transmission. The high resistance to isoniazid and ethionamide aligns with global trends, confirming their reduced efficacy. MDR-TB and Pre-XDR-TB cases spread within the population rather than arising independently due to new mutations. Findings emphasize the need for drug-resistance surveillance, and targeted TB control.

FROM GENOMICS TO NEUROANATOMY: HOW CNVs CONTRIBUTE TO RISK OF NEURODEVELOPMENTAL DISORDERS AND BRAIN STRUCTURE ALTERATIONS

Sara Azidane[1,2], Xavier Gallego[1], Lynn Durham[1], Mario Cáceres[2], Emre Guney[1], Laura Pérez-Cano[1]

[1]STALICLA, Discovery and Data Science Unit, World Trade Center, Moll de Barcelona, Edif Este, Barcelona, Spain, [2]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra, Bellaterra, Spain

Copy-number variants (CNVs) are genome-wide structural variations involving the duplication or deletion of large nucleotide sequences. While commonly found in humans, large and rare CNVs are known to contribute to the development of neurodevelopmental disorders (NDDs), including autism spectrum disorder (ASD). Given that these NDD-risk CNVs cover broad genomic regions, pinpointing the critical gene(s) responsible for the phenotype remains a challenge. In this study, we performed a meta-analysis of CNV data from 11,614 affected individuals with NDDs and 4,031 control individuals from SFARI database to identify 41 NDD-risk CNV loci, including 24 novel regions. These loci contain dosage-sensitive genes significantly enriched for known NDD-risk genes and pathways, many of which converge in protein-protein interaction networks, show high expression in the brain across all developmental stages, and are over-transmitted in multiplex ASD families from the iHART cohort. Burden analysis using 4,281 NDD cases and 2,504 controls validated 162 dosage-sensitive genes driving risk for NDDs, including 22 novel NDD-risk genes. Moreover, using MRI and genetic data from over 30,000 individuals from the UK Biobank, we found that the volumen of several brain areas, such as the amygdala, caudate, and cerebellum, were associated with specific CNVs, further linking these genomic variations to neuroanatomical changes and NDD risk. These findings reinforce the role of dosage-sensitive genes in NDD pathology and highlight CNVs as key contributors to both genetic risk and brain structural abnormalities in NDDs.

ESTIMATING GENE MEAN PATHOGENICITY WITH PREDICTION-POWERED INFERENCE

Ayesha Bajwa[1], Ruchir Rastogi[1], Nilah M Ioannidis[1,2,3]

[1]UC Berkeley, Electrical Engineering and Computer Sciences, Berkeley, CA, [2]UC Berkeley, Center for Computational Biology, Berkeley, CA, [3]Chan Zuckerberg Biohub, San Francisco, CA

Gene-wide mean pathogenicity estimates for missense variants can be used as per-gene or aggregated into whole-genome prior probabilities of pathogenicity for clinical calibration of missense pathogenicity predictors or for downstream applications such as gene prioritization or essentiality tasks. These gene-wide means can be understood as the estimated proportion of pathogenic missense variants per gene. Prediction-powered inference (PPI) is a statistical method that combines gold standard labeled data with predictions of unlabeled data points, giving updated point estimates and also allowing shrinking of the confidence intervals for population statistics such as means. It is designed to improve statistical power in the setting of small labeled data and large unlabeled data, both with associated predictions from an orthogonal predictor. Here we apply PPI using ClinVar labeled data and predictions from AlphaMissense, which provides an evolutionary-informed prediction of missense variant pathogenicity that avoids training on ClinVar labels directly, instead using a small subset of ClinVar variants for hyperparameter tuning and calibration. After using PPI to combine gold standard labeled data and pathogenicity predictions, we explore whether updated gene-wide pathogenicity estimates can provide additional performance or insight into downstream tasks such as gene prioritization. PPI's statistical guarantees rely on the untestable and relatively strong assumption that the observed distribution of the labeled data is close enough to the unobserved true distribution of the unlabeled data. Here we do note significant concordance between the predicted distributions for labeled and unlabeled data for many genes. We recognize the persistent issue of data circularity and attempt to reason carefully about individual genes or sets of genes. We also consider orthogonal data and validation strategies when possible.

REVEALING THE EXTENSIVE ALLELIC HETEROGENEITY AND IMPACT OF TRANSPOSABLE ELEMENTS ACROSS 130 DIVERSE HUMAN HAPLOTYPES

Parithi Balachandran[1], Mark Loftus[2], Tylor L Brewster[3], Ryan E Mills[4], Weichen Zhou[4], Miriam K Konkel[2], Christine E Beck[1,3]

[1]The Jackson Laboratory, Genomic Medicine, Farmington, CT, [2]Clemson University, Department of Genetics and Biochemistry, Clemson, SC, [3]UConn Health Center, Department of Genetics and Genome Sciences, Farmington, CT, [4]University of Michigan Medical School, Department of Computational Medicine & Bioinformatics, Ann Arbor, MI

Transposable elements (TEs) diversify human genomes through retrotransposition and recombination. Polymorphic TE Insertions (TEIs) contribute to both inter- and intra-individual genetic variation, leading to ongoing mutagenesis and disease. In this study, we have conducted a thorough analysis of TE-driven variation across sequence-resolved assemblies of 130 phased human haplotypes from diverse individuals.

Across these 130 haplotypes, we identified 12,984 non-redundant polymorphic TEIs, including 1,604 LINE-1s, 10,270 *Alu* elements, 764 SVAs, 3 HERV-Ks, 1 snRNA and 72 processed pseudogenes (PPGs) comprising 10.4 Mbp of sequence. Nearly half (47.6%) of these insertions fell within existing repetitive regions, making them difficult to identify from short-read approaches. Additionally, we identified 2,478 reference TEs, including 288 LINE-1s, 2,051 *Alu* elements, 111 SVAs, and 28 PPGs comprising 1.8 Mbp of sequence that exhibited polymorphic deletions, indicating that individuals in our study lacked these reference insertions. Our assembly data have enabled the interrogation of allelic heterogeneity associated with each sequence-resolved TEI locus. For 2,148 reference and polymorphic full-length L1PA2 and L1HS sequences we have determined the extent of allelic heterogeneity across the 130 haplotypes and found that the amount of heterogeneity was positively correlated with insertion allele frequency. Allelic differences frequently impact the coding and mobilization potential for LINE-1s, including hot (i.e. highly active) elements. Among SVAs we observed the greatest variability in lengths due to allelic variation within the hexameric and VNTR regions.

In addition to insertion variation, two homologous TEs can act as substrates of ectopic DNA repair leading to structural variant formation. We identified 2,459 *Alu*, 817 LINE-1 and 10 SVA mediated rearrangements that affect nearly 21 Mbp of human reference genome (T2T-CHM13). A majority (74.7%) of TE-mediated rearrangements were driven by *Alu* elements and more than half of them occur within genes. Upon further inspection, we found that TE-mediated inversions harbor small deletions or duplications at the junctions of 78% of *Alu* driven and 12% of LINE-1 driven rearrangements, indicating the role for replication-based mechanisms.

In summary, sequence-resolved genomic assemblies of diverse individuals have enabled a comprehensive understanding of TE-derived variation between genetically diverse individuals, highlighting extensive differences caused by transposon mobility, allelic heterogeneity, and TE-driven rearrangements.

FROM MILK TO MICROBES: A TARGETED SEQUENCING APPROACH TO DAIRY CATTLE HEALTH

Vanessa A. Barbosa, Andrew Wallace, Hong Ling, Martina Franz, Sean Gatenby, Catherine Neeley, John Williamson, Chad Harland, Christine Couldrey

Livestock Improvement Corporation, Research & Development, Newstead, New Zealand

Milk contains a complex microbial ecosystem providing valuable insights into dairy cow health. With genome reference sequences available for many pathogenic species, next-generation sequencing (NGS) methods have revolutionised genomic surveillance, enhancing detection and tracking of pathogens.

Targeted sequencing (TS) has emerged as a powerful NGS technique by enriching for particular DNA regions and reducing the sequencing of non-relevant DNA, offering high accuracy, shorter turnaround times, lower cost, and reduced computational demand. While TS provides high sensitivity, species-level classification can be influenced by specificity and coverage of targeted regions, requiring careful interpretation. As part of the MilkOmics® project, a TS panel was developed to detect dairy cattle pathogens in bulk milk samples from nearly 350 farms across New Zealand. Among 189 taxa, the panel targets 13 key mastitis pathogens, responsible for the most common disease in dairy cattle, which leads to potentially fatal mammary gland infections, costing farmers around NZ$ 10,000 annually. Over the past four years, approximately 3,000 samples have been sequenced using Shotgun sequencing, and around 2,200 samples have been sequenced using TS, providing insights into species presence, abundance, and their correlation with animal health and on-farm practices. TS reduced the proportion of cow DNA reads from an average 85% to an average 33% compared to shotgun sequencing. Comparative analyses of culture-based methods with Shotgun sequencing demonstrated that TS can be used to reliably identify and quantify key dairy cattle mastitis pathogens. Findings reveal on-farm prevalence and regional and seasonal differences in pathogen presence and abundance.

During a mastitis prototype trial, baseline species profiles were established, enabling the ranking of new samples relative to these baselines. Reports summarizing pathogen profiles were shared with veterinarians to assess the usefulness of the data. These findings underscore the potential of the data to reflect past management practices and support targeted interventions to mitigate the risk of infection spread.

Future directions include defining microbial load thresholds associated with infection risk and refining diagnostic tools for proactive disease management, illustrating the promise of genomics in reshaping pathogen diagnostics and dairy farm health practices.

VISUALIZE COMPLEX STRUCTURAL VARIANTS IN HIFI DATA WITH SVTOPO

Jonathan R Belyeu, William J Rowell, Juniper Lake, James M Holt, Zev Kronenberg, Christopher T Saunders, Michael A Eberle

Pacific Biosciences, Computational Biology, Menlo Park, CA

Method: Structural variants (SVs), defined as genomic variants that impact at least 50 nucleotides, are common in the human genome, and play major roles in diversity and human health. Many SVs are simple deletions or duplications, which can often be represented effectively by existing visualization tools optimized for small variants. Other SVs are less easily categorized and are challenging to represent visually. SVTopo uses high accuracy long reads to identify genome alignment break locations relative to a reference genome, connects breaks into complex multi-locus SVs via chimeric alignments, and presents the supporting evidence in easily understood figures. SVTopo shows breakpoint evidence in ways that aid reasoning about the impact of large, multi-breakpoint events such as inversions, translocations, and combinations of simpler SVs.

Results: SVTopo was applied to six unrelated HiFi samples from the publicly available Platinum Pedigree cohort. 160 events were found, including 35 simple inversions, 34 translocations, and 34 multi-deletion SVs. The remaining 58 variants fit more classical definitions of complex variation. 46 were complex variants with an inverted component, including 41 inversions with a flanking deletion on one or both sides. These deletions ranged in size from 11 bp to 20 kbp, with the inverted sequences ranging from 50 bp to 19 kbp. The other five inversions were sequences duplicated from another location (non-tandem duplication inversions). The complex SVs found also included three non-tandem duplications with deletions, one non-tandem duplication without a deletion, and a duplication-deletion pair.

Conclusions: These insights gained from SVTopo visualization provide a significant enhancement for analysis of complex SVs through simple and complete representation of the rearrangement structures, not available via other tools for genome or SV visualization. The identification of many complex SVs within this small sample set also highlights the importance of applying analytic methods designed for complex SVs, as the unusual structures of these SVs could be easily missed in analysis with general methods. By efficiently demonstrating the structures of complex SVs, SVTopo enables users to maximize the advantages of high-accuracy long-read genome sequencing.

RFMIX-READER: ACCELERATED READING AND PROCESSING FOR LOCAL ANCESTRY STUDIES

Kynon J Benjamin

Northwestern University Feinberg School of Medicine, Stephen M Stahl Center for Psychiatric Neuroscience, Department of Psychiatry and Behavioral Sciences, Chicago, IL

Background: Human genetic diversity, shaped by history and drift, results in varying allele frequencies across populations. While population genetic differences are small, disease prevalence often varies with ancestry due to gene-environment factors, notably in admixed populations like Black/Hispanic Americans. These large, underserved groups present unique opportunities to study complex diseases but require analytical methods accounting for population structure. Global ancestry, representing overall ancestral proportions, lacks locus specificity. Local ancestry inference (LAI) offers granular views of ancestral origin at specific genomic locations, enhancing population history insights and genetic association studies, especially molecular QTL mapping in admixed populations over global ancestry. Despite LAI's value and RFMix's common use, large-scale studies face processing bottlenecks. RFMix output processing is memory and time intensive, hindering large-scale analysis. Integrating RFMix data with genotype data and visualization also pose hurdles.
Results: Here, we present RFMix-reader, a new Python-based parser designed to streamline large-scale LAI data analysis by prioritizing both computational efficiency and memory optimization. The software leverages GPUs when available, offering substantial speed boosts in addition to efficient CPU processing. Crucially, RFMix-reader introduces loci imputation to align local ancestry loci with genotype variant locations for both RFMix and FLARE output, improving data integration and downstream analyses. To facilitate efficient storage and downstream processing, RFMix-reader also supports writing both loci and admixture haplotype data to a BED format, enabling compression and indexing with specialized genetic software like bgzip and tabix. Finally, RFMix-reader offers visualization for both RFMix and FLARE data.
Conclusions: By overcoming critical data processing hurdles associated with local ancestry data, RFMix-reader empowers researchers to unlock the full potential of this information for understanding human health and health disparities. Its advanced features, such as loci imputation, BED file output for compression and indexing, and visualization capabilities, further enhance its utility for large-scale genomic studies and facilitate deeper insights into the genetic basis of complex diseases. The improved efficiency and scalability offered by RFMix-reader will enable more effective utilization of local ancestry data in population genetics and genomic medicine, ultimately contributing to a more comprehensive understanding of human health.

THE ROLE OF MATERNAL CHOLINE SUPPLEMENTATION ON OFFSPRING BEHAVIORAL OUTCOMES AND GENE ACCESSIBILITY IN THE FRONTAL CORTEX

Naomi Boldon[1], Bo Shui[2], Jen Grenier[3], Brian D Cherrington[4], Jill Keith[5], Barbara Strupp[6], Paul Soloway[7]

[1]University of Wyoming, Biomedical Sciences, Laramie, WY, [2]Cornell University, Molecular Genetics, Ithaca, NY, [3]Cornell University, Genomics Innovation Hub, Ithaca, NY, [4]University of Wyoming, Zoology and Physiology, Laramie, WY, [5]University of Wyoming, Family and Consumer Sciences, Laramie, WY, [6]Cornell University, Nutritional Sciences, Ithaca, NY, [7]Cornell University, Molecular Genetics, Ithaca, NY

Choline is an essential vitamin critical for neurocognitive development. It is vital for production of acetylcholine (neuronal signaling), formation of phosphatidyl choline (cell membranes), and is the primary dietary methyl donor (DNA methylation). Although widely distributed in foods, 90% of pregnant women do not consume Adequate Intake levels. Despite the increased need for choline during pregnancy, it is not included in prenatal vitamins. Evidence shows maternal choline supplementation (MCS) during fetal development lessens offspring cognitive dysfunction in the Ts65Dn mouse model of Down syndrome (DS) and exerts lifelong cognitive benefits for neurotypical offspring in both humans and rodent models, yet we know very little about the underlying mechanisms. This research aims to investigate the underlying genomic mechanisms that contribute to the molecular influence of MCS on offspring behavioral outcomes through exploration of chromatin accessibility. Our hypotheses are that MCS during pregnancy and lactation leads to chromatin accessibility states in specific cell types of offspring, and that these changes are at least in part responsible for the beneficial neurocognitive and behavioral outcomes. MCS normalizes epigenetic programming in certain neural cell types, bringing DS mice cognitively closer to 2N controls. To test our hypotheses, we conduct computational and statistical analyses of frontal cortex tissue from 4 groups of mice (n=56) who completed series of attention tasks using treatment groups assigned by genotype (2N or Ts65Dn) and diet (normal lab chow or 4.5 times control). Our results identify significant correlations between chromatin accessibility states and complex behaviors. We provide a candidate list of genes and peaks that mediate the effects of trisomy and MCS in trisomy and/or neurotypical offspring. By correlating candidate mediators with behavioral indices of attentional and affective functioning, we provide preliminary information about the likelihood of each as a mediator. Our findings inform future efforts to directly test if these associations reflect causal links to improve cognitive function in DS, as well as provide widespread neuroprotection and improved cognitive function in the population at large.

FUNCTIONAL ANNOTATION WORKFLOW FOR GENOME EDITING OF NOVEL MODEL ORGANISMS

Hidemasa Bono[1,2,3]

[1]Hiroshima University, Graduate School of Integrated Sciences for Life, Higashi-Hiroshima, Japan, [2]Hiroshima University, Department of Biological Science, School of Science, Higashi-Hiroshima, Japan, [3]Hiroshima University, Genome Editing Innovation Center, Higashi-Hiroshima, Japan

For model organisms such as humans and mice, progress has been made in genome sequencing and gene function analysis, and annotation information for these genes has been developed as a database, enabling integrated data analysis. Genome sequencing is becoming possible even for other novel model organisms with high-quality long-read sequencing using PacBio HiFi reads.

However, the application is limited to acquiring sequences of target and related genes, and much of the acquired sequences remain unused because other available sequence information is not sufficiently annotated. In other words, even after genome sequencing is completed, knowledge of the functional information of all the genes of newly sequenced organisms remains incomplete.

This information is becoming increasingly important in determining target genes for genome editing. We therefore have developed Fanflow as a workflow (a sequence of data analysis by a computer program) for gene function annotation using transcript sequence information and reference genome and protein sequence databases (DOI: 10.3390/ insects13070586). In this presentation, we report an extended version of Fanflow. The pathway information required for novel model organisms is often not included in existing pathway databases and must be created by the user. We integrated Quest for Pathways for eXpression (QPX) as a downstream analysis using the WikiPathways system, which allows users to create their pathways (DOI: 10.37044/osf.io/4uskb). We have already applied Fanflow starting from the genome sequence assembly of PacBio HiFi reads in several collaborative studies. We will introduce an example applied to insecticide resistant common bed bug, *Cimex lectularius* (DOI: 10.3390/insects15100737). We plan to apply the extended Fanflow to useful substance-producing species such as plants and fungi in collaboration with private companies as well as academia.

UNDERSTANDING HOW URBAN, INDUSTRIALIZED LIFESTYLES MODULATE THE IMMUNE RESPONSE IN THE ORANG ASLI OF MALAYSIA

Layla Brassington[1], Grace Rodenberg[1], Audrey M Arner[1], Tan Bee Ting A/P Tan Boon Huat[2], Kee Seong Ng[2], Yvonne Ai Lian Lim[2], Vivek V Venkataraman[3], Ian Wallace[4], Thomas S Kraft[5], Amanda J Lea[1]

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

A rapid innate immune response in the form of inflammation is an evolved, adaptive process to clear infections and protect the host. However, aspects of modern, industrialized environments are thought to perturb this process: for example, low-pathogen environments, diets high in processed foods, physical inactivity, and central adiposity are all thought to create an unresolved state of chronic inflammation that predisposes individuals to myriad non-communicable diseases. However, the degree to which these lifestyle factors directly modulate immune function remains poorly understood due to a lack of studies that measure lifestyle on a continuous gradient from non-industrial to industrial within a group of individuals with the same genetic background. To address this gap, the Orang Asli Health and Lifeways Project (OA HeLP) has collected data with the Orang Asli, the Indigenous people in Peninsular Malaysia, who live across an extreme lifestyle gradient of rural and non-industrial to urban and industrialized. We performed ex vivo stimulation of white blood cells with lipopolysaccharide (LPS) in 209 Orang Asli individuals to observe changes in gene expression in response to infection (i.e., comparing LPS+ to control samples). We observed a strong transcriptional response to infection: 5,031 out of 11,429 tested protein coding genes were upregulated in response to LPS (FDR < 1%), and these genes were enriched for expected processes such as regulation of innate immune response, T cell activation, and response to bacteria (GO and GSEA). At least part of this response is modified by lifestyle, such that individuals with greater exposure to industrialization have a stronger transcriptional response to LPS (n=458 genes). These genes are enriched for key immune processes such as cytokine signaling, response to stress, and immune cell differentiation. Ongoing analyses are focused on teasing apart the contribution of individual lifestyle factors (i.e., physical activity vs diet), as well as incorporating DNAm data (n=88) to ask whether lifestyle modulates the immune response through regulation of the epigenome. By working with a unique group that spans a major lifestyle gradient, this work begins to address the degree to which immune function is directly perturbed by urban, industrialized environments.

LONG-READ ASSEMBLY OF THE PLACENTA TRANSCRIPTOME REDUCES INFERENTIAL UNCERTAINTY AND UNVEILS NOVEL ISOFORMS ASSOCIATED WITH GESTATIONAL DIABETES MELLITUS

Sean T Bresnahan[1], William Wu[1], Jonathan Huang[2], Arjun Bhattacharya[1]

[1]The University of Texas MD Anderson Cancer Center, Epidemiology, Houston, TX, [2]The University of Hawai‘i at Mānoa, Epidemiology, Honolulu, HI

Transcriptome quantification using short-read RNA-seq often faces challenges in accurately detecting isoform diversity due to read-to-transcript mapping ambiguity and reliance on species-level references like GENCODE or adult tissue references like GTEx. These limitations result in significant uncertainty in isoform-level quantification, especially in developmental tissues like the placenta. Despite its critical role in metabolic programming, the placenta remains understudied in large-scale genetic consortia, with existing datasets focusing on gene-level expression and overlooking isoform variations that are crucial for understanding complex traits. Long-read RNA-seq offers a solution by directly sequencing full-length transcripts, providing greater resolution and uncovering isoforms missed by short-read approaches. To improve the catalog of placental isoform diversity, we present a long-read assembly from term placental samples (N = 72) using Nanopore sequencing. The assembled isoforms were supported by ENCODE datasets, including DNase-seq (N = 22), CAGE-seq (N = 8), and 3'-RNA-seq (N = 2), and splice junctions were validated using short-read RNA-seq (N = 200). We find that, although the placenta exhibits lower transcriptional complexity compared to other tissues, it has expanded isoform diversity in pregnancy-specific genes, including chorionic somatomammotropin hormone 1 (CSH1) and pregnancy specific beta-1 glycoproteins 1-6 (PSG1-PSG6). This refined catalog enhances our understanding of biological mechanisms underlying complex traits like gestational diabetes mellitus (GDM), a pregnancy complication with significant metabolic consequences. By integrating long- and short-read data, we demonstrate that long-read assembly reduces inferential uncertainty, improves the consistency of differential expression results across cohorts, including the Growing Up in Singapore Towards healthy Outcomes (GUSTO) cohort (N = 200) and the Genetics of Glucose Regulation in Gestation and Growth (Gen3G) cohort (N = 150), and enriches for gene ontology terms related to GDM. Our approach underscores the importance of precise transcriptomic profiling to uncover subtle isoform-level variations that may inform early intervention strategies targeting maternal and offspring health disparities.

CHROMATIN PROFILING FROM FORMALIN-FIXED PARAFFIN-EMBEDDED SAMPLES FOR BIOMARKER DISCOVERY

Eva Brill, Emily A Madden, Alysha E Simmons, Vishnu U Sunitha Kumary, Martis W Cowles, Bryan J Venters, Michael-Christopher Keogh

EpiCypher, Inc., Durham, NC

Understanding and profiling the epigenetic landscape is increasingly important for clinical research as the medical field moves towards a precision medicine model of patient care. Cell function and fate are regulated at the chromatin level by chromatin structure (e.g., accessibility) and epigenomic features, including histone post-translational modifications (PTMs) and chromatin-associated proteins (CAPs). Mapping the location of these features provides a powerful approach to study epigenomic mechanisms underlying health and disease. The development of tools for genomic mapping has ushered in a new era of epigenetic research, highlighting how histone PTMs and CAPs can provide deep mechanistic insights into gene regulation and be leveraged as novel biomarkers and therapeutic targets. However, widely used chromatin mapping assays, such as ATAC-seq and ChIP-seq, are technically challenging to perform with formalin-fixed paraffin-embedded (FFPE) tissue, the gold-standard method for preservation and storage of clinical samples. Banked samples from cancer clinical trials could be a resource for retrospective biomarker studies due to their association with clinical response and disease progression data. EpiCypher® is developing a modified CUT&Tag workflow that is compatible with FFPE samples (CUT&Tag-FFPE). In CUT&Tag, antibodies localize pAG-Tn5 to specific targets on chromatin. Next, Tn5 is activated to ligate sequencing adapters directly into chromatin at antibody-directed targets. This immune-tethering based assay yields exquisite sensitivity compared to ChIP-seq, the previous gold-standard for chromatin profiling. Here, we have optimized key steps in our CUT&Tag workflow to improve mapping sensitivity in heavily fixed and FFPE samples. We have successfully mapped RNA polymerase II (RNAPII), the polymerase that transcribes mRNA and non-coding RNAs, in heavily fixed cell lines, and generated chromatin maps comparable to native controls. We are currently applying our modified workflow to optimize RNAPII mapping in FFPE tissue samples such as liver, colon, and brain; reliable genomic profiles have been obtained from material as thin as 5um sections. Future work will be done to validate additional chromatin targets using matched FFPE and fresh frozen tissue samples from various tissues. In parallel, we are optimizing workflows to pair with downstream single cell and spatial platforms to expand the impact of this technology. We envision that CUT&Tag-FFPE will accelerate the discovery of epigenetic drug targets and biomarkers to advance precision medicine.

CHARACTERIZING EXTRACHROMOSOMAL DNA IN THE MALARIA PARASITE AND ITS RELATIONSHIP TO CHROMOSOMAL COPY NUMBER VARIATIONS

Noah J Brown, Caroline F Webb, Julia A Zulawiniska, Jennifer L Guler

University of Virginia, Biology, Charlottesville, VA

Malaria continues to represent a large global disease burden, infecting more than 240 million people every year. The protozoan parasite responsible for the most malaria deaths is *Plasmodium falciparum*; its persistence has largely been attributed to the ability to develop antimalarial resistance. The *P. falciparum* genome boasts one of the highest AT contents of any organism (>81%), which may contribute to its adaptive success. Our prior work has suggested that highly repetitive AT-rich regions are prone to form DNA hairpins, which increase the frequency of double strand breaks across the parasite genome. Ultimately, error-prone break repair by microhomology-based pathways increases rates of copy number variations (CNVs), particularly tandem head-to-tail amplifications. These CNVs contribute to the evolution of the parasite including the acquisition of drug resistance and ability to infect new hosts. Subsequently, large stretches of homology between tandem CNVs may trigger the formation of extrachromosomal DNA (ecDNA), which provides additional benefits such as facile tuning and enhanced expression. Recently, we identified ecDNA in *P. falciparum* parasites that harbor resistance-conferring tandem genomic CNVs at one locus on chromosome 6. We determined that *P. falciparum* ecDNA is complex in nature, containing both single and double stranded elements, but limitations in gel-based ecDNA purification impeded further study of ecDNA characteristics and frequency. Here, we established an gel-free enrichment pipeline capable of isolating ecDNA molecules. Using endogenous non-linear organellar genomes as positive controls (mitochondrial and apicoplast genomes), we assessed the sensitivity of the assay using parasite lines with differing levels of chromosome 6-derived ecDNA. Further, we investigated whether ecDNA can be formed from other tandem chromosomal CNVs in the parasite genome. Finally, we sequenced enriched ecDNAs to identify characteristics that contribute to their generation and maintenance in the *P. falciparum* parasite. Defining the characteristics of *P. falciparum* ecDNA is critical for understanding its generation and function in this eukaryotic microbe as well as other contexts, such as cancer progression. Ultimately, a better understanding of genomic instability and adaptation can lead to novel ways to prevent and control many diseases.

MOLECULAR QTL ANALYSIS OF EXPRESSION, SPLICING, AND CHROMATIN ACCESSIBILITY IN HUMAN CHONDROCYTE IDENTIFY NOVEL PUTATIVE OSTEOARTHRITIS RISK GENES

Seyoun Byun[1], Nicole E Kramer[1], Philip Coryell[1], Susan D'Costa[1], Eliza Thulson[1], Susanna Chubinskaya[2], Karen L Mohlke[1], Brian O Diekman[1], Richard F Loeser[1], Douglas H Phanstiel[1]

[1]University of North Carolina, Chapel Hill, NC, [2]University of Texas Medical Center-Galveston, Galveston, TX

Osteoarthritis (OA) is a leading cause of disability affecting over 500 million people worldwide. Despite extensive studies, current treatments remain limited due to the poorly understood molecular mechanisms driving the disease. Genome-wide association studies (GWAS) have identified over 100 loci associated with OA, but their functional interpretation remains challenging as most variants are in non-coding regions. The present study aimed to decipher the molecular mechanisms underlying OA by generating and integrating multiple layers of genomic regulation, including gene expression (eQTLs), RNA splicing (sQTLs), and chromatin accessibility (caQTLs), along with 3D chromatin architecture.

We established an ex vivo model using primary human chondrocytes isolated from 126 tissue donors. Cells were treated for 18 hours with either PBS (resting) or fibronectin fragment (FN-f) - a known OA trigger that mimics cartilage degradation products. Multi-omic profiling included RNA-seq (101 donors), ATAC-seq (21 donors), and Hi-C (4 donors). Differential analysis between FN-f and PBS conditions revealed 1,435 differentially expressed genes (padj < 0.01, $|\log_2 FC| > 0.2$), 974 differentially spliced intron junctions (padj < 0.05, $|\Delta PSI| > 0.15$), 22,232 differentially accessible regions (padj < 0.05, $|\log_2 FC| > 0.15$), and 53 newly gained chromatin loops ($p < 0.05$). These changes were enriched in pathways critical to OA pathogenesis, including extracellular matrix organization, inflammatory response, and cartilage development. Key OA-related genes such as *IL6*, *MMP13*, and *JUN* showed significant changes across multiple molecular layers.

Integration of genetic variation with molecular phenotypes identified 330,945 eQTLs (3,782 eGenes), 7,188 sQTLs (3,056 sGenes), and 2,655 caQTLs. Notably, genotype-treatment interaction analysis revealed condition-specific effects: 262 eGenes and 200 sGenes were uniquely associated with the FN-f condition. Colocalization analysis (PP4 > 0.7) with OA GWAS loci identified 13 high-confidence eQTLs, including novel associations with *ABCA5*, *ABCA9*, *ABCA10*, and *PAPPA*. Of particular interest, PAPPA demonstrated both eQTL colocalization and long-range chromatin interactions. Additional analysis of splicing regulation revealed 6 colocalized sQTLs, including the novel candidate gene *PBRM1*.

Taken together, our FN-f model system and comprehensive molecular QTL analysis provide novel insights into OA pathogenesis and regulatory mechanisms. This multi-omic resource identifies potential therapeutic targets and establishes a framework for future mechanistic studies in OA drug development.

DISCOVERY OF RNA DOMAINS THAT HARBOR RELATED FUNCTIONS USING hmSEEKR

Shuang Li[1,2], Quinn E Eberhard[1,2], J. Mauro Calabrese[1,2]

[1]RNA Discovery Center, Chapel Hill, NC, [2]University of North Carolina, Pharmacology, Chapel Hill, NC

Long noncoding RNAs (lncRNAs) play critical roles in gene regulation across kingdoms of life. However, lncRNAs with related functions often lack linear sequence similarity, making it challenging to leverage studies of one lncRNA to inform the understanding of others. We describe a k-mer-based hidden Markov model, hmSEEKR, that enables scanning of transcriptomes for regions of non-linear sequence similarity to a query domain, without prior knowledge of where within the transcriptome the similarities may be located. When individual lncRNA domains are used as search features, hmSEEKR successfully identifies regions in other RNAs that harbor non-linear sequence similarity and bind similar sets of proteins. hmSEEKR can accelerate the functional characterization of noncoding transcriptomes by enabling the a priori discovery of RNA domains that may encode related functions.

IDENTIFICATION OF OPTIMAL EXPERIMENTAL CONDITIONS BY ESTABLISHMENT OF SINGLE-POT AUTOMATED (SPA)-ChIP-SEQ

Yuwei Cao[1], Lauren Patel[1], Lauren Alcoser[2], Eric Mendenhall[3], Christopher Benner[1], Sven Heinz[1], Alon Goren[1]

[1]UC San Diego, Department of Medicine, La Jolla, CA, [2]Agilent, Technologies, Santa Clara, CA, [3]HudsonAlpha, Institute for Biotechnology, Huntsville, AL

ChIP-seq is a well-established method for studying genomic localization of histone modifications and DNA-associated proteins. While ChIP-seq is highly useful, the overall experimental workflow involves multiple steps that increase the risk of introducing inconsistency within experiments and between groups. These challenges were partially addressed by the incorporation of robotic liquid handlers to improve the robustness of the process. Yet, some of these protocols have automated only a subset of the steps. Moreover, most of ChIP-seq protocols, both automated and manual, were limited in their ability to efficiently map non-histone DNA-associated proteins, such as chromatin regulators that indirectly interact with the genome.

Recently, we developed a single-pot ChIP-seq that incorporates an on-bead library preparation step, reducing both the workflow time and and costs (~$60 per sample) to allow scaling up the number of conditions tested. Here, we adapted this single-pot ChIP-seq protocol to be operated by a liquid handler and created an end-to-end fully automated version, namely Single-Pot-Automated ChIP-seq (SPA-ChIPseq). We benchmarked the automated protocol by performing parallel manual histone modification ChIP-seq and demonstrated a high reproducibility and nearly indistinguishable signal-to-noise ratio between manual and automated workflows. As a proof of principle, we used SPA-ChIPseq to evaluate multiple parameters, including shearing and crosslinking conditions, buffer compositions, and antibody ratios. SPA-ChIPseq allowed us to identify optimal conditions for dual crosslinked chromatin, enhancing the ability to capture weak interactions between DNA-associated proteins. Using SPA-ChIPseq, we robustly titrated amounts of antibodies to chromatin cell equivalent and identified that the major variations between these ratios could have an impact on the signal to noise ratio.

Together, in this study we demonstrate the ability of automating a multi-step protocol to systematically survey various parameters and identify optimal conditions. Our SPA-ChIPseq protocol will be publicly available and will include specific deck setups, software files and parameters. Note, while this protocol is defined for a specific liquid handler, this approach is applicable to any programmable liquid handler. Lastly, we envision that our robust protocol can advance research by enabling core facilities to provide ChIP-seq as a service and can potentially be incorporated into compound screening and diagnostic frameworks.

ESTIMATING *CIS* AND *TRANS* CONTRIBUTIONS TO DIFFERENCES IN GENE REGULATION

Ingileif Hallgrimsdottir[1], Maria Carilli[1], Lior Pachter[1,2]

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

In 1961, Jacob and Monod developed a theory of gene regulation in which they distinguished local effects (*cis*) from distal regulation (*trans*), immediately raising the question of the relative contributions of these two regulatory strategies. One approach to assessing *cis* or *trans* contributions to differences in gene expression between strains or species is to compare gene expression ratios in parents to allele-specific ratios in F1 hybrids. This approach was used to explore gene regulation differences between mouse (Cowles et al., 2002), fly (Massouris et al., 2012), and yeast strains (Tsouris et al., 2024), and even between humans and chimpanzees (Barr et al., 2023).

We show that a linear transformation is needed to account for an asymmetry between *cis* and *trans* and explain how this leads naturally to a hypothesis testing framework for single and multi-sample experiments. Using our approach, we obtain markedly different quantitative and qualitative results than prior testing methods.

COMMON VARIATION IN CORE MEIOSIS GENES SHAPES HUMAN RECOMBINATION PHENOTYPES AND ANEUPLOIDY RISK

Sara A Carioscia, Arjun Biddanda, Margaret Starostik, Rajiv C McCoy

Johns Hopkins University, Department of Biology, Baltimore, MD

Only approximately half of human conceptions survive to birth. The leading cause of pregnancy loss is aneuploidy, often tracing to chromosome missegregation in female meiosis and increasing with maternal age. While it has long been known that abnormal meiotic crossover recombination confers aneuploidy risk, limited data have precluded a more complete understanding of the links between these fundamental phenotypes and their potential shared genetic basis. To address this gap, we performed retrospective analysis of preimplantation genetic testing data from 139,416 in vitro fertilized embryos from 22,850 sets of patients and partners. Using a Bayesian approach to trace the transmission of haplotypes from parents to offspring, we identified 3,656,198 (2,226,218 maternal, 1,429,980 paternal) crossovers and 115,507 chromosomal aneuploidies of maternal meiotic origin. Patterns of crossovers were altered in aneuploid versus euploid embryos, consistent with their role in chromosome cohesion. Our analyses further revealed that a common recombination-associated haplotype spanning the meiotic cohesin SMC1B is significantly associated with maternal meiotic aneuploidy ($P = 3.2 \times 10^{-8}$), driven by two independent causal cis-regulatory mutations, including one within the SMC1B promoter. Beyond SMC1B, predicted expression of synaptonemal complex component C14orf39 and condensin component NCAPD2 were also associated with aneuploidy incidence. More broadly, we found that recombination phenotypes and meiotic aneuploidy possess a partially shared genetic basis that also overlaps with other female reproductive traits such as ages at menarche and menopause. These fitness-related traits are together depleted of heritability, consistent with quantitative genetic theory. Our findings highlight the dual role of meiotic recombination in generating genetic diversity, while ensuring fidelity of chromosome segregation.

CHARACTERIZATION OF HAIRPIN LOOPS AND CRUCIFORMS ACROSS 118,065 GENOMES SPANNING THE TREE OF LIFE

Nikol Chantzi[1], Camille Moeckel[1], Candace Chan[1], Akshatha Nayak[1], Guliang Wang[2], Ioannis Mouratidis[1], Dionysios Chartoumpekis[3], Karen M Vasquez[2], Ilias Georgakopoulos-Soares[1]

[1]The Pennsylvania State University, College of Medicine, Department of Biochemistry and Molecular Biology, Hershey, PA, [2]The University of Texas at Austin, Dell Pediatric Research Institute, Division of Pharmacology and Toxicology, College of Pharmacy, Austin, TX, [3]School of Medicine, University of Patras, Division of Endocrinology, Department of Internal Medicine, Patras, Greece

Inverted repeats (IRs) can form alternative DNA secondary structures called hairpins and cruciforms, which have a multitude of functional roles and have been associated with genomic instability. However, their prevalence across diverse organismal genomes remains only partially understood. Here, we examine the prevalence of IRs across 118,065 complete organismal genomes. Our comprehensive analysis across taxonomic subdivisions reveals significant differences in the distribution, frequency, and biophysical properties of perfect IRs among these genomes. We identify a total of 29,589,132 perfect IRs and show a highly variable density across different organisms, with strikingly distinct patterns observed in Viruses, Bacteria, Archaea, and Eukaryota. We report IRs with perfect arms of extreme lengths, which can extend to hundreds of thousands of base pairs. Our findings demonstrate a strong correlation between IR density and genome size, revealing that Viruses and Bacteria possess the highest density, whereas Eukaryota and Archaea exhibit the lowest relative to their genome size. Additionally, the study reveals the enrichment of IRs at transcription start and termination end sites in prokaryotes and Viruses and underscores their potential roles in gene regulation and genome organization. Through a comprehensive overview of the distribution and characteristics of IRs in a wide array of organisms, this largest-scale analysis to date sheds light on the functional significance of inverted repeats, their contribution to genomic instability, and their evolutionary impact across the tree of life.

ANCIENT GENE DESERTS AND CONSERVED MICROSYNTENY SURROUNDING MAMMALIAN NEURODEVELOPMENTAL GENES

Margaret Chapman[1], Eirene Markenscoff-Papadimitriou[2], E. Josephine Clowney[3]

[1]University of Michigan Medical School, Neuroscience Graduate Program, Ann Arbor, MI, [2]Cornell University College of Agriculture and Life Sciences, Molecular Biology and Genetics, Ithaca, NY, [3]University of Michigan, Molecular, Cellular, and Developmental Biology, Ann Arbor, MI

Mammalian genomes have striking variation in AT/GC content and in the linear distance between neighboring genes along the chromosome. As we showed recently, the human genome houses genes of different kinds in two distinct environments: 1) GC-rich regions with genes that retain pronounced CpG islands in their promoters and that fulfill housekeeping roles inside the cell (e.g. actin, transcription factors), versus 2) AT-rich regions with genes that lack CpG islands in their promoters and that fulfill functions outside of the cell (e.g. immune receptors, chemosensors, digestive enzymes). Here, we identify a third class: enormous stretches of AT-rich content housing only 1 gene that retains its CpG island and is extremely isolated from other genes. In mammalian genomes, the vast majority of genes are within 100 kb of the nearest protein-coding gene, but roughly 20 genes are over 1.5 Mb away from any other protein-coding gene in the human genome. Interestingly, all but a few of the genes therein are implicated in neural development, with several explicitly fulfilling cell adhesion functions. How did these genes come to inhabit such exotic genomic environments, and what does this mean for brain development and evolution at large? Currently, it is not known what sequences besides protein-coding genes reside within these isolated arrays and how evolution of the genomic organization of the constituent genes has impacted their transcriptional regulation. Recent work on chromatin regulators, like POGZ and KDM6B, has demonstrated that neurodevelopmental genes with this kind of genomic organization require unique activation mechanisms to be transcribed, suggesting an explicit functional purpose for these seas of AT content. We have used synteny of the neurodevelopmental genes, the nearby gene deserts, and flanking genes that lie on either side of the deserts, to track this extreme genomic organization over evolutionary time. Despite the ambiguous function of the “empty” deserts, their presence is surprisingly ancient and pervasive. In ongoing studies, we continue to interrogate the origin, conservation, and function of the “lonely” organization of these neurodevelopmental genes in animals.

LONG-READ *DE NOVO* ASSEMBLY AND COMPARATIVE ANALYSIS OF SIX HOWLER MONKEY GENOMES WITHIN GENUS *ALOUATTA*

Bide Chen[1], Patrícia Domingues de Freitas[2], Ellie Armstrong[3], Bernard Kim[4], Luana Portela[2], Amy Goldberg[1]

[1]Duke University, Evolutionary Anthropology, Durham, NC, [2]Federal University of São Carlos, Genetics and Evolution, São Paulo, Brazil, [3]University of California Riverside, Evolution, Ecology, and Organismal Biology, Riverside, CA, [4]Princeton University, Ecology and Evolutionary Biology, New Haven, CT, [5]Federal University of São Carlos, Genetics and Evolution, São Paulo, Brazil, [6]Duke University, Evolutionary Anthropology, Durham, NC

As our closest relatives, primate genomics and evolution inform human genetic interpretation. Though only one species in the genus Homo exists today, multiple monkeys have wide radiations, allowing us to compare closely related species within a genus. Despite a growing number of primate genomes available, most genera are represented by a single reference genome. Here, we generate de novo reference assemblies using long-read sequencing for 6 Howler monkeys (genus *Alouatta*) to understand within-genus, between-species evolution. With between 9 to 14 species howler monkeys are some of the most widely distributed platyrrhines, extending from southern Mexico to northern Argentina, inhabiting a range of environments. Prior phylogenetic investigations into howler monkeys have exhibited discrepancies across studies or remained inconclusive, based a mix of on cytogenetic, morphological traits, and a restricted set of molecular markers. Attempts to address these questions have been limited because only a single poor-quality reference genome of mantled howler monkey (*Alouatta palliata*, Mexico) is currently available (assembly contig count: 1152695; N50: 51.304 Kbp). Our assemblies contain roughly ~2000 contigs, with N50 of ~15Mb and L50 of ~60. Based on phylogeny constructed with OrthoFinder, IQtree, and ASTRAL, we find the clade grouping *A. belzebul* and *A. belzebul /discolor* as sister taxa, with *A. belzebul/ululate*, *A. guariba*, *A. seniculus*, and *A. caraya* as respective sister lineages to this clade. We identify loci that show species-specific signatures of selection, as well as adaptive loci shared across the genus. In addition, we are catalog species-specific and genus-specific structural variation that may be associated with howler-specific phenotypic traits, such as immunity, olfaction, and visual perception.

EFFICIENT TELOMERE-TO-TELOMERE ASSEMBLY OF ONT SIMPLEX READS USING HIFIASM (ONT)

Haoyu Cheng[1], Heng Li[2,3]

[1]Yale School of Medicine, Department of Biomedical Informatics and Data Science, New Haven, CT, [2]Dana-Farber Cancer Institute, Department of Data Science, Boston, MA, [3]Harvard Medical School, Department of Biomedical Informatics, Boston, MA

High-quality telomere-to-telomere (T2T) assembly is the ultimate goal of de novo genome assembly. Existing T2T assembly algorithms achieve this by leveraging both highly accurate PacBio HiFi reads and longer, but less accurate, ONT ultra-long simplex reads, combining the strengths of both technologies. However, generating ONT ultra-long reads is costly and requires a substantial DNA input, making it impractical for many clinical and large-scale studies. Moreover, due to recurrent, non-random sequencing errors in ONT simplex reads, existing T2T assembly algorithms cannot directly assemble them in a de novo manner, thereby underutilizing their potential. The recent deep learning-based error correction tool, HERRO, can correct ONT simplex reads for de novo assembly. However, HERRO requires extensive computational resources and high-end GPUs, making it impractical for many genome assembly projects.

Here, we propose hifiasm (ONT), an ultra-fast algorithm that directly corrects and assembles ONT simplex reads without relying on deep learning. Our results demonstrate that hifiasm (ONT) is an order of magnitude faster than HERRO-based assembly while eliminating the need for high-performance GPUs. Most importantly, we show that hifiasm (ONT) enables T2T assembly using only non-ultra-long, standard ONT reads, which are significantly more cost-efficient and easier to obtain. These improvements greatly expand the applicability of T2T assembly to real-world clinical settings and large-scale genomic studies.

FOXO1 REGULATES INTESTINAL TISSUE-RESIDENT MEMORY CD8 T CELL BIOLOGY IN AN ANATOMIC COMPARTMENT- AND CONTEXT-SPECIFIC MANNER

Paul Hsu[2], Eunice Choi[1], William Wong[2], Yun Hsuan Lin[2], Sara Vandenburgh[2], Yi Chia Liu[2], Priscilla Yao[2], Cynthia Indralingam[2], Gene Yeo[4], Elina Zuniga[3], Ananda Goldrath[3], Wei Wang[1,4], John Chang[2,5]

[1]University of California San Diego, Chemistry and Biochemistry, La Jolla, CA, [2]University of California San Diego, Medicine, La Jolla, CA, [3]University of California San Diego, Biological Sciences, La Jolla, CA, [4]University of California San Diego, Cellular and Molecular Medicine, La Jolla, CA, [5]Jennifer Moreno Department of Veteran Affairs Medical Center, Medicine, San Diego, CA

Tissue-resident memory CD8 T cells (TRM) serve as a frontline defense against microbial pathogens in barrier and mucosal tissues. Predicting the tissue-specific roles of transcription factors (TFs) that regulate TRM remains a challenge. While some TFs are broadly required for TRM formation, others have specialized functions depending on the microenvironment. By integrating gene expression and chromatin access, we predicted Foxo1 as a key compartment-specific regulator of intestinal TRM. Foxo1 is well known for supporting circulating memory T cells, but its role in tissue-resident subsets remains less understood. Our computational predictions, validated through Foxo1 knockout models, revealed an unexpected dual role in intestinal TRM regulation. We found that Foxo1 represses early maintenance of small intestinal intraepithelial TRM (siIEL), despite its broader function in promoting intestinal TRM formation. This contrasts with its role in the small intestinal lamina propria and colon, where it sustains TRM populations. These findings suggest Foxo1's function is highly context-dependent, regulated by factors beyond expression levels alone. Potential mechanisms include differential co-factor interactions, chromatin accessibility, and post-translational modifications. Our study also highlights an underappreciated role for CD103 in siIEL TRM survival. While CD103 is typically linked to TRM retention in epithelial niches, our findings suggest it may also provide critical survival cues in certain tissue compartments, expanding our understanding of integrin-mediated TRM persistence. These insights reveal that TRM transcriptional regulation is more nuanced than previously appreciated, emphasizing the need to consider tissue-specific regulatory networks rather than assuming uniform TF functions across compartments. Understanding these context-dependent mechanisms has broad implications for immunotherapy and vaccine design, particularly in targeting TRM responses in barrier tissues. By refining our ability to modulate TRM function through transcriptional regulation, these findings may inform strategies to enhance protective immunity or mitigate immune-related pathology.

PRIORITIZING NONCODING VARIANT-GENE PAIRS IN PSORIASIS USING COUPLED MATRIX-MATRIX COMPLETION.

Elysia Chou[1], Andre Guerra[2], Zhaolin Zhang[3], Tingting Qin[1], Shiting Li[1], Kai Wang[1], James T Elder[3,4], Lam C Tsoi[1,2,3], Maureen A Sartor[1,2]

[1]University of Michigan Medical School, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, [2]University of Michigan School of Public Health, Department of Biostatistics, Ann Arbor, MI, [3]University of Michigan Medical School, Department of Dermatology, Ann Arbor, MI, [4]Ann Arbor VA Hospital, Department of Dermatology, Ann Arbor, MI

Psoriasis is a complex inflammatory skin disease with a prevalence of ~3.0% among US adults.
GWAS signals for psoriasis are highly enriched in gene regulatory regions in CD4+ and CD8+ T cells, suggesting that linking these GWAS variants to their target genes will further our understanding of the molecular mechanisms driving psoriasis. Traditional approaches such as eQTL first link variants to genes and require additional ad-hoc steps to nominate variant-gene pairs in association with a specific disease, thus not using disease information to inform variant-gene pair prioritization. Disease-specific methods such as fine-mapping and MAGMA prioritize either only variants or genes, respectively. Here, we developed a novel data fusion approach, DisCO-VG (Disease-specific CMMC Optimization for Variant-Gene pairs), that prioritizes cell type-specific, disease-associated noncoding variant-gene pairs. DisCO-VG leverages Coupled Matrix-Matrix Completion (CMMC), an algorithm that has proven effective in predicting drug-target interactions. DisCO-VG scores, which represent the strength of a disease-associated variant-gene pair, were computed by integrating various information sources: variant-variant associations from ATAC-Seq, gene-gene associations based on Gene Ontology (GO) term similarities, and variant-gene associations from eQTLs and enhancer-gene links, while considering GWAS significance and distance between variant and gene. We have applied DisCO-VG to study psoriasis, using data from peripheral blood mononuclear cells of both healthy and psoriatic samples. Thus far, we have identified psoriasis-associated variant-gene pairs in CD4+ T cells, rigorously evaluating the predictions using independent datasets at multiple levels: at the gene level with MAGMA-prioritized psoriasis genes, at the variant level with RegulomeDB annotations, and at the variant-gene level using ENCODE-rE2G data. Results demonstrated that GWAS lead variants and known psoriasis genes tended to score higher, with DisCO-VG at times linking known psoriasis genes to variants with more functional evidence rather than the lead SNVs. These findings highlight DisCO-VG as a robust tool for nominating cell type-specific, putatively causal variant-gene pairs for experimental validation and as a basis for potential drug targets.

SPATIAL DRIVERS OF RESPONSE TO CANCER IMMUNOTHERAPY

Francesca D Ciccarelli[1,2]

[1]The Francis Crick Institute, Cancer Systems Biology, London, United Kingdom, [2]Barts Cancer Institute, Centre for Cancer Evolution, London, United Kingdom

The use of immune checkpoint inhibition (ICI) therapy has revolutionised the treatment of cancer. However, not all cancer patients are eligible to receive ICI and only a fraction of eligible patients will respond to it. For example, ICI is used as first line treatment of mismatch repair deficient (dMMR) or microsatellite instability high (MSI-H) colorectal cancer (CRC) at advanced or metastatic stage. These tumours have high tumour mutational burden (TMB) that is thought to favour response to ICI because a high load of somatic mutations leads to increased production of peptide neoantigens able to initiate an immune response. However, fewer than 50% of dMMR CRC patients respond to ICI Moreover, dMMR CRCs constitute less than 10% of all CRCs, leaving the vast majority of patients with a proficient mismatch repair (pMMR) CRC ineligible to ICI treatment. Identifying and expanding the patient population benefitting from ICI remains a pressing clinical need.

We and others have previously shown that TMB does not correlate with immune infiltration while the type and abundance of immune populations that infiltrate the tumour play a major role in determining response. dMMR CRCs responding to ICI treatment are enriched in tumour-associated macrophages (TAMs) interacting with T cells. Yet, the factors controlling immune infiltration remain largely unknown.

In this study, we compared the tumour and immune landscapes of dMMR and pMMR CRCs to gain a deeper understanding of immune infiltration is regulated and what factors determine response to ICI. We performed multi-regional transcriptomics of tumour epithelium (Te) and tumour-associated stroma (Ts) separated by laser capture microdissection (LCM). This enabled comprehensive comparison of intrinsic (i.e., constitutive of the tumour) and extrinsic (i.e., related to the TIME) properties between and within regions of the two CRC subtypes. Using single-cell spatial transcriptomics, in vitro cell co-cultures and mouse models, we defined correlates of ICI response that stratify patient outcomes independently on TMB or MMR status, thus in principle simultaneously enhancing targeted intervention and enlarging the patient population benefitting from ICI.

DECIPHERING THE AUTISM-ASSOCIATED GENE REGULATORY LANDSCAPE

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

[1]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, NJ, [2]Graduate Program in Cell and Developmental Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, [3]Department of Neuroscience and Cell Biology, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ, [4]Department of Biochemistry and Molecular Biology, Rutgers, The State University of New Jersey, Piscataway, NJ

Autism spectrum disorder (ASD) is a neurodevelopmental condition with complex genetic underpinnings. To elucidate the regulatory mechanisms contributing to ASD, we performed a multi-omics analysis integrating ATAC-seq, ChIP-seq for H3K27ac, and RNA-seq across induced pluripotent stem cells (iPSCs), neural progenitor cells (NPCs), and induced neurons (iNs) derived from NIH-originated control and idiopathic control samples. Our principal component analysis revealed distinct molecular signatures that distinguish cell types within and across origins. Moreover, differential accessibility and differential expression analyses identified cell type and origin-specific chromatin and transcriptional landscapes. Finally, functional enrichment analyses highlighted developmental pathways specific to each cell type, with idiopathic control samples exhibiting unique regulatory features.

In addition, we constructed enhancer-promoter interaction (EPI) networks and mapped ASD-associated de novo variants (DNVs) to these networks. ASD-associated DNVs were significantly enriched in NPC enhancers, implicating key transcription factor binding motifs such as RORA and ZNF423 in neurodevelopmental regulation. Motif disruption analyses suggest that ASD-associated variants may perturb regulatory interactions essential for neuronal differentiation. We observed enrichment for ASD-associated EPI interactions in iNs. Altogether, these findings provide insight into the epigenetic and transcriptional mechanisms underlying ASD and demonstrate the potential impact of genetic variation on neurodevelopmental regulatory networks.

In summary, our findings reveal fundamental differences in chromatin accessibility, transcriptional regulation, and EPIs across distinct cell types and sample origins. The enrichment of ASD-associated variants and genes in iN and NPC regulatory regions suggests a potential mechanistic link between chromatin dynamics and ASD susceptibility. By integrating multi-omics data with genetic variation analysis, this study provides a framework for identifying key regulatory elements and molecular pathways disrupted in ASD. Future investigations should explore how these epigenetic alterations contribute to ASD phenotypes and assess whether they can be targeted for therapeutic interventions.

REFERENCE-QUALITY GENOMES OF HUMAN CELL LINES FOR PRECISION OMICS

Luca Corda[1], Emilia Volpe[1], Alessio Colantoni[1], Elena Di Tommaso[1], Franca Pelliccia[1], Riccardo Ottalevi[2], Danilo Licastro[3], Giulio Formenti[4], Mattia Capulli[5], Andrea Guarracino[6], Evelyne Tassone[1], Simona Giunta[1]

[1]Sapienza University of Rome, Department of Biology and Biotechnology "Charles Darwin", Rome, Italy, [2]Dante Genomics Corp Inc., Department of Bioinformatic, New York, NY, [3]Area Science Park, Genomics and Epigenomics Laboratory, Trieste, Italy, [4]The Rockefeller University, The Vertebrate Genome Laboratory, New York, NY, [5]University of L'Aquila, Department of Biotechnological and Applied Clinical Sciences, L'Aquila, Italy, [6]University of Tennessee, Department of Genetics, Genomics and Informatics, Memphis, TN

The human pangenome draft revealed extensive sequence polymorphism and variation between individuals, with DNA changes affecting over 40% of all coding regions and peaking within repetitive elements. Given this genomic diversity, the current reliance on a single reference genome for multi-omics analyses introduces significant biases.
Here, we present a novel approach in human genomics that improves read alignment by using matched reference genomes to generate "isogenomic" alignments of sequencing data. As a proof of concept, we applied this strategy using a newly phased assembly of RPE-1, one of the most widely used non-cancer cell lines. Aligning reads to the RPE1v1.1 genome significantly enhances alignment quality, enables accurate haplotype-specific read assignment, and improves peak calling accuracy. This approach uncovers inter-haplotype variation and provides unprecedented precision in CRISPR guide design, which is not achievable using CHM13 or other non-matched reference genomes.
This study underscores the value of matched reference genomes for multi-omics analyses and highlights the need for comprehensive assembly of experimentally relevant cell lines to enable widespread adoption of isogenomic reference genomes.

SIMPHENY: INTEGRATING LARGE-SCALE GENOMIC AND PHENOTYPIC DATA FOR RARE DISEASE VARIANT PRIORITIZATION

Isabelle B Cooperstein[1], Alistair Ward[1,2], Shilpa N Kobren[3], Barry Moore[1], Undiagnosed Diseases Network[4], Gabor Marth[1]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]Frameshift Genomics, Boston, MA, [3]Harvard Medical School, Biomedical Informatics, Boston, MA, [4]National Institutes of Health, Bethesda, MD

Background: Despite advances in next-generation sequencing, over 50% of rare disease cases remain unresolved. The primary challenge has shifted from variant detection to accurate variant prioritization and interpretation. In many cases, identifying a phenotypically and genotypically similar patient is crucial for diagnosis and advancing knowledge of a rare disease. We introduce SimPheny, a novel phenotype-first algorithm that automates patient matching to prioritize genetic variants for diagnostic review. We validated this approach using harmonized, jointly-called whole-exome (WES) and whole-genome sequencing (WGS) data from thousands of families in the Undiagnosed Diseases Network (UDN) cohort. Methods: Pairwise phenotypic similarity scores were calculated using the Human Phenotype Ontology (HPO) terms curated by UDN clinical teams. Multi-sample VCF files for each case - including affected probands and relevant affected and unaffected family members - were extracted from the cohort-level WES or WGS variant datasets. We ran Exomiser on 404 diagnosed UDN probands to generate gene lists that mimic candidate gene lists containing variants of uncertain significance (VUS) typically encountered in undiagnosed cases, but with knowledge of the true diagnostic gene. To refine these Exomiser-generated gene lists, we identified overlaps with the diagnostic genes of phenotypically similar individuals in a background dataset of 768 diagnosed UDN participants and ~10,000 pathogenic or likely pathogenic ClinVar submissions. Gene matches were classified as true or false positives based on known diagnoses. Statistical analyses were performed to assess the relevance of phenotypic similarity scores and gene hits between these matches, optimizing false discovery rate (FDR) thresholds to distinguish true positives. We then applied this methodology to 1,445 unsolved UDN cases to identify candidate diagnostic genes. Results: SimPheny significantly reduced the number of candidate genes per participant, achieving a median of one prioritized gene with high precision. Under a strict significance threshold, SimPheny identified 44 true positive gene matches, eight false positive, and seven candidate gene matches in undiagnosed cases. To date, two candidate genes have been confirmed as diagnostic, while others are under clinical review. Incorporating ClinVar data provided additional support for these matches, suggesting further diagnostic evidence. Conclusions: By integrating large-scale harmonized genomic datasets with phenotype-driven computational approaches, SimPheny enables scalable, cohort-wide variant prioritization and has the potential to uncover novel phenotype-to-gene associations in rare monogenic diseases. SimPheny is publicly available as a web-based tool, facilitating seamless integration into both research and clinical diagnostic workflows.

TRANSCRIPTION START SITES EXPERIENCE A HIGH INFLUX OF HERITABLE VARIANTS FUELLED BY EARLY DEVELOPMENT

Miguel A Cortes-Guzman[1,2], David Castellano[1,3], Claudia Serrano-Colome[1,2], Vladimir Seplyarskiy[4,5], Donate Weghorn[1,2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Department of Medicine and Life Sciences, Barcelona, Spain, [3]University of Arizona, Molecular and Cellular Biology, Tucson, AZ, [4]Harvard Medical School, Division of Genetics, Brigham and Women's Hospita, Boston, MA, [5]Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Mutations drive evolution and genetic diversity, but the impact of transcription on germline mutagenesis remains poorly understood. We identified a hypermutation phenomenon at transcription start sites in the human germline, spanning several hundred base pairs in both directions. We link this TSS mutational hotspot to divergent transcription, RNA polymerase II stalling, R-loops, and mitotic—but not meiotic—double-strand breaks, revealing a recombination-independent mechanism distinct from known processes. Notably, the hotspot is absent in *de novo* mutation data. We reconcile this by showing that TSS mutations are significantly enriched with early mosaic variants often filtered out in *de novo* mutation calls, indicating that the hotspot arises during early embryogenesis. Mutational signature analysis reinforces these findings and implicates alternative non-homologous end joining and maternal mutation clusters. Our study provides the first detailed description of a germline TSS mutation hotspot, with broad evolutionary and biomedical implications.

GENETIC EFFECTS ON THE TRANSCRIPTIONAL RESPONSE TO IMMUNE CHALLENGE IN THE RHESUS MACAQUE

Christina E Costa[1,2], Mitchell R Sánchez Rosado[3], Rachel M Petersen[4], Marina M Watowich[4], Josue E Negron-Del Valle[5], Daniel Phillips[5], Michael Platt[6], Michael J Montague[6], Lauren J N Brent[7], James P Higham[1,2], Noah Snyder-Mackler*[5], Amanda J Lea*[4]

[1]New York University, Anthropology, New York, NY, [2]New York Consortium in Evolutionary Primatology, New York, NY, [3]University of Puerto Rico, Microbiology and Medical Zoology, San Juan, PR, [4]Vanderbilt University, Biological Sciences, Nashville, TN, [5]Arizona State University, School of Life Sciences, Tempe, AZ, [6]University of Pennsylvania, Neuroscience, Philadelphia, PA, [7]University of Exeter, Psychology, United Kingdom

*Authors contributed equally

Individuals differ in their ability to respond to pathogens. While factors like age and sex influence these differences, a large proportion of immune variation is under genetic control, with non-coding variants playing a significant role. Work in humans has found that some genetic effects on gene regulation are only visible when cells mount an immune response. These "context-dependent" effects are particularly relevant to disease, but remain poorly studied in natural primate populations, limiting our understanding of their evolutionary conservation and contribution to phenotypic variation. Here, we investigated *cis* genetic effects on gene expression (eQTL) in unstimulated (n=247) and lipopolysaccharide (LPS; n=252), dexamethasone (Dex; n=267), Zika virus (n=184), and *F. hepatica* parasite (n=168) stimulated peripheral blood mononuclear cells from the free-ranging rhesus macaque population of Cayo Santiago, Puerto Rico. We hypothesized that eQTL linked to immune response genes would show stronger effects after stimulation. We tested for effects of 7,494,456 SNPs within 200kb of 11,212 genes in the RNA-seq dataset (22,379,904 SNP-gene pairs). Treatments resulted in significant transcriptional change; 4,143 genes were differentially expressed after LPS (FDR < 0.01), and showed expected inflammatory pathway enrichment (e.g., cAMP signaling). We found 16,272 *cis* eQTL (FDR < 0.05) affecting the expression of 424 genes across conditions (Control: 9,873; LPS: 8,974; Dex: 7,366; *F. hepatica*: 143; Zika: 0). Only 115 eQTL (3 genes) were shared across four conditions, while 6,134 eQTL (51 genes) were shared across at least two, indicating that the majority of SNP-gene associations (62.3%) were condition-specific and likely reflect context-dependent effects on immune regulation. We expanded our dataset (n=3,012 samples total) to include six additional stimuli that will allow us to tease apart specificity of genetic effects in different immune states. Ongoing work uses an empirical Bayes approach to jointly estimate genotype effect size and characterize ubiquitous versus condition-specific eQTL.

ASSEMBLING UNMAPPED READS REVEALS MISSING VARIATION IN SOUTH ASIAN GENOMES

Arun Das[1], Arjun Biddanda[2], Rajiv C McCoy[2], Michael C Schatz[1,2]

[1]Johns Hopkins University, Computer Science, Baltimore, MD, [2]Johns Hopkins University, Biology, Baltimore, MD

The rapid growth in genomics has not been uniform across human genetic ancestry groups, contributing to systemic biases that can impact biological and clinical interpretation. In this work, we examine unmapped reads from South Asian (SAS) genomes against linear and pangenome references with the goal of revealing hidden diversity and understanding its implications, origins and distribution.

Using high-coverage (30×) short-read data from 640 SAS individuals in the 1000 Genomes Project and Simons Genome Diversity Project, we investigated the variation between these individuals relative to linear and pangenome references. Using an approach modeled after the African Pan-Genome project (Sherman et al. 2019), we assembled contigs from unmapped reads and attempted to place these larger contigs in the reference, identifying variants and novel sequences.

We repeated this approach across various reference genomes including GRCh38, T2T-CHM13, and two Human Pangenome Reference Consortium (HPRC) pangenome references. Using CHM13, we observed improved alignment rates (+0.5%) relative to GRCh38. However, even for this more complete reference, we assembled ~550 kbp of sequence per individual from unmapped reads (compared to ~2 Mbp against GRCh38), totalling 400 Mbp in 200K contigs across all SAS samples. Much of this sequence is shared across individuals, with 42% of contigs present in >50% of samples, resulting in a total of ~50 Mbp of newly resolved sequence across the set. For 21 of these individuals, we validated 93% of their assembled contigs using long read data from the same individual. Against the HPRC references we assemble ~500 kbp of sequence per individual from unmapped reads, despite slight improvements in alignment rate (+0.3-1.0%). Importantly, most of these contigs are not detectable using standard alignment-based SV discovery methods.

We then attempted to place the unmapped read contigs against CHM13. Across 20,000 placed contigs, we found 8,215 intersections with 106 protein coding genes, including genes in which mutations have been associated with eye-related conditions, facial dysmorphia, ciliary function and cancers, and >15,000 placements within 1 Kbp of a known GWAS hit. Linkage disequilibrium between placed and unplaced contigs helps place a further 4,100 contigs, many of which are also close to annotated genes, regulatory elements, and GWAS hits. We also aligned RNA-seq data from 140 SAS individuals against the assembled contigs to investigate their transcriptional potential. Across the set, we found >200 placed and unplaced contigs with a high density of RNA-seq alignments.

We find that a substantial amount of sequence in SAS populations remains absent from even the most complete reference genomes. This includes both sequences private to a single SAS individual and shared sequences present in both SAS and non-SAS populations. Our efforts highlight the limitations of reference genomes and provide a model for understanding the distribution of hidden variation in any human population.

SLIDING WINDOW INTERACTION GRAMMAR (SWING): A GENERALIZED INTERACTION LANGUAGE MODEL FOR PEPTIDE AND PROTEIN INTERACTIONS

Jane Siwek, Alisa Omelchenko, Prabal Chhibbar, Alok Joglekar, Jishnu Das

University of Pittsburgh, Immunology, Pittsburgh, PA

The explosion of sequence data has allowed the rapid growth of protein language models (pLMs). pLMs have now been employed in many frameworks including variant-effect and peptide-specificity prediction. Traditionally, for protein-protein or peptide-protein interactions (PPIs), corresponding sequences are either co-embedded followed by post-hoc integration or the sequences are concatenated prior to embedding. Interestingly, no method utilizes a language representation of the interaction itself. We developed an interaction LM (iLM), which uses a novel language to represent interactions between protein/peptide sequences. Sliding Window Interaction Grammar (SWING) leverages differences in amino acid properties to generate an interaction vocabulary. This vocabulary is the input into a LM followed by a supervised prediction step where the LM's representations are used as features.

SWING was first applied to predicting peptide:MHC (pMHC) interactions. With over 10,000 MHC I and 3,000 MHC II alleles, the possible pMHC combinations are vast, making it infeasible to experimentally identify all potential pMHC interactions. SWING was not only successful at generating Class I and Class II models that have comparable prediction to state-of-the-art approaches, but the unique Mixed Class model was also successful at jointly predicting both classes. Further, the SWING model trained only on Class I alleles was predictive for Class II, a complex prediction task not attempted by any existing approach. For de novo data, using only Class I or Class II data, SWING also accurately predicted Class II pMHC interactions in murine models of SLE (MRL/lpr model) and T1D (NOD model), that were validated experimentally.

To further evaluate SWING's generalizability, we tested its ability to predict the disruption of specific edges in protein interactome networks by missense mutations. Although modern methods like AlphaMissense and ESM1b can predict interfaces and variant effects/pathogenicity per mutation, they are unable to predict edge-specific disruptions in protein networks. Predicting which missense mutations can lead to the disruption of specific protein interactions provides a fundamental genotype to phenotype link (edgotype) at a molecular level. SWING was successful at accurately predicting the impact of both Mendelian mutations and population variants on PPIs. This is the first generalizable approach that can accurately predict interaction-specific disruptions by missense mutations with only sequence information. When benchmarked against other PPI methods such as passively using protein embeddings, using only the interaction encoding, and alternative iLM architectures, only SWING was able to learn enough information to perform well across prediction tasks for missense mutation perturbation prediction and pMHC binding. Overall, SWING is a first-in-class generalizable zero-shot iLM that learns the language of PPIs.

The corresponding manuscript is currently in press at Nature Methods

UNCOVERING NOVEL CELLULAR PROGRAMS AND REGULATORY CIRCUITS UNDERLYING BIFURCATING HUMAN B CELL STATES

Zarifeh Rarani*, Swapnil Keshari*, Akanksha Sachan, Nicholas Pease, Jingyu Fan, Peter Gerges, Harinder Singh#, Jishnu Das#

University of Pittsburgh, Immunology, Pittsburgh, PA
*=co-first, #=co-corresponding

B cells upon antigen encounter undergo activation followed by a bifurcation either into extrafollicular plasmablasts (PB) that rapidly secrete low-affinity antibodies, or into germinal center (GC) cells that over a longer timeframe generate a higher affinity long-lived humoral response. We have assembled gene regulatory networks (GRNs) underlying this bifurcation using temporally resolved single cell multiomics and neural network modeling. To complement and extend this framework we analyzed transcriptomic states of GC and PB cells using SLIDE, a novel interpretable machine learning approach method to infer a small set of cellular programs (latent factors/LFs) necessary and sufficient to distinguish GC and PB cells. These LFs provide stronger discrimination between the two emergent cell states, than differential gene expression (DEG) or topic modeling. Further, LFs inferred from scRNA-seq alone or scRNA-seq in a multi-ome assay were equally robust. Interestingly, when the LF genes were cross-referenced with state-specific GRN gene linkages, the LFs recapitulated aspects of GRN architecture orchestrating the bifurcation. We also evaluated the ability of LFs to capture cellular programs that are induced in activated B cells (ABCs) and could predict acquired cell fate bias before the bifurcation. Intriguingly, the LFs captured gene programs reflective of cell-fate propensity prior to the bifurcation. These programs were validated using perturbation of key TFs including PRDM1, IRF4, IRF8, SPIB and BATF which drive the bifurcation.
To move beyond high-resolution static state-specific GRNs, we used a stochastic ODE-based framework to construct a dynamic GRN across the 5 states. In addition to recapitulating previously known lineage-defining TFs and their regulons, we identify novel regulons (e.g., those involving PAX5 and CREB3L2) as driving divergent gene activity across the bifurcation trajectory. We also combined the dynamic GRN with the inferred cellular programs (SLIDE LFs) to predict TF pairs that combinatorically control B cell fate dynamics. The inferred pairs included known TF pairs that are known to coordinately regulate extrafollicular PBs e.g., IRF4 and PRDM1 as well as novel pairs including TCF4 and BACH2. Intriguingly, several of these inferred TF pairs are not detected by conventional network topological metrics. Overall, our framework is generalizable and applicable across contexts to identify cellular programs and regulatory circuits underlying diverse cell fate bifurcations.
This work was conducted as part of the Impact of Genomic Variation in Function Consortium funded by the NHGRI.

A GENOME-WIDE VIEW OF TANDEM REPEAT VARIATION IN HUMANS AND CHIMPANZEES

Carolina de Lima Adam[1], Joana L Rocha[2], Peter H Sudmant[2], Rori Rohlfs[1,3]

[1]University of Oregon, Institute of Ecology and Evolution, Eugene, OR, [2]University of California Berkeley, Department of Integrative Biology, Berkeley, CA, [3]University of Oregon, School of Computer and Data Science, Eugene, OR

Tandem repeats (TRs) are hypothesized to be key contributors to genome evolution, driving phenotypic diversity and adaptation as ubiquitous variants with high mutation rates. Genes containing TRs in their promoter regions exhibit disproportionately high expression divergence between humans and chimpanzees, suggesting a role for TR-mediated selection. However, previous studies have been limited by short-read sequencing and accompanying genotyping tools, which cannot reliably identify TRs larger than 300bp or genotype TRs with motif lengths larger than 6bp, allowing the analysis of only a fraction of genomic TRs. We have characterized previously unidentified TR loci in humans and chimpanzees by leveraging high-fidelity long-read sequencing data and newly developed genotyping tools. We developed the pipeline TRACK to create catalogs of homologous TRs with genotypes across individuals. We recovered 967,710 TRs between human and chimpanzee T2T reference genomes, with 4.2% overlapping exons. Applying TRACK to long-read sequences, we genotyped 953,525 homologous TR with no missing data from 47 humans and 23 chimpanzees. After filtering genotypes based on read depth and motif constancy, we kept 294,506 shared TR loci, of which 93,023 were variable between species. Among these shared TRs, 50,923 have motif lengths > 6bp. Mean allele lengths are remarkably conserved between species, particularly in exonic TRs (r^2=0.976). Both species displayed substantial variation in genetic diversity estimates across motif sizes, with higher diversity in mono- and di-nucleotide repeats. Similar results were observed in short-read data, suggesting differences in mutation rates and/or selective pressures associated with motif sizes. TR loci identified in exonic regions also exhibit depleted heterozygosity relative to those in non-exonic regions, consistent with stabilizing selection on exonic TRs. Furthermore, chimpanzees exhibit higher TR heterozygosity than humans, consistent with SNP and short-read STR data trends.

To identify loci under selection, we analyzed allele lengths of TRs shared within and between species, detecting loci with a disproportionate divergence between species (suggesting recent directional selection) or high within-species diversity (suggesting balancing selection). We identified 32,162 TRs with significant length differences between humans and chimpanzees, 2.5% of which intersected exons. We are further characterizing these candidate loci based on their genomic location, motif and total length, sequence composition, and repeat constancy. These findings provide insights into TR evolution and their role in human-chimpanzee divergence.

CONSTRUCTING CELL TYPE-SPECIFIC ENHANCER-PROMOTER REGULATORY INTERACTION NETWORKS WITH MASSIVELY PARALLEL REPORTER ASSAYS

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

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

Enhancers are cis-regulatory elements, non-coding sequences of DNA, pivotal to cell type-specific gene regulation. While a consensus that enhancers are hubs for disease-associated variants has been reached, little is known about the mechanisms through which these elements mediate their effects on gene expression. Additionally, the map of these enhancers' locations and their target genes remains partial. Computational models, which extrapolate epigenetic markers to regulatory activity, have proven immensely successful in predicting enhancer-promoter interactions (E-P-Is). Still, these models require further improvements to better capture cell type specificity and generate more accurate E-P-I predictions.

Massively parallel reporter assays (MPRAs) are a powerful technique for assessing the functionality of regulatory elements and their perturbations under varied conditions (e.g., different cell types). However, computational methods for analyzing MPRA data in a cell type-specific manner remain limited. MPRA provides a unified approach for quantifying the regulatory activity of enhancers in E-P-I networks. Using a convolutional neural network trained on MPRAs from K562, HepG2, and WTC11 cell lines in combination with epigenetic datasets, we defined and scored enhancer regions and linked them to target genes. The K562 E-P-I network was benchmarked against existing models, using a CRISPR interference dataset as a ground truth. We performed a series of analyses on these three cell type-specific networks, including mapping eQTLs and GWAS-identified variants, dissecting regulatory substructures, and integrating transcription factor interactions into the network. Our framework demonstrated improved accuracy in predicting cell type-specific E-P-Is compared to existing approaches.

The incorporation of MPRA in E-P-I prediction models has the potential to enable the development of high-accuracy models that rely on fewer multi-omic datasets. Disseminating our model could encourage wider adoption and further development of MPRA, a breakthrough technology, within the scientific community. Notably, this approach is generalizable and can be applied to diverse cellular contexts.

GENOMIC ANALYSES OF HYBRIDS INDICATE CHROMOSOMAL INVERSIONS MAINTAIN GENETIC DIFFERENCES BETWEEN FIRE ANT SPECIES *SOLENOPSIS INVICTA* AND *S. RICHTERI*

Allyson Dekovich[1], Sydney Eriksson[2], Lydia Uptain[3], Margaret Staton[1], Sean Ryan[4], Kenneth G Ross[5], DeWayne Shoemaker[1]

[1]University of Tennessee, Department of Entomology and Plant Pathology, Knoxville, TN, [2]Hamilton College, Departments of Mathematics and Computer Science, Clinton, NY, [3]University of North Alabama, Departments of Biology, Chemistry, and Physics, Florence, AL, [4]Exponent, Department of Ecological and Biological Sciences, Menlo Park, CA, [5]University of Georgia, Department of Entomology, Athens, GA

Understanding the genetic basis of speciation is a fundamental goal of evolutionary biology. For speciation to occur, gene flow between two populations must be reduced or restricted, allowing for the accumulation of reproductive barriers, often resulting in complete reproductive isolation. However, studying speciation empirically remains challenging and has sparked extensive debate among evolutionary biologists. One approach to investigating reproductive isolation is through analyses of hybrid zones, geographical areas where the boundaries between two genetically distinct populations weaken, resulting in the production of viable hybrid offspring. These "natural laboratories" facilitate the exploration of reproductive isolation at the genomic level through next-generation sequencing, providing insights into the formation, genomic architecture, and maintenance of genetic barriers between species. Here, we examine a hybrid zone between the invasive fire ants, *Solenopsis invicta* and *S. richteri*, in the southeastern United States. Previous genetic studies found no evidence of hybridization in overlapping areas of their native range in South America. By utilizing reduced representation sequencing (RADSeq) and genomic cline models, we identified nearly 7000 SNP loci that exhibited impeded introgression (i.e., genomic signatures of reduced gene flow) within hybrid genomic backgrounds. Interestingly, most of these SNPs were co-localized in narrow regions on at least eight different chromosomes. By comparing measurements of genetic differentiation (FST), linkage disequilibrium (LD), and SNP association tests, we have determined that inversions—rearrangements in which DNA segments are reversed within a chromosome—likely contribute to the maintenance of species differences between *S. invicta* and *S. richteri*. Our study highlights the role of genomic rearrangements in shaping species divergence, as well as the interaction between genome architecture and reproductive isolation in two natively parapatric species that are known to successfully hybridize only in their invaded range.

ONE YEAR AFTER THE ALL OF US RESEARCH PROJECT: REFLECTIONS ON VISUALIZING HUMAN GENETIC DATA IN BIOBANKS

Alex Diaz-Papkovich[1], Shevaughn Holness[1], Sohini Ramachandran[1,2]

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

In February 2024, the All of Us Research Program Genomics Investigators published a manuscript describing the program's genomic data[1]. Meant to celebrate the data from one of the most diverse biobanks in the world, the fanfare was quickly overshadowed by controversy over a figure combining uniform manifold approximation and projection (UMAP), model-based clustering (ADMIXTURE), and questionnaire data about race and ethnicity. A debate over visualization of human population genetic variation immediately followed, largely over the use of UMAP, and how genetics visualizations can foment misinterpretation of relationships among race, ethnicity, and genetic variation and enable bad faith use of research. One year later, we reflect on the publication of the All of Us manuscript, the recent historical and social context of genetics visualizations published using biobank datasets, and what lessons have been learned from last year's debate.

To study how human genetic data from biobanks is visualized, we use the Global Biobank Meta-Analysis Initiative (GBMI), which includes 23 biobanks covering over 2.2 million individuals. We collect manuscripts that use data from the GBMI and analyze the different methods and variables used for visualization and contexts that motivated their usage. We show that visualizations that use race and ethnicity are mostly limited to American biobanks and that Bick et al. (2024), rather than being aberrant in its presentation, continued a well-established pattern of using race and ethnicity variables in its visualizations. We argue that, rather than critiquing low-dimensional projections of genetic data purely based on the underlying methodology, our research community should focus on best practices for visualizing genetic data today and in the future.

Finally, we offer recommendations on best practices for visualizations. We discuss which visualization practices from the early part of the genomic era should be abandoned by authors and/or critiqued heavily by reviewers moving forward. We offer examples of alternative visualizations of genomic data in All of Us, including using UMAP. With these figures we emphasize the many different definitions of diversity, highlight how they can be used to study genetics in social and environmental contexts, search for subtle signals in complex biobank data, and act as accurate distillations of data for researchers and readers alike.

[1] Bick et al. "Genomic data in the all of us research program." Nature 627.8003 (2024): 340-346.

ORIGIN AND MAINTENANCE OF A SHARED SEXUAL MIMICRY POLYMORPHISM

Tristram O Dodge, Molly Schumer

Stanford University & HHMI, Biology, Stanford, CA

How functional variation persists within species is a great mystery of evolutionary biology. In particular, the importance of long-term balancing selection (LTBS)—which maintains variation over long evolutionary timescales—remains controversial. Because neutral polymorphisms rarely coalesce deeper than speciation, genetic trans-species polymorphisms are often used to identify loci under LTBS, such as MHC and ABO in primates. This model assumes that variation arose in the common ancestor and has been maintained in the daughter lineages after speciation. Yet, several alternative scenarios—including introgression, convergent evolution, or turnover of genetic control—could also explain the persistence of shared polymorphic phenotypes and influence their detectability as genetic trans-species polymorphisms. While diverse shared polymorphic phenotypes are sometimes observed across species radiations (e.g., mating system type, coloration, social behavior, etc.), it is unknown how often these reflect LTBS or other evolutionary forces. Clarifying these mechanisms will improve understanding of the role of LTBS in the maintenance of variation.

We studied a shared male-limited coloration trait, the “false gravid spot,” segregating in a dozen *Xiphophorus* fish species diverged by ~6 million generations. This trait mimics the female pregnancy spot, and our behavioral experiments suggest it is balanced by male aggression and female preference. To reconstruct its evolutionary history, we conduct GWAS and QTL mapping in five distantly related *Xiphophorus* species. In four species, we find the false gravid spot maps to a narrow region upstream of *kit ligand a* (*kitlga*), a conserved pigmentation gene. Combining long-read sequencing with allele-specific expression data, we uncover distinct structural variants associated with *cis*-regulatory changes in *kitlga* expression. These haplotypes show phylogenetic discordance suggesting multiple introgression events. However, this trait is not underpinned by structural variation upstream of *kitlga* in all species. In one clade, the false gravid spot maps to a repetitive region on a distinct chromosome enriched with proteins containing zinc-finger domains. Instead of representing a convergent origin, our data suggest that this locus interacts with *kitlga* in *trans* leading to turnover of genetic control.

Together, our findings indicate that a simple model of LTBS maintaining ancestral variation is inaccurate. Instead, introgression and turnover contribute substantially to the evolution and maintenance of this shared polymorphic phenotype. These mechanisms can influence the ability to detect LTBS using traditional molecular approaches, and potentially led to a misestimate of its importance.

GREATER OVERLAP OF caQTLs THAN eQTLs WITH GWAS-IMPLICATED GENES

Max F Dudek[1,2], Brandon M Wenz[3], Laura Almasy[3,4], Struan F Grant[1,3]

[1]Children's Hospital of Philadelphia, Center for Spatial and Functional Genomics, Philadelphia, PA, [2]Perelman School of Medicine, University of Pennsylvania, Graduate Group in Genomics and Computational Biology, Philadelphia, PA, [3]Perelman School of Medicine, University of Pennsylvania, Department of Genetics, Philadelphia, PA, [4]Children's Hospital of Philadelphia and Perelman School of Medicine, University of Pennsylvania, Lifespan Brain Institute, Philadelphia, PA

Genome-wide association studies (GWAS) have revealed numerous non-coding loci associated with common traits, most of which likely exert their effect via gene expression. To uncover the effector genes at such signals, follow-up studies have tested for association with gene expression i.e., quantitative trait loci (eQTLs). However, a study by the Genotype-Tissue Expression (GTEx) Consortium in 49 tissue types found that only 43% of GWAS signals yielded colocalization with eQTLs. This "colocalization gap" presents a major challenge for understanding the mechanisms of genetic trait association. Recently, a study by Mostafavi et al. [PMID:37857933] showed that GWAS and eQTL discoveries are systematically biased towards different types of variants. In contrast to eQTLs, GWAS loci were found to be enriched near genes involved in more fundamental "important" biology i.e., those with more functional annotations, complex regulatory landscapes, and under stronger selective constraint. The authors proposed a model where natural selection drives down the frequency of expression-associated variants to a greater extent at these more fundamental genes, hindering their discovery as eQTLs.
Here, we expand this analysis by hypothesizing that effects of many expression-associated variants are mediated via epigenetic regulation, such as chromatin accessibility and methylation. We extend the variant-disease association model to include a chromatin accessibility component, hypothesizing that GWAS hits mediated via epigenetic effects i.e., chromatin accessibility QTLs (caQTLs), are more powered than eQTLs to detect colocalizations. Specifically, caQTLs could be found more frequently at such "important" genes than with eQTL colocalizations.
To test this hypothesis, we aggregated caQTL data from across 5 studies derived from different tissue types, representing a total sample size of 3,097 and consisting of 41,269 lead caQTLs. By mirroring the assessment of variant properties by Mostafavi et al. with this new dataset, we observed that caQTL-implicated genes reveal more genes closer to fundamental biology. Specifically, properties of these caGenes lay between those of eGenes and GWAS genes (eGenes < **caGenes** < GWAS) robust to MAF-matched SNPs – for example, the proportion of highly conserved genes (12% < **20%** < 26%) and the average number of promoters (4.4 < **6.0** < 6.4). Our results suggest that even with a limited sample size, epigenetic association signals can provide complimentary information to eQTLs by implicating functional mechanisms of additional disease-associated loci.

THE SHIFTING TEMPO OF EVOLUTION: MAPPING HETEROTACHY ACROSS THE TREE OF LIFE

Muhammed Rasit Durak, Julien Dutheil

Max Planck Institute for Evolutionary Biology, Department of Theoretical Biology, Plön, Germany

Phylogenetic models traditionally assume that while evolutionary rates vary across sites, they remain constant at each site over time. It is now well established that substitution rates can vary over time and across lineages -a phenomenon known as heterotachy, which challenges the traditional assumption of site-specific rate constancy. This temporal variation in evolutionary rates raises fundamental questions: How widespread is heterotachy across the tree of life? Is it confined to specific lineages, or is it a universal feature of protein evolution? Despite the significance of within-site rate variation, our understanding remains limited by the absence of large-scale, phylogenetically diverse studies. To address this gap, we use the OMA (Orthologous MAtrix) database, which clusters orthologous protein sequences from over 2900 species into phylogenetically coherent groups spanning all domains of life. We built alignments of these orthologous protein sequences spanning the tree of life and used model-based inference to systematically detect and quantify site-specific rate shifts across branches of the phylogenetic tree. By mapping these shifts onto the tree of life, we aim to uncover the prevalence and distribution of heterotachy, assessing whether it aligns with lineage-specific evolutionary dynamics or represents a more stochastic, intrinsic feature of molecular evolution. Additionally, we explore how gene- and site-specific properties influence rate shifts, investigating whether certain genes are more prone to heterotachy and whether positional factors play a role. By integrating large-scale phylogenomic data with analyses of rate heterogeneity, this study provides new insights into the mechanisms underlying rate variation. A better understanding of these dynamics can improve evolutionary models, with implications for phylogenetic inference, molecular dating, and the broader study of evolutionary processes.

EVOLUTIONARY INSIGHTS FROM GERMLINE-SPECIFIC CHROMOSOMES OF LAMPREY AND HAGFISH GENOMES

Kaan I Eskut[1], Nataliya Timoshevskaya[1], Vladimir A Timoshevskiy[1], Jeramiah J Smith[1,2]

[1]University of Kentucky, Biology, Lexington, KY, [2]University of Kentucky, Markey Cancer Center, Lexington, KY

In most well studied species, the primary sequence of the genome is nearly invariant across all of the cells in an organism's body (with the exception of somatic mutations and local rearrangement of immune receptors). However, several species depart from this general pattern and undergo physical elimination of large portions of their genome from somatic cell lineages as a normal part of their development, only retaining and transmitting these portions of the genome through the definitive germline. Among the vertebrates, this pattern of developmentally programmed DNA elimination is known to occur in songbirds, and jawless vertebrates (hagfish and lampreys), having independently evolved at least twice in the last 500 million years. To better understand how the germline-specific (somatically eliminated) chromosomes evolve we generated germline assemblies for several lamprey and hagfish species and performed resequencing studies to identify germline-specific chromosomes (GSCs). These analyses permitted the identification of 113 expanded gene families that have been recruited to GSCs in one or more species. Analysis of gene trees resolves the timing of recruitment of individual genes to the GSCs and changes that cumulate after recruitment. These analyses reveal a stereotypical evolutionary path wherein GSC genes are copied from somatic chromosomes (typically via segmental duplication), subsequently experience alternate fates of loss vs expansion in copy number, and apparent subsequent tuning of function to perform distinct roles in primordial germ cells, oocytes and spermatocytes. As such, GSCs of various species provide overlapping and unique insights into the long-term roles of several gene families with respect to germline functions and highlight the diversity of silencing mechanisms that exist to mediate genetic conflicts between germline and soma.

GENETIC AND EPIGENETIC SELECTION SIGNATURES FROM POOL SEQUENCING EXPERIMENT

Sonia E Eynard[1], Cécile Donnadieu[2], Loïc Flatres-Grall[3,4], Carole Iampietro[2], Sandrine Lagarrigue[5], Sophie Leroux[1], Joanna Lledo[2], Marie-José Mercat[6], Juliette Riquet[1], Céline Vandecasteele[2], Frédérique Pitel[1], Bertrand Servin[1]

[1]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet-Tolosan, France, [2]GeT-PlaGe, Genotoul, France Génomique, Université de Toulouse, INRAE, Castanet-Tolosan, France, [3]AXIOM, Azay sur Indre, France, [4]Alliance R&D, Le Rheu, France, [5]PEGASE, INRAE, Institut Agro, Saint Gilles, France, [6]IFIP, Institut du porc, Le Rheu, France

It is known that natural or artificial selection drives genome evolution. However, little is known about the effect of selection on epigenetic marks through time. It appears crucial to be able to integrate factors driven by the environment in our standard selection models, in the current context : adaptation to climate change, evolution of breeding conditions to better address contemporary challenges encompassing animal resilience, health, welfare, reduced resource use and environmental impact. Livestock species, bred under controlled conditions, traced throughout history and of large population size across many generations, offer a unique opportunity to trace the evolutionary trajectory of genetic and epigenetic patterns over time. For this study, we have access to 15 generations, representing over two decades, of selection for a sino-european pig breed. Sperm samples were collected and sequenced for more than 150 individuals, this tissue contributing significantly to the passing of genetic and epigenetic information between generations. To obtain both genetics and epigenetics information in a unique experiment and without DNA denaturation, samples were sequenced using the Oxford Nanopore Technology (ONT) PromethION instrument. We thus obtained an extensive data set providing information for more than 20 millions genetic variants with on average 30X coverage and for about 30 millions CpGs sites. Using statistical approaches based on Hidden Markov Models of the evolution of allele frequencies we identified genetic regions for signature of selection in the breed. In addition, using standard packages (e.g. DSS and edgeR), we identified differentially methylated regions across generations. We are currently developing a statistical framework to identify and cluster evolution patterns of CpGs sites and islands throughout the genome, with the aim to identify selection signatures on epigenetic marks. We aim to also correlate genetic and epigenetic selection signatures to identify regions where epigenetic changes are driven by genetic changes or independent of them. In fine we aspire to contribute to a better understanding of the impact of environmental changes on epigenetic marks throughout time and its relationship with selection undergone by the population. In the long term, our results will contribute to a more accurate accounting for the missing, non genetic, heritability in selection decisions.

THE FARM ANIMAL GENOTYPE-TISSUE EXPRESSION (FARMGTEX) PROJECT

Lingzhao Fang

Aarhus University, Center for Quantitative Genetics and Genomics, Aarhus, Denmark

Natural and human-mediated selection and migration, coupled with genetic mutation and drift, have produced a wide variety of genotypes and phenotypes in farmed animals. These diverse genetic resources thus provide unparalleled opportunities to address fundamental gaps in our biological knowledge, such as the intricate pathways linking genome to phenome within and across species. Beyond their key roles in agriculture, several farmed animals have substantial potential as biomedical models for in vivo elucidation of human biology and diseases, due to their higher similarities to humans in anatomical size and structure, development, physiology, and immunology, compared to the widely adopted rodent model. The Farm animal Genotype-Tissue Expression (FarmGTEx) Project (including 90 universities and research institutes worldwide to date) have been established to elucidate the genetic and evolutionary determinants of gene expression across 16 terrestrial and aquatic domestic species under diverse biological and environmental contexts. Since its inception in 2018, the FarmGTEx Consortium has completed several milestones in cattle, pigs, and chickens, which have already offered valuable insights into the genetic, molecular and evolutionary basis of complex phenotypes. It is now expanding this pioneering work into the next decade, aiming to serve as a platform for investigating context-specific regulatory effects. In each species, we aim to collect multi-omics data, particularly genomics and transcriptomics, from 50 tissues of 1,000 healthy adults and 200 additional animals representing a specific context. The knowledge and insights provided by FarmGTEx will deepen our understanding of molecular mechanisms underlying complex phenotypes and environmental adaptation, contributing to improving sustainable agriculture-based food system, comparative biology, and eventual human biomedicine.

T2T PRIMATE GENOMES REVEAL 60 MILLION YEARS OF STRUCTURAL VARIATION AND KARYOTYPE EVOLUTION

Scott Ferguson[1], Glennis Logsdon[2], Erik Garrison[3], Matthew Mitchell[4], Peter Sudmant[1]

[1]UC Berkeley, Integrative Biology, Berkeley, CA, [2]University of Pennsylvania, Department of Genetics, Philadelphia, PA, [3]University of Tennessee, Department of Genetics, Genomics & Informatics, Memphis, TN, [4]Coriell Institute for Medical Research, Camden, NJ

High-quality reference genomes are essential for identifying the causal genetic mechanisms underlying phenotypic differences between species. The initial sequencing and assembly of the human genome, followed by several additional primate genomes, significantly advanced our understanding of primate genetics and diversity. These genomes ushered in the era of primate comparative genomics, providing valuable insights into the evolution and diversity of our species and our close relatives. However, of the approximately 500 extant primate species, only about fifteen high-quality reference genomes currently exist. We are constructing complete and accurate, near telomere-to-telomere (T2T) reference-quality genomes for 50 diverse primate species, significantly expanding the number of available high-quality primate genomes. For each genome, high molecular weight DNA is obtained from a curated set of Coriell cell lines, from which PacBio HiFi reads, ultra-long Oxford Nanopore Technologies (ONT) reads, and Hi-C reads are generated. Additionally, to enable annotation, we are generating long-read RNA (PacBio Kinnex). To date, we have generated 10 near-T2T primate genomes. We are using these genomes to investigate the evolution of primate karyotypes over 60 million years. A key focus of this project is the identification of structural variants (SVs) within and between primate species. The high-quality, haplotype-resolved nature of our genomes will allow for comprehensive detection of SVs, including insertions, deletions, inversions, duplications, and translocations. We will investigate the association of these SVs with genes, exploring their potential impact on gene expression and function. Furthermore, we will analyse the role of SVs in primate adaptation and divergence by examining their distribution across species and searching for signatures of conservation and selection.Our future work with these genomes and resources will enable the analysis of evolutionary relationships, including the identification of incomplete lineage sorting and large-scale karyotype changes. Additionally, the resources created by this project will provide a valuable resource for comparative genomics, evolutionary studies, and biomedical research.

DETECTING RARE SOMATIC CELL TYPE SPECIFIC DRIVER MUTATIONS IN AUTOIMMUNE DISEASE USING SINGLE-CELL MULTI-OMIC TECHNOLOGIES

Matt A Field[1,2], Mandeep Singh[2], Fabio Luciano[3], Dan Suan[2], Chris Goodnow[2]

[1]James Cook University, Centre for Tropical Bioinformatics, Cairns, Australia, [2]Garvan Institute of Medical Research, ImmunoGenomics, Sydney, Australia, [3]University of New South Wales, Kirby Institute, Sydney, Australia

Background: Autoimmune diseases (AID) affect 10% of the population and have double the economic burden of cancer. However, the mechanisms of how rare AID-causing self-reactive lymphocytes escape tolerance checkpoints remain unclear. Current treatments work by broadly suppressing the immune system, increasing infection risk. Single-cell multi-omic technologies offer a promising alternative by identifying and characterising self-reactive "rogue" lymphocytes and their mutations earlier in disease progression. We've recently trialled technologies including G&TSeq, MissionBio Tapestri and Twist UMI to deeply characterise these clonal lineages and have identified cell type specific somatic pre-lymphoma driver mutations driving clonal expansion. Here I present an overview of the workflows developed and their application across AIDs.
Novel Methods: Each technology requires custom workflows to detect cell-type-specific somatic variants. While some technologies require modest modifications to traditional somatic variant workflows, others like Tapestri require substantial development. Our new Tapestri workflow uses a supervised learning approach for cell annotation to better identify duplicates and dead cells, refining the gating thresholds compared to unsupervised methods. To identify true variants, additional filters based on cell counts, allele frequencies and cell-type enrichment statistics have been critical to identify somatic drivers. We are currently benchmarking the Twist UMI Adaptor system with promising results.
Novel Applications: We've successfully applied these workflows across three diseases. In vasculitis, we combined single-cell DNA, RNA, and antibody sequencing to identify pre-lymphoma driver mutations in rare B lymphocytes producing pathogenic autoantibodies (Singh / Cell). Next, in self-reactive B cells driving a virus-induced autoimmune disease, we used G&T-Seq and cell-type-specific WGS to identify known somatic driver mutations including SNVs, indels, SVs and even trisomy (Young / Immunity). Lastly, in celiac disease we developed custom cell surface marker and gene panels for the Tapestri platform to uncover driver somatic mutations (In press SciTransMed).
Conclusion: We can now identify rare cell type specific driver mutations in as few as 0.1% of total cells, often before clinical diagnosis. The ability to target rogue lymphocytes containing pre-lymphoma driver mutations paves the way for earlier personalized treatment.

COMPENSATORY COPY NUMBER VARIATIONS IN THE MALARIA PARASITE GENOME REVEAL METABOLIC INTERPLAY BETWEEN ANTIMALARIAL TARGETS

Kwesi Akonu Adom Mensah Forson[1], Shiwei Liu[2], Julia Zulawinska[4,1], Jennifer L Guler[1,3]

[1]University of Virginia, Biology, Charlottesville, VA, [2]Indiana University, Radiology & Imaging Sciences, Indianapolis, IN, [3]University of Virginia, Infectious Diseases and International Health, Charlottesville, VA, [4]University of Virginia, Biomedical Sciences, Charlottesville, VA

Malaria remains a major global health challenge, with *Plasmodium falciparum* developing resistance to frontline antimalarials, including newly developed inhibitors of key biochemical pathways. Copy number variations (CNVs) have emerged as a key adaptive strategy to develop drug resistance, yet their metabolic impacts remain poorly understood. Here, we investigate the relationship between CNVs of genes involved in folate and pyrimidine metabolism—two interconnected pathways targeted by distinct antimalarials. By studying a family of resistant *P. falciparum* parasites, we observed that parasites that already have extra copies of a folate biosynthesis gene (chr 12) are more likely to acquire resistance-conferring CNVs on chr 6. We hypothesize that the pre-existing gene amplification and its impacts on folate biosynthesis facilitates the subsequent amplification of the pyrimidine biosynthesis gene, potentially increasing flux through the two pathways to elevate nucleotide availability for the parasite. Using single-cell RNA sequencing, we are currently investigating the transcriptional consequences of these extra gene copies. Additionally, using isogenic parasite lines with varying copies of the chr 12 CNV, we are exploring necessity of the preexisting CNV for parasite fitness and resistance evolution. Finally, while this compensatory mechanism was uncovered in laboratory parasite strains, we are using comparative genomics across Plasmodium species and clinical isolates to understand the relationship between the two pathways in natural parasites across various environments. Our findings highlight a novel instance of metabolic adaptation that facilitates drug resistance in malaria and may serve as a target for future antimalarial strategies.

ANCESTRY-DRIVEN METHYLATION DIFFERENCES IMPACT IMMUNE FUNCTION REGULATION IN BREAST CANCER

Kyriaki Founta[1,2], Nyasha Chambwe[2,1]

[1]Zucker School of Medicine at Hofstra/Northwell, Molecular Medicine, Hempstead, NY, [2]Feinstein Institutes for Medical Research, Northwell Health, Molecular Medicine, Manhasset, NY

Black women of predominantly African descent face a disproportionately higher incidence of aggressive breast cancer subtypes, leading to worse outcomes across all racial and ethnic groups in the Unites States. This disparity is likely due to the complex interplay between genetics, sociocultural, and environmental factors. DNA methylation (DNAm), an epigenetic mechanism regulating gene expression, has been proposed as a means to capture the molecular impact of adverse environments on tumor biology. However, although DNAm levels can be influenced by environmental stimuli, they are also regulated by individual genotypes at methylation quantitative trait loci (meQTLs). We hypothesized that ancestry differential DNAm in breast cancer is a result of differential genotype frequencies of SNPs at meQTL loci that can result in phenotypic differences between African and European breast cancer patients through differential gene expression regulation. Analysis of 578 breast tumor samples from individuals of African and European descent in The Cancer Genome Atlas cohort, revealed 757 differentially methylated sites associated with genetic ancestry. Methylation quantitative trait loci (meQTL) mapping showed that the majority of these sites are regulated by multiple SNPs with differential frequencies by ancestry. 91% of the differentially methylated sites influenced nearby gene expression via expression quantitative trait methylation (eQTMs), indicating potential impact on gene regulation. 69% of ancestry eQTMs were themselves regulated by meQTLs, underscoring the potential of genetic variation influencing expression via DNAm. Ancestry eQTM target genes were predominantly involved in immune functions like MHC class II protein complex assembly and antigen processing. Additionally, 24% of differentially expressed genes between groups defined by genetic ancestry were regulated by ancestry eQTMs. These findings highlight the gene regulatory potential of ancestry-associated DNAm, providing new insights into how population level genetic variation can influence epigenetic changes that can affect immune regulation in breast cancer.

SEGMENTAL DUPLICATION-MEDIATED REARRANGEMENTS ALTER THE LANDSCAPE OF MOUSE GENOMES

Eden R Francoeur[1,2], Ardian Ferraj[1], Peter A Audano[1], Parithi Balachandran[1], Christine R Beck[1,2]

[1]The Jackson Laboratory for Genomic Medicine, Genetics and Genome Sciences, Farmington, CT, [2]University of Connecticut Health Center, Genetics and Genome Sciences, Farmington, CT

Segmental duplications (SDs) are among the most rapidly evolving regions of mammalian genomes and can generate significant variation within the species that harbor them. These duplications are defined as large (≥1 kb) and highly homologous DNA sequences (≥90%) that are not mobile elements. SDs constitute over 5% of both human and mouse genomes, and often contain genes. Due to the homology and length of SDs, they can undergo ectopic rearrangements resulting in large structural variants (SVs) such as deletions, duplications, and inversions ≥ 50bp, depending on the sequence orientation of the SD paralogs. Rearrangements between SDs can lead to changes in gene dosage, which may result in fitness changes subject to natural selection, such as the copy number differences of amylase genes within mammalian genomes. We generated SD annotations for the GRCm39 reference genome using SEDEF and examined potential mechanisms of the origin of these duplications. Previously studies have shown that human SDs are enriched for flanking Alu transposable elements (TEs) that may have mediated rearrangements that result in large duplicated sequences. We find that the flanks of SDs in mouse genomes are significantly enriched for LINE-1 TEs, in particular high copy number LINE-1 families that are relatively young in mouse genomes. To predict loci in the mouse genome that may be subject to SD-mediated instability, we first identified intrachromosomal SD paralogues that have >95% sequence identity, are longer than 1kb, and have <10Mb of sequence between paralogs. To test our predicted loci for variation between mice, we used a combination of PacBio long-read sequencing, optical mapping, and copy number (CN) estimation from Illumina short-read sequencing to identify putative SD-mediated rearrangements in 8 diverse strains of laboratory mice that are derived from three subspecies of Mus musculus spanning ~1/2 a million years of evolution. With our analyses, we identify 223 SD-mediated rearrangements (146 deletions, 66 duplications, and 11 inversions) that contribute to over 14Mb of variation when compared to GRCm39. We additionally trained a machine learning model using the identified events to predict regions where rearrangements are likely to occur. Finally, we have examined potential transcriptomic differences due to SD-mediated rearrangements using VEP and RNA sequencing. We find that SD-mediated rearrangements often affect transcription factors, immune related genes, and most abundantly transmembrane signal receptors.

MULTI-OMIC GENOMIC MAPPING WITH LONG READ SEQUENCING

Connor Frasier[1], James T Anderson[1], Eva Brill[1], Paul W Hook[2], Allison Hickman[1], Vishnu Kumary[1], Anup Vaidya[1], Jamie Moore[2], Ryan Ezell[1], Jonathan M Burg[1], Zu-wen Sun[1], Martis W Cowles[1], Winston Timp[2], Bryan J Venters[1], Michael-Christopher Keogh[1]

[1]Epicypher Inc., Genomics, Durham, NC, [2]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD

Gene transcription is regulated by the complex interplay between histone post-translational modifications (PTMs), chromatin associated proteins (CAPs), and DNA methylation (DNAme). Mapping their genomic locations and examining the relationships between these chromatin elements is a powerful approach to decipher mechanisms of disease, thereby enabling discovery of novel biomarkers and therapeutics. Leading epigenomic mapping technologies (e.g., ChIP-seq, CUT&RUN) rely upon DNA fragmentation to isolate regions of interest for sequencing on short read platforms (e.g., Illumina). This strategy leads to substantial loss of contextual information regarding the surrounding DNA, precluding the identification of multiple co-occurring epigenomic features on a single DNA molecule. By contrast, long-read sequencing (LRS) platforms are capable of sequencing very long reads from a single molecule (typically >10kb), allowing relationships between features on a single molecule to be used to resolve heterogeneity within mixed populations.

Here we report a robust multi-omic method that leverages LRS to simultaneously profile histone PTMs (or CAPs), DNAme, and parental haplotype in a single assay. This nondestructive, epigenomic mapping approach leverages a novel DNA methyltransferase fusion protein (pAG-M.EcoGII) to label DNA underneath antibody-targeted chromatin features, thereby marking sites of interest while preserving DNA molecules intact for LRS. Inspired by our work with state-of-the-art immunotethering-based approaches (CUT&RUN / CUT&Tag), nuclei are bound to magnetic beads to streamline and automate sample processing. Next, adenosines nearby antibody-targeted chromatin features are methylated with pAG-M.EcoGII, which are then directly read from genomic DNA using Oxford Nanopore Technologies or Pacific Biosciences LRS platforms. Importantly, this method is highly reproducible across biological replicates, and highly concordant with orthogonal SRS assays (e.g., CUT&RUN). Further, we showed that this method is a true multi-omic approach by simultaneously profiling histone PTMs, native DNAme (5mC), and parental single-nucleotide variants from single DNA molecules within a single reaction. Finally, this workflow preserves chromatin integrity for LRS, revealing heterogeneity (e.g., haplotype or paternal origin) within / between data types and providing access to previously unmappable genomic regions (e.g., centromeres).

REPEATED EVOLUTION OF REPRODUCTIVE ISOLATION IN A MONKEYFLOWER SPECIES COMPLEX

Megan Frayer, Hagar Soliman, Pia Schwarz, Jenn Coughlan

Yale University, Ecology and Evolutionary Biology, New Haven, CT

In populations where multiple paternity is common, alleles may arise to upset the balance of resource allocation to developing offspring, followed by the appearance of new alleles that counteract those effects. This evolutionary arms race, known as parental conflict, can provide a general mechanism for reproductive isolation when mismatched pairs of alleles lead to extreme phenotypes in developing seeds. Given the importance of multiple demographic and life history traits in determining the conditions for this conflict, the "strength" of conflict may vary across related populations. In this study, we describe the repeated evolution of "high conflict" lineages in the *Mimulus guttatus* complex. We demonstrate that these lineages follow the same, predicted patterns of crossing when crossed to "low conflict" lineages, and to each other. We provide evidence that the conflict is driven by incompatibilities in the endosperm; a primary seed tissue that facilitates nutrient exchange between mothers and offspring. We then use multiple genomic approaches to test whether this repeated evolution is the result of parallel evolution or introgression between "high conflict" lineages. We find that the rapid evolution of strong isolation between sister lineages northern *M. decorus* and southern *M. decorus* is likely driven by introgression from a distantly related lineage. Our results support growing evidence that parental conflict may be a general mechanism for reproductive isolation within the *Mimulus guttatus* complex and highlights a potential role of introgression of incompatibility alleles in the origin of new species.

HYBRID SHORT AND LONG-READ SEQUENCING AFFORDABLY ENHANCES GENOME CHARACTERIZATION IN DIFFICULT REGIONS

Don Freed, Frank Hu, Hanying Feng, Haodong Chen, Hong Chen, Zhipan Li, Brendan Gallagher, Louqi Chen

Sentieon Inc., San Jose, CA

Whole-genome sequencing (WGS) is increasingly being adopted as a first-line diagnostic tool, offering comprehensive genomic coverage and superior detection of structural variants (SVs), copy number variants (CNVs), and repeat expansions compared to whole-exome sequencing (WES). While short-read sequencing remains the standard for high-throughput WGS, long-read sequencing, such as PacBio HiFi, provides key advantages in resolving complex genomic regions, improving SV detection, methylation analysis, and haplotype phasing. However, the higher cost and lower throughput of long-read platforms have limited their widespread use.

Here, we introduce a hybrid secondary analysis pipeline that integrates high-coverage short-read WGS with low-coverage PacBio HiFi sequencing to enhance genome characterization. By leveraging long-read haplotypes to refine short-read alignment, our approach significantly improves variant calling accuracy, particularly in difficult-to-map regions and medically relevant genes such as *SMN1*, *GBA*, and *PMS2*.

We benchmark our method using well-characterized genome datasets, including the Genome in a Bottle (GIAB) v4.2.1 and the draft Q100 benchmarks. With 35× short-read (NovaSeq) and 10× HiFi coverage, our pipeline achieves a 76% reduction in variant calling errors compared to DeepVariant with 35× short-read data alone and a 58% improvement over DeepVariant with 30× HiFi data alone. The method particularly excels in resolving long homopolymer tracts and tandem repeats, surpassing the accuracy of single-technology pipelines. Importantly, our hybrid approach enables the detection of pathogenic variants in *SMN1*, *CYP21A2*, *OPN1LW*, and other medically relevant genes that are challenging or impossible to resolve with short-read sequencing alone.

While demonstrated here on human genomes, this method generalizes to non-human species, making it a versatile tool for genome characterization. By combining the strengths of short- and long-read sequencing, our pipeline achieves variant detection accuracy comparable to full-coverage long-read approaches at a fraction of the cost, providing a scalable solution for comprehensive genomic analysis.

GLOBAL CIS-REGULATORY LANDSCAPE OF DOUBLE-STRANDED DNA VIRUSES

Tommy Taslim[1], Youssef A Finkelberg[2], Susan Kales[3], Luis Soto-Ugaldi[4], Elvis Morara[5], Jacob Purinton[5], Harshpreet Chandok[3], Jaice Rottenberg[5], Rodrigo Castro[3], George Munoz[6], Lucia Martinez-Cuesta[5], Matias Paz[7], Beedetta D'Elia[1], Ryan Tewhey[3], Juan Fuxman Bass[1,2,5]

[1]Boston University, MCBB Program, Boston, MA, [2]Boston University, Bioinformatics Program, Boston, MA, [3]The Jackson Laboratory, Bar Harbor, ME, [4]Rockefeller University, Tri-Institutional Program in Computational Biology and Medicine, New York, NY, [5]Boston University, Biology Department, Boston, MA, [6]National University of San Marcos, Lima, Peru, [7]University of Buenos Aires, Faculty of Medicine, Buenos Aires, Argentina

Most double-stranded DNA viruses use the host transcriptional machinery to express viral genes at different stages of viral replication or in response to cellular signals. This process is mediated by viral cis-regulatory elements (CREs) that bind host and viral transcription factors (TFs). Although some viral CREs and their regulatory mechanisms have been determined, most remain unidentified. Here, we used massively parallel reporter assays to identify ~3,000 CREs across 27 dsDNA viruses from the Adenovirus, Herpesvirus, Polyomavirus and Papillomavirus families. Most of these CREs are promoter-like elements and are located close to transcription start sites. Viral genomes have a higher density of CREs than the human genome with most viral CREs overlapping coding sequences. Using a combination of saturation mutagenesis, machine learning models, and yeast one-hybrid assays, we report regulators of viral CREs, including SP, ETS factors, bZIPs, and TFs acting downstream of signal- or ligand-activated pathways. Altogether, we present a comprehensive functional CRE map of human infecting viruses that serves as a blueprint for further studies in viral regulation, reactivation, evolution, and viral vector design.

ANONYMIZED SOMATIC TUMOR TWINS (STTs) FOR OPEN DATA SHARING IN CANCER GENOMIC RESEARCH

Nicolás Gaitán[1], Rodrigo Martín[1], David Torrents[1,2]

[1]Barcelona Supercomputing Center (BSC), Life Sciences Department, Barcelona, Spain, [2]Institució Catalana per la Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Analyzing the somatic variation landscape of cancer genomes is a cornerstone of modern oncology. Access to somatic genome data from real clinical tumor cohorts is crucial for advancing research, and for developing and benchmarking emerging infrastructures for the management and analysis of cancer data. However, stringent data protection frameworks regulate and restrict the sharing of identifiable information, such as germline variants, limiting the exchange and availability of high-quality cancer genome samples. To address these challenges, while ensuring compliance with data protection laws, we developed and validated a DNA sequence anonymization strategy that removes all detectable germline variation from normal-tumor sequenced genome pairs while preserving the tumor's original somatic information. We demonstrate that the resulting sequences, Somatic Tumor Twins (STTs), retain full utility for somatic variation analysis, discovery, and benchmarking without compromising the donor's privacy. By applying this strategy to existing normal-tumor datasets, we generated a cohort of STTs that can be shared openly across projects and centers worldwide. This innovative approach will allow the exchange of cancer somatic data for discovery in oncology research and enable robust benchmarking of large-scale infrastructures for cancer genomic data management, analysis and clinical application.

PRECISECALLER: A COMPREHENSIVE, SCALABLE, USER-FRIENDLY AND OPEN-SOURCE PLATFORM FOR GENOMIC VARIANT DETECTION IN ONCOLOGY AND PRECISION MEDICINE.

Thiago L Miller, Gabriela D Guardia, Pedro A Galante

Hospital Sirio-Libanes, Centro de Oncologia Molecular, Sao Paulo, Brazil

Precision medicine aims to personalize medical treatment based on patients' genetic characteristics, being especially relevant in oncology. For this purpose, accurate detection of genetic variants such as single nucleotide variants (SNVs) and structural variations is crucial. However, many current bioinformatics tools have limitations, including high false-positive rates and extensive computational resource requirements. In this work, we developed an open-source tool called PreciseCaller to identify the main types of genetic variants in tumor samples. This tool can detect SNVs, small insertions and deletions (Indels), and mobile element insertions (retroelements). PreciseCaller is user-friendly, well-documented, simple to install, and scalable for execution in cloud computing environments. From a technical perspective, the tool is being implemented in Nextflow, a workflow management system that enables scalable and reproducible scientific workflows. Nextflow provides platform independence, containerization support, and efficient pipeline execution across diverse computing infrastructures, including high-performance computing clusters and cloud environments. PreciseCaller aims to offer not only exceptionally high precision in variant calling but also reduces operational costs, making this process accessible to professionals without specialized technical knowledge in bioinformatics and viable in locations with limited computational infrastructure by optimizing resource utilization, implementing efficient algorithms, and providing intuitive result and interpretation tools that facilitate clinical decision making.

THE GENE EXPRESSION LANDSCAPE OF DISEASE GENES

Judit García-González[1], Alanna C Cote[1], Saul Garcia-Gonzalez[1,2], Lathan Liou[1], Paul F O'Reilly[1]

[1]Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences, New York City, NY, [2]Icahn School of Medicine at Mount Sinai, Center for Excellence in Youth Education, New York City, NY

Fine-mapping and gene-prioritization techniques applied to the latest Genome-Wide Association Study (GWAS) results have prioritized hundreds of genes as causally associated with disease. Here we leverage these recently compiled lists of high-confidence causal genes to interrogate where in the body these genes operate. By integrating GWAS summary statistics, gene prioritization results, and RNA-seq data from 46 tissues and 204 cell types, we systematically analyze their relationship with 11 common major diseases and cancers. In tissues and cell types with established disease relevance, prioritized genes show higher and more specific expression compared to control genes. We also detect elevated expression in tissues and cell types without known disease links. While some of these results may be explained by cell types that span multiple tissues, such as macrophages in brain, blood, lung and spleen in relation to Alzheimer's disease (P-values < $1x10^{-3}$), the cause for others is unclear and motivates further investigation. To support functional follow-up studies, we identified key predictors of gene expression for disease-associated genes, highlighted disease genes with the highest expression in relevant tissues, and explored how gene expression influences the likelihood of being included in drug development programs. We present our systematic testing framework as an open-source, publicly available tool to provide novel insights into the genes, tissues and cell types involved in disease that could inform drug target and delivery strategies.

REFERENCE-FREE, HAPLOTYPE-RESOLVED NOMINATION OF CRISPR OFF-TARGETS ACROSS GLOBAL AND INDIVIDUAL GENOMIC DIVERSITY

Erik Garrison[1], Farnaz Salehi[1], Linda Lin[2], Haarika Kathi[3], Daniel E Bauer[3], Luca Pinello[4]

[1]University of Tennessee Health Science Center, Department of Genetics, Genomics, and Informatics, Memphis, TN, [2]Yale University, Department of Genetics, New Haven, CT, [3]Boston Children's Hospital, Hematology/Oncology, Boston, MA, [4]Massachusetts General Hospital, Molecular Pathology Unit, Boston, MA

Traditional CRISPR off-target analysis inherits fundamental biases from its reliance on reference genomes—missing sequences absent from the reference, ignoring variant phasing, overlooking repetitive regions, and failing to capture complex structural variation. This reference bias becomes particularly concerning as CRISPR therapeutics enter clinical use, where undetected off-targets could have serious consequences.

CRISPR therapeutics like Casgevy demonstrate that genome editing has entered mainstream medical practice. To improve the safety of CRISPR therapeutics, we developed a method to search for candidate off-targets both in high-quality complete genome assemblies from the human pangenome (HPRC) or the unassembled reads of whole genome shotgun sequencing.

We present CRISPRapido, a framework that leverages fast exact alignment with the Wavefront Algorithm (WFA) to find all candidate off-targets in complete genome assemblies in minutes and high-coverage sequencing datasets in hours—a modest computational investment given the stakes of therapeutic genome editing. Downstream, predictive scoring like Cutting Frequency Determination (CFD) allows users to prioritize off-target sites based on their cleavage potential.

Our validation demonstrates two critical applications: population-scale screening across the HPRC pangenome to assess off-target frequency, and patient-specific profiling using standard short-read sequencing. In both contexts, we identify candidate off-target sites for guides for the treatment of sickle cell disease and β-thalassemia that are invisible to reference-based methods. Given the transformative potential and significant cost of CRISPR therapeutics, we argue that such comprehensive off-target screening should become a standard pre-treatment safety measure.

By eliminating reference bias and enabling both population-level and patient-specific analysis, this approach establishes a more rigorous framework for therapeutic safety assessment. This breakthrough becomes especially crucial as CRISPR therapies expand, where the consequence of undetected off-targets must be weighed against their remarkable potential for treating previously intractable diseases.

DETECTION OF EARLY METABOLIC STRESS MECHANISMS DRIVING RISK FOR CARDIOMETABOLIC DISORDERS IN AN URBAN-TRANSITIONING KENYAN POPULATION

Kristina M Garske[1], Thomas Atkins[1], Emma Gerlinger[1], Julie Peng[1], Matthew Chao[1], John C Kahumbu[2,3], Varada Abhyankar[1], Benjamin Muhoya[2,3], Charles M Mwai[2,3], Patricia Kinyua[3], Anjelina Lopurudoi[3], Francis Lotukoi[3], Boniface Mukoma[3], Dino Martins[3,4], Sospeter Njeru[3,5], Amanda J Lea[6], Julien F Ayroles[1,2]

[1]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [2]Princeton University, Ecology and Evolutionary Biology, Princeton, NJ, [3]Turkana Health and Genomics Project, THGP, Nairobi, Kenya, [4]Stony Brook University, Turkana Basin Institute, Stony Brook, NY, [5]Kenya Medical Research Institute, Centre for Community Driven Research, Nairobi, Kenya, [6]Vanderbilt University, Biological Sciences, Nashville, TN

Modern, industrialized, urban lifestyles are associated with myriad cardiometabolic disorders (CMDs) such as obesity, dyslipidemia, and hyperglycemia. One hallmark of CMDs is chronic inflammation, and interactions between metabolism and the immune system have been proposed to drive pathological mechanisms underlying these disorders. We hypothesize that we can detect urban lifestyle-driven, CMD-related gene regulatory alterations in circulating classical CD14+ monocytes, a cell-type that can contribute to metabolic dysfunction through tissue infiltration and differentiation into resident macrophages. The Turkana people of northern Kenya, historically a subsistence-level pastoralist community, have recently begun migrating toward urban centers, which coincides with metabolic disturbances linked to poor cardiovascular health. We partnered with this community to investigate how urban living and metabolic stress manifest at the molecular level. Using single-cell RNA-seq (scRNA-seq) analysis on peripheral blood mononuclear cells (PBMCs) from 250 Turkana individuals from rural and urban settings, we examined CD14+ monocyte gene expression patterns. Genes associated with serum triglycerides (TGs) exhibit greater co-expression loss in urban relative to rural participants. We have termed this phenomenon ‘decoherence’ – a breakdown in coordinated gene regulation that unmasks potentially harmful genetic effects – and we have previously shown this to be predictive of metabolic syndrome. Hub genes with extensive decoherence highlight pathways in growth factor and cytokine signaling, as well as iron homeostasis. Importantly, lifestyle-associated genes as a whole do not exhibit such strong decoherence, pinpointing monocyte gene regulation in response to serum TGs as a key axis in lifestyle-mediated CMD risk. Through integrating chromatin accessibility information and genetic associations with serum TG levels and decoherent genes, our ongoing work aims to prioritize candidate drivers of metabolic stress, shedding light on regulatory mechanisms underlying CMDs.

CHARACTERIZING GENETIC ANCESTRY ASSOCIATED VARIATION IN LYNCH SYNDROME GENES

Devin A Gee[1], Nyasha Chambwe[1,2]

[1]Feinstein Institutes for Medical Research, Institute of Molecular Medicine, Northwell Health, Manhasset, NY, [2]Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY

Despite being one of the most common cancer predisposition syndromes, most individuals with Lynch Syndrome (LS) remain undiagnosed. LS increases an individual's lifetime risk of colorectal and endometrial cancer to 40-60%, representing a population in critical need of increased clinical surveillance. Under diagnosis of LS is partly due to low rates of genetic testing, an issue especially relevant for historically underrepresented groups who experience structural barriers to accessing testing. Previous studies identified higher rates of variants of uncertain significance (VUS) among groups of non-European ancestries, limiting the interpretation of genetic testing results in these populations. By characterizing ancestry associated variation in LS genes, we aim to identify populations associated with an increased prevalence of LS and annotate VUS for pathogenicity to better inform testing strategies to diagnose high risk individuals.

In this study, we examined germline whole genome sequencing data from 384,246 individuals in the All of Us biobank. 45% of this cohort were of predominantly non-European ancestry (18% African, 17% American Admixed, 2% East Asian, 8% other) representing one of the most genetically diverse cohort studies of LS prevalence. In preliminary analyses, we analyzed single nucleotide polymorphisms, insertions, and deletions to identify carriers of pathogenic variants in LS genes. 1,196 individuals had pathogenic or likely pathogenic variants previously annotated in ClinVar. Using these variants, we demonstrate the distribution of pathogenic LS variants across genetic ancestries. We also demonstrate the distribution of VUS by ancestry and calculate the probability of pathogenicity using computational variant annotation tools.

In conclusion, our study aims to pinpoint the genetic ancestries most in need of germline genetic testing for LS and to characterize the impact of VUS within these groups. By achieving these objectives, we anticipate enriching the understanding of LS across diverse genetic backgrounds and informing targeted testing strategies.

THE REPERTOIRE OF SHORT TANDEM REPEATS ACROSS THE TREE OF LIFE

Nikol Chantzi, Ilias Georgakopoulos-Soares

The Pennsylvania State University College of Medicine, Department of Biochemistry and Molecular Biology, Hershey, PA

Short tandem repeats (STRs) are widespread, dynamic repetitive elements with a number of biological functions and relevance to human diseases. However, their prevalence across taxa remains poorly characterized. Here we examined the impact of STRs in the genomes of 117,253 organisms spanning the tree of life. We find that there are large differences in the frequencies of STRs between organismal genomes and these differences are largely driven by the taxonomic group an organism belongs to. Using simulated genomes, we find that on average there is no enrichment of STRs in bacterial and archaeal genomes, suggesting that these genomes are not particularly repetitive. In contrast, we find that eukaryotic genomes are orders of magnitude more repetitive than expected. STRs are preferentially located at functional loci at specific taxa. Finally, we utilize the recently completed Telomere-to-Telomere genomes of human and other great apes, and find that STRs are highly abundant and variable between primate species, particularly in peri/centromeric regions. We conclude that STRs have expanded in eukaryotic and viral lineages and not in archaea or bacteria, resulting in large discrepancies in genomic composition.

DISCOVERING LOW-FREQUENCY SOMATIC MUTATIONS IN A CRANIOFACIAL MICROSOMIA PATIENT USING RUFUS: A REFERENCE-FREE, KMER-GUIDED DETECTION ALGORITHM

Stephanie J Georges[1], Nancy Parmalee[2], Lila Sutherland[2], James T Bennett[2], Gabor T Marth[1]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]Seattle Children's Research Institute, Center for Developmental Biology and Regenerative Medicine, Seattle, WA

Comprehensive detection of somatic mutations proves to be a challenging problem for existing variant calling tools; signal to noise ratios are high, as somatic variant allele frequencies are often less than or equal to those of the sequencing-technology error rates (<1%). Many variant calling tools, thus, struggle to retain high sensitivity while also eliminating false positive calls for somatic data sets; indeed, the reported variant set may be prohibitively large to validate or interrogate further. To address the critical need for accurate and precise detection tools, we present the RUFUS algorithm. RUFUS is a kmer-guided detection algorithm which is free of reference bias associated with many canonical variant calling pipelines. Originally developed for germline de novo mutation detection, we have adapted RUFUS as part of the Somatic Mosaicism in Human Tissues (SMaHT) initiative to identify somatic variants uniformly across mutation type, length, and allele frequency spectrums.

In collaboration with Seattle Children's Research Institute (SCRI) and the laboratory of James Bennett, we have called variants using RUFUS on a patient with craniofacial microsomia (CFM). The genetic underpinnings of CFM are currently under investigation, but some evidence suggests somatic variation may play a role. From deeply-sequenced exome data derived from preauricular tag tissue, RUFUS called approximately ten variants around 1-2% allele frequency. This number starkly contrasts the almost 1000 variants called by the Dragen pipeline on the same data. Subsequent ddPCR experiments have shown that four out of four RUFUS-unique variant calls have validated, while zero of four Dragen-unique calls have validated.

A containerized version of RUFUS is available for download and use at https://github.com/marthlab/RUFUS.

IMPROVED SPIKE-IN NORMALIZATION CLARIFIES THE RELATIONSHIP BETWEEN ACTIVE HISTONE MODIFICATIONS AND TRANSCRIPTION

Lauren Patel[1,2,3], Yuwei Cao[3], Tamar Dishon[3], Tianyao Xu[3], Eric Mendenhall[4], Itamar Simon[5], Christopher Benner[2], Alon Goren[3]

[1]UCSD, Department of Bioengineering, Jacobs School of Engineering, La Jolla, CA, [2]UCSD, Department of Medicine, Division of Endocrinology & Metabolism, La Jolla, CA, [3]UCSD, Department of Medicine, Division of Genomics & Precision Medicine, La Jolla, CA, [4]HudsonAlpha, Institute for Biotechnology, Huntsville, AL, [5]The Hebrew University, Department of Microbiology and Molecular Genetics, Institute of Medical Research Israel-Canada, Jerusalem, Israel

Spike-in normalization enables quantitative analysis of ChIP-seq signal. Here, we describe ChIP-wrangler – an improved spike-in normalization approach for ChIP-seq – and its use to revisit previous conclusions regarding the link between histone acetylation and transcription.
In our recent study we noted that there is a widespread misuse of spike-in normalization (Patel et al, *Nat. Biotech.* 2024). Here, we present a novel approach that minimizes such misuse cases by implementing additional guardrails and QC steps. Our method employs the addition of exogenous chromatin from two species at a predefined ratio to assess the measurement error associated with spike-in normalization of ChIP-seq data. We developed a matching computational analysis framework that enables inferring confidence intervals based on the variation between the signal obtained from the two species. We anticipate that our new method will be highly beneficial in helping researchers catch problems with their spike-in analysis and avoiding erroneous conclusions.
As a case for our new approach we revisited the relationship between marks of active chromatin and transcription. Specifically, recent studies have suggested that acetylation is a consequence of transcription, contradicting the long-standing model that active histone modifications function in regulatory mechanisms that occur upstream of transcription. Our reanalysis of the corresponding data generated in these recent studies indicated high variability in the spike-in chromatin used for normalization (detailed in Patel et al, 2024), prompting us to use our ChIP-wrangler approach to revisit these findings. Employing two approaches to deplete RNA polymerase II in conjunction with accurate dual species we observed an intermittent slight decrease and, in contrast to recent findings, an overall slight increase of histone acetylation at late time points relative to the steady state.
Together, our innovative ChIP-seq normalization approach provides increased rigor and "guardrails" for successful spike-in normalization, and as applied here refines the understanding of the intricate crosstalk between RNA polymerase II activity and histone marks associated with transcription, contributing to resolving a controversy in the field of transcription regulation and epigenomics that has lingered for years.

GENETIC AND MULTI-OMIC INSIGHTS INTO INFLAMMATION AND METABOLISM IN A FRENCH POLYNESIAN COHORT

Olivia A Gray[1], Anne-Katrin Emde[1], Iman Hamid[1], Megan Leask[2,3], Jaye Moors[1], Baptiste Gerard[4], Melissa Hendershott[1], Sarah LeBaron von Baeyer[1], Tehani Mairai[1], Vehia Wheeler[5,6], Tony Merriman[2,3], Kaja Wasik[1], Keolu Fox[7,8], Tristan Pascart[9], Laura Yerges-Armstrong[1], Stephane Castel[1]

[1]Variant Bio, Seattle, WA, [2]University of Alabama, Division of Clinical Rheumatology and Immunology, Birmingham, AL, [3]University of Otago, Department of Microbiology and Immunology, Dunedin, New Zealand, [4]Centre Hospitalier Universitaire Rouen, Service de Rhumatologie, Rouen, France, [5]Australian National University, Canberra, ACT, Australia, [6]Sustainable Oceania Solutions, Afareaitu, Mo'orea, French Polynesia, [7]University of California San Diego, Global Health Program, Department of Anthropology and Indigenous Futures Institute, Division of Design and Innovation, San Diego, CA, [8]Native BioData Consortium, Eagle Butte, SD, [9]Hôpital Saint-Philibert, Lille, France

Multi-omic approaches are powerful tools for interpreting the results of genome-wide association studies and enabling therapeutic development. However, the vast majority of both GWAS and functional genomics studies to date have been carried out in predominantly European-ancestry populations, limiting our ability to identify novel genetic effectors of disease. Across Polynesia, there have been increasing efforts to address this lack of diversity by conducting genomic research with Pasefika cohorts to identify population-enriched variants associated with health-relevant traits. Yet, these previous genomic studies have often been too narrowly focused on one type of disease and lack paired functional genomic data to aid interpretation. Here, we aim to fill this gap by performing the largest integrative analysis in a Polynesian cohort to date as part of the Ma'i u'u Survey, which focused on gout and related inflammatory and metabolic diseases. We performed whole genome sequencing and comprehensive phenotyping in a cohort of ~1,100 French Polynesian adults paired with whole blood transcriptomics and untargeted metabolite profiling in representative subsets of the cohort. Using these data we created a comprehensive map of genetic variants that impact phenotypes (GWAS), metabolites (mGWAS), gene expression (eQTL), and splicing (sQTL). Through colocalization analysis, we characterized the impact of these variants and identified putatively causal genes and mechanisms, demonstrating the value of our paired multi-omic approach. This genomic resource not only deepens our understanding of the genetic architecture of health and disease in Mā'ohi Nui, French Polynesia but also serves as an important genomic reference for future studies with Polynesian populations.

REPLICATION STRESS INCREASES DE NOVO CNVs ACROSS THE MALARIA PARASITE GENOME

Noah Brown[1], Aleksander Luniewski[1], Xuanxuan Yu[2,3], Michelle Warthan[1], Shiwei Liu[1], Julia Zulawinska[1], Syed Ahmad[1], Feifei Xiao[2], Jennifer L Guler[1]

[1]University of Virginia, Biology, Charlottesville, VA, [2]University of Florida, Statistics, Gainsville, FL, [3]University of Florida, Surgery, Gainsville, FL

Advances in genomics allow researchers to appreciate mutations that contribute to cellular phenotypes, especially when they are present in the majority of cells of a population. However, we miss rare genetic changes that also contribute to important phenotypes. This is especially true for structural variations (or copy number variations, CNVs) that arise in few genomes. The analysis of single genomes has led to progress in detecting these infrequent CNVs, but most of this work has focused on large genomes with large variations. Our group aims to study the impact of CNV diversity in microbial populations, where those present in a few cells are likely to have enormous implications for rapid evolution. We study the protozoan malaria parasite, Plasmodium falciparum, which has a small, highly AT-rich genome; CNV evolution in this organism is especially interesting because of its unique genome architecture and alternative repertoire of CNV-generating pathways. To identify rare CNVs in the P. falciparum genome, we previously modified a whole genome amplification method to reduce amplification bias, limit human host contamination, and facilitate CNV detection. Additionally, we developed two computational methods to identify and count CNVs in complex datasets (HapCNV and SVCROWS). Using these improved tools, we showed a dramatic increase in our ability to detect “known” CNVs, or those that we know are present in ALL parasite genomes. With confidence in our methodology, we sought to identify newly arising CNVs across the malaria parasite genome (termed “de novo” CNVs) and evaluate conditions that affect their evolution. We found that de novo CNVs increased with replication stress. These novel CNVs arose at random locations across the genome and encompassed genes that participate in diverse cellular pathways important for parasite survival. This advance shows that single-cell genomics is accessible for challenging microbes and highlights how this deadly protozoan parasite has adapted to encourage CNV accumulation across its genome. Insights on CNV dynamics moves the field towards a mechanistic understanding of genome plasticity and improves our understanding of how microbial populations expand their host range, tolerate new environments, and survive new antibiotics.

DIFFERENTIAL cfDNA ENRICHMENT IN OPEN CHROMATIN ENHANCES CANCER PREDICTION AND BIOMARKER DISCOVERY.

Sakuntha D Gunarathna, Paige Bonnet, Regina Nguyen, Aerica Nagornyuk, Motoki Takaku

University of North Dakota, School of Medicine and Health Sciences, Department of Biomedical Sciences, Grand Forks, ND

Cell-free DNA (cfDNA) has emerged as a promising non-invasive biomarker that reflects the genetic and epigenetic landscapes of its cells of origin. In cancer, cfDNA fragments exhibit distinct nucleosome positioning patterns, which could be harnessed for cancer prediction. However, a complete understanding of these patterns and their utility in developing computational models for cancer prediction remains elusive. To address this, we isolated cfDNA from plasma samples to establish a streamlined protocol for capturing unique genomic signatures enriched in breast cancer patients. Our analysis confirmed that cfDNA is preferentially enriched in regulatory open chromatin regions, displaying tissue-specific characteristics. By focusing on open chromatin loci relevant to breast cancer and immune cells, we identified 2,804 genomic regions that exhibited differential enrichment in cfDNA from breast cancer patients compared to healthy individuals. To validate the predictive relevance of these differentially enriched loci, we implemented an XGBoost machine learning model using publicly available data to assess their ability to distinguish breast cancer cfDNA patterns. The model demonstrated strong predictive performance, achieving an overall accuracy of 85.29% and a 3-fold cross-validation score of 84.43%, surpassing models trained on randomly selected genomic regions. Furthermore, by expanding our analysis to include all previously defined ATAC-seq peaks in luminal breast cancer and CD4-positive T cells, our optimized machine learning model achieved an even higher accuracy of 92.06%, with a 3-fold cross-validation score of 89.04%. Notably, the established XGBoost model provides interpretable outputs, enabling the identification of key genomic regions essential for cancer prediction. Our findings underscore the potential of cfDNA as a non-invasive screening tool for cancer detection and demonstrate an effective strategy for pinpointing critical genomic loci that distinguish cancer patients from healthy individuals, laying the foundation for a robust non-invasive diagnostic approach.

COMPACT NATIVE PROMOTER DESIGN WITH MACHINE LEARNING-GUIDED MINIATURIZATION

Laura Gunsalus, Avantika Lal, Tommaso Biancalani, Gokcen Eraslan

Biology Research | AI Development, gRED Computational Sciences, Genentech, South San Francisco, CA

Compact and functional cell type-specific cis-regulatory elements (CREs) are essential for gene therapy applications that rely on size-limited vectors. Existing miniaturized sequences have been hand-selected and curated, relying on costly experimental iteration to maintain potency and specificity. We present a method for designing compact and specific regulatory elements by nominating and iteratively editing endogenous elements with state-of-the-art DNA sequence-to-function models. Our approach involves scoring elements in silico, removing subsequences with limited predicted impact, and introducing minimal mutations to increase specificity. We demonstrate the effectiveness of our approach by reducing a 10kb heart-specific locus with proximal regulatory elements to under 300bp. Furthermore, we investigate the minimum viable element size by applying our method at scale across cardiomyocyte-specific enhancers as an alternative application to distal elements. Our method offers a generalizable framework for engineering mini-elements across diverse target cell types. More broadly, we identify core sequence features within larger regulatory elements sufficient to determine cell-type specific expression patterns, advancing our understanding of the mechanisms underlying precise control of gene expression.

GEMINI: A BREAKTHROUGH SYSTEM FOR ROBUST GENE REGULATORY NETWORK DISCOVERY, ENABLING THE APPLICATION OF GENE REGULATORY NETWORKS TO INDUSTRIAL LEVEL GENETIC ENGINEERING

Ridhi Gutta

Curabitrix LLC, Computer Science, Brambleton, VA

In order to resolve crucial global issues, the widespread application of genetic engineering at an industrial level is key. Effective genetic engineering at an industrial scale hinges heavily on precise cellular control of the organism at hand. However, the majority of synthetically engineered strains fail at the industrial level due to disruptions in gene regulation. This stems from a lack of understanding and usage of gene regulatory networks (GRNs), which control cellular processes and metabolism. Research shows that effective manipulation of host GRNs and effective introduction of synthetic GRNs can improve product yield and functionality significantly. However, current GRN inference tools are extremely slow, inaccurate, and incompatible with industrial scale processes, because of which there are no complete expression based GRNs for any commonly used organism, limiting the application of GRNs as a practical tool in genetic engineering at the industrial level. This research proposes a novel computational system, GEMINI, to enable fast and efficient GRN inference for integration into industrial scale pipelines. GEMINI consists of two main parts. First, we create a novel information theoretic algorithm that replaces traditional sequential inference and calculation methods, ensuring compatibility with parallel processing. Second, we integrate a novel GNN architecture based on spectral convolution to bypass intensive eigenvalue computation and efficiently learn global and local regulatory structures. On the DREAM4 and DREAM5 in silico benchmarks, GEMINI outperforms all industry leaders in terms of AUROC and AUPRC, achieving a nearly 300% increase in AUPRC compared to the industry leading method, GENIE3. When applied on a real biological *E. coli* dataset, GEMINI not only recovered 98% of existing interactions, but discovered 468 novel candidate interactions, which were validated against literature. Thus, GEMINI was able to construct the most complete expression based GRN of *E. coli* to date, providing a novel biological blueprint for genetic engineers to use at the industrial level. GEMINI removes reliance on expensive computing equipment and enables fast and accurate GRN inference for the first time, opening doors to more efficient gene expression control and metabolic pathway manipulation for more effective application of genetic engineering at an industrial level.

UNVEILING NON-CODING REGULATORY DRIVER MUTATIONS IN METASTATIC MELANOMA THROUGH ALLELE-SPECIFIC TRANSCRIPTION FACTOR FOOTPRINTING

Jessica Hacheney*[1], David van Bruggen*[1], Muyi Yang[1], Suzanne Egyhazi Brage[1], Hildur Helgadottir[1,2], Martin Enge[1]

[1]Karolinska Institutet, Department of Oncology-Pathology, Stockholm, Sweden, [2]Karolinska University Hospital, Theme Cancer, Stockholm, Sweden

*Authors contributed equally

Large-scale population studies have established that most cancer-predisposing genetic variants reside in cis-regulatory elements (cREs) rather than protein-coding regions. However, even expansive whole-genome sequencing (WGS) efforts have struggled to identify non-coding driver mutations, leading to the conclusion that they are less frequent than their coding counterparts. We hypothesize that these regulatory drivers have remained largely undetected due to technical limitations and propose a novel approach to uncover them using allele-specific transcription factor footprinting derived from genome-wide chromatin accessibility profiling paired with transcriptional profiling of the same samples.

As a proof of concept, we retrospectively analyzed 130 fresh-frozen metastatic melanoma patient samples and cell lines. We developed a state-of-the-art deep learning model capable of predicting the effects of non-coding sequence variation on chromatin accessibility and gene expression. Using this model, we identify recurrent functional non-coding mutations in regions potentially regulating metastatic melanoma-associated genes, such as *STAT1*. Furthermore, we detect convergent evolution of *de novo* cREs, where distinct non-coding mutations create new transcription factor binding sites for the same regulator, alongside recurrently targeted proximal genes.

Our findings suggest widespread positive selection of somatic variants during melanoma progression. The approach is broadly applicable to large-scale studies of gene regulatory mutations across tumor types and holds the potential to significantly advance our understanding of cancer biology.

EXTENSIVE MODULATION OF A CONSERVED CIS-REGULATORY CODE ACROSS 625 GRASS SPECIES

Charles O Hale[1], Sheng-Kai Hsu[2], Jingjing Zhai[2], Aimee J Schulz[1,3], Taylor AuBuchon-Elder[4], Germano Costa-Neto[2], Matthew B Hufford[5], Elizabeth A Kellogg[4], Thuy La[2], Alexandre P Marand[6], Arun Seetharam[5], Armin Scheben[7], Michelle C Stitzer[2], Travis Wrightsman[1], M Cinta Romay[2], Edward S Buckler[1,2,3]

[1]Cornell University, Section of Plant Breeding and Genetics, Ithaca, NY, [2]Cornell University, Institute for Genetic Diversity, Ithaca, NY, [3]University of Minnesota, Department of Agronomy and Plant Genetics, St. Paul, MN, [4]Donald Danforth Plant Science Center, Donald Danforth Plant Science Center, St. Louis, MO, [5]Iowa State University, Department of Ecology, Evolution, and Organismal Biology, Ames, IA, [6]University of Michigan, Department of Molecular, Cellular, and Development Biology, Ann Arbor, MI, [7]Cold Spring Harbor Laboratory, School of Biological Sciences, Cold Spring Harbor, NY, [8]USDA, ARS, Ithaca, NY

The growing availability of genomes from non-model organisms offers unprecedented opportunities to use comparative genomics to pinpoint functional loci underlying trait variation. Cis-regulatory regions drive much of phenotypic evolution, but linking these sequences to specific functions remains a major challenge. We identified a set of 496 cis-regulatory sequence motifs enriched in the regulatory regions of diverse grass species. 82% of motifs were consistently enriched across all species, suggesting the presence of a deeply conserved regulatory code. We then quantified conservation of specific motif instances across 625 grass species. We uncovered widespread gain and loss of cis-regulatory motifs across species, with a nonlinear decay in motif conservation over increasing evolutionary time scales. Approximately 50% of maize motif instances were found to be conserved in rice across roughly 70 million years of divergence. Motif conservation varied widely across genes. We observed subtly higher motif conservation at transcription factor genes compared to downstream target genes, suggesting that the regulatory regions of highly pleiotropic genes may be under stronger constraint. We then tested for adaptive cis-regulatory changes, using phylogenetic mixed models to detect motif gains and losses associated with environmental niche transitions. Our results revealed polygenic patterns of adaptation, with weak but significant convergence at several hundred individual motif instances. These findings support a model in which cis-regulatory evolution occurs primarily via extensive turnover of a conserved regulatory code. Regulatory changes at hundreds if not thousands of genes appear to underpin environmental adaptation. Our findings underscore the potential of comparative genomics and phylogenetic mixed models to discover genetic loci underlying trait variation.

GENETIC VARIATION AND DNA METHYLATION ASSOCIATED WITH LOCAL ADAPTATION IN GROWTH RATE OF ATLANTIC SILVERSIDES (*MENIDIA MENIDIA*)

Søren B Hansen[1], Jessica Rick[2], Michael L Pepke[1], Kasper D Hansen[3], Nina O Therkildsen [4], Morten T Limborg[1]

[1]University of Copenhagen, Center for Evolutionary Hologenomics, Copenhagen, Denmark, [2]University of Arizona, School of Natural Resources & the Environment, Tucson, AZ, [3]Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD, [4]Cornell University, Department of Natural Resources and the Environment, Ithaca, NY

Through millions of years of evolution, organisms have adapted to biotic and abiotic environmental conditions via local adaptation. One example of this is the Atlantic silverside, an annual pelagic fish, whose growth rate is locally adapted to the varying length of the growth season along the North American East Coast. Studies of wild and laboratory-reared individuals have identified that the higher growth rate in northern individuals is associated with genetic variance on several large genomic linkage blocks and chromosome inversions. While chromosome inversions are increasingly appreciated for their importance in adaptation and speciation in the presence of gene flow, the exact mechanisms remain unclear.

In this study we analyzed differences in DNA methylation as a potential factor influencing the evolution of growth rate differences in Atlantic silversides. Using Whole Genome Bisulfite Sequencing (WGBS) on 68 individuals from a size selection experiment led by David Conover and Stephen Munch, we identified Differentially Methylated Regions (DMRs) among selection groups and replicate lines. These regions were often consistently differentially methylated across different tissues and located in highly divergent regions including inversions segregating in the wild.

A concern in WGBS studies of DNA methylation is bias from sample-specific genomic variation in CpG sites, which could lead to similar but false conclusions. We adjusted for genetic variation using genotype calls from low-coverage whole genome sequencing and additionally sequenced 9 individuals carrying different orientations of chromosome inversions using Oxford Nanopore sequencing. By combined sequencing of DNA and DNA methylation analyzed using our likelihood-based framework for sample-specific genetic variation, this confirmed high nucleotide diversity and the presence of specific DMRs between the inversions. The longer reads further helped to resolve structural variation near DMRs, potentially acting as meQTLs, while haplotype phasing of heterozygous individuals elucidated strong allele-specific patterns.

Our results suggest that meQTL-driven DMRs play a role in the local adaptation of Atlantic silversides and demonstrate the advantage of new sequencing platforms for combined DNA and DNA modification analysis.

EVALUATING THE DYNAMICS OF GERMLINE MUTATION AT HOMOPOLYMERS WITH AVITI SEQUENCING IN A LARGE, MULTI-GENERATIONAL PEDIGREE

Hannah C Happ[1,2], Thomas A Sasani[1,2], Derek Warner[1,3], Deb Neklason[2,4], Aaron R Quinlan[1,2]

[1]University of Utah, Department of Human Genetics, Salt Lake City, UT, [2]University of Utah, Utah Center for Genetic Discovery, Salt Lake City, UT, [3]University of Utah, DNA Sequencing Core, Salt Lake City, UT, [4]University of Utah, Huntsman Cancer Institute, Salt Lake City, UT

De novo mutations (DNMs) are a fundamental source of genetic diversity. By sequencing thousands of familial genomes, including those of 33 large, multi-generation CEPH/Utah pedigrees, we understand many of the factors that influence single-nucleotide and structural mutation. However, nearly all studies ignore the most mutable loci in the human genome: short tandem repeats (STRs) and homopolymers. The high mutability is largely understood to be due to high rates of polymerase slippage during DNA replication. This same mechanism plagues Illumina short-read sequencing, leading to an elevated error rate near repetitive sequences that overwhelms the signal of true mutations and hinders studies of STR mutation and homopolymers in particular.

To overcome this limitation, we performed PCR-free genome sequencing of 49 members of a four-generation CEPH/Utah pedigree with the Element Biosciences' new AVITI sequencing technology. We genotyped each individual at ~800k homopolymer loci to identify putative DNMs in 39 trios within the pedigree. Notably, we can genotype 66% more homopolymer loci using Element data compared to Illumina data. This substantial increase is largely due to a lower sequencing error rate, which yields more aligned reads for genotyping.

We observe a median of 35 (range: 15-59) homopolymer DNMs per individual and a median mutation rate of 4.89 x 10e-5 (range: 2.73 x 10e-5 - 8.65 x 10e-5) DNMs per locus per generation. For seven individuals in the pedigree for whom we have both parents and offspring, we can identify the parent-of-origin for DNMs by assessing transmission to the offspring of the focal individual. From this set of high-confidence DNMs, we observe a ratio of paternal to maternal DNMs of 2.89:1 and find that 73% of DNMs are paternal in origin. Analyses to interrogate how homopolymer mutation rates vary as a function of length, sequence context, coding or noncoding context, and parental age and sex are ongoing and progress will be presented. Together, this work will advance our understanding of the fundamental genome biology underlying these hypermutable genomic elements and create important references of variation. Finally, this technology is a promising substrate for characterizing variation in homopolymer and other STR types at disease loci.

EFFECTS OF TRANS-ACTING REGULATORY MUTATIONS ON GENE EXPRESSION PLASTICITY AND FITNESS

Taslima Haque[1], Patricia J Wittkopp[1,2]

[1]University of Michigan, Department of Ecology and Evolutionary Biology, Ann Arbor, MI, [2]University of Michigan, Department of Molecular, Cellular, and Developmental Biology, Ann arbor, MI

Variation in gene expression is a major driver of phenotypic evolution. Both cis- and trans-regulatory variants contribute to gene expression differences. Empirical studies have shown that trans-acting changes tend to have more wide-spread effects on gene expression and larger pleiotropic effects, suggesting that trans-acting changes are more likely to be under purifying selection. Moreover, in different environments, relationships in gene regulatory networks can change, causing genetic variants to have different effects on both gene expression and fitness, which is a phenomenon known as plasticity. We hypothesized that gene expression plasticity can modify pleiotropic effects associated with trans-acting mutations and allow deleterious mutations to persist in populations evolving in heterogeneous environments. To test this hypothesis, we used a TDH3 reporter gene system in the baker's yeast Saccharomyces cerevisiae to estimate expression plasticity and fitness of 51 strains, each carrying a single regulatory mutation (12 cis- and 39 trans-regulatory mutants), in four different growth environments (glucose, galactose, glycerol, and ethanol). We observed a significant plastic response of TDH3 reporter gene expression in a subset of 6 trans-regulatory mutants among different environments. Subsequently, we estimated the fitness of these plastic trans-regulatory mutants using a competitive fitness assay. Our preliminary results demonstrated fitness variation in some trans-regulatory mutants across different environments, underscoring the potential of this framework to test our hypothesis. We are now using the plastic and non-plastic trans-regulatory mutants we identified to test our hypothesis that environmental changes mediate pleiotropic effects and potential evolutionary fates of different regulatory variants. Ultimately, understanding how plasticity in pleiotropy and gene regulatory networks is impacted by-regulatory mutations will allow us to test ideas about environmental heterogeneity facilitating the persistence of deleterious mutations in populations.

COMPLETE ASSEMBLIES AND PANGENOME REFERENCE REVEAL UNIQUE FEATURES IN THE COMPLEX GENOMIC REGIONS OF TIBETAN HIGHLANDERS

Yaoxi He[1], Kai Liu[1], Leyan Mao[1], Dongya Wu[2], Yafei Mao[3], Bing Su[1]

[1]Kunming Institute of Zoology, Chinese Academy of Sciences, State Key Laboratory of Genetic Evolution & Animal Models, Kunming, China, [2]Zhejiang University, Center for Evolutionary & Organismal Biology, Hangzhou, China, [3]Shanghai Jiao Tong University, Bio-X Institutes, Shanghai, China

Pangenome references and complete genome assemblies provide unprecedented opportunities to gain in-depth insights into the human genomic complexity and diversity. Here, we present a pangenome reference of Tibetan populations by sequencing 105 Tibetan genomes from 35 complete trios. We integrated 70 haplotype-resolved assemblies in the pangenome graph, including 8 nearly complete genomes and 27 high-quality diploid genomes. We identified ~15 million small variants and 223,151 structural variants (SVs), of which 5 million small variants and 53,829 SVs have not been reported in the current global pangenome reference. We added 122 million base pairs of euchromatic polymorphic sequences relative to the reference genome of T2T-CHM13. Furthermore, we characterized a large number of novel genomic variants and architectures in the complex genomic regions using fully-sequence-resolved Y chromosomes and T2T autosomes of the Tibetan genomes, especially the distinctive repeat motifs in telomeres and centromeres. Additionally, we uncovered novel inversions and variable number tandem repeats (VNTRs) enriched in Tibetan populations with potential contribution to high-altitude adaptation. We also identify 1,911 archaic-introgressed segments and reconstructed the evolutionary history of the Denisovan-EPAS1 haplotype in the Tibetan genomes. Our study demonstrates the efficacy of pangenome analysis in resolving complex regions of the human genome shaped by population history and biological adaptation.

FOLDBACK READ ARTIFACTS IN OXFORD NANOPORE DATASETS

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

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

Cancer genomes undergo significant and complex genomic rearrangements, which long-read technologies from Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PB) can help elucidate at the DNA and RNA levels. For long-read structural variation (SV) calling, we typically look for read "breakpoints," where fragments of the same read map to different locations on a reference genome. When doing so, we discovered an elevated number of ONT reads supporting foldback or inverted duplication SVs throughout the genome. In ONT direct-cDNA samples of the HCC1395, HCT116, and K562 cell lines, approximately 9-14% of all reads supported a foldback event, while in matched ONT direct-RNA samples, at most one read supported a foldback event. We suspect that the elevated rate of foldbacks is not an actual biological event but rather a technical artifact. We analyzed cDNA samples from mouse brain and liver samples, the HCT116 and K562 cell lines (SG-NEx consortium), and gDNA metagenomic samples to explore this technical artifact further. We found numerous reads had known adaptor sequences between the foldback alignments. Foldback artifacts were most prevalent in direct-cDNA libraries (9-14%), observed at lower rates in metagenomic gDNA libraries (0.5-3%) and standard cDNA libraries (~0.1%), and absent in direct-RNA libraries. Unidentified foldback artifacts can lead to specificity issues in SV calling and tangles in assembly graphs. Here, we propose a quality control tool to identify these foldback artifacts by leveraging their palindromic nature.

THE IGVF CATALOG

Ben Hitz, The IGVF Consortium

Stanford University, Genetics, Palo Alto, CA

Our genomes influence nearly every aspect of human biology from molecular and cellular functions to phenotypes in health and disease. Human genetics studies have now associated hundreds of thousands of differences in our DNA sequence ("genomic variation") with disease risk and other phenotypes, many of which could reveal novel mechanisms of human biology and uncover the basis of genetic predispositions to diseases, thereby guiding the development of new diagnostics and therapeutics. Yet, understanding how genomic variation alters genome function to influence phenotype has proven challenging. To unlock these insights, we need a systematic and comprehensive catalog of genome function and the molecular and cellular effects of genomic variants. Toward this goal, the Impact of Genomic Variation on Function (IGVF) Consortium will combine approaches in single-cell mapping, genomic perturbations, and predictive modeling to investigate the relationships among genomic variation, genome function, and phenotypes. We present a first look at the IGVF Variant-Element-Phenotype Catalog, a sophisticated Knowledge Graph and User interface which puts the work of the consortium in the context of human biology. https://catalog.igvf.org

CHARACTERIZING THE DEMOGRAPHIC HISTORY OF THE ECOLOGICALLY IMPORTANT ACROPORA GENUS OF STONY CORALS

Carla R Hoge*[1,2], Arjun S Krishnan*[2], Daria Bykova[2], Ana Pinharanda[2], Zachary Fuller[2], Veronique Mocellin[3], Line Bay[3], Peter Andolfatto+[2], John Novembre+[1], Molly Przeworski+[2]

[1]University of Chicago, Department of Human Genetics, Chicago, IL,
[2]Columbia University, Department of Biological Sciences, New York, NY,
[3]Australian Institute of Marine Science, Townsville, Queensland, Australia

* contributed equally
+ co-supervised this work

Species in the Acropora genus are some of the most abundant reef-building corals around the world, and therefore central to a critical marine ecosystem. Many species in the genus have been the focus of ecological studies and reef restoration efforts, enabled by their rapid growth rate and broadcast spawning mode of reproduction. Yet despite their ecological significance, relatively little is known about Acropora demographic histories, and the few studies conducted have often considered one species at a time, despite evidence for introgression. We collected high coverage whole genome resequencing (WGS) data from *A. millepora* colonies sampled from reefs across the Great Barrier Reef (GBR) and analyze these data alongside recently published WGS datasets from four additional Acropora species sampled in the wild and high-quality reference genomes from eight Acropora species. This collation includes members of the genus that share a common ancestor ~50 Mya but are now found across the world, as well as closely related species with ranges overlapping on the GBR. To integrate the datasets, we generated a multi-species alignment and called variants using the same bioinformatic pipeline for all species. Using new approaches to visualize and study historical connectivity patterns, we characterized the population structure within each species across their ranges, identifying signals of asymmetric and long-range historical migration. Additionally, we inferred a phylogeny for the group using whole genome data and characterized introgression among species. Together, these analyses provide a comprehensive picture of the evolutionary history of a key coral genus in the response to climate change.

RARE PREDICTED LOSS-OF-FUNCTION AND DAMAGING MISSENSE VARIANTS IN *CFHR5* ASSOCIATE WITH PROTECTION FROM AGE-RELATED MACULAR DEGENERATION

Aaron M Holleman, Aimee M Deaton, Rachel A Hoffing, Lynne Krohn, Philip LoGerfo, Paul Nioi, Mollie E Plekan, Sebastian Akle Serrano, Simina Ticau, Tony E Walshe, Anna Borodovsky, Lucas D Ward

Alnylam Pharmaceuticals, Human Genetics, Cambridge, MA

Age-related macular degeneration (AMD) is a leading cause of blindness among older adults worldwide, but treatment options are limited. Genetics studies have implicated the *CFH* locus, containing *CFH* and five *CFHR1-5* genes, in AMD. While the *CFH* gene has been robustly linked with AMD risk, potential additional roles for the *CFHR* genes remain unclear, obscured by strong linkage disequilibrium across the locus. Investigating rare coding variants can help to identify causal genes in such regions. We used whole exome sequencing data from 406,952 UK Biobank participants to examine AMD associations with genes at the *CFH* locus. For each gene, we used burden testing to examine associations of rare (MAF<1%) predicted loss-of-function (pLOF) and predicted-damaging missense variants with AMD. We considered 'broadly defined AMD' (ICD-10 35.3; n_{cases}=10,700) and 'strictly defined AMD' (dry or wet AMD; n_{cases}=346). Adjusting for *CFH*-region variants known to independently associate with AMD, we find that *CFHR5* rare variant burden significantly associates with decreased risk of broadly defined AMD (OR=0.75, $p=7x10^{-4}$), with this association primarily driven by pLOF variants. Furthermore, the association of *CFHR5* rare variants with AMD protection is estimated as stronger for carriers of the *CFH* Y402H AMD risk allele (interaction p=0.04). Corresponding analyses of strict AMD were underpowered. However, we observe that thinning of the photoreceptor layer outer segment strongly predicts strict AMD, and find that *CFHR5* rare variant burden significantly associates with increased thickness of this retinal layer (+0.34 SD, $p=4x10^{-4}$, n=45,365). These findings suggest *CFHR5* inhibition as a potential therapeutic approach for AMD.

AFCONVERGE: MAPPING THE HIDDEN REGULATORY LANDSCAPE OF CONVERGENT EVOLUTION

Rezwan Hosseini[1], Elysia Saputra[1,2], Nathan Clark[3], Maria Chikina[1]

[1]Joint Carnegie Mellon University - University of Pittsburgh Program in Computational Biology, Department of Computational and Systems Biology, Pittsburgh, PA, [2]Merck & Co., Inc., Department of Data, AI and Genome Sciences, Rahway, NJ, [3]University of Pittsburgh, Department of Biological Sciences, Pittsburgh, PA

Understanding the relationship between genotype and phenotype is a central challenge in genomic science. Comparative genomics offers a powerful lens for this inquiry, as phenotypic variation across species far exceeds that within populations or what can be achieved through direct manipulation. While studies of individual species with extreme traits have yielded insights, convergent traits provide a statistically robust foundation for pinpointing trait-associated genetic changes entirely through computational means.
Sequence-based convergence analysis has yielded insights into diverse phenotypes, from limb morphology to hair density and life history traits. However, much of mammalian phenotypic diversity is thought to arise from regulatory changes rather than coding variation, complicating the link between genetic elements and functional outcomes. Regulatory elements, composed of transcription factor (TF) binding motifs, can undergo rapid turnover while preserving function, and even single-nucleotide polymorphisms can drive dramatic phenotypic shifts in the right context.
We introduce AFconverge, an alignment-free framework for uncovering associations between convergent traits and non-coding elements. AFconverge projects syntenic sequences into a functional layer—either through TF binding motifs or more complex regulatory models—and applies a phylogenetically calibrated statistical approach to detect high-confidence trait associations. Benchmarking experiments on the classical case of vision loss in mammals demonstrate that AFconverge outperforms alignment-based methods in correctly predicting the convergent degradation of ocular functions. We further apply AFconverge to study body mass variation, a highly variable mammalian trait, leveraging a diverse set of sequence features, including mono- and di-nucleotide content, TF motifs, and higher-level functional predictions. We find that as previously reported, body size correlates with shifts in nucleotide and dinucleotide composition in promoters. However, applying a factor analysis model across our multi-scale functional feature set reveals independent and specific regulatory pathways with plausible mechanistic associations with the evolution of large body size.

*Conflict of Interest Disclosures:
Author Elysia Saputra is currently employed by Merck Sharpe and Dohme LLC, a subsidiary of Merck & Co., Inc. Merck Sharpe and Dohme LLC was not involved in funding the presented research, nor play any role in the study design, collection, analysis, interpretation of data, and writing of the manuscript. All other authors declare no competing interests.

BENCHMARKING POOLED CELL CULTURE AND EXPERIMENTAL PERTURBATIONS FOR EXAMINING REGULATORY RESPONSES ACROSS EVOLUTIONARY SCALES IN PRIMATES

Christian Gagnon[1], Amy Longtin[2], Kathrin Köhler[1], Audrey Arner[2], Jenny Tung[1,3], Amanda Lea[2], Genevieve Housman[1]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Primate Behavior and Evolution, Leipzig, Germany, [2]Vanderbilt University, Department of Biological Sciences, Nashville, TN, [3]Duke University, Departments of Evolutionary Anthropology and Biology, Durham, NC

Genetically distinct individuals often respond to the same environmental stimuli in different ways. However, these genotype-by-environment interactions (GxE) are poorly understood from an evolutionary perspective. Interrogating GxE in human populations has become more accessible by combining perturbation screens and cell culture methods, including "village-in-a-dish" approaches where cell lines derived from different individuals are pooled and cultured together. Here, we evaluate whether a similar experimental approach – including a "phylogeny-in-a-dish" approach – can be applied across nonhuman primate species to study GxE across evolutionary timescales.

We pooled lymphoblastoid cell lines derived from multiple individuals and species (5 chimpanzees, 2 bonobos, 2 gorillas, and 2 orangutans) into village-in-a-dish and phylogeny-in-a-dish experiments, exposed these cells to a maximum of 45 control and perturbation conditions, and measured single-cell gene expression patterns (scRNA). Altogether, our data set consisted of scRNA profiles for 231 individual-conditions. We compared our results to parallel experiments using a traditional unpooled approach and bulk gene expression measurements (bulkRNA).

Pooled experiments recovered expression data from over 50,000 individually-assignable cells. Expression patterns were highly correlated between scRNA and bulkRNA data for matched individual-conditions (median r=0.904), which is much greater than correlations across mismatched samples ($p<2.2e-16$). This result is robust across control and perturbed conditions. However, different perturbations produced variable regulatory responses. Of particular interest, immune-, hormone-, and drug-related perturbations produced a mixture of conserved and divergent patterns across species.

Altogether, these findings suggest that village-in-a-dish and phylogeny-in-a-dish approaches are appropriate strategies for scaling-up the number and phylogenetic distribution of cell lines used to study GxE in culture. Our results also provide preliminary insight into which perturbation categories may be of interest for comparative research in primates. Continued work with these systems at larger scales will improve our understanding of GxE variation in primates and the evolution of environmental sensitivity.

DALE-EVAL: A COMPREHENSIVE CELL TYPE-SPECIFIC EXPRESSION DECONVOLUTION BENCHMARK FOR TRANSCRIPTOMICS DATA

Mengying Hu[1], Martin Zhang[2], Maria Chikina[1]

[1]University of Pittsburgh, Department of Computational and Systems Biology, Pittsburgh, PA, [2]Carnegie Mellon University, Ray and Stephanie Lane Computational Biology Department, Pittsburgh, PA

Bulk transcriptomic data represents a composite of signals from multiple cell types. The growing interest for extracting cell-type-specific insights from such data has fueled the rapid development of cell-type expression deconvolution methods. These approaches aim to infer cell-type-specific gene expression (CTSE) profiles from bulk RNA-seq data, providing a means to achieve cell-type resolution expression without the need for single-cell profiling. The deconvolved profiles hold promise in wide-ranging applications such as cell-type-specific expression quantitative trait loci (eQTL) mapping and differential expression (DE) analysis. However, the extent to which these results outperform simple baselines on capturing inter-sample variation, which is most relevant to downstream interpretation, has been largely unexplored.

To address these gaps, we present **DALE-Eval** (Deconvolution Assessment in Real Environments), a first documented comprehensive benchmarking study to systematically evaluate 8 CTSE deconvolution methods. Leveraging large-cohort single-cell RNA-seq (scRNA-seq) datasets, we construct benchmark datasets spanning diverse biological contexts—including tumor, blood, and brain tissues. Our study advances previous benchmarks by: **(1)** emphasizing inter-sample correlation as a biologically meaningful performance metric, **(2)** comparing against simple baselines, and **(3)** assessing the cell-type specificity of deconvolved expression.

Overall, we find that current CTSE deconvolution methods fall short of recapitulating cell-type-resolved profiles achieved by single-cell data, instead providing only partial gene-level deconvolution. Even the most promising methods accurately predict only 5% of genes per cell type and the predictions are only reliable for the cell type with the highest expression. Moreover, most methods fail to capture cell-type specificity, often inflating cell-type marker expression across all cell types, making it difficult to distinguish deconvolved profiles between cell types.

Finally, we provide practical guidelines for applying CTSE deconvolution, recommendations for interpreting results across methods, and post hoc strategies to identify accurately predicted genes. Together, these insights will inform real-world applications and drive the development of improved deconvolution approaches.

MICROBIOME-ASSOCIATED HOST VARIANTS ACT IN TISSUES BEYOND SAMPLING SITES

Naomi E Huntley[1,2], Emily R Davenport[1,2]

[1]The Pennsylvania State University, Biology, University Park, PA, [2]The Pennsylvania State, One Health Microbiome Center, University Park, PA

The microbiome plays a vital role in maintaining our health, with its composition influenced both by environmental and host genetic factors. Genome-Wide Association Studies (GWAS) have identified hundreds of host genetic variants associated with the relative abundance of gut bacteria, referred to as microbiome associated host variants (MAVs). However, most MAVs are intergenic, leaving the genes, pathways, and tissues involved unclear. Identifying the tissues and cell-types these variants act in is a crucial first step for understanding the physiological mechanisms by which the host regulates microbial abundance via genetic mechanisms. To address this, we took an unbiased approach to identify microbiome-relevant organs and cell-types across the body. Specifically, using the single-cell disease relevance score (scDRS) method, we integrated publicly available microbiome GWAS summary statistics for 52 bacterial genera generated from 18,340 individuals and single-cell transcriptomic data from 120 cell types collected from 23 organ systems. We found organs and cell types outside of the gut were implicated, suggesting genetic mechanisms can occur beyond the tissues directly adjacent to the sampling site. Of the 52 bacterial genera tested, five were significantly associated with a variety of organs, including heart, mammary gland, gonadal adipose tissue, lung, and tongue. For example, we found that MAVs for the genus *Intestinimonas* are predicted to function in the tongue ($P<0.001$), while MAVs associated with *FamilyXIIIAD3011* are predicted to function in adipose tissue ($P<0.001$). By analyzing the data at the single-cell level, we further identified associations that were not apparent when analyzing the data at the organ level, highlighting the importance of considering cellular heterogeneity within tissues. We found that 12 bacterial genera show cell-type level results. For example, while no organs were implicated in the scDRS analysis of the MAVs for genus *Streptococcus*, at the cell-level, pancreatic polypeptide cells were significantly associated ($P<0.001$). Additional associations included *Eubacterium* in smooth muscle cell of the pulmonary artery ($P<0.001$), *Intestinimonas* in keratinocytes (epidermis) ($P<0.001$), and Ruminococcaceae NK4A214 group in astrocyte cells (central nervous system) ($P<0.001$). The diverse tissues and cell types implicated here suggests that host genetic regulation of the microbiome extends beyond local interactions to distant organs and specialized cells, underscoring its systemic nature. This aligns with evidence linking gut microbiome composition to non-gastrointestinal conditions like heart disease and depression. By pinpointing the tissues and cell types where MAVs act, these findings lay the groundwork for future mechanistic studies into how host genetic variation shapes the gut microbiome and its systemic effects on health.

A STUDY OF SINGLE-CELL MULTIOMICS BASED ON THE ANALYSIS OF CANCER DRIVER MUTATION DIVERSITY

Tadashi Imafuku[1], Kyohei Matsumoto[1], Shigeyuki Shichino[2], Shinichi Hashimoto[1]

[1]Wakayama Medical University, Department of Molecular Pathophysiology, Wakayama, Japan, [2]Tokyo University of Science, Division of Molecular Regulation of Inflammatory and Immune Disease, Research Institute for Biomedical Sciences, Tokyo, Japan

Somatic mutations are a primary cause of cancer and have extensively been investigated using bulk DNA-seq. Recent advancements in single-cell genome mutation analysis have revealed the presence of cancer cells with distinct mutation patterns in the tumor microenvironment. However, the relationship between these diverse mutation patterns and their respective contributions to cancer pathogenesis remains unclear. In this study, we have developed a micro-well device capable of single-cell genomic mutation analysis by incorporating oligonucleotide sequences from a cancer panel targeting driver mutations onto microbeads. Furthermore, since these microbeads also enable concurrent analysis of single-cell gene expression and chromatin accessibility, they facilitate comprehensive single-cell multiomics analysis. At first, we aimed to establish a protocol for single-cell genomic mutation analysis and single-cell multiomics analysis using pancreatic cancer cell lines harboring distinct genetic mutations. As a result, we successfully identified cancer cells with different genetic mutations and analyzed their unique transcriptional activities. Subsequently, the analysis was extended to clinical pancreatic cancer tissue samples, which revealed the presence of multiple mutation patterns of cancer cells in the tumor microenvironment. The technology developed in this study contributes to the understanding of the role of genomic mutations in cancer cells and the identification of the novel targets for prediction, diagnosis, and treatment of cancer progression.

A GENOME-TO-PROTEOME ATLAS CHARTS NATURAL VARIANTS CONTROLLING MOLECULAR AND PHENOTYPIC DIVERSITY

Christopher Jakobson[1], Johannes Hartl[2,3], Pauline Trébulle[4], Michael Mülleder[3], Daniel Jarosz[1], Markus Ralser[2,3]

[1]Stanford University School of Medicine, Chemical & Systems Biology, Stanford, CA, [2]Berlin Institute of Health, Berlin, Germany, [3]Charité - Universitätsmedizin Berlin, Department of Biochemistry, Berlin, Germany, [4]University of Oxford, Centre for Human Genetics, Nuffield Department of Medicine, Oxford, United Kingdom

Understanding the genotype-phenotype relationship remains one of the central challenges in genetics, with implications from evolution to health and disease. We reasoned that connecting individual genetic changes to their functional consequences across scales from proteins to their effects on growth could shed light on this problem. By studying the genetic variation in two budding yeast strains adapted to very different ecological niches, we hoped to survey the array of mechanisms that drive adaptation via the proteome. Using a combination of fast and precise mass-spectrometry proteomics and super-resolution genetic mapping, we charted a genome-to-proteome atlas consisted of over 6,400 genotype-protein associations–1,600 of which were resolved to a single underlying polymorphism. This nucleotide-resolution map allowed us to pinpoint causal variants within and outside the core functional domains of proteins, with coding variation playing a critical role in proteomic diversification. Genotype-protein connections revealed coherent physiological regulatory relationships not evident from genetic and physical interaction databases, and potent trans-regulatory effects often originated from enzymes and other factors not conventionally associated with gene regulation. Pairing our aligned genotype-protein and genotype-phenotype atlases identified causal variants and molecular mechanisms underlying drug resistance and revealed fitness effects disguised by epistasis. Finally, the genome-to-proteome map we charted in the absence of stress persisted across environments and forecasted the effects of genetic variation under diverse insults, suggesting that this latent molecular layer is broadly predictive of consequences for the organism.

STRUCTURAL VARIANT DISCOVERY AND CHARACTERIZATION FROM *DE NOVO* ASSEMBLY OF KHOE-SĀN GENOMES

Zoeb N Jamal*[1], Daniela C Soto*[1], Kristin Hardy*[1], Mohamed Abuelanin[1], William Palmer[1], Javier Prado-Martinez[2], Paul Norman[3], Marlo Moller[4], Brenna M Henn[1], Megan Y Dennis[1]

[1]University of California, Davis, Genome Center, Davis, CA, [2]Institute of Evolutionary Biology, PRBB, Barcelona, Spain, [3]University of Colorado, Biomedical Informatics, Aurora, CO, [4]Stellenbosch University, Molecular Biology and Human Genetics, Stellenbosch, South Africa

*Authors contributed equally

Long-read sequencing (LRS) of large cohorts is enhancing understanding of the genomic variation landscape. While West African populations have been included in LRS efforts, many African regions are underrepresented. As genetic diversity within African populations can exceed that between continental groups, this gap suggests ample genetic diversity is missing in large datasets. The Khoe-Sān indigenous peoples of southern Africa carry some of the most divergent haplotypes in extant human lineages. We performed 10X Genomics long-range sequencing and de novo assembly of three Khoe-Sān individuals (1 ≠Khomani San, two Nama), obtaining diploid pseudo haplotypes (N50: 54, 88, and 155 kbp). We identified 12,784 contigs (~12 Mbp) that initially did not map to T2T-CHM13 or GRCh38, of which roughly 2,500 are over 1 kbp, contamination-free, and lack matches in the HPRC and human whole genome sequencing (WGS) samples in the SRA. These contigs are either novel, or only partially match repetitive/complex regions in T2T-CHM13. Focusing on structural variants (SVs) (>50 bp), we identified 34,164 deletions, 6,650 insertions, and 341 inversions against T2T-CHM13, of which 17,104 (50%), 1,971 (30%), and 230 (67%) were unique to Khoe-Sān assemblies when compared with published assemblies from other ancestries generated with 10X long-range data (n=8) and the HPRC year one release (n=44). Of discovered insertions, those with <10% repetitive sequence and absent from T2T-CHM13/GRCh38 were classified as non-reference unique insertions (NUIs), totaling 146 kbp, with 43 kbp being Khoe-Sān specific. NUIs range in size from 50 bp to 1.5 kbp and affect 12 protein-coding genes with moderate to high impact. Additionally, 223 deletions and 17 inversions unique to Khoe-Sān assemblies are predicted to have high impacts on protein-coding genes. Using WGS data from 93 Khoe-Sān individuals, we genotyped discovered SVs and identified genes with divergent copy-number, some of which were not annotated as segmental duplications in T2T-CHM13. To find SVs and copy-number variations with divergent frequencies compared to other ancestries, we are processing HGDP WGS data. Finally, we leveraged long haplotypes to investigate structurally complex loci, including the HLA locus, and assess improved mappability on reference bias in Khoe-Sān individuals.

ROBUST INFERENCE OF CO-REGULATED GENE AN PEAK MODULES FROM CELL TYPE SPECIFIC SINGLE CELL DATA

Benjamin T James[1], Carles A Boix[2], Manolis Kellis[1]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, [2]Harvard Medical School, Department of Genetics, Cambridge, MA

We present an framework, scdemon, for detecting cell-type–specific gene and chromatin accessibility modules from single nucleus RNA sequencing (snRNA-seq) and ATAC sequencing (snATAC-seq) data. Building on our earlier methods we had made to characterize snRNA-seq and snATAC-seq atlases, which introduced this method, our framework introduces significant methodological refinements to robustly identify co-regulated gene or peak programs in a unified framework, with a focus on applications to Alzheimer's disease.
Our approach begins by conceptually modeling the single-cell data through principal component analysis, from which we derive optimal gene–gene and peak–peak weights in a fitting-free procedure directly through principal component regression. To determine significant gene-gene and peak-peak pairs, we subsequently apply a Wald test for each gene (or peak) pair, computing standard errors using asymptotic distributions of these weights derived from matrix perturbation theory through the delta method. Because of this, we mitigate several issues that commonly plague single-cell coexpression methods --- such as sequencing depth and noisy measurements --- while keeping biologically meaningful networks by using the top principal components themselves, which are typically uncorrelated with depth in snRNA-seq or or filtered based on correlation with depth in snATAC-seq.

We then group genes or peaks into modules from either graph-based or hierarchical clustering methods. This integrated strategy builds on top of traditional differential expression (DEG) and differential accessibility (DAP) analyses by aggregating summary statistics from these analyses using these modules to ascertain each module's significance while controlling for the family-wise error rate in multiple comparisons.

We applied our method to a collection of hundreds of postmortem human brain samples spanning multiple brain regions. The results demonstrate that our module-based approach not only scales efficiently but also reliably uncovers cell-type–specific gene programs and chromatin accessibility patterns relevant to human disease. Moreover, this framework facilitates easy integration with standard DEG/DAP analyses, offering a complementary perspective that is particularly beneficial in studies affected by substantial batch effects.

Our tools are implemented as open source software in packages for both R and Python, and they seamlessly integrate with major single-cell analysis platforms including AnnData, SingleCellExperiment, and Seurat.

DRUGSAGE: AN INTERPRETABLE METHOD FOR DRUG RESPONSE IMPUTATION

Peilin Jia

Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

Accurate drug response prediction can help to optimize clinical applications of cancer drugs. However, due to the heterogeneity of genetics and genomics features and the complexity of the tumor microenvironment, cancer samples may respond dramatically differently, even though they carry the same driver mutations or genes. In this work, we developed a new method, DrugSAGE, to enable response prediction based on not only transcriptome features of the sample itself but also its most similar counterparts. DrugSAGE integrates the advantages of the GraphSAGE method to generate embeddings using an aggregation-based strategy. It also implemented a customized linear layer to incorporate gene-pathway annotations, thus enabling interpretability, which helps identify informative pathways critical for drug response prediction. We benchmarked DrugSAGE using the TCGA bulk data, six bulk datasets from GEO, and five scRNA-seq data. Predictions by DrugSAGE showed significant associations between drugs and their known target genes (e.g., HER2+ inhibitor, MET inhibitors, BRAF inhibitors) or significant differences between patient groups stratified by their known treatment (e.g., Erlotinib, PLX4720, 5-Fluorouracil). Applied in scRNA-seq data, we demonstrated that DrugSAGE can be successfully applied in drug response prediction in single-cell transcriptome data. By further exploring the pathway-based embeddings, we identified pathways that play critical roles for the models, including pathways in cancer, MAPK signaling pathway, endocytosis, and JAK-STAT signaling pathway, among others. Compared to previous methods, DrugSAGE showed superior or comparable performance, providing a unique method for drug response prediction.

GLOBAL ACTIVITIES OF THE RNA-DEPENDENT ATPASE DDX41 IN HEMATOPOIESIS AND CANCER

Christina M Jurotich[1], Jeong-Ah Kim[1], Siqi Shen[2], Kirby D Johnson[1], Sunduz Keles[2], Emery H Bresnick[1]

[1]University of Wisconsin School of Medicine and Public Health, Wisconsin Blood Cancer Research Institute, Department of Cell and Regenerative Biology, Carbone Cancer Center, Madison, WI, [2]University of Wisconsin School of Medicine and Public Health, Department of Biostatistics and Biomedical Informatics, Carbone Cancer Center, Madison, WI

The extensive information content of the human genome is considerably amplified by splicing factors, which mediate alternative splicing to generate vast numbers of RNA transcripts. Splicing factors are essential for controlling genome stability, and mutations in genes including *SRSF2*, *SF3B1*, *ZRSR2*, and *DDX41* disrupt pre-mRNA splicing, predisposing to and/or promoting myeloid neoplasms. Heterozygous *DDX41* germline mutations have been identified in familial myelodysplastic syndrome/acute myeloid leukemia and acute erythroid leukemia. DDX41 encodes a multi-functional RNA-dependent ATPase regulating RNA splicing, the innate immune cGAS-Sting pathway, and genome stability, but a detailed mechanistic understanding is lacking. Over 3,000 *DDX41* variants are reported in GnomAD, while only 294 have clinical documentation in ClinVar. Despite many variants altering conserved sequences, how variants impact DDX41 functions genome-wide and confer pathogenesis is unknown. Recent initiatives to inform population genomics led to the *All of Us* consortium, and analysis of the *DDX41* variants reported in this data yielded 120+ novel variants. Further computational analyses of these variants revealed patterns of protein alterations unique to different regions of the gene. We established an innovative genetic rescue system to discriminate pathogenic from benign *DDX41* variants and elucidate mechanisms governing the complex process of hematopoiesis. Analysis of the activities of known disease-associated variants in this assay revealed that mutant proteins lose almost all DDX41 activities, thus unveiling functional defects that inform pathogenesis. By incorporating an ensemble of activity metrics, including capacities to control proliferation, differentiation, gene repression, gene activation, and splicing, with computational analyses, including, but not limited to, Differential Transcript Usage (DTU) and Differential Transcript Expression (DTE), we are transforming the classification system. Using DDX41-specific criteria will enable high-fidelity clinical curation and reduce the number of variants of uncertain significance to advance genomic medicine.

SINGLE-CELL GENOMICS OF PERIPHERAL IMMUNE CELLS REVEALS ANTI-INFLAMMATORY GENE REGULATORY MECHANISMS IN OLDER ADULTS WITH POSITIVE PSYCHOSOCIAL EXPERIENCES

Ali Ranjbaran[1], Cynthia Kalita[3], Julong Wei[1], Julian Bruinsma[2], Henriette Mair-Meijers[1], Sam Zilioli[2,4], Roger Pique-Regi[1], Francesca Luca[3]

[1]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [2]Wayne State University, Department of Psychology, Detroit, MI, [3]University of Chicago, Department of Human Genetics, Chicago, IL, [4]Detroit, Department of Family Medicine and Public Health Sciences, Detroit, MI

Adverse psychosocial factors, like high psychological stress, are linked to chronic inflammation and cardiovascular disease. Conversely, psychosocial resources, such as strong social connections and positive coping strategies, mitigate chronic inflammation. However, the molecular mechanisms for these associations remain largely unknown. To address this gap, we performed single-cell RNA sequencing (scRNA-seq) and ATAC sequencing (scATAC-seq) on peripheral blood mononuclear cells from 165 African American adults (50-89 years), and clustered cells into 6 immune cell types: CD4+ and CD8+ T cells, NK cells, monocytes, B cells and dendritic cells. Participants answered validated questionnaires about positive strategies for coping with stress, sense of control, psychological well being, and self-esteem (psychological resources); completed a well-established scale of perceived social support; and filled out daily diaries to report levels of psychological stress. Social support and psychological resources impacted gene expression across all cell types, with the most differentially expressed genes (DEGs, FDR 10%) in CD4+ T cells, while psychological stress showed the most DEGs in monocytes. Social support and psychological resources shared gene expression changes ($r>0.51$), which negatively correlated with stress-associated changes ($r<-0.38$). Stress upregulated genes enriched for interferon signaling in CD4+ T cells and monocytes, whereas social support and psychological resources downregulated genes in pro-inflammatory pathways. Furthermore, gene expression changes associated with levels of pro-inflammatory cytokines were positively correlated with stress-associated changes and negatively correlated with positive psychosocial factor-associated changes. Using scATAC-seq chromatin accessibility data, we identified differentially active transcription factor motifs (DAMs): 36 DAMs linked to social support in CD4+ T cells, 119 DAMs to psychological resources in monocytes, and 95 DAMs to psychological stress in CD4+ T cells. The antagonistic patterns between positive and negative psychosocial factors observed in gene expression were also reflected in transcription factor motif activity. Interestingly, we found high expression of pro-inflammatory regulators (STAT1, STAT2, and IRF2) in individuals with elevated psychological stress, and low expression in those with high psychological resources in CD4+ T cells. These differences in RNA abundance of pro-inflammatory regulators were reflected in their motif activities. Overall, DAMs were enriched near DEGs, suggesting a mechanism where positive and negative psychosocial experiences affect both transcription factor activities and gene expressions, antagonistically. Our findings offer molecular insights into the gene regulatory mechanisms through which psychosocial determinants influence inflammatory pathways in humans.

MYC AND AP-1 ONCOGENES SYNERGISTICALLY BIND ENHANCERS TO TRANSCRIPTIONALLY REWIRE CELLS

Reshma Kalyan Sundaram[1], Ravi Radhakrishnan[1,2], Bomyi Lim[1]

[1]University of Pennsylvania, Chemical and Biomolecular Engineering, Philadelphia, PA, [2]University of Pennsylvania, Bioengineering, Philadelphia, PA

The transcription factor c-Myc (Myc) is known to regulate a multitude of genes and cellular processes. Myc is deregulated in 70% of human cancers, making it a potent oncogene. We hypothesized that increased Myc levels can alter Myc's transcriptional function by changing Myc-DNA interactions and hence Myc regulated gene expression. Myc deregulation through increased Myc levels has been suggested to cause cells to experience stoichiometric stress which may result in stress-induced cellular reprogramming. To test this hypothesis, we first analyzed the DNA binding patterns of Myc by performing de novo motif discovery analysis on Myc ChIP-seq datasets of several cancer cell lines obtained from the ENCODE database. Our analysis revealed the expected enrichment of Myc's canonical binding motif (EBOX) in all the cell lines examined. Surprisingly, we also observed a cell-type specific co-enrichment of an additional motif, the TRE motif. Since the TRE motif is a well-known binding site for the AP-1 family of transcription factors, we hypothesized that TRE motifs could represent an indirect binding site for Myc occupied in synergy with AP-1. Subsequent analysis integrating Myc and various AP-1 transcription factors ChIP-seq datasets validated this hypothesis. Due to Myc's well-established role in gene regulation through binding both enhancers and promoters, we also studied the difference in enrichment of Myc-bound TRE sites across these regulatory regions in multiple cell lines. Our analysis revealed that the TRE motifs were preferentially enriched at Myc-bound enhancers over promoters, indicating that TRE functions as an enhancer-specific Myc binding site. Given that the Myc oncogene is overexpressed in multiple human cancers, we further examined the effect of increase in Myc levels on Myc binding to TRE sites at enhancers. De novo motif discovery performed on Myc ChIP-seq datasets with low and high Myc expression levels revealed that when Myc levels increase, Myc preferentially occupies TRE sites over EBOX sites at enhancers. This indicates that at high Myc levels, such as in cancerous conditions, the TRE enhancer binding sites could be co-opted for gene regulation by Myc. We also identified Myc regulated genes by integrating ChIP-seq and RNA-seq datasets with differential Myc expression. The results revealed that the TRE enhancer binding site is frequently associated with transcriptional repression by Myc. Gene Ontology on Myc target genes revealed that Myc utilizes TRE binding sites at enhancers to transcriptionally rewire cells, including by modulating several cancer hallmarks such as proliferation, apoptosis, and cell adhesion. Findings from our work will potentially aid in development of new treatment strategies for targeting Myc, particularly approaches that focus on cancer-specific functions of Myc and its co-regulators.

MECHANISMS UNDERLYING CHROMOSOME END-SPECIFIC TELOMERE LENGTH REGULATION IN HUMANS

Rebecca Keener[1], Hyun Joo Ji[2], Aljona Groot[3], Andreas Rechtsteiner[3], Steven Salzberg[1,2], Carol Greider[3], Alexis Battle[1,2]

[1]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]University of California Santa Cruz, Molecular Cell and Developmental Biology, Santa Cruz, CA

Telomeres are repetitive DNA sequences at the ends of linear chromosomes and when dysregulated, can lead to diseases such as idiopathic pulmonary fibrosis and predisposition to cancer. Here we analyzed the regulation of telomere length across individuals and chromosome ends using recent Nanopore long-read DNA sequencing data[1] where reads captured the full telomere length and some proximal sequence, referred to as the subtelomere. For 30 years the telomere length equilibrium model has held that all chromosome ends are regulated around a shared equilibrium length. In our recent study[1] we observed that across 147 individuals some chromosome ends had consistently shorter or longer telomere length. Furthermore, sequencing from HG002 cells showed that for a given chromosome end, the haplotypes may have significantly different telomere length. We hypothesized that sequences in the subtelomere region regulate telomere length.

To begin identifying genomic elements associated with telomere length we leveraged the HG002 T2T diploid reference and paired telomere length sequencing data. We observed that an increase of interstitial telomere repeats within 50 kilobases of the telomere was associated with shorter telomere lengths. In addition, higher copy number of certain tandem repeats was associated with longer telomere lengths. We are now extending these analyses to our 147 individual dataset and will further conduct a subtelomere association study. To circumvent the need for de novo haplotype resolved genome assembly, we developed a method for simulating telomere sequencing data to improve the accuracy of chromosome end assignments via alignment to a reference genome. Our simulation accounts for genetic variation observed from pairwise alignments of HG002 haplotypes and varied read lengths based on empirical data. This simulation will provide guidance for future work leveraging sequencing data to examine the understudied subtelomere region. Following the protocol nominated by our simulation, we will realign the 147 individual dataset to assign reads to chromosome ends, haplotype resolve the data, call genetic variants, and conduct association testing with chromosome end-specific telomere length. Overall, we provide evidence that chromosome end-specific telomere length regulation is driven, in part, by sequence variation. This work will lay the groundwork for a new telomere length regulation model.

1. Karimian et al. Science 2024, PMID: 38603523

TISSUE-SPECIFIC EPIGENOMIC PROFILES INFORM PLEIOTROPIC PARTITIONING OF DISEASE LOCI

Gaspard Kerner[1], Alkes L Price[1,2,3]

[1]Harvard T. H. Chan School of Public Health, Epidemiology, Boston, MA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, [3]Harvard T. H. Chan School of Public Health, Biostatistics, Boston, MA

Genome-wide association studies (GWAS) identify disease associated loci spanning heterogeneous biological processes, but distinguishing these processes remains challenging. We propose Joint Pleiotropic and Epigenomic Partitioning (J-PEP), a method that integrates pleiotropic SNP effects and tissue-specific epigenomic data to partition disease loci into biologically distinct clusters. To benchmark J-PEP against other approaches, we developed a new validation metric—Pleiotropic and Tissue Prediction Accuracy (PTPA)—that evaluates how well clusters predict SNP-to-trait and SNP-to-tissue associations. Applying J-PEP to GWAS summary statistics for 164 diseases/traits (average N=378K), we attained 40-45% higher PTPA than simpler approaches, consistent with simulations. Notably, J-PEP refines previous partitioning efforts, including those applied to type 2 diabetes (T2D), by capturing biological processes—such as developmental pathways—that were previously underemphasized, thereby offering new insights into disease mechanisms. By linking GWAS results to biological processes and their tissue contexts, J-PEP supports the prioritization of therapeutic targets by pinpointing contexts where specific interventions are likely to be most effective.

DECIPHERING PEDIATRIC GLIOMA SUBTYPES: SUPER-ENHANCER DYNAMICS AND (EPI)GENOMIC INSIGHTS INTO CELL OF ORIGIN

Devishi Kesar[1,2,3], Michaela K Keck[1,2,3], Robert J Autry[1,2,3], David T Jones[1,2,3]

[1]Hopp Children's Cancer Center Heidelberg (KiTZ), Division of Pediatric Glioma Research, Heidelberg, Germany, [2]National Center for Tumor Diseases (NCT), Heidelberg, Germany, [3]German Cancer Research Center (DKFZ), Heidelberg, Germany

Pediatric gliomas are a heterogeneous group of brain tumors with distinct molecular subtypes and diverse cells of origin. While genetic alterations have been extensively studied, the regulatory mechanisms underlying glioma subtype specification, particularly for histone H3 WT subtypes, remain poorly understood. Thus, we systematically investigated the role of super-enhancers in shaping the epigenomic landscape of pediatric gliomas, providing insights into their cell of origin and oncogenic drivers.

Using chromatin immunoprecipitation sequencing (ChIP-Seq) data, we identified super-enhancer regions across different glioma subtypes, revealing key regulatory elements and transcription factors that likely contribute to tumor initiation and progression. By integrating these data with multi-omic datasets, including RNA-Seq and copy number alterations, we hope to delineate subtype-specific super-enhancer networks and their association with tumor heterogeneity.

Our findings will offer a comprehensive framework for understanding how super-enhancer dynamics define pediatric glioma subtypes and their potential cells of origin. The ultimate goal of this work is not only to uncover novel biomarkers and insights into developmental biology, but also to provide a foundation for future precision medicine approaches by identifying key regulatory circuits that may serve as therapeutic targets.

BAYESIAN POLYGENIC PREDICTION WITH A NON-PARAMETRIC FUNCTIONALLY INFORMED PRIOR IMPROVES PREDICTION OF COMPLEX TRAITS FROM GENOTYPES

April Kim*[1], Joshua Weinstock*[2], Alexis Battle[3]

[1]Johns Hopkins University, Department of Computer Science, Baltimore, MD, [2]Emory University, Department of Human Genetics, Atlanta, GA, [3]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD

Polygenic risk scores (PRS) estimate an individual's genetic predisposition to complex traits by aggregating effects from multiple single-nucleotide polymorphisms (SNPs). Given their potential clinical and translational impact, numerous statistical methods have been developed to optimize PRS weight estimation, including pruning and thresholding, Bayesian hierarchical models, and penalized regression. While a subset of methods incorporate genomic and regulatory annotations, previous methods assume a linear relationship between annotations and SNP weights, which may limit their ability to capture complex functional effects on genetic architecture.

Here, we introduce the Polygenic Risk Score Functionally-informed Neural Network (PRSFNN), a Bayesian method that uses a neural network to learn the relationship between the parameters of the prior distribution over SNP weights and functional annotations of the SNPs. PRSFNN extends the widely used spike-and-slab prior framework by parameterizing both the prior inclusion probability and slab variance as functions of genomic annotations. This enables the model to learn complex relationships between hundreds of SNP annotations and their effects on SNP importance without restriction to a particular parametric form.

We curated a diverse set of 242 genomic and functional annotations for each SNP, including single-cell chromatin accessibility profiles spanning hundreds of adult and fetal human cell types, regulatory element annotations from ENCODE4, conservation metrics from Zoonomia, and missense variant effect predictions from AlphaMissense. We infer the posterior with a scalable implementation of coordinate ascent variational inference.

We benchmarked PRSFNN against state-of-the-art PRS methods using quantitative traits from the UK Biobank, demonstrating its improved predictive performance, which we evaluated on a test set of individuals that were not included in the original GWAS. Across six phenotypes, inclusion of a functional prior improved out-of-sample prediction accuracy by up to 2.9%, with the greatest performance gain observed for red blood cell distribution width. Notably, PRSFNN achieves the highest predictive accuracy across multiple quantitative traits, highlighting the advantage of non-linear functional priors over traditional linear models. Overall, PRSFNN advances our ability to predict complex traits from genotypes.

MicroRNA PERTURB-SEQ REVEALS GENOME-WIDE FUNCTIONAL TARGETS AND DELETERIOUS 3'UTR VARIANTS

Eyal Ben-David*[1], Doyeon Kim*[1], Wayne Xianding Deng*[1], Zakaria Louadi*[1], Robin Bombardi[1], Marcos Assis Nascimento[1], Thy Pham[1,2], Kyle Kai-How Farh[1]

[1]Illumina, Foster City, CA, [2]Massachusetts Institute of Technology, Department of Biology, Cambridge, MA

MicroRNAs (miRNAs) are essential regulators of post-transcriptional gene expression, yet our understanding of their targets and mechanisms remains incomplete. To obtain a comprehensive catalog of miRNA-transcript target pairs, we performed miRNA perturb-seq with a library of 400 miRNAs and siRNAs in five human cell lines, profiling 2.2 million cells using Fluent PIP-seq V 3' single cell RNA-seq. We assayed all conserved mammalian miRNA families, and used five diverse cell lines (HEK293T, A549, RPE-1, K562, SH-SY5Y) to ensure that ~85% of the human transcriptome was expressed in at least one cell line. We identified a total of ~240,000 significantly downregulated miRNA-transcript target pairs and verified that these followed established rules of miRNA targeting. Leveraging this dataset, we used deep learning to identify rare 3'UTR variants that disrupt gene expression across human genetics cohorts. Together, our study provides a comprehensive resource for deciphering functional miRNA targets and understanding the consequences of non-coding 3' UTR genetic variation.

INTEGRATING KNOWLEDGE GRAPH-BASED DRUG REPRESENTATION WITH CANCER OMICS DATA TO IMPROVE DEEP LEARNING MODELS FOR DRUG RESPONSE PREDICTION

Taeho Kim[1,2], Casey Sederman[1,2], Tonya Di Sera[1,2], Gabor T Marth[1,2]

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

Deep learning models have been increasingly utilized in precision oncology to predict drug responses for targeted cancer therapy. These models primarily rely on two key features: tumor omics data to represent cancer and chemical structure-based vectors (e.g., Morgan fingerprints) to represent drugs. However, chemical structure alone does not fully capture the biological context of drug-target interactions, particularly in the limited chemical space of approved cancer drugs. This limitation hinders the ability of cancer drug response prediction (CDRP) models to generalize to (1) never-before-seen (NBS) drugs and (2) combination therapies. To address this, we generated a knowledge graph-based drug representation that integrates drug-target interactions. We constructed a heterogeneous knowledge graph using drug and protein nodes from the STRING database and the Therapeutic Target Database, then applied *node2vec* (biased random walks and *word2vec*) to learn drug embeddings. We validated these embeddings using t-SNE plots, which revealed superior clustering of cancer drugs targeting the same pathways to Morgan fingerprints. We evaluated the impact of these embeddings in our deep learning-based CDRP model, *ScreenDL*, which predicts ln(IC50) values, and observed improved predictive performance. Moving forward, we plan to explore attribute-assisted node representation learning to further enhance drug response prediction.

TRANSPOSABLE ELEMENT MEDIATED REARRANGEMENTS ACROSS GREAT APE GENOMES

Magda Kmiecik[1], Parithi Balachandran[1], Jessica M Storer[2,3], Rachel J O'Neill[2,3,4], Christine R Beck[1,2,4]

[1]The Jackson Laboratory, Genomic Medicine, Farmington, CT, [2]University of Connecticut, Institute for Systems Genomics, Storrs, CT, [3]University of Connecticut, Department of Molecular and Cell Biology, Storrs, CT, [4]University of Connecticut Health Center, Department of Genetics and Genome Sciences, Farmington, CT

Transposable elements (TEs) are segments of DNA capable of moving from one genomic location to another. The mobility and high inter-element homologous nature of TEs can result in structural variant (SV) formation, leading to TE mediated rearrangements (TEMRs). These rearrangements can have implications for disease, evolution, adaptation and genomic diversity, and can both act as markers of evolutionary relationships and lead to regulatory differences. For this study, we used the recently published T2T genomes of the five great ape species (Bornean orangutan, Sumatran orangutan, bonobo, chimpanzee, gorilla), and generated an SV callset using the human T2T-CHM13 genome as the reference. We used PAV as our primary variant detection tool for deletions and inversions, and pbsv for duplications. Further, we improved the quality of our callset by adding orthogonal support from multiple variant calling tools, such as Sniffles2, Delly2, cuteSV, HiFiCNV. Using this callset, we identified 20,269 TEMRs in regions with <90% tandem repeat content and primarily focused on Alu and L1 elements due to their abundance in primate genomes. Overall, we found that Alu elements and L1 elements drive 8.6% and 5% of the total 127,821 rearrangements, respectively (deletion: 116,372, duplication: 10,874, and inversion: 575). We found that median size of Alu-driven TEMRs was smaller than L1-driven TEMRs for deletions (626bp vs 1.3kb) and inversions (2kb vs 11.7kb) but had the opposite trend for duplications (319bp vs 128bp). The majority of deletions and duplications involve TEs in the same orientation, 88% and 90% respectively, while elements flanking inversions are largely in the opposite orientation (71%). In total, 156 Mb of sequence is impacted by TEMR differences across the five primates. Our goal is to use these T2T great ape genomes to further our understanding of how TEs shape the recent dynamics of primate genomes and the mechanisms behind these genetic rearrangements.

ANNOTATING EUKARYOTIC GENOMES AT NCBI AND BEYOND

Vamsi K Kodali, Terence D Murphy, Francoise Thibaud-Nissen, NCBI Eukaryotic Genome Annotation Team

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

The NIH Comparative Genomics Resource (CGR) project led by NCBI aims to maximize the utility of the high-quality genome assemblies made possible by recent advances in genome sequencing technologies. Central to the CGR's organism-agnostic approach to facilitate and advance comparative genomics analyses of eukaryotes is the NCBI genomics toolkit, comprised of high-quality data and sophisticated analytical tools. One such tool is the Eukaryotic Genome Annotation Pipeline (EGAP): a modular, evidence-based, iterative pipeline to perform structural and functional annotation of eukaryotic genomes.

Over the past two decades, the RefSeq team at NCBI has developed and refined EGAP and successfully used it to annotate over 2000 genome assemblies across 1300 distinct species. As an evidence-based pipeline, EGAP relies on alignments of transcripts, transcriptomics data and proteins to the genome to construct accurate, full-length gene models. These structural annotations are further enhanced by multiple layers of functional information such as gene nomenclature, Gene Ontology terms and calculation of orthologs. EGAP annotations are seamlessly accessible through core NCBI resources such as Gene, Nucleotide, and Protein, and can be readily explored using powerful visualization tools like the Genome Data Viewer (GDV), Comparative Genome Viewer (CGV), and the newly released Multiple Comparative Genome Viewer (MCGV). To further enhance data accessibility, EGAP annotation products are available for bulk download from the browser, via the command-line interface, and through an API using NCBI Datasets.

Given that less than a fifth of the eukaryotic genomes submitted to GenBank are accompanied by gene model annotations, the wealth of genome data becomes significantly less accessible for meaningful biological analyses. Recognizing the need for high-quality annotation tools that researchers can deploy within their own compute infrastructure, we have developed EGAPx. This tool is a containerized Nextflow adaptation of the EGAP pipeline, designed for seamless execution in cloud environments or local high-performance computing clusters. EGAPx is currently available for download at https://github.com/ncbi/egapx and produces annotation in a format suitable for direct submission to GenBank. As we continue to actively incorporate new features and enhancements into EGAPx, we strongly encourage feedback from the scientific community to ensure it meets the evolving needs of genomic researchers.

This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

HAPLOTYPE PHASING AND COMPARATIVE GENOMICS OF ALGAE *NANNOCHLORIS DESICCATA, SCENEDESMUS OBLIQUUS, TETRASELMIS STRIATA*

Samuel I Koehler, Taehyung Kwon, Yuliya Kunde, Taraka Dale, Claire Sanders, Erik R Hanschen

Los Alamos National Laboratory, Bioscience division, Los Alamos, NM

Many challenges and subsequent technological innovations of the 21st century are based in biology, with modern agriculture requiring genomic approaches to yield solutions. Accurate and accessible nucleic acid data is the currency of genetics research, mandating continued work to assemble agronomic crop genomes. Biofuel research is exploring high-lipid algal species for fuel synthesis, with quality reference sequences driving crop improvements: increasing lipid content, resisting pathogen infection, and optimizing growth conditions. Currently, few algal genomes have been published compared to the known taxonomic diversity of algae. As the number of published algal sequences has increased, algal ploidy evolution is revealing itself to be more complicated that initially expected, with many species being diploid instead of haploid. Much ploidy ambiguity is caused by difficulties in rebuilding haplotype-resolved genomes. Diploid species are commonly assembled and reported as “collapsed”, with the assembly created as a haploid unintentionally, as a haploid “consensus” and variable loci represented by higher-confidence alleles, as hybridized chimeras of alleles, or with alternative alleles included as a fractured set of sequences. Collapsed haploid assemblies mis-represent many aspects of genome structure and composition that are crucial to facilitating agronomic improvement.

LANL Bioscience Division continues to generate many haplotype-phased diploid genomes for agronomically-promising algal species, reporting the haplotype-specific properties that are uncovered. Recent work phasing *Tetraselmis striata* identified hundreds of features and transcript motifs unique to a single haplotype. Subsequent haplotype-specific gene enrichment identified over-representation of membrane-localized transcription regulation and cell motility functions from gene ontology annotations. Similarly, phasing analyses for two strains of *Nannochloris desiccata*, UTEX 2437 and 2525, identified diploid genomes with similar assembly sizes but functional annotation differences of 89 and 1,788 haplotype-specific genes, respectively. The phased, telomere-complete diploid genome of the green alga *Scenedesmus obliquus* UTEX 3031 exhibits structural characteristics suggesting long-term existence as a recombinant diploid while simultaneously encoding a single chromosome exhibiting much more homozygosity than all others. The *S obliquus* assembly also contained haplotype-specific genes, with particular chromosomal regions being enriched for constituents of specific metabolic functions like chromatin and structural binding. Phylogenetic analyses exploring evolutionary patterns of diploidy in algae are ongoing, with preliminary average nucleotide identity analysis of sister chromosomes for 8 phased algal assemblies suggesting that homozygosity may correlate with genus.

INVESTIGATING AIS ASSOCIATED GWAS VARIANT DISRUPTIONS TO GENE TRANSCRIPTION

Justin Koesterich*[1,2,3], Darius Ramkhalawan*[4], Nadja Makki[4,5], Anat Kreimer[1,2,3]

[1]Rutgers The State University of New Jersey, Graduate Programs in Molecular Biosciences, Piscataway, NJ, [2]Rutgers The State University of New Jersey, Department of Biochemistry and Molecular Biology, Piscataway, NJ, [3]Rutgers The State University of New Jersey, Center for Advanced Biotechnology and Medicine, Piscataway, NJ, [4]University of Florida, College of Medicine, Department of Anatomy and Cell Biology, Gainesville, FL, [5]University of Florida, Genetics Institute, Gainesville, FL

* These authors contributed equally to this work.

Adolescent Idiopathic Scoliosis (AIS) is a severe disease that affects up to 3% of children around the world. AIS is diagnosed when a patient displays a greater than 10 degree spinal curvature from the coronal plane, often occurring during growth spurts in adolescence of otherwise healthy individuals. Individuals who develop AIS often require invasive surgeries to straighten the spine. Due to the severity of the symptoms, it is paramount that the genetic causes of AIS are well understood, so that the disease can be diagnosed early to develop potential drug therapies and administer preventative measures.

Recent Genome Wide Association Studies (GWAS) have found novel loci that have been associated with AIS. Our study takes 1,664 positions of noncoding variants identified in these GWAS and near AIS associated genes and analyzes their ability to disrupt transcriptional activity. Utilizing the Massively Parallel Reporter Assay (MPRA), we calculated the regulatory activity of candidate regulatory sequences harboring the reference and alternate alleles of these variants to identify which variants are significantly disrupting downstream gene transcription. We focused our MPRA analysis on chondrocytes, as this is one of the major cell types implicated in AIS pathogenesis.

Out of the variants tested in chondrocytes, five percent have a significantly different transcriptional activity compared to the wild type regulatory sequences. Additionally, these variants are located in regions of accessible DNA near H3K27 acetylated histone markers of active enhancers. Furthermore, we observe these variants disrupting transcription factor binding sites, suggesting a mechanism in which the variant leads to disrupted transcription.

Our findings highlight a subset of noncoding variants that are significantly disrupting transcriptional activity of AIS associated genes and provide better understanding of the mechanisms that are involved in AIS pathogenesis. Such variants are candidates for future experimental validation.

DESIGNING DNA WITH TUNABLE REGULATORY ACTIVITY USING DISCRETE DIFFUSION

Anirban Sarkar, Yijie Kang, Nirali Somia, Peter K Koo

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

Engineering regulatory DNA sequences with precise activity levels in specific cell types hold immense potential for medicine and biotechnology. However, the vast combinatorial space of possible sequences and the complex regulatory grammars governing gene regulation have proven challenging for existing approaches. Supervised deep learning models that score sequences proposed by local search algorithms ignore the global structure of functional sequence space. While diffusion-based generative models have shown promise in learning these distributions, their application to regulatory DNA has been limited. Evaluating the quality of generated sequences also remains challenging due to a lack of a unified framework that characterizes key properties of regulatory DNA. Here we introduce DNA Discrete Diffusion (D3), a generative AI framework based on score-entropy discrete diffusion for learning the data distribution and conditionally sampling regulatory sequences with targeted functional activity levels. We develop a comprehensive suite of evaluation metrics that assess the functional similarity, sequence similarity, and regulatory composition of generated sequences. By benchmarking on high-quality functional genomics datasets spanning human promoters and fly enhancers, we demonstrate that D3 outperforms existing diffusion-based models in capturing the diversity of cis-regulatory grammars and generating sequences that more accurately reflect the properties of genomic regulatory DNA. Furthermore, we show that D3-generated sequences can effectively augment supervised models and improve their predictive performance, even in data-limited scenarios. To further demonstrate D3's capabilities, we highlight how a multi-task D3 model can be utilized to design sequences with desired task-specific activity levels by imposing constraints during the generation process, such as small sequence lengths, which is an important objective when considering utility for targeted therapies. Moreover, we also highlight how D3's generative capabilities can be leveraged to interpret rules of cis-regulatory mechanisms it learns. Our results demonstrate D3 is a powerful generative AI approach for exploring and engineering regulatory DNA sequences, opening new avenues for functional genomics research and the development of precision genetic therapies.

POPULATION GENOMICS OF THE STONY CORAL *ACROPORA MILLEPORA* ACROSS THE GREAT BARRIER REEF

Arjun S Krishnan*[1], Carla Hoge*[2], Daria Bykova*[1], Ana Pinharanda[1], Zachary Fuller[3], Josephine Nielsen[4], Veronique Mocellin[4], Line Bay[4], Peter Andolfatto+[1], Molly Przeworski+[1]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]University of Chicago, Department of Human Genetics, Chicago, IL, [3]Bristol Myers Squibb, Cambridge, MA, [4]Australian Institute of Marine Science, Townsville, Queensland, Australia

* contributed equally
+ co-supervised this work

Coral reefs are home to more than 25% of global marine biodiversity. Under stress, corals lose their dinoflagellate symbionts, or "bleach", putting them at increased risk for starvation, disease and ultimately death. Rising ocean temperatures have led to increasingly frequent bleaching, triggering mass mortality events on the Great Barrier Reef (GBR) and other reefs worldwide. However, some reef colonies are less affected or recover faster, with differences seen both among and within species. To better understand variation in bleaching susceptibility, we focused on *Acropora millepora*, a stony coral found throughout the Indo-Pacific, in which inter-individual variation has been shown to be partly heritable. Samples of this species were collected from 1081 colonies in 40 reefs distributed across more than 1600 km of the GBR, and characterized for multiple phenotypes related to bleaching. We sequenced whole genomes from 125 samples to high coverage (>15X) and imputed genotypes for the remaining 952 low coverage (~2X) samples. We used these data to infer the population structure and demographic history of this species, as well as to identify the symbiont composition of each colony. We also identified genomic targets of natural selection, in which allele frequency differences are associated with local environmental conditions or maintained by balancing selection. These analyses provide an unprecedented characterization of demographic and selective pressures shaping genetic diversity in a coral species and its symbionts.

CHARACTERIZATION OF ARCHAIC ANCESTRY IN 63,000 JAPANESE INDIVIDUALS FROM THE TOHOKU MEDICAL MEGABANK

Mikel Lana Alberro[1], Stéphane Peyrégne[1], Shu Tadaka[2], Fuji Nagami[2,3], Makiko Taira[2,4,5], Kengo Kinoshita[3,6,7], Nobuo Fuse[2,3,5], Masayuki Yamamoto[2], Svante Pääbo[1,8], Janet Kelso*[1], Hugo Zeberg*[1,9]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, Germany, [2]Tohoku University, Tohoku Medical Megabank Organization, Sendai, Japan, [3]Tohoku University, The Advanced Research Center for Innovations in Next-Generation Medicine, Sendai, Japan, [4]Tohoku University, Tohoku University Hospital, Sendai, Japan, [5]Tohoku University, Tohoku University Graduate School of Medicine, Sendai, Japan, [6]Tohoku University, Tohoku University Graduate School of Information Sciences, Sendai, Japan, [7]Tohoku University, Institute of Development, Aging, and Cancer, Sendai, Japan, [8]Okinawa Institute of Science and Technology, Human Evolutionary Genomics Unit, Okinawa, Japan, [9]Karolinska Institutet, Department of Neuroscience, Stockholm, Sweden

*Authors contributed equally and share last authorship

Biobanks that provide genetic information and phenotypes for large cohorts have been used to explore the impact of introgressed Neandertal DNA on phenotypic diversity and disease susceptibility in present-day people. To date most studies have focused primarily on European populations who carry Neandertal introgressed DNA. We have characterised Neandertal and Denisovan introgression in 63,000 Japanese individuals from the Tohoku Medical Megabank Organization (ToMMo) biobank using a reference-free Hidden Markov Model (*hmmix*). We describe the set of Denisovan and Neandertal haplotypes in ToMMo and assess their relationship to the high-quality genomes of four Neandertals and two Denisovans. We confirm that Denisovan ancestry in Japan derives from two pulses of introgression from genetically distinct Denisovan populations. In agreement with previous studies, we identify extended genomic regions where Neandertal and Denisovan introgressed DNA is rare or absent. We also identify regions enriched for Neandertal and Denisovan DNA and describe high-frequency Neandertal and Denisovan haplotypes in the Japanese population that have significant replicable associations to human health outcomes in multiple East Asian biobanks.

INVESTIGATING NEGLECTED HUMAN MALARIA PARASITES FROM NATURAL INFECTIONS WITH SINGLE CELL AND LONG READ APPROACHES

Sunil Dogga[1], Seri Kitada[1], Jesse Rop[1], Antoine Dara[2], Abdoulaye Djimde[2], Mara Lawniczak[1]

[1]Wellcome Sanger Institute, Tree of Life, Hinxton, United Kingdom, [2]Malaria Research and Training Center, Faculty of Pharmacy, Bamako, Mali

Single-cell RNA sequencing applied to single-celled organisms provides a window into understanding individual organism variability that is not possible with bulk approaches. We have been using scRNAseq to study individual *Plasmodium* parasites from the blood of malaria-infected children in Mali. While *P. falciparum* accounts for the majority of deaths and cases, *P. ovale* and *P. malariae* also contribute significantly to the health burden. Despite recent studies indicating an increasing prevalence of these non-falciparum parasites, little is known about their genomic and transcriptomic diversity. To bridge this knowledge gap, using natural infections, we generated single-cell transcriptomic atlases and produced greatly improved reference genomes for both of these species. We used short and long-read scRNAseq to investigate the transcriptional, isoform, and genotypic diversity of these parasites within and between 20 young malaria-infected study participants from Faladie, Mali. Combining data from across participants for each species, we captured the entire intraerythrocytic cycle including the sexual and asexual stages allowing exploration of gene expression signatures along developmental trajectories. In order to improve reference genomes for both species, we also sequenced high molecular weight DNA directly extracted from parasites enriched from blood donations and generated chromosomal-level genome assemblies for *P. ovale wallikeri* and *P. malariae* using PacBio's Ultra-Low Input High-Fidelity long-read sequencing and Hi-C. Both species contain a high concentration of genes from the *Plasmodium* interspersed repeat (PIR) family at their subtelomeric regions. We find that this gene family is vastly expanded in *P. ovale*, with ~2000 PIRs, accounting for ~30% of its total genes, present almost exclusively at the subtelomeric ends. Consistent with the proposed role of PIRs in facilitating the recombination of heterologous chromosomes, we observed considerable recombination among the chromosomal ends of the sequenced specimens, likely contributing to the diversity of this surface gene family. The high-resolution transcriptomic cell atlases are made accessible via the Malaria Cell Atlas website at www.malariacellatlas.org.

UCSC GENOME BROWSER: HUBSPACE TRACK STORAGE

Christopher M Lee, Jairo N Gonzalez, Lou R Nassar, Jonathan Casper, Maximilian Haeussler

University of California Santa Cruz, Genomics Institute, Santa Cruz, CA

The UCSC Genome Browser has long been able to display custom annotations submitted by users, although for certain track types like bigBed, bigWig, and BAM, we can only display these types of data if they are hosted in web accessible locations. Even when a users' institution offers web-accessible storage, actually getting data there and then over to UCSC is a major pain point if you are not already a system administrator. Furthermore, most of our new display features and track configurations are only available for these web-accessible data formats. Here we present a user interface for direct upload and storage of these tracks, so users do not need to find web accessible hosting space anymore. Track data are stored directly on UCSC servers, providing a free, easy to use, and speed optimized solution over the traditional third party server process.

UCSC TRACK HUB BROWSER ADOPTION AND RECENT IMPROVEMENTS

UCSC Genome Browser Group, Maximilian Haeussler, Christopher Lee

UC Santa Cruz, Genomics Institute, Santa Cruz, CA

Most research groups at some point want to extend the annotations found in the main web-based genome browsers with custom data, add a custom assembly or want to make available their own annotation data through genome browsers for a manuscript. Setting up a custom genome browser or servers with special APIs that send data to existing browsers works, but is cumbersome and requires future updates and support. Uploading the data to a genome browser is an alternative but means that the data must be stored there and is locked to a particular genome browser.

To solve this problem, we developed UCSC Track Data Hub system in 2013. It allows hosting genome annotations as simple static data files located on any HTTP server, effectively creating a federated data ecosystem for genome assemblies and annotations. Today, the standard has been adopted by around 100 research groups: they created and submitted their track hubs to us and we allow to connect to these hubs via our "public track hubs" list at https://genome.ucsc.edu/cgi-bin/hgHubConnect. In addition, all other major genome browsers have implemented support for the standard, IGV, IGB, Ensembl, Jbrowse2 and to some extent NCBI's GDV, which means that track hubs made by research groups can be opened by almost any browser. And since UCSC's newer assemblies are released as assembly track hubs, users can open UCSC-hosted assemblies and annotations in other browsers. Around 4000 of these hubs are available at the time of writing and are already loaded by other genome browsers, such as IGV.

We will present an overview of the current track hub specification, how it has evolved over the last decade, a tool to simplify their creation by research groups and recent additions to the hub specification, mostly in the area of alignments.

MECHANISTICALLY INTERPRETABLE CNNs FOR DISENTANGLING GENOMIC INTERACTIONS

Marta S Lemanczyk[1,2], Chandana Rajesh[1], Peter K Koo[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Hasso-Plattner-Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany

Convolutional neural networks (CNNs) have emerged as powerful tools for analyzing genomic sequences, particularly in learning motifs in cis-regulatory elements (CRE) and capturing the complex interplay between regulatory components. However, traditional convolutional filters have difficulties disentangling interactions at different levels of biological organization. In DNA sequences, individual nucleotides display local dependencies within CRE motif positions and their flanking regions, while higher-order interactions among motifs play a critical role in gene regulation.

To address these challenges, we propose a mechanistically interpretable CNN architecture that explicitly models both intra-motif and inter-motif interactions. Our approach employs pairwise convolutional kernels designed to capture complete motif binding sites and their local dependencies. These pairwise kernels explicitly model the interactions between nucleotides, effectively disentangling local effects from broader motif-level interactions. Subsequent layers are structured to progressively integrate and model the higher-order interactions between motifs across the entire sequence. This design not only improves interpretability—by enabling direct visualization of the learned parameters as biologically meaningful features—but also overcomes limitations associated with post-hoc attribution methods, which often struggle to resolve spurious importance scores and complex interaction patterns.

We evaluate our architecture on transcription initiation, demonstrating its effectiveness in revealing both motif identities and their interaction partners. By providing a structured framework that mirrors the hierarchical organization of genomic regulation, our method paves the way for more biologically grounded representations in deep learning models.

DISCOVERY OF CELL TYPE-SPECIFIC REGULATORY NETWORKS USING SINGLE-CELL MULTI-OMIC ANALYSIS OF HUMAN HETEROGENOUS DIFFERENTIATING CULTURES (HDC)

Taibo Li[1], Kenneth Barr[2], Radhika Jangi[3], Katherine Rhodes[2], Josh Popp[1], Mingyuan Li[3], Hsing-Chiao Huang[2], Yoav Gilad[2], Alexis Battle[1]

[1]Johns Hopkins School of Medicine, Biomedical Engineering, Baltimore, MD, [2]The University of Chicago, Department of Medicine, Section of Genetic Medicine, Chicago, IL, [3]Johns Hopkins University, Department of Biology, Baltimore, MD

Recent advances in the generation and interpretation of atlas-scale multi-omic studies at the single cell resolution have provided important insights into regulatory networks in diverse tissues. However, the context-specificity of genetic regulation of transcription in early human development and how they contribute to risks of complex diseases remains incompletely understood. We previously developed a scalable and efficient protocol to generate human HDC from induced pluripotent stem cells (iPSCs) and demonstrated its ability to identify a multitude of dynamic cellular states. Here, we performed single-cell multi-omic sequencing on HDC from 17 Yoruba (YRI) individuals with paired whole-genome sequencing, and analyzed transcriptome and chromatin accessibility in 209,734 single cells. We applied MultiVI to jointly embed both modalities and mitigate sequencing batch effects. In total, we identified 29 cell clusters and developed an automated pipeline to combine unsupervised analysis by Celltypist and marker-gene based enrichment analysis to annotate each cell cluster. We identified cell types from all three germ layers, including intermediate cell types such as cardiac and hepatic progenitors, which were challenging to resolve with transcriptome data alone. We observed that intermediate cell states are characterized by distinct transcription factors (TFs) such as EN1 and GBX1 in neuroblasts, OTX1 in developing astrocytes, and TCF23 in cardiac progenitors. Using tensorqtl on pseudobulk-level ATAC expression profiles, we identified 116 unique caQTL SNPs which were significantly enriched in enhancer regions of multiple cell types in the Roadmap consortium, suggesting that caQTL in HDC capture relevant regulatory mechanisms in diverse tissue contexts. Lastly, to prioritize transcription factors with key regulatory functions in specific cellular contexts, we developed a new method, Single-Cell SPArse Modeling (SC-SPAM), which can efficiently decompose ATAC count matrix using expression level of TFs as dictionary with L0 sparsity constraints, thereby mapping TFs to their target peaks. This analysis enabled us to identify cell type-specific TFs not previously identified by ATAC-seq data alone, such as IRF9 in fibroblasts, and FOXO1 and ZIC5 in neural crest cells. We also identified TF-interacting caQTLs, where caQTL effects were modulated by TF expression, revealing interactions with known pioneer factors such as NFY and SALL2. This suggests that pioneer factors in addition to genetics play important modulatory functions on inter-individual variability of chromatin accessibility. Overall, our work presents a novel computational framework and demonstrates the potential for the human HDC system to dissect cell type-specific genetic regulatory networks across diverse contexts using single-cell multi-omic data.

FUNCTIONAL PLASTICITY AND TUNABILITY IN THE EVOLUTION OF DEVELOPMENTAL ENHANCERS

Tony Li[1], Jean-Benoît Lalanne[1,2], Emma A Kajiwara[1], Shruti Jain[1], Xiaoyi Li[1], Samuel G Regalado[1], Riza M Daza[1], Beth K Martin[1], Choli Lee[1], Jay A Shendure[1,3,4,5]

[1]University of Washington, Department of Genome Sciences, Seattle, WA, [2]Université de Montréal, Department of Biochemistry, Montréal, Canada, [3]Allen Institute, Seattle Hub for Synthetic Biology, Seattle, WA, [4]Howard Hughes Medical Institute, Seattle, WA, [5]University of Washington, Brotman Baty Institute for Precision Medicine, Seattle, WA

Cis-regulatory elements (CREs, commonly 'enhancers') are essential for orchestrating species-specific development, yet our understanding of their evolutionary trajectories and functional diversification remains limited. Here, as a case study on five high-activity mouse developmental enhancers whose activity is exquisitely specific to parietal endoderm, we comprehensively map mammalian regulatory evolution by functionally profiling enhancer orthologs from 480 extant & ancestrally reconstructed mammalian genomes using massively parallel reporter assays.

Leveraging this unprecedented dataset, we observe substantial differences in phylogenetic distribution of activity for these enhancers across extant and ancestral orthologs. Some elements, such as the *Gata4* enhancer, exhibit rodent-specific activity and others, like *Epas1* and *Lama1* elements, display high-activity across distal mammalian clades, with only loose correlation to sequence conservation.

To unravel the precise base-pairs underlying functional changes, we perform 'model-driven reconstitution', wherein we sequentially introduce mutations one-by-one, generating trajectories in sequence space connecting inactive ancestral sequences to their active descendants, profiling activity along the full path. Using deep learning models trained on ATAC-seq data, we reintroduce mutations by the predicted optimal order, enabling functional recovery of otherwise inactive enhancers. Through this approach, we elucidate the reactivation pathways of these elements in mouse endodermal cells, identifying key mutational requirements, e.g. the restoration of critical transcription factor binding sites necessary for activity. Further analysis of randomly-ordered reconstitution trajectories reveals significant epistatic interactions among reactivating mutations.

Finally, we apply this deep learning framework to iteratively introduce *de novo* mutations that optimize or disrupt activity within the existing mouse elements using a gradient-based greedy search. Using accessibility-predicting models, we observe a 5-fold median gain in activity compared to endogenous levels within 8 steps, suggesting chromatin accessibility can be used as a proxy for functional activity. Conversely, predicted-disrupting mutations typically completely abrogate activity below baseline within 3 steps. As we apply gradient ascent/descent up to 50 steps, we start observing divergence in measured function relative to the training objective after 20 steps, suggesting model-driven mutagenised sequences start to depart from the training support.

In summary, by densely sampling the CRE sequence landscape through both evolutionary data and deep learning, we systematically re-establish and optimize CRE activity through phylogenetic and synthetic approaches.

INFERENCE OF GENETIC ANCESTRY FROM CHALLENGING MOLECULAR DATA ACROSS MULTIPLE EXPERIMENTAL STRATEGIES

Xintong Li[1], Pascal Belleau[1,2], Astrid Deschênes[2], Laine Marrah[1], David A Tuveson[2], Alex Krasnitz[1,2]
[1]CSHL, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]CSHL, Cancer Center, Cold Spring Harbor, NY

Knowledge of the donor's genetic ancestry is a prerequisite to any study of human disease and condition phenotypes against the ancestral background. In order to overcome the lack of ancestral metadata for most molecular data in the public domain, we have developed a versatile, thoroughly validated set of tools termed RAIDS (Robust Ancestry Inference using Data Synthesis). RAIDS is currently able to handle sequence data originating from a broad variety of experimental strategies, ranging from the whole-genome to ChIP-derived sequences, and including cancer-derived and -altered sequences. From all these, RAIDS accurately infers both global ancestry and ancestral admixtures. RAIDS is publicly available on Bioconductor, Galaxy and Github. Following thorough validation, RAIDS has been used to elucidate ancestral effects on the phenotypes and somatic genotypes of pancreatic and colorectal cancers.

A HIGH-RESOLUTION PANGENOME STRUCTURAL VARIANT RESOURCE INCREASES SENSITIVITY FOR PATHOGENIC VARIANT DETECTION

Jiadong Lin[1], Jonas A Gustafson[2], Yang Sui[1], Danny E Miller[2,3], Evan E Eichler[1,4]

[1]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, [2]University of Washington, Department of Pediatrics, Seattle, WA, [3]University of Washington School of Medicine, Department of Laboratory Medicine and Pathology, Seattle, WA, [4]University of Washington, Howard Hughes Medical Institute, Seattle, WA

Long-read sequencing (LRS) and diploid genome assembly have enabled nearly complete structural variant (SV) discovery making it valuable for clinical research, such as characterization of unsolved cases after routine genetic testing. However, the lack of a high-resolution, openly accessible reference control set for identifying rare variants has complicated pathogenic variant interpretation, especially at complex and variable regions.
We generated and combined LRS data of 1650 diverse samples from the 1000 Genomes Project (1KG) produced in part from the Human Pangenome Reference (HPRC), Human Genomic Structural Variant (HGSVC), and 1KG sequencing consortiums. To harmonize SV discovery across different platforms and sequence coverages, we coupled state-of-the-art genome assembly methods with a machine-learning guided cross-datasets SV integration approach. Specifically, we showed our diploid assemblies built from high-quality ONT are comparable to telomere-to-telomere (T2T) assemblies in terms of genome completeness and SV detection performance. We developed BoostSV to assess SV quality based on a model trained from validated SV transmission data from T2T genomes constructed from a four-generation human pedigree. Using BoostSV, a confidence score was assigned to each SV making it possible to quantify and minimize the batch effects during SV integration. Finally, we applied a quality and sequence identity-based integration approach after all-by-all alignment to resolve complex tandem repeats to construct a nonredundant pangenome reference of ~850,000 SVs for both GRCh38 and T2T-CHM13 reference genomes.
Our first fully integrated callset constructed from 570 haplotypes contains 327,531 SVs on GRCh38, including 217,723 insertions and 109,807 deletions. Using the Adotto catalog, we found that 75,934 insertions were located at 7.6Mb coding gene tandem repeat regions, and 583 insertions intersected 85 pathogenic or phenotypic genes. We then applied this pangenome SV callset to LRS data generated for 192 samples from 52 unresolved autism families and aimed to increase sensitivity for rare pathogenic variant discovery. The pangenome SVs reduced the number of SVs by >98% from 24,500 SVs to <225 rare SVs per patient genome. Among this subset were new disease-causing (n=2) and likely pathogenic variants (n=8 or 5-15% yield) that were missed by short-read sequencing of the same genomes. Together, this pangenome SV callset and the methods we developed will enhance the discovery of pathogenic variants from LRS of patients and further our understanding of the missing heritability of human genetic disease.

INTEGRATION AND ANNOTATION OF SPATIAL MULTI-OMIC DATA WITH DIRAC HIGHLIGHTS SPATIAL ORGANIZATION OF LYMPHOID ORGANS

Chang Xu[1], Shibo Liu[1], Yang Heng[2], Yuan Cao[1], Junbin Gao[1], Dongmei Jia[2], Diyan Liang[2], Chen Yang[1], Yong Ma[2], Siok-Bian Ng[3], Ao Chen[2], Xun Xu[2], Sha Liao[2], Qinghua Jiang[4], Boxiang Liu[1,5]

[1]National University of Singapore, Department of Pharmacy and Pharmaceutical Sciences, Singapore, [2]BGI Research, BGI Research Institute, Shenzhen, China, [3]National University Hospital, Department of Pathology, Singapore, Singapore, [4]Harbin Institute of Technology, School of Life Science and Technology, Harbin, China, [5]National University of Singapore, Department of Biomedical Informatics, Singapore

Spatial multi-omics technologies enable simultaneous measurements of multiple omics modalities. Integration across spatial multi-omics modalities and cell-type annotation are two fundamental tasks for downstream analysis. We present Domain Invariant Representation through Adversarial Calibration (DIRAC), a geometric deep learning model that unifies both tasks by treating horizontal integration (different cells/spots, same omics modality) and vertical integration (same cells/spots, different omics modalities) under a generalized domain adaptation framework. DIRAC uses an adversarial domain discriminator to integrate multiple spatial omics modalities into a unified domain-invariant embedding space and to automate cell-type annotation by transferring labels from reference multi-omic data.

Using DBiT-seq data on E10 mouse embryo, we compared DIRAC against SOTA vertical integration methods, SpatialGLUE, MaxFuse, and MultiVI. DIRAC significantly outperformed the runner-up in terms of batch correction (two-sided t-test $P < 1.31e-4$) and preservation of biological signals (two-sided t-test $P < 4.24e-5$). Further, DIRAC outperformed the runner-up (two-sided t-test $P < 8.67e-5$) when compared against nine SOTA single-cell multi-omic integration methods, including Seurat, GLUE, LIGAR, iNMF, bindSC, MMD-MA, UnionCom, and Pamona. Using 10x Visium data on human dorsolateral prefrontal cortex, we assessed DIRAC's horizontal label transfer capability against three SOTA methods, STELLAR, scmap, and scANVI, and three strong baselines, random forest, AdaBoost, and SVM. DIRAC performed 11% better than the runner-up (two-sided t-test $P < 2.33e-8$). DIRAC demonstrated excellent generalizability when tested across multiple omics modalities (histone marks, chromatin accessibility, RNA, and protein) and technology platforms (sequencing and imaging-based). Lastly, DIRAC is substantially faster than existing methods and scales to millions of cells. On vertical integration tasks, DIRAC finished processing spatial data of 2M spots/cells within 1 hour, whereas the runner-up failed to process spatial data past 63k spots/cells.

We used DIRAC to build cellular-resolution spatial multi-omics atlases of mouse spleen and thymus. We measured the whole transcriptome and 128 proteins for an estimated 173,649 single-cell at 500-nm resolution with Stereo-CITE-seq to reveal the spatial developmental trajectory of T cells within the mouse thymus. Using the same spatial assay, we measured an estimated 120,117 single cells of the mouse spleen, revealing clear differences in T cell expression patterns between the thymus and the spleen, as well as T cell communication differences between the two key lymphoid organs.

MULTI-LINEAGE TRANSCRIPTIONAL AND CELL COMMUNICATION SIGNATURES DEFINE PATHWAYS IN INDIVIDUALS AT-RISK FOR DEVELOPING RHEUMATOID ARTHRITIS THAT INITIATE AND PERPETUATE DISEASE

Cong Liu[1], Wei Wang[1], Gary Firestein[2], Peter Skene[3], Kevin Deane[4], Jane Buckner[5]

[1]Department of Chemistry and Biochemistry, University of California San Diego, San Diego, CA, [2]Division of Rheumatology, Autoimmunity, and Inflammation, University of California San Diego, San Diego, CA, [3]Allen Institute for Immunology, Seattle, WA, [4]Division of Rheumatology, University of Colorado Anschutz Medical Campus, Aurora, CO, [5]Center for Translational Research, Benaroya Research Institute, Seattle, WA

Elevated anti-citrullinated protein antibodies (ACPA) levels in the peripheral blood are associated with an increased risk for developing rheumatoid arthritis (RA). Currently, no treatments are available that prevent progression to RA in these at-risk individuals. In addition, diverse pathogenic mechanisms underlying a common clinical phenotype in RA complicate therapy as no single agent is universally effective. We propose that a unifying set of transcription factor and their downstream pathways regulate a pro-inflammatory cell communication network, and that this network allows multiple cell types to serve as pathogenic drivers in at-risk individuals and in early RA. To test this hypothesis, we identified ACPA-positive at-risk individuals, patients with early ACPA-positive RA and matched controls. We measured single cell chromatin accessibility and transcriptomic profiles from their peripheral blood mononuclear cells. The datasets were then integrated to define key TF, as well as TF-regulated targets and pathways. A distinctive TF signature was enriched in early RA and at-risk individuals that involved key pathogenic mechanisms in RA, including SUMOylation, RUNX2, YAP1, NOTCH3, and β-Catenin Pathways. Interestingly, this signature was identified in multiple cell types, including T cells, B cells, and monocytes, and the pattern of cell type involvement varied among the at-risk and early RA participants, supporting our hypothesis. Similar patterns of individualized gene expression patterns and cell types were confirmed in single cell studies of RA synovium. Cell communication analysis provided biological validation that diverse lineages can deliver the same core set of pro-inflammatory mediators to receiver cells that subsequently orchestrate rheumatoid inflammation. These cell-type-specific signature pathways could explain the personalized pathogenesis of RA and contribute to the diversity of clinical responses to targeted therapies. Furthermore, these data could provide opportunities for stratifying individuals at-risk for RA, and selecting therapies tailored for prevention or treatment of RA. Overall, this study supports a new paradigm to understand how a common clinical phenotype could arise from diverse pathogenic mechanisms and demonstrates the relevance of peripheral blood cells to synovial disease.

COMPUTATIONAL FRAMEWORK FOR PREDICTING THE EFFECT OF NON-CODING VARIATION

Jiayi Liu[1,2,3], Justin Koesterich[1,2,3], Anat Kreimer[2,3]

[1]Rutgers University, Graduate Programs in Molecular Biosciences, Piscataway, NJ, [2]Rutgers University, Department of Biochemistry and Molecular Biology, Piscataway, NJ, [3]Rutgers University, Center for Advanced Biotechnology and Medicine, Piscataway, NJ

It is well established that disease-associated variants in non-coding regions of the genome play a critical role in complex human disorders. This study focuses on predicting the regulatory effects of autism-associated non-coding variants using machine learning (ML) applied to two independent datasets, both experimentally tested with Massively Parallel Reporter Assays (MPRA) and annotated with extensive sequence-specific features.

The primary dataset included 3,090 non-coding variants from autism patients and healthy control siblings, tested in neural progenitor cells. It comprised 165 high-confidence variants (HCVs) and 2,925 non-HCVs annotated with 30,770 features, such as transcription factor (TF) motif occurrences, epigenetic marks, DNA shape features, polyA/T lengths, and k-mer frequencies. The secondary dataset consisted of 84 expression quantitative trait loci (eQTLs) in open chromatin regions linked to autism-associated variants, tested in mid-gestation cortex tissue. These variants were categorized as functional (n=35) or non-functional (n=49) and annotated with 31,157 features each.

To address high-dimensional feature space and overfitting risks, we applied principal component analysis (PCA) to delta values between reference and alternative alleles, selecting the top 1,500 principal components (PCs) for downstream modeling. Feature importance was determined by the highest loading scores across selected PCs, highlighting key contributors to variant functionality.

Six ML classification models were trained, including linear SVM with stochastic gradient descent (SGD), C-Support Vector Classifiers (SVC), k-nearest neighbors (KNN), ExtraTrees (ET), histogram-based gradient boosting (HGB), and multilayer perceptron (MLP). Performance was evaluated via area under the receiver operating characteristic curve (AUROC) to ensure consistent comparisons.

For intra-dataset predictions, ET and SVC achieved AUROC scores exceeding 0.75 on the primary dataset, identifying deltas of motif count, regulatory scores and k-mer frequency as top predictors. Similarly, robust performance was observed for the secondary dataset.

In inter-dataset predictions, the ET model achieved an AUROC of 0.79 when predicting the primary dataset using the secondary dataset and 0.64 in reverse, demonstrating the robustness of our approach.

Our novel ML framework effectively prioritizes functional non-coding variants and elucidates their regulatory mechanisms, demonstrating generalizability across common and rare variants. By introducing a novel feature importance scoring method, we highlight key genomic features driving variant functionality. Future work aims to integrate multi-omics datasets and refine regression models for improved predictions of quantitative variant impacts.

THERMODYNAMIC SURROGATE MODELS FOR INTERPRETING GENOMIC DEEP NEURAL NETWORKS

Kaiser Loell, Zhihan "Leo" Liu, Evan Seitz, David McCandlish, Peter Koo, Justin Kinney

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

Deep neural networks (DNNs) trained on genome-scale datasets have attained high levels of performance at predicting a wide range of functional readouts from sequence. They are, however, generally considered "black boxes" from which it is difficult to extract mechanistic insight. The most widely used methods for interpreting genomic DNNs generate attribution maps, which reveal the contributions of individual bases to DNN predictions but are noisy and do not provide insight into the interactions between binding motifs. In prior work, we have developed methods to perform in-silico mutagenesis and either fit generalized linear models as DNN surrogates or classify the mutant sequences based on their attribution maps. We demonstrate here that combining these techniques allows genomic DNN predictions to be interpreted in terms of explicit molecular mechanisms. We have created an automated pipeline to segment loci into transcription factor binding sites based on the patterns of nucleotide substitutions that define clusters of attribution maps. Using these segmentations, we can build thermodynamic state models and fit them to DNN predictions. We demonstrate that these thermodynamic surrogate models can identify the transcription factors and the interactions between them that underlie the regulatory activity predicted by DNNs. By further automating and optimizing this process, we hope to create genome-wide maps of functional transcription factor binding sites and their interactions.

TISSUE-DEPENDENCY OF meQTLs IN A FREE-RANGE POPULATION OF RHESUS MACAQUES

Amy Longtin[1], Rachel M Petersen[1], Baptiste Sadoughi[2,3], Christina E Costa[4], Josue E Negron-Del Valle[2,3], Daniel Phillips[2,3], Cayo Biobank Research Unit[5], Michael L Platt[5], Michael J Montague[5], Lauren J Brent[6], James P Higham[4], Noah Snyder-Mackler[2,3,7], Amanda J Lea[1]

[1]Vanderbilt University, Biological Sciences, Nashville, TN, [2]Arizona State University, School of Life Sciences, Tempe, AZ, [3]Arizona State University, Center for Evolution and Medicine, Tempe, AZ, [4]New York University, Anthropology, New York, NY, [5]University of Pennsylvania, Psychology, Philadelphia, PA, [6]University of Exeter, Psychology, Exeter, United Kingdom, [7]Arizona State University, School of Human Evolution and Social Change, Tempe, AZ

DNA methylation (DNAm) is an epigenetic gene regulatory mechanism that plays a crucial role in biological processes such as development, aging, and tissue differentiation. However, because genetic variation can also impact DNAm, a major area of interest has been understanding how genotype interacts with these processes to shape interindividual variation–particularly in ways that contribute to complex trait differences. Previous work mapping genetic effects on DNAm levels–in the form of cis methylation quantitative trait loci (or meQTL)–has focused on peripheral tissues, primarily blood. As a result, the extent to which meQTLs are tissue-dependent remains severely understudied, especially in natural non-human populations that can provide insight into the evolution of gene regulatory variation and whether patterns observed in humans are unique. Here, we leveraged a genotype and DNAm dataset spanning 14 tissues collected from 209 free-ranging rhesus macaques from the island of Cayo Santiago (n=2,640 total samples). We tested for genetic effects on DNAm at 6.65 million single nucleotide variants and an average of 1.12 million CpG sites per tissue. Then we used an empirical Bayes approach (mashR) to compare effect sizes across tissues. We identified both tissue-dependent and shared meQTLs, compared their patterns to those previously reported in humans, and investigated mechanisms of tissue-dependency (e.g., transcription factor binding). Together, this unique dataset provides a novel, multi-tissue map of genetic effects on DNAm, and more generally expands our understanding of the genetic architecture of DNAm across primate species.

ANCIENT CENTROMERE SPANNING HAPLOTYPES PROVIDE INSIGHT INTO HUMAN CENTROMERIC SATELLITE EVOLUTION IN TELOMERE-TO-TELOMERE (T2T) GENOMES

Hailey Loucks*[1], Sasha Langley*[2,3], Fedor Ryabov[1], Julian K Lucas[1], Julian Menendez[1], Viviane Slon[4], Gary Karpen [3], Ivan A Alexandrov[4], Charles Langely[2], Karen H Miga[1]

[1]University of California, UC Santa Cruz, Biomolecular Engineering, Santa Cruz, CA, [2]University of California, Davis, Evolution and Ecology, Davis, CA, [3]University of California, Berkeley, Molecular and Cell Biology, Berkeley, CA, [4]Tel Aviv University, Department of Anatomy and Anthropology & Department of Human Molecular Genetics and Biochemistry, Tel Aviv, Israel

Centromeric satellite arrays play an important role in nuclear organization, kinetochore recruitment, and division fidelity. Genetic and epigenetic variation within centromeric regions can lead to chromosome instability, which contributes to aneuploidies and cancers, defects in early development, and decreased fertility. Despite their importance, we are only beginning to understand the architecture of these regions and the patterns of evolution across human populations. Previous efforts to study linkage blocks spanning these regions identified centromere-spanning haplotypes, or 'cenhaps' which can be traced as centromeric population lineages. Meiotic recombination is rare in these regions allowing for intact transmission of Mb-sized centromere regions. Dating based on datasets from the 1000 Genome Consortium (1000G) provided evidence for ancient cenhaps within modern humans. Some of these sequences can be traced back to archaic hominins, and appear to be the result of Neanderthal and Denisovan introgression.

Complete telomere-to-telomere (T2T) reference genomes from humans and non-human primate assemblies offer an unprecedented look into haplotypes and variation within and between cenhaps. Here we present an initial analysis of satellite evolution in the context of cenhap structure using T2T assemblies generated by the Human Pangenome Reference Consortium. In doing so, we have established a panel of homolog-phased and assembled centromeric regions representing archaic cenhaps on chromosomes 10 and 12. Our initial study revealed satellite variants of higher order repeats (HOR) that differ in structure and organization from modern cenhaps. The dynamics of repeat expansion, degradation, and mutation are not well understood in these complex regions, and this study provides a new system for studying repeat evolution in the context of shared haplogroup structures. Additionally, this work aims to provide millions of bases of archaic-derived genomic sequence information to advance our understanding of structural variation in humans.

ASSESSING THE VALUE OF THE HUMAN PANGENOME REFERENCE FOR TRAIT ASSOCIATION ANALYSES

Shuangjia Lu[1], Wen-Wei Liao[1], Marianne D Gorter[2], Page C Goddard[2], Stephen B Montgomery[2,3], Ira M Hall[1]

[1]Yale University School of Medicine, Department of Genetics, New Haven, CT, [2]Stanford University School of Medicine, Department of Pathology, Stanford, CA, [3]Stanford University School of Medicine, Department of Genetics, Stanford, CA

The human pangenome reference consortium (HPRC) is generating hundreds of high quality reference genomes from diverse populations, and describing the variation within them in pangenome graphs that have the potential to improve read alignment and variant calling. Here, we examine the extent to which using the human pangenome reference may improve the power of human trait association studies, using expression quantitative trait locus (eQTL) mapping as a proxy for what we might expect from future pangenome-based GWAS.

We conducted eQTL analysis using RNA-seq data of 430 individuals from 5 African subpopulations generated by the African Functional Genomics Resource, leveraging deep short-read whole genome sequencing (WGS) data from the 1000 Genomes Project (1KG). We developed a novel method to measure variant genotype information based on realignment of WGS reads to pangenome graphs, followed by cross-sample normalization of read-depth at pangenome graph features including edges, nodes, and combined sets of adjacent nodes. We compared our methods to three published variant genotyping pipelines: (1) a standard small variant callset generated using GATK and the GRCh38 reference genome; (2) a structural variant (SV) callset generated using multiple traditional SV detection methods and GRCh38; (3) a combined SNP/indel/SV callset generated using Pangenie, a graph-based genotyping method utilized in the HPRC Freeze 1 publication.

Using the graph-based reference generated by HPRC as part of Freeze 1 (including 94 haplotypes), our edge-based method ("EDGE") discovered 26.5 million variants, among which 1.2M SNPs (8.5%) and 4.8M indels (42.0%) were not identified by GATK. We then performed a competitive joint eQTL analysis including all variants and methods included in this study and identified 10,263 eQTLs. At 47.1% of these eQTLs, the lead marker was derived from our pangenome graph-based methods, and 7.3% of these were not found at genome-wide significance by any of the other three methods, meaning the association would otherwise have been missed. In contrast, 35.3% of lead markers were from GATK, 0.2% from the 1KG SV callset, and 17.4% from Pangenie. We further investigated the 4,558 eQTLs where the lead marker was identified by our EDGE method: in 3,724 cases EDGE outperformed GATK by improving genotyping quality; in 792 cases EDGE discovered novel trait-associated variants missing by GATK, whereas GATK found more weakly associated linked variants; and in 38 cases EDGE discovered novel trait associated variants that were not found by any GATK variants. These results indicate that implementation of sensitive pangenome graph-based variant analysis methods will increase the power of molecular trait association studies and, by extension, complex trait genetics projects as well.

THE LANDSCAPE OF GERMLINE AND SOMATIC CANCER VARIANTS IN TUMOR SUPPRESSOR GENES.

Suhasini D Lulla[1], Deborah I Ritter[1], Chimene Kesserwan[2], Sharon E Plon[1]

[1]Baylor College of Medicine, Pediatrics, Houston, TX, [2]NYU Langone Health, Pathology, New York, NY

Background: Though germline and somatic cancer variants in TSGs share LoF mechanisms, genes like *DICER1* and *CEBPA* show different profiles by variant type and location. Hence, we systematically analyzed germline and somatic cancer variant data across TSGs to understand the mutational mechanisms leading to cancer.
Methods: We selected 41 TSGs with established germline cancer predisposition and sufficient data. An analysis pipeline was built to extract high quality pathogenic/likely pathogenic (P/LP) germline variants from ClinVar and oncogenic/likely oncogenic (O/LO) somatic variants in tumors from cBioPortal, excluding tumors with high mutation burden. Variants were harmonized on the MANE Select transcript and annotated using Variant Effect Predictor v112. We calculated total occurrence of each variant and compared germline and somatic distributions by variant type using chi-squared tests (FDR 0.05) allowing for multiple testing and estimating residuals. Location patterns of variants were assessed by a moving difference in fraction of germline and somatic events along the coding sequence, capturing preferential clustering of germline or somatic events (>+/-2.5 SD).
Results: From 41 TSGs, we obtained 33,023 P/LP germline and 12,986 O/LO somatic variants. However, only 3,871 (9.2%) unique variants were shared. In total, 19 TSGs have significantly different variant type distributions replicated using non-overlapping data in COSMIC. Only *DICER1*, *TP53*, and *SMAD4* have excess somatic missense events, while 17 TSGs (including *RB1*, *APC*) have excess somatic stop-gain events seen in multiple cancer tissue types. Comparison by location revealed 106 regions in 40 TSGs with clustering of events (79 somatic, 27 germline). Strikingly, 23 somatic clusters comprise recurring frameshift variants in homopolymer runs in bowel and uterine cancers.
Conclusion: Despite the shared LoF mechanism in TSGs, a minority of variants overlap between germline and somatic datasets for 41 TSGs. Additionally, the distribution patterns by variant type and location differ substantially. Many TSGs have excess somatic stop-gain events that rarely cluster, suggesting distinct germline and somatic mutational mechanisms, independent of tissue type. Unlike missense hotspots in oncogenes, somatic missense clusters in TSGs are few, but we identified clusters of frameshift variants in homopolymers, in tumors with microsatellite instability. Altogether, the mutational mechanisms underlying germline and somatic cancer variants in TSGs are fundamentally different, adding an unexpected layer of complexity to the two-hit hypothesis. Supported by 2U24HG009649.

EXPLORING HYBRIDIZATION PERSISTENCE: GENE REGULATORY DYNAMICS AND SEX-SPECIFIC RECOMBINATION LANDSCAPES IN LEPIDOPTERA

Ava Mackay-Smith[1,2], Gregory A Wray[1]

[1]Duke University, Department of Biology, Durham, NC, [2]Duke University Medical Center, University Program in Genetics and Genomics, Durham, NC

Lepidopteran insects, such as butterflies and moths, are frequently in the scientific limelight due to their charismatic color patterns. These patterns are not only frequently aesthetically pleasing to us, but also have ecological relevance: in many cases, color patterning plays an important role in camouflage or mimicry. However, despite the evolutionary and ecological relevance of having the right coloration pattern, it has been documented for centuries that many lepidopteran groups stably hybridize with closely related species. If an ecologically relevant phenotype is evolutionarily important to maintain, why then are such stable patterns of hybridization possible? We propose a model for understanding persistent hybridization in Lepidoptera that incorporates knowledge from gene regulatory models, ancestry reconstruction, and reproductive genetics to explain why these patterns may be more frequent than expected under standard evolutionary assumptions.

The bulk of the proposed concept rests on the consequences of sex-specific achiasmy during recombination in many female Lepidoptera. The ramifications of this interesting sexually dimorphic recombination have not been explored through the perspective of its impacts on gene regulation and selection in hybrid lineages. First, species that regulate their genes predominantly *in cis* may be able to support hybridization longer than otherwise identical species with predominant regulation *in trans*, because *cis*-regulatory structures for a given chromosomal copy are retained in linkage in hybrid females in the first (and possibly subsequent) hybrid generations. Alternatively, under other conditions, this non-recombination could provide grounds for mutational variation to occur and subsequently be rapidly purged or selected for, to either accelerate or dampen sequence divergence. Second, genetic tracing of hybridization via approaches like trio binning in Lepidoptera are sensitive to, and may incorrectly measure, hybridization ancestry because female non-recombination phenomenon in later generations can appear to ‘replicate’ earlier hybrid genotypes in a subset of cases. This may result in ambiguity or mis-assignment of hybrid event timing from genomic information. The proposed framework opens several interesting doors for improved ancestry inference in cases of admixture and recombination modeling, and an interesting natural case study for genetic linkage during hybridization.

TAD-INDEPENDENT CHANGES IN CHROMOSOME-SPECIFIC SPATIAL AND GEOMETRIC CHARACTERISTICS OCCUR DURING MYOGENIC DIFFERENTIATION

Andrew Skol*[1], Lucas M Carter*[2,3], Tyler Hershenhouse[4], Joe Ibarra[4], Luay Almassalha[2,5], Kyle L MacQuarrie[4,6]

[1]Stanley Manne Children's Research Institute, Pathology, Chicago, IL, [2]Northwestern University, Center for Physical Genomics and Engineering, Evanston, IL, [3]Northwestern University, Interdisciplinary Biological Sciences Graduate Program, Evanston, IL, [4]Stanley Manne Children's Research Institute, Pediatrics, Chicago, IL, [5]Northwestern Memorial Hospital, Gastroenterology and Hepatology, Chicago, IL, [6]Northwestern University, Feinberg School of Medicine, Chicago, IL

*: equal co-authors

Myogenesis, the process by which proliferative myoblasts terminally differentiate into myotubes, is characterized by widespread gene expression changes and changes in genome organization. We have recently demonstrated myogenic chromosome-specific differences in chromosomal positioning relative to the nuclear axis that are absent in rhabdomyosarcoma (RMS) cells, a pediatric tumor of skeletal muscle, that are partially rescued when RMS cells undergo induced myogenesis. Specifically, chromosome 2 exhibits differentiation-dependent positioning along the major nuclear axis, while chromosome 18 exhibits no such preference. We hypothesized that this geometric spatial difference would result in differences in other chromosomal characteristics, such as accessibility and DNA contacts, that would correlate to the restricted geometric positioning in space. Analyses including ATAC-Seq, Hi-C, and partial wave spectroscopic (PWS) imaging performed on myogenic cells in vitro demonstrate multiple changes in chromosomal characteristics independent of changes in TADs (topologically associating domains). Restricted geometric positioning of chromosome 2 correlates with changes in cis-chromosomal contact scaling, sequence-dependent changes in trans-chromosomal contact frequency, and differences in DNA accessibility when compared to chromosome 18, while there is no significant difference in the pattern of changes of TADs. Coupling single-cell PWS imaging to chromosome painting, we find no evidence of chromosome-specific changes in chromatin packing domains, despite there being changes in chromatin packing in the nucleus as a whole. Taken together, our data identify 1) the presence of chromosome-specific preferential geometric positioning during myogenesis, 2) changes in DNA contact scaling, accessibility, and trans-chromosomal contacts and 3) no evidence for chromosome-specific changes in chromatin packing domains or TADs. This suggests both that these geometric-associated changes have functional consequences on chromosome function, and the potential to use next-generation sequencing data to infer geometric characteristics for other chromosomes and in other cell types.

DEEP LEARNING PREDICTS CIS-REGULATORY TURNOVER IN HUMAN EVOLUTION

Riley J Mangan[1,2,3], Nikitha Thoduguli[1], Jayashabari Shankar[1], Yuru Lin[1], Zunpeng Liu[1,2], Manolis Kellis[1,2]

[1]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Lab, Cambridge, MA, [2]The Broad Institute of MIT and Harvard, Cambridge, MA, [3]Harvard Medical School, Genetics Training Program, Boston, MA

Identifying recent gene regulatory innovations in the human lineage is challenging, as only a small subset of the millions of derived alleles are expected to contribute to human-specific phenotypic differences. Previous approaches have prioritized highly-divergent regions, following the idea that rapid molecular evolution aligns with functional adaptation. However, individual regulatory mutations that influence gene expression, traits, and disease risk may evade detection in divergence-based scans. Meanwhile, comprehensive multi-species epigenomic comparisons across millions of derived variants in diverse cell and tissue contexts are infeasible, particularly for extinct hominins like Neanderthals and Denisovans, where direct experimental profiling is inaccessible due to the absence of preserved tissues. To address this, we leverage advances in deep learning to infer chromatin accessibility directly from genomic sequence, reconstructing genome-wide regulatory evolution across modern humans, archaic hominins, and great apes. Because sequence-to-function models require long input sequences, we generated personalized reference-guided genome assemblies from short read sequencing, integrating called variants into reference synteny across 10 modern human, 8 archaic hominin, and 16 great ape haplotypes. We predicted 37,548 differentially accessibility (DA) events across 10,918 distinct regions between hominins and great apes, and 695 DA events across 231 regions between modern humans and archaic hominins. To evaluate the validity of our approach, we confirmed that regions with predicted regulatory divergence aligned with interspecies differences measured experimentally with ChIP-seq and massively parallel reporter assays. Remarkably, we find that differentially accessible elements are enriched for neurodevelopmental, locomotor, and developmental pathways, consistent with major phenotypic transitions in human evolution. Lastly, we used in silico saturation mutagenesis to predict how individual nucleotide changes impact chromatin accessibility, identifying key bases essential for regulatory function. We then integrated transcription factor binding site analysis, weighting motifs by their predicted regulatory importance, to prioritize human-derived mutations with candidate functional consequences. Overall, our comprehensive deep learning-based framework for predicting cis-regulatory evolution opens new avenues for identifying functional divergence shaping human-specific traits and disease risk.

IDENTITY-BY-DESCENT CAPTURES SHARED ENVIRONMENTAL FACTORS AT BIOBANK SCALE

Franco Marsico[1], Silvia Buonaiuto[1], Ernestine Amos-Abanyie[1], Lokesh Chinthala[2], Akram Mohammed[2], Terri Finkel[3], Robert Davis[2], Chester Brown[1,3], Robert Williams[1], Pjort Prins[1], Vincenza Colonna[1,3]

[1]UTHSC, Dept of Genetics, Genomics and Informatics, Memhpis, TN, [2]UTHSC, Center for Biomedical Informatics, Memhpis, TN, [3]UTHSC, Dept of Pediatrics, Memhpis, TN

"The apple doesn't fall far from the tree" is an old idiom that encapsulates a key concept: being related extends beyond only sharing genetic material. It often implies sharing the same environment, like culture, language, dietary habits, and geographical location. Environment, much like DNA, is a crucial determinant of health and plays a significant role in evolutionary processes. Genetic relatedness can also be used to cluster individuals when discrete groups are required.

In this study, we hypothesize that beyond direct study of genetic mechanisms, analyzing distant DNA-based relationships can inform on health conditions and evolution, as indicators of a shared environment. We also investigate to what extent clustering by genetic relatedness overlaps with ancestry-based groups.

To test our hypothesis, we analyzed genomic data from 13k individuals from the Biorepository of Integrative Genomics, which enrolled over 35k participants with electronic health records, together with individuals from 1000 Genomes and HGDP. We used an IBD–based hierarchical community detection algorithm to characterize population structure. Communities were mapped to neighborhood-level geographical data, census-based metrics, and biogeographical ancestry inference.

At the broadest scale, we detected four major communities that align with continental ancestries: predominantly European (C1), African (C2), American (C3), and East Asian (C4). Within communities, we detected sub-communities. For example, within C3, one sub-community (C3_2) is strongly associated with the Puerto Rican population (1000 Genomes), whereas another (C3_1) shows closer genetic ties to Peruvian and Mexican (HGDP) groups.

We observe an inverse relationship between IBD sharing and geographic distance. This is particularly pronounced in the C2 community, while sub-communities within C1 are more widely dispersed. Furthermore, we observed significant differences in environmental exposures across sub-communities. Variations in air pollution, smoking prevalence, urbanization, and socioeconomic vulnerability were notable. Particularly, C2 sub-communities were overrepresented in regions with elevated environmental stressors, potentially contributing to health risk.

These findings suggest that genetic relatedness, even at distant levels, may serve as an indicator of shared environmental and social factors affecting health while inherently capturing ancestral population structure without using potentially biased reference populations.

EXPLAINING THE MECHANISTIC FRAMEWORK OF SPLICEAI USING PHANTOMFOREST

Cristina Martin Linares, Jonathan Ling

Johns Hopkins School of Medicine, Pathology, Baltimore, MD

Deep neural networks have revolutionized biological modeling, yet they often function as opaque black box models. Post hoc methods for interpreting black boxes excel at local explanations but often struggle to extend these insights to the model's global behavior, whereas mechanistic interpretability methods such as sparse autoencoders fail to capture the hierarchical structure of learned features. Here we introduce PhantomForest, a global interpretability framework that builds deep decision trees by recursively generating synthetic data to identify important local features at each node. As a proof-of-concept, we applied PhantomForest to SpliceAI, a convolutional neural network for mRNA splicing prediction. We found that SpliceAI's underlying structure is simpler than expected, relying primarily on core consensus splice site motifs and a cooperativity model in which neighboring splice sites reinforce one another. To further investigate SpliceAI's internal representations, we trained a sparse autoencoder on layers of the model and found that the features driving each neuron's activation corresponded to a mix of consensus motifs. In certain cases, neurons exhibited sensitivity not only to the evaluated splice site but also to neighboring splice sites, mirroring the cooperativity patterns found with PhantomForest. Taken together, these findings demonstrate how blending explainable AI with mechanistic interpretability can yield a deep understanding of black box models used for biology.

HAPLOTYPE-BASED FINE-MAPPING OF VARIANT ASSOCIATIONS FOR COMPLEX TRAITS

Arya R Massarat[1], Utkarsh Jain[2], Jonathan Margoliash[2], Michael Lamkin[2], Yang Li[3], Melissa Gymrek[2,3]

[1]Bioinformatics and Systems Biology Graduate Program, Bioengineering, San Diego, CA, [2]CSE, Department of Computer Science and Engineering, San Diego, CA, [3]Medicine, Department of Medicine, San Diego, CA

Genome-wide association studies (GWAS) have identified tens of thousands of associations between genomic regions and complex traits in humans, but identifying the causal variants underlying these associations is challenging. Statistical fine-mapping techniques attempt to quantify the probability of causality of individual variants while accounting for the correlation, or linkage disequilibrium (LD), between nearby variants. Existing fine-mapping solutions are limited to considering additive and independent contributions of individual variants, and also assume the causal variant is included in the analysis. Here, we introduce a novel fine-mapping method, happler, which models the effects of variant *haplotypes* (sets of alleles of one or more variants inherited on a single chromosome) in addition to individual variants. By identifying causal haplotypes, happler can identify additional effects including those driven by combinations of variants in cis and causal variants, such as complex structural variants, not included in the genotyped variants but that are well tagged by one or more haplotypes. Simulation experiments show that whereas happler can correctly identify causal haplotype effects, traditional methods based on individual variants can produce inconclusive results in such cases. Finally, we apply happler to fine-map expression quantitative trait loci (eQTLs) and identify 42 signals that are fine-mapped to a candidate causal haplotype, rather than an individual genotyped variant.

DEVELOPMENT OF A NEW METHOD FOR IDENTIFICATION OF ANCESTRAL ALLELES FROM WHOLE GENOME SEQUENCE DATA

Hunter L McConnell[1], Caleb M Stull[1], Jenna A Kalleberg[1], Cody W Edwards[2,3], Budhan S Pukazhenthi[4], Klaus-Peter Koepfli[3], Robert D Schnabel[1]

[1]University of Missouri, Division of Animal Sciences, Columbia, MO, [2]George Mason University, Department of Biology, Fairfax, VA, [3]George Mason University, Smithsonian-Mason School of Conservation, Front Royal, VA, [4]Smithsonian's National Zoo and Conservation Biology Institute, Front Royal, VA

For numerous extant species, modern cattle included, their ancestral species are known but extinct. This renders a vast amount of information, including which alleles are ancestral, not directly observable. Knowledge of ancestral alleles is critical for reconstructing evolutionary histories and understanding population genetic patterns and processes. This data gap has led to multiple methods for estimating ancestral alleles, but existing methods are limited by their focus on biallelic positions only. Here we present a method that makes predictions for multiallelic as well as biallelic alleles based on coalescent theory. Specifically, our method assumes that after population branching, the ancestral allele at any given position is retained in all descendant lineages and, that the ancestral allele is still the major allele in all those lineages. To determine the ancestral allele, our method first compares allele frequencies in at least two outgroup species. Then, it identifies the allele with the highest combined frequency across those outgroups. The level of consensus that can be achieved between the outgroups correlates with a consistency value used to assign quality to the estimate. We used a cattle (*Bos taurus*) genomic dataset along with four outgroup species for which we had adequate data to test our method. The results of the estimates showed that 91% of the ancestral alleles were the reference allele. This suggests that modern cattle, specifically Hereford cattle which is the breed used as the reference genome, have accumulated few high-frequency derived alleles since the divergence of this lineage. When comparing ancestral alleles based on our method against those predicted using the est-sfs method with the Kimura two-parameter model, the same allele was identified at 93% of biallelic positions genome-wide. For a final experiment, 89 genes were identified as having at least two standard deviations above the average number of variants present in this dataset. For 35 of these genes the derived allele was the major allele for most of the variants. We also evaluated our method using a genome dataset for scimitar-horned oryx (*Oryx dammah*), a species that differs from our cattle dataset in sample size and phylogenetic relatedness of available outgroups. Despite limitations in sample size and the more distant relatedness of the outgroups, ancestral alleles were confidently estimated at 79% of variable positions, with the reference allele identified as ancestral in 49% of those cases. Given the substantial differences and respective limitations of the input datasets used, we believe our method can be adopted by anyone using data from a mammalian species, and that these new ancestral allele datasets will be a useful resources for their respective research communities.

A SINGLE CELL APPROACH TO STUDY COCAINE USE DISORDER

Cecilia McCormick[1,2,4], Nathan Nakatsuka[1,2,4], Lauren Wills[3,1], Eric Nestler[3,1], Paul Kenny[3,1], Rahul Satija[1,2]

[1]New York Genome Center, Core Member, New York, NY, [2]New York University, Center for Genomics and Systems Biology, New York, NY, [3]Icahn School of Medicine at Mount Sinai, Nash Family Department of Neuroscience and Friedman Brain Institute, New York, NY, [4]New York University Grossman School of Medicine, Department of Psychiatry, New York, NY

While the physical health and financial burden of substance use disorders (SUDs) cannot be overstated, the cellular mechanisms underlying SUDs have remained incompletely understood, hindering the development of clinical treatments. Recently, intravenous self-administration (IVSA) procedures in rodents have been shown to recapitulate behavioral abnormalities often observed in humans with SUDs, including drug consumption resistant to negative outcomes. At the same time, advancements in single-nucleus RNA sequencing (snRNA-seq) technologies have rapidly accelerated, allowing for higher-throughput experimentation to ascertain cell type-specific transcriptional changes for millions of nuclei. In this work, we combine developments in both fields by applying EasySci to assess the cell type-specific transcriptional responses to cocaine IVSA across mice brains, with an emphasis on subcortical regions implicated in SUD-related behaviors (e.g., nucleus accumbens, dorsal striatum, ventral tegmental area). This whole-brain atlas provides, for the first time, the means to assess coordinated transcriptional changes between brain regions in response to cocaine addiction in an unbiased, cell type-specific manner.

Here we subdivided 30 mice into a control group with IVSA saline access, a short-term access group (10 sessions of 1-hour access to IVSA cocaine administration), and long-term access group (10 sessions of 5-hour access to IVSA cocaine administration). This experimental setup is designed to allow the identification of a gradient of transcriptional changes related to the mouse's degree of addiction, which can be correlated with behavioral analysis. Following IVSA, we performed snRNA-seq using EasySci in a regional specific manner for 10 regions associated with the mesolimbic pathway and several cortical regions to generate a large (~1.5 million nuclei) dataset for downstream analysis. We identify cell types of interest in each region through a combination of reference mapping and unsupervised clustering before comparing cell types across regions. This approach allows us to assess drug-induced transcriptional remodeling in a brain-wide, unbiased manner as a method for unraveling the molecular basis for addiction. We also use the newly developed PASTA algorithm to assess differential 3' UTR and correlate this with RNA binding protein expression.. Lastly, we assess the cell type-specific expression of genetic variants predisposing humans to addiction using GWAS data and find cell types and brain regions with GWAS heritability enrichment.

DEVELOPING BADGERSEQ, AN AI-ASSISTED MODEL FOR ULTRA-RAPID LONG-READ GENOME SEQUENCING FOR CRITICALLY ILL INFANTS

M Stephen Meyn[1,2], Jessica M Chen[1,2], Derek Pavelec[3], Brian Ross[1], Jadin Heilmann[1], Leah A Frater-Rubsam[4], Hieu Nguyen[4], Xiangqiang Shao[4], Vanessa Horner[5], Bryn D Webb[1,2], April L Hall[1,2]

[1]University of Wisconsin - Madison, Center for Human Genomics, Madison, WI, [2]University of Wisconsin - Madison, Pediatrics, Madison, WI, [3]University of Wisconsin - Madison, Biotechnology Center, Madison, WI, [4]Wisconsin State Laboratory of Hygiene, Madison, WI, [5]Nationwide Children's Hospital, Columbus, OH

Background/Objectives: Rare genetic disorders affect >20% of infants in level IV neonatal intensive care units (NICUs), where rapid genome sequencing is becoming the genetic test of choice. BadgerSeq is designed to improve availability, quality, and speed of sequencing, with a goal of 96 hours from clinical presentation to laboratory results.
Methods: BadgerSeq selects patients for sequencing prior to a genetics consult by evaluating all NICU patients daily using artificial intelligence (AI) software. FastHPOCR software text mines their EMR records for Human Phenotype Ontology (HPO) terms, which are input into the Mendelian Phenotype Search Engine, a machine-learning based software trained to mimic the decisions of clinical geneticists regarding which infants to sequence, Infants whose scores exceed a threshold are offered rapid trio long-read Nanopore genome sequencing followed by AI-assisted tertiary variant analysis.
Results to Date: We are testing BadgerSeq's patient selection workflow on a retrospective cohort of 339 NICU patients, of whom 33% had genetic consults and 20% underwent rapid exome/genome testing. At the same time, we are piloting BadgerSeq's sequencing workflow using trios (proband + parents) from our Undiagnosed Disease Program.
In order to focus on speed and identification of disease-causing variants in known disease genes rather than discovery of novel variants and genes, we sequence each genome using two PromethION flow cells and then evaluate the resulting data using a standard Nanopore human genome analysis workflow combined with Fabric Genomics tertiary variant analysis. Importantly, we achieve Q27 raw read accuracy using Nanopore's SUP v5 base caller. The first 10 trio genomes averaged 46-48 hours from DNA to completion of initial tertiary analysis, including 36 hours sequencing. The workflow correctly identified pathogenic SNVs and structural variants in the 3 positive controls and suggested candidate variants for 3 of 7 unknowns.
Once we complete validating our workflows, we will run a prospective trial that will carry out trio genome sequencing on 120 NICU patients who meet our screening criteria for sequencing. Outcomes and performance of BadgerSeq's selection and sequencing workflows will be assessed by comparing these patients to a cohort of historical controls – NICU patients hospitalized in 2023-2024.
Conclusions: Our initial results demonstrate the feasibility of BadgerSeq's AI-assisted ultra-rapid sequencing workflow, a workflow that will offer hospitals a faster, more comprehensive long-read alternative to the current model of centralized high volume rapid short-read genome sequencing.

BUFFERING AND NON-MONOTONIC BEHAVIOR OF GENE DOSAGE RESPONSE CURVES FOR HUMAN COMPLEX TRAITS

Nikhil Milind[1], Courtney J Smith[1], Huisheng Zhu[2], Jeffrey P Spence[1], Jonathan K Pritchard[1,2]

[1]Stanford University, Department of Genetics, Stanford, CA, [2]Stanford University, Department of Biology, Stanford, CA

The genome-wide burdens of deletions, loss-of-function mutations, and duplications correlate with many traits. Curiously, for most of these traits, variants that decrease expression have the same genome-wide average direction of effect as variants that increase expression. This seemingly contradicts the intuition that, at individual genes, reducing expression should have the opposite effect on a phenotype as increasing expression. To understand this paradox, we introduce a concept called the gene dosage response curve (GDRC) that relates changes in gene expression to expected changes in phenotype. We show that, for many traits, GDRCs are systematically biased towards one trait direction relative to the other and, surprisingly, that as many as 40% of GDRCs are non-monotone, with large increases and decreases in expression affecting the trait in the same direction. We develop a simple theoretical model that explains this bias towards one trait direction. We also discuss an extension of our analysis to common variants with methods such as the transcriptome-wide association study (TWAS). Our results have broad implications for complex traits, drug discovery, and statistical genetics.

ARCHAIC ADMIXTURE REFUTES THE CURRENT PARADIGM OF TWO INDEPENDENT CATTLE DOMESTICATIONS

J L Miraszek[1], R D Schnabel[1], B Llamas[2], K Chen[2], A van Loenen[2], Y Souilmi[2], H J Rowan[1], P J Wrinn[2], S Vasil'ev[2], N D Ovodov[2], M Sinclair[3], J F Taylor[1], A Cooper[3], J E Decker[1]

[1]University of Missouri, Division of Animal Sciences, Columbia, MO, [2]The University of Adelaide, Australian Centre for Ancient DNA (ACAD), Adelaide, Australia, [3]Charles Sturt University, Gulbali Institute, Albury, Australia

Domesticated cattle and their wild relatives, including yak, bison, gaur, banteng and aurochs, are members of a clade with a rich evolutionary history, and cosmopolitan success, which has been shaped by anthropogenic forces. Domestication has been the most profound phenotypic change within this clade, however, migration, isolation, admixture and selection have driven evolution within this clade pre- and post-domestication. Previous studies of the evolution of the cattle genome either lacked the species diversity or sample sizes necessary to accurately recreate a global population history. Consequently, signatures of selection during domestication have been obscured. We used variation detected in 5606 whole genome sequences representing more than 200 breeds of wild and domestic animals from the genus *Bos*, to establish demography, ancestry and historical population sizes. Using f_4-ratios and qpGraphs, we identified gene-flow from wild species into key domestic populations, most notably gaur into the ancestor of indicine cattle at ~40% ancestry. Models support historic gene flow between aurochs and a gaur-like population in South Asia creating a progenitor of indicine cattle, *Bos namadicus*. Later admixture between *Bos namadicus* and domestic taurines transported from the Fertile Crescent created the iconic humped *Bos indicus* cattle from the Indus Valley. This rejects the widely-accepted paradigm of two independent cattle domestications, because indicine cattle were established through introgression with already domesticated *Bos taurus* cattle. We used D-statistic based methods, namely f_{dM}, to identify regions of modern cattle genomes with selected introgressed segments. This identified selection on 126 genes originating in *Bos taurus* and introgressed into *Bos indicus* as domestication loci, including olfaction (*OR2AD1, OR8D6, OR2H10*) and intelligence (*NELL1, PRKN*) genes. We also identified 104 genes within *Bos indicus* that originated in the *Bos namadicus* gaur ancestor, implicated in environmental adaptation including cardiovascular associated (*DDR1, TFF1, CAMK2D*) and immune system (*HLA-DRA, HLA-DOB, VARS2*) loci. Leveraging a globally representative, whole genome data, we identified the evolutionary forces and population sources that gave rise to one of the most economically important agricultural species, rewriting the evolutionary history of cattle in the process.

MULTI-STUDY FINE-MAPPING ENABLES IDENTIFICATION OF SHARED AND ANCESTRY-SPECIFIC SIGNALS DRIVING COMPLEX TRAITS

Tara Mirmira[1], Nichole Ma[2], Jonathan Margoliash[1], Wilfredo Gabriel Gonzalez Rivera[3,4], Tiffany Amariuta[4,5], Kelly Frazer[6,7], Alon Goren[2], Melissa Gymrek[1,2]

[1]University of California, San Diego, Computer Science and Engineering, La Jolla, CA, [2]University of California, San Diego, Department of Medicine, La Jolla, CA, [3]University of California, San Diego, Bioinformatics and Systems Biology Graduate Program, La Jolla, CA, [4]University of California, San Diego, Department of Medicine, Division of Biomedical Informatics, La Jolla, CA, [5]University of California, San Diego, Halıcıoğlu Data Science Institute, La Jolla, CA, [6]University of California, San Diego, Department of Pediatrics, La Jolla, CA, [7]University of California, San Diego, Institute of Genomic Medicine, La Jolla, CA

Genome-wide association studies (GWAS) quantify associations between variants and traits but do not directly identify causal variants. Statistical fine-mapping aims to identify candidate causal variants from GWAS, but struggles to distinguish between variants in high linkage disequilibrium (LD). Multi-study fine-mapping methods can improve the resolution of causal variant identification by leveraging population-specific LD patterns. Although prior work suggests many GWAS signals are shared across populations, the growing size and diversity of GWAS datasets provides increasing evidence that a subset of signals are specific to certain ancestries. Yet, existing multi-study fine-mapping methods typically assume causal variants are shared and polymorphic across studies. Here, to overcome this limitation, we introduce PIPSORT, a multi-study fine-mapping method that leverages study-specific LD patterns and simultaneously detects both shared and ancestry-specific signals. We apply PIPSORT to perform multi-ancestry fine-mapping in individuals of primarily African vs. European ancestry for platelet count, LDL cholesterol, and estimated glomerular filtration rate in both the UK Biobank (UKB) and All of Us (AoU) datasets. Due to the bias of these datasets toward Europeans (94% in UKB and 49% in AoU), most trait-associated regions identified have strong signals in Europeans. Of these regions, we estimate 89%-99% are shared with the African population, but detect dozens of examples of ancestry-specific signals. These include 17 that could only be confidently detected in AoU. These results highlight the advantage of performing fine-mapping in the more diverse AoU dataset, which enabled fine-mapping with 1.5-3x more samples of African ancestry for each phenotype compared to UKB. Finally, we leverage the high degree of admixture within AoU to identify ancestry-specific signals that are potentially driven by local vs. global ancestry. Overall, our results emphasize the value in accounting for ancestry-specific causal variants, a finding with important implications for applications beyond fine-mapping including construction of polygenic risk scores.

MtDNA MUTATIONS DIFFERENTIALLY AFFECT MITOCHONDRIAL TRANSCRIPTION

Yuval Caruchero, Sarah Dadon, Dan Mishmar

Ben-Gurion University of the Negev, Life Sciences, Beer-Sheva, Israel

Mitochondrial DNA (mtDNA) transcription is polycistronic and is governed by separate strand-specific promoters. MtDNA disease-causing mutations have been mapped to the mtDNA coding regions, while mutations in the regulatory mtDNA regions were overlooked. We previously showed that ancient mtDNA mutations affect mtDNA transcription and transcription factors binding capacity. Nevertheless, the principles that govern the phenotypic impact of regulatory region mtDNA mutations are poorly understood. Current transcriptional assays provide only qualitative assessments of mtDNA transcription with poor detailed information regarding transcriptional initiation and elongation. Here, we used the Oxford Nanopore MinION sequencing platform to quantify the impact of mtDNA mutations on transcriptional initiation, pausing and elongation in a cell free assay. Our results show that point mutations and deletions, some within known regulatory elements and some outside of such, alter transcriptional pausing and initiation. Interestingly, not all mutations within known regulatory elements affected transcription, thus pointing towards the potential discovery of novel regulatory elements. Our approach provides proof of concept that paves the path towards high throughput functional analysis of both ancient and recently accumulated mtDNA mutations.

DYNAMIC CLASSIFICATION OF CIS-REGULATORY ELEMENTS ACROSS DIVERSE CELLULAR CONTEXTS

Gregory Andrews[1], Nicole Shedd[1], Vivekanandan Ramalingam[2], Anshul Kundaje[2,3], Zhiping Weng[1], Jill E Moore[1]

[1]University of Massachusetts Chan Medical School, Genomics and Computational Biology, Worcester, MA, [2]Stanford University, Department of Genetics, Stanford, CA, [3]Stanford University, Department of Computer Science, Stanford, CA

Transcriptional regulation is a critical determinant of cellular identity and function. Central to this regulation are cis-regulatory elements (CREs), which orchestrate gene expression through interactions with transcription factors and chromatin-associated proteins. We previously developed the ENCODE Registry of candidate CREs (cCREs), which comprises 2.3 million such elements across the human genome and provides a comprehensive resource for studying transcriptional regulation. Traditionally, cCREs are classified into exclusive categories such as promoters, enhancers, insulators, or silencers based on biochemical features; however, evidence suggests that their functional roles can be highly context-dependent. For example, our recent work identified cCREs with context-dependent enhancer/silencer activities, demonstrating that regulatory function is not static but can change depending on cellular context. Here, we extend this analysis to systematically investigate the dynamic classification of cCREs across a broad compendium of 170 diverse cell and tissue types.

To identify cCREs that switch between functional categories, we developed an entropy-based method to quantify functional variability. While the majority of cCREs exhibit consistent functional classifications, thousands of promoters and enhancers exhibit cell-type-specific transitions. Notably, promoters that change classification are frequently associated with non-coding genes, particularly lncRNAs, and have lower GC content compared to stable promoters. Similarly, we found that multi-function enhancers display widespread enhancer activity across multiple tissues but are notably depleted for enhancer activity in embryonic stem cells and immune cells.

To dissect the regulatory mechanisms underlying these functional transitions, we used deep learning-based methods to identify putative transcription factor binding sites. This analysis revealed that alternative transcription factor binding patterns, rather than changes in chromatin accessibility at the macroscopic level, are a primary driver of functional plasticity. For example, we identified an enhancer of the brain-specific MAST1 gene that has enhancer activity exclusively in brain tissues and acts as a CTCF-loop anchor in all other cell types. We attribute this difference in activity to the binding of FOS family transcription factors, which was only observed in brain tissues and absent in cell types exhibiting the CTCF-only function.

Our findings provide new insights into the dynamic nature of cis-regulatory elements and their role in transcriptional regulation. By expanding functional annotations beyond static classifications, we highlight the need for context-aware models of gene regulation.

COMPLETE CHARACTERIZATION OF HUMAN POLYMORPHIC INVERSIONS AND OTHER COMPLEX VARIANTS FROM LONG READ DATA

Ricardo Moreira-Pinhal[1,2], Konstantinos Karakostis[2], Illya Yakymenko[1,2], Odei Blanco-Irazuegui[2], Maria Díaz-Ros[2], Marta Puig[1,3], Mario Cáceres[1,2,4]

[1]Universitat Autonoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Cerdanyola del Vallès, Spain, [2]Hospital del Mar Research Institute, Research Programme on Biomedical Informatics, Barcelona, Spain, [3]Universitat Autonoma de Barcelona, Departament de Genètica i de Microbiologia, Cerdanyola del Vallès, Spain, [4]ICREA, -, Barcelona, Spain

Inversions are a special type of structural variants (SVs) whose study has lagged behind due to their balanced nature and the presence of large inverted repeats (IR) at their breakpoints. New techniques are finally allowing us to identify the full spectrum of human inversions. However, in most cases, only a limited number of individuals has been studied, which precludes the analysis of the effects of the detected variants. Here, we use Oxford Nanopore Technologies (ONT) long read data of multiple individuals to generate and characterize the most exhaustive catalogue of IR-mediated inversions in human genomes. First, we have developed GeONTIpe, a bioinformatic package to genotype accurately inversions from long reads. Using this method, we have interrogated 612 candidate inversions from different studies, ranging from 197 bp to 4.4 Mb and flanked by up to 190-kb long IRs, in a diverse set of 1211 samples. We detected both orientations in 327 inversions, validating 273 novel inversions. Among the polymorphic inversions, we observed a wide range of frequencies (0.04-50%), with 161 inversions having more than 1% of frequency. Moreover, we showed that ONT genotypes were highly accurate, matching previous experimental genotypes of 54 inversions. By comparing the obtained genotypes with four recent SV prediction studies and with the new Pangenome reference. We observed that some of these regions present discrepancies which suggest that they might not be well resolved. Importantly, 428 inversion regions present also additional SVs that were identified on the predictions, revealing the complexity of repeat-rich sequences. Finally, we calculated the linkage disequilibrium (LD) of the polymorphic inversions with nearby SNPs to determine their origin and assess if they can be imputed in other datasets. Long reads therefore have a great potential for the characterization of currently missed inversions and other complex genomic regions at a large scale, opening the door to determining their real functional impact.

ENABLING GENOMIC RESEARCH AT SCALE WITH NHGRI ANVIL: A CLOUD PLATFORM FOR GENOMIC DATA ANALYSIS

Stephen L Mosher[1], Michael C Schatz[1,2], Jonathan Lawson[3], Robert Carroll[4]

[1]Johns Hopkins University, Biology, Baltimore, MD, [2]Johns Hopkins University, Computer Science, Baltimore, MD, [3]Broad Institute of MIT and Harvard, Data Science Platform, Cambridge, MA, [4]Vanderbilt University Medical Center, Biomedical Informatics, Nashville, TN

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

The rapid expansion of human genomics presents enormous opportunities for advancing human health. However, traditional models of genomic data sharing—where datasets are downloaded to local infrastructure for analysis—are increasingly unsustainable, cost-prohibitive, and pose challenges for security and compliance.

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL) (https://anvilproject.org/) addresses these challenges by providing a secure, cloud-based platform for genomic data storage, management, and analysis. Instead of requiring data downloads, AnVIL allows researchers to bring their analyses to the data, leveraging scalable computing and security monitoring.

The platform hosts over 600,000 genomes from NHGRI-supported projects such as GREGoR Consortium (Genomics Research to Elucidate the Genetics of Rare diseases), GTEx (Genotype-Tissue Expression Project), T2T (Telomere-to-Telomere), Human Pangenome Reference Consortium (HPRC), eMERGE (Electronic Medical Records and Genomics), CCDG (Centers for Common Disease Genomics), CMG (Centers for Mendelian Genomics) and more.

AnVIL is built on a foundation of established components that support flagship scientific initiatives. Terra provides a secure and scalable computing environment, while the Terra Data Repository enables efficient data storage and access management. Dockstore facilitates standards-based sharing of containerized tools and workflows, ensuring reproducibility. For interactive data exploration, AnVIL integrates Jupyter, R/Bioconductor, and Galaxy, offering users access to thousands of analytical tools. Together, these components create a collaborative ecosystem for managing, analyzing, and sharing genomic data and workflows at scale.

AnVIL serves as a unified platform for integrating current and future genomic and related datasets. By streamlining access to protected data, AnVIL significantly lowers barriers for investigators, making it easier to conduct large-scale analyses across datasets and fully harness the value of genomic data production efforts.

IDENTIFICATION OF THE SHORTEST SPECIES-SPECIFIC OLIGONUCLEOTIDE SEQUENCES

Ioannis Mouratidis*[1,2], Maxwell A Konnaris*[2], Nikol Chantzi*[1,2], Candace S Chan*[1], Michail Patsakis[1,4], Kimonas Provatas[1,4], Austin Montgomery[1], Fotis Baltoumas[5], Congzhou M Sha[1], Manvita Mareboina[1], Georgios A Pavlopoulos[5,6], Dionysios V Chartoumpekis[7], Ilias Georgakopoulos-Soares[1,2]

[1]The Pennsylvania State University College of Medicine, Department of Molecular and Precision Medicine, Hershey, PA, [2], The Pennsylvania State University, Huck Institutes of the Life Sciences, University Park, PA, [3]University of California San Francisco, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA, [4]National Technical University of Athens, School of Electrical and Computer Engineering, Athens, Greece, [5]BSRC “Alexander Fleming", Institute for Fundamental Biomedical Research, Vari, Greece, [6]National and Kapodistrian University of Athens, Center for New Biotechnologies and Precision Medicine, Athens, Greece, [7]Lausanne University Hospital, Service of Endocrinology, Diabetology and Metabolism, Lausanne, Switzerland

Despite the exponential increase in sequencing information driven by massively parallel DNA sequencing technologies, universal and succinct genomic fingerprints for each organism are still missing. Identifying the shortest species-specific nucleotide sequences offers insights into species evolution and holds potential practical applications in agriculture, wildlife conservation, and healthcare. We propose a new method for sequence analysis termed nucleic “quasi-primes,” the shortest occurring sequences in each of 45,076 organismal reference genomes, present in one genome and absent from every other examined genome. In the human genome, we find that the genomic loci of nucleic quasi-primes are most enriched for genes associated with brain development and cognitive function. In a single-cell case study focusing on the human primary motor cortex, nucleic quasi-prime genes account for a significantly larger proportion of the variation based on average gene expression. Nonneuronal cell types, including astrocytes, endothelial cells, oligodendrocytes, and vascular and leptomeningeal cells, exhibit significant activation of quasi-prime-containing gene associations related to cancer, whereas simultaneously suppressing quasi-prime-containing genes are associated with cognitive, mental, and developmental disorders. We also show that human disease–causing variants, eQTLs, mQTLs, and sQTLs are 4.43-fold, 4.34-fold, 4.29-fold, and 4.21-fold enriched at human quasi-prime loci, respectively. These findings indicate that nucleic quasi-primes are genomic loci linked to the evolution of species-specific traits, and in humans, they provide insights in the development of cognitive traits and human diseases, including neurodevelopmental disorders.

*authors contributed equally

NONA: A UNIFYING MULTIMODAL MASKED MODELING FRAMEWORK FOR FUNCTIONAL GENOMICS

Surag Nair, Nathaniel Diamant, Alex Tseng, Ehsan Hajiramezanali, Avantika Lal, Tommaso Biancalani, Gabriele Scalia, Gokcen Eraslan

ReLU, BRAID, Genentech, South San Francisco, CA

We present Nona, a unifying multimodal masked modeling framework for functional genomics. Nona is a neural network model that operates on both DNA sequence and epigenetic tracks such as DNase-seq, ChIP-seq, and RNA-seq at base-pair resolution. By leveraging a flexible masking strategy, Nona can predict any subset of masked DNA and/or tracks from the unmasked subset. Nona supports existing sequence-to-function models such as Enformer and BPNet, and their applications such as variant effect prediction. Beyond this, Nona enables multiple novel application modes that we highlight below:

- Sequence generation mode: In this mode, Nona is trained as a masked language model conditioned on epigenetic profiles. Upon training, this model can be used to generate novel and synthetic regulatory elements that have both desired activity as well as shape (such as broad vs narrow ATAC-seq peak) across multiple cell states.

- Functional genotyping mode: Surprisingly, we show that commonly used ATAC-seq fragment files, though devoid of variant level information, leak private information. Using the conditional masked language mode of Nona, we predict genotypes solely from base-pair resolution ATAC-seq profiles. We then perform a linking attack that identifies the sample donors with near perfect accuracy. The linking is robust to read subsampling, and is highly accurate at read depths as low as 5M.

- Sequence+context mode: In this mode, Nona predicts epigenetic tracks in a local genomic window by taking into account the observed epigenetic tracks in adjacent windows, in addition to the DNA sequence. We show that this mode allows the model to integrate broader chromatin context information in its predictions. We demonstrate that the sequence+context mode improves prediction of expression of randomly integrated reporters across diverse genomic backgrounds, as measured by TRIP-seq experiments.

- Denoising mode: In this mode, epigenetic tracks are randomly masked at a subset of bases. Nona predicts the tracks at these masked bases from DNA sequence and the observed epigenetic tracks at unmasked bases. We show that this method is superior to denoising with DNA-only sequence-to-function models.

Altogether, Nona is a versatile framework that extends sequence-to-function and masked language modeling to novel applications in regulatory genomics. We anticipate that Nona can also be used for studying epigenetic perturbations and extensions of language modeling, and that interpretation of Nona can shed more light on context-specific dependencies between epigenetic tracks and regulatory elements.

ARAB PANGENOME REFERENCE: UNCOVERING NOVEL SEQUENCES

Nasna Nassir[1,2], Mohamed Almarri[2,3], Muhammad Kumail[1], Nesrin Mohamed[1], Bipin Balan[1], Shehzad Hanif[1], Maryam AlObathani[1], Bassam Jamalalail[1], Hanan Elsokary[1], Dasuki Kondaramage[1], Suhana Shiyas[1], Hamda H Khansaheb[1,2], Alawi Alsheikh-Ali[1,2], Mohammed Uddin[1,2,4]

[1]Mohammed Bin Rashid University of Medicine and Health Sciences, Center for Applied and Translational Genomics (CATG), Dubai, United Arab Emirates, [2]Mohammed Bin Rashid University of Medicine and Health Sciences, College of Medicine, Dubai, United Arab Emirates, [3]Dubai Police GHQ, Genome Center, Department of Forensic Science and Criminology, Dubai, United Arab Emirates, [4]GenomeArc Inc., GenomeArc Inc., Mississauga, Canada

Pangenomes offer a comprehensive view of genetic diversity, yet Arab populations remain underrepresented in current human genomic references. We present the Arab Pangenome Reference (APR), constructed from 53 individuals representing diverse Arab ethnicities. Using 35.27X high-fidelity long reads, 54.22X ultralong reads, and 65.46X Hi-C reads, we generated haplotype-phased *de novo* assemblies with exceptional quality, achieving an average N50 of 124.28 Mb. This effort revealed 111.96 million base pairs of novel euchromatic sequences absent in current global human references, including the T2T-CHM13, GRCh38, and other public datasets. We identified 8.94 million population-specific small variants and 235,195 structural variants, which were absent from linear references and existing pangenomes. Notably, we discovered 883 gene duplications, including a unique duplication of the TATA-binding protein gene *TAF11L5*, consistently found across all Arab individuals. Interestingly, 15.06% of the duplicated genes are associated with recessive diseases, underscoring their clinical relevance. Our mitochondrial analysis uncovered 1,436 bp of novel sequences, contributing additional insights into population-specific variation. Our study presents the first Arab pangenome, uncovering previously uncharacterized genomic features. The APR serves as a foundational resource for advancing genetic research, understanding disease predispositions, and enabling precision medicine initiatives for Arab populations and others with shared genetic backgrounds.

THE IMPACT OF PASSENGER MUTATIONS ON CANCER DEVELOPMENT

Akshatha Nayak, Ioannis Mouratidis, Ilias Georgakopoulos-Soares

Pennsylvania State University College of Medicine, Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, Hershey, PA

Cancer progression is characterized by the accumulation of somatic mutations. The prevailing dichotomy categorizes mutations as either “drivers” that directly contribute to cancer development or “passengers” considered incidental byproducts of the genomic instability of cancer cells. Studies suggest that about 30% of cancer patients have no identifiable driver mutations. While this could indicate the presence of undiscovered drivers, it could also suggest that passenger mutations play a driver-like role in these patients. Previous research has examined the role of weak drivers and deleterious passenger mutations, along with their additive effects on cancer progression. In this study, we present a comprehensive analysis of the cumulative impact of passenger mutation clusters on cancer development using WGS tumor samples from the Pan-Cancer Analysis of Whole Genomes project. We found that when driver mutations are absent in a driver gene, it exhibits a higher density of passenger mutations compared to when a driver mutation is present. Moreover, these passenger mutation clusters that emerge in the absence of drivers have higher CADD scores, indicating that these mutations are more deleterious. We also employed AI models, including Borzoi, to predict the consequences of passenger mutations. Further clinical analysis revealed genes in which these passenger mutation clusters are associated with a decrease in both survival probability and progression-free interval, even in the absence of driver mutations in the gene. These findings highlight the need for a holistic model of cancer development that moves beyond the traditional dichotomy of driver and passenger mutations, potentially leading to more accurate prognosis for cancer patients and improved healthcare outcomes.

RESOLVING GENE-ALTERING SVs IMPROVES THE QUANTIFICATION OF TRANSCRIPT ABUNDANCES

Bohan Ni[1], Alexis Battle[2], Michael C Schatz[1]

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

Structural Variants (SVs) have greater potential impact on gene function compared to individual SNVs. Common SVs have been analyzed in eQTL studies, and rare SVs have been connected to extreme changes in gene expression through methods such as our recent approach Watershed-SV, enabling functional characterization. However, estimates of expression levels are usually based on fixed reference transcriptome annotations, neglecting the inter-individual variation in gene structures. As a result, variation in expression level may be attributed to changes in transcript abundance or to unaccounted gene elongations and ablations. The entangled signals of altered gene structure and altered transcript abundance could compromise identifications of both gene-altering SVs and variants regulating transcript abundance.

Addressing this need, we present a transcript-aware graph-based algorithm to genotype SVs from RNA-seq and to derive putative transcript models. For this, we extended the interval-adjacency graph (IAG) [1] to create a transcript-aware IAG (T-IAG) with personal SV genotypes. Re-aligning RNA-seq data against T-IAG reduces the number of paths for a gene and estimates the extent of disruption of an SV on the expressed transcripts in a tissue. By performing an Eulerian Decomposition of T-IAG, it also extracts the most likely transcript structures accounting for variant impact.

Among 688,356 sample-gene pairs with at least 1 SV affecting the gene body from the Genotype-Tissue Expression Project (GTEx), we investigated the expression of 636,299 pairs with 1 or 2 SVs to find widespread evidence of gene-altering SVs. T-IAGs were constructed on a per-gene basis for > 79% of the cases, allowing parallel graph construction and analysis. Among them, a rare duplication in *CMTR2* not only duplicates affected exons but also leads to intron retention at the SV boundary. Notably, this gene was previously identified as an overexpression outlier although our method resolves it as normal transcript abundance, albeit of a previously unknown isoform. We also apply this analysis to datasets like Undiagnosed Disease Network to assess its ability to detect disease variants. We anticipate that accounting for personalized genomes will improve the accuracy of transcriptomic analysis and shed light on regulatory mechanisms of functional variants.

NEW PATHWAY ANALYSIS ENVIRONMENT USING WIKIPATHWAYS MECHANISM

Ryo Nozu[1], Naoya Oec[2], Shota Matsumoto[2], Alexander R Pico[3], Hidemasa Bono[1,4]

[1]Hiroshima University, Graduate School of Integrated Sciences for Life, Higashi-Hiroshima, Japan, [2]dogrun Inc., Shizuoka, Japan, [3]Gladstone Institutes, Data Science and Biotechnology, San Francisco, CA, [4]Hiroshima University, Genome Editing Innovation Center, Higashi-Hiroshima, Japan

The development of high-accuracy long-read sequencing technologies has expanded de novo genome sequencing for non-model organisms. Additionally, improvements in gene model prediction accuracy allow the retrieval of comprehensive gene sets, including transcript isoform information, for genome-sequenced species. One of the major motivations for genome analysis in non-model organisms is the understanding of biosynthetic pathways of species-specific metabolites. To achieve this, pathway analysis based on genome annotations is essential. Currently, multiple pathway databases serve as the basis of such analyses, but challenges remain when applying them to non-model organisms.
First, most pathway data originate from model organisms. Consequently, sequence IDs generated from de novo genome annotations of the target species often need to be converted to reference species IDs. Furthermore, pathway diagrams generally do not account for transcript isoforms, necessitating the summarization of information at the gene level during ID conversion. This raises concerns that crucial information about specific isoforms involved in metabolite biosynthesis may be lost, potentially obscuring the underlying biosynthetic mechanisms in the target species.
To address these issues, we developed QPX (Quest for Pathways with eXpression), a pathway analysis environment designed for non-model organisms. For pathway visualization, QPX supports the GPML (Graphical Pathway Markup Language) format, adopted by WikiPathways, an open-source biological pathway database. The main reason for adopting this format in our system is that it allows users to create custom pathway diagrams from scratch or modify existing ones in WikiPathways, enabling flexible pathway analysis for non-model species. For expression profile visualization, QPX requires an input table linking enzyme-coding genes (nodes) in the pathway diagram to transcript expression values via xref_id. This relation enables transcript expression levels to be displayed within selected gene nodes. Furthermore, attribute information of selected nodes can be retrieved as data frames for further downstream analyses.
In this way, our system provides an integrated pathway analysis environment, covering various aspects from species-specific pathway diagram creation to transcript isoform expression data visualization. We expect that this will significantly accelerate pathway analysis in non-model organisms.

GENOME-WIDE ASSOCIATION MAPPING OF DROUGHT-INDUCED OXIDATIVE STRESS RESPONSES IN INDICA RICE REVEALS STRUCTURAL VARIATION IN *OSAAO2* AS A KEY REGULATOR OF ASCORBATE REDOX STATE

Chosen E Obih[1], Yong Zhou[3,4], Dario Copetti[3], Lin-Bo Wu[2], Rod Wing[3,4], Giovanni Melandri[1]

[1]University of Arizona, School of Plant Sciences, Tucson, AZ, [2]Justus Liebig University, 2Department of Agronomy and Crop Physiology, Institute for Agronomy and Plant Breeding, Giessen, Germany, [3]University of Arizona, Arizona Genomics Institute (AGI), Tucson, AZ, [4]King Abdullah University of Science and Technology (KAUST), Center for Desert Agriculture (CDA), Biological and Environmental Sciences & Engineering Division (BESE), Thuwal, Saudi Arabia

Crop genetic diversity provides a reservoir of beneficial alleles that can be leveraged to enhance abiotic stress tolerance. In this study, we conducted a genome-wide association mapping using 271 indica rice accessions to identify genetic loci associated with the variation of oxidative stress-related traits measured from flag leaf tissues under well-watered and water-limited field conditions. We identified a major-effect quantitative trait locus (QTL) on chromosome 6 that was simultaneously associated with drought-induced variation in dehydroascorbate reductase (DHAR), monodehydroascorbate reductase (MDHAR), and total non-enzymatic antioxidant capacity (TAC). To fine-map this QTL, we generated high-coverage (30×) *de novo* genome assemblies for 17 rice accessions representing two distinct QTL haplotypes. These assemblies enabled the discovery of previously undocumented causal variants within OsAAO2, a gene encoding ascorbate oxidase, a key regulator of the molecular antioxidant ascorbic acid. Comparative sequence analysis and protein modeling suggest that these variants lead to a partial loss of function in OsAAO2, enhancing drought tolerance. Furthermore, we analyzed a 10K rice accession dataset to assess the frequency and distribution of OsAAO2 variants across global rice diversity. Our findings provide a valuable genetic resource for marker-assisted selection (MAS). Introgressing these variants, into the genetic background of commercial rice varieties could help breeders improve their resilience to drought stress.

DEVELOPING FLEXIBLE AND SCALABLE VISUALIZATION OF WHOLE GENOME ALIGNMENTS AT NCBI

Dong-Ha Oh, Andrea Asztalos, Evgeny Borodin, Vladislav Evgeniev, Raymond Koehler, Vadim Lotov, Marina Omelchenko, Dmitry Rudnev, Joël Virothaisakun, Sanjida H Rangwala

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

Genome alignments offer essential frameworks for interpreting genomic data on scales ranging from variations among pangenomes to conservation across clades in the tree of life. Vastly expanding genome resources from concerted efforts such as the Vertebrate Genome Project and the Earth Biogenome Project expose the need for tools to visualize and analyze genome alignments that are both flexible and scalable.

As part of the NIH Comparative Genome Resource (CGR) initiative, our team at NCBI provides a set of interactive visualization approaches. Comparative Genome Viewer (CGV; https://ncbi.nlm.nih.gov/cgv) explores structural variations and synteny for a pair of genomes. CGV is interconnected with Genome Data Viewer (GDV; https://ncbi.nlm.nih.gov/gdv), which enables users to browse any number of pairwise genome alignment tracks in detail, along with other tracks such as gene expression, variation, and custom data. We accept user requests to generate pairwise genome alignments or adapt existing alignment data for CGV and GDV.

Our newest interactive genome browser, the Multiple Comparative Genome Viewer (MCGV; https://ncbi.nlm.nih.gov/mcgv), navigates multiple genome alignments and helps researchers trace evolutionarily related regions. Pre-calculated conservation graphs summarize the multiple genome alignment and allow users to identify and zoom into genome regions that contain variations and conservations that may be biologically interesting. Gene models with exon structures are displayed for all annotated genomes to aid in assessing orthologous regions. The genomic region and gene models included in an alignment can be further investigated in GDV through the link available for each alignment block. Users can customize the view to include only the subset of species or clades of interest.

MCGV, under active development, is now available to view EBI's EPO mammalian alignment and a Cactus alignment of eight NHGRI telomere-to-telomere primate assemblies. We are particularly interested in collaborating with researchers to display multiple genome alignments for their study species or clade.

NCBI's genome browsers, GCV, CGV, and MCGV, each with complementary functions, aim to meet user needs in accessing genome data. We strive to make our tools more useful to the research community and welcome feedback.

This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

WE INTERPRET CONGENITAL HEART DISORDERS BY COMPREHENSIVE ANALYSIS OF CELL TYPE-SPECIFIC GENE REGULATORY PROGRAM IN THE EARLY DEVELOPING HUMAN HEART

Sungryong Oh[1], Kevin Child[2], Justin Cotney[1]

[1]Children's Hospital of Philadelphia, Plastic Surgery, Philadelphia, PA, [2]University of Connecticut Health Center, Genetics and Genome Sciences, Farmington, CT

Congenital heart diseases (CHD) are among the most common defects found in newborns. However, around 50% of CHD cases remain unexplained by gene mutations or other large copy number variations. It has been proposed that defective gene regulatory sequences could be at fault in many of these unexplained cases, especially many tissue-specific defects during the organogenesis. Heart is one of the earliest organs to form, which precludes comprehensive profiling of active regulatory sequences and their targets. To address this gap, we focused on uncovering the gene regulatory dynamics of the developing human heart at the single-cell level by integrating primary tissue data and in vitro organoid models. We performed joint profiling of gene expression and chromatin accessibility in single nuclei (snMultiome) from early heart tissues, ranging from CS(Carnegie Stage) 13 to 16. By integrating with other single-cell RNA sequencing from fetal human heart, we identified 14 distinct clusters of 86,549 cells based on transcriptomes. These clusters include previously unknown small cell types enriched in early developing heart cells, as well as most of cell types found in the fetal heart. This encompasses the key structural milestones of heart development, when many critical cell types are specified. This encompasses most of the structural milestones of heart development and when many critical cell types are specified. We also profiled three-dimensional chromatin dynamics from 99,969 cells in developing human heart. With this integration analysis, we could build physical links between regulatory sequences and their gene targets in cardiac relevant cell types. This not only helped us understand how cell types are specified during organogenesis, but also allowed us to identify highly accessible genomic regions that may be linked to risk factors for CHDs. To model the disease, we optimized heart organoid models from hESCs, which exhibited morphological and structural features of the developing heart. Using snMultiome analysis on this model, we found that it includes most of the cell types present in primary tissues, specifically derived from a lineage of human cardiac progenitors. Our organoid model also incorporated an inducible KRAB-dCas9 system to suppress the activity of specific regions during differentiation, thereby expected to mimic individual unexplained CHD cases. Overall, this approach provides a comprehensive understanding of human heart organogenesis and offers valuable genetic insights into unexplained CHD cases.

SYSTEMATIC DISCOVERY OF DIRECTIONAL REGULATORY MOTIFS ASSOCIATED WITH HUMAN INSULATOR

Naoki Osato

Institute of Science Tokyo, School of Life Science and Technology, Tokyo, Japan

Insulator proteins act as a barrier of enhancer–promoter interactions. The main insulator protein in vertebrates is CTCF, a DNA-binding protein. However, other DNA-binding proteins associated with insulators of enhancer–promoter interactions are still unclear.

Hence, we developed a systematic, comprehensive deep learning-based approach for identifying the DNA motifs of DNA-binding proteins associated with insulators. We discovered 97 directional and minor nondirectional motifs in human fibroblast cells that corresponded to 23 DNA-binding proteins related to insulator function, CTCF, and/or other types of chromosomal transcriptional regulation reported in previous studies [1][2][3]. The DNA binding sites of the motifs were located at insulator sites identified from chromatin interaction data. The estimated CTCF orientation bias was consistently proportional to CTCF orientation bias observed in chromatin interactions. The motifs were significantly more abundant at insulator sites separated by repressive and active regions, at boundary sites identified by chromatin interaction data, and at splice sites than other DBPs. For instance, MyoD forms chromatin loops in muscle cells [4]. We found that the DNA-binding site of the chromatin loop is located at an insulator site near a gene involved in skeletal muscle differentiation and function. Furthermore, we observed that the motifs potentially regulate transcribed regions differentially repressed in alternative transcripts. Finally, the insulator-pairing model explains that homologous and heterologous insulator-insulator pairing interactions are orientation-dependent [5][6][7]. These findings contribute to elucidate novel transcriptional regulations.

[1] Osato N. and Hamada M. Systematic discovery of directional regulatory motifs associated with human insulator sites. bioRxiv (2024) doi: https://doi.org/10.1101/2024.01.20.573595
[2] Xiao T. et al. The Myc-associated zinc finger protein (MAZ) works together with CTCF to control cohesin positioning and genome organization. PNAS (2021) doi: 10.1073/pnas.2023127118
[3] Ortabozkoyun H. et al. CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. Nature Genetics (2022) doi: 10.1038/s41588-021-01008-5
[4] Wang R. et al. MyoD is a 3D genome structure organizer for muscle cell identity. Nature Communications (2022) doi; 10.1038/s41467-021-27865-6
[5] Fujioka M. et al. Determinants of Chromosome Architecture: Insulator Pairing in cis and in trans. PLoS Genetics (2016) doi: 10.1371/journal.pgen.1005889
[6] Bing X. et al. Chromosome structure in Drosophila is determined by boundary pairing not loop extrusion. eLife (2024) doi: 10.7554/eLife.94070
[7] Ke, W. et al. Stem-loop and circle-loop TADs generated by directional pairing of boundary elements have distinct physical and regulatory properties. eLife (2024) doi: 10.7554/eLife.94114

FUNCTIONAL PREDICTION OF DNA/RNA-BINDING PROTEINS BY DEEP LEARNING FROM GENE EXPRESSION CORRELATIONS

Naoki Osato

Institute of Science Tokyo, School of Life Science and Technology, Tokyo, Japan

Experimental identification of DNA/RNA-binding sites of proteins is difficult for some proteins. Deep learning methods have been developed to accurately predict gene expression levels from DNA-binding sites of transcription factors identified by ChIP-seq assay. Instead of the ChIP-seq data, if the interactions between transcription factors and other genes are obtained from the correlation of mRNA expression levels in different cell and tissue types, the interactions can be used as the input data for the deep learning. The deep learning method would select important interactions to accurately predict gene expression levels by learning from the data. The selected interactions potentially include the same or similar interactions obtained from experimental assays such as ChIP-seq.

I replaced some of the input data for deep learning with the sites of the same number of different DNA/RNA-binding proteins based on co-expression data. Interestingly, the correlation coefficient between actual and predicted gene expression levels increased from 0.70 to 0.80 using co-expression data. The contribution scores of the input data selected from the gene expression correlation were higher than the original, and showed high positive or low negative contribution scores. Some functional annotations that were statistically overrepresented in the predicted target genes correspond to the known and associated functions of the DNA/RNA-binding proteins.

EVOLUTION OF MUC1 EXONIC VARIABLE NUMBER TANDEM REPEATS.

Petar Pajic[1], Bida Gu[2], Stacy Malaker[3], Mark J Chaisson[2], Omer Gokcumen[1]

[1]University at Buffalo, Biological Sciences, Buffalo, NY, [2]University of Southern California, Quantitative and Computational Biology, Los Angeles, CA, [3]Yale University, Chemistry, New Haven, CT

Exonic variable number tandem repeats (eVNTRs) in mucin genes directly influence their function by altering their encoded glycosylated domains. Mucin-1 (MUC1) exhibits remarkable eVNTR diversity, marked by extreme variation in repeat copy numbers among humans. However, the absence of haplotype-level sequence resolution and accurate repeat copy number estimation has limited evolutionary investigations. Using long-read sequencing data from over 700 human haplotypes across six major consortia, we comprehensively characterized MUC1 eVNTRs and corrected for PacBio-specific sequencing errors found in recent assembly releases. MUC1 eVNTR copy numbers ranged from 25 to 116 repeats across six continental populations. Population-level analysis revealed a higher frequency of low-copy MUC1 alleles (<50 repeats) in Asia, while medium-copy alleles (50–75 repeats) were more prevalent in Europe and Africa. Comparisons with long-read sequenced non-human primates indicated human-specific expansions of MUC1 eVNTRs. Sequence-level analysis revealed evidence of at least two mechanisms of repeat expansion: stepwise repeat gain and duplication of large repeat blocks, suggesting non-homologous recombination as a driver of rapid repeat acquisition. Bioinformatic prediction of glycosylation levels is significantly associated with additive repeats. These findings underscore the evolutionary significance of eVNTR dynamics in shaping functionally critical genes and highlight the importance of understanding mechanisms underlying extensive structural variation.

SPERM COMPETITION INTENSIFIES PURIFYING SELECTION ON SPERMATOGENESIS-RELEVANT GENES IN PRIMATES

Vasili Pankratov, Bjarke Meyer Pedersen, Mengjun Wu, Juraj Bergman, Mikkel Heide Schierup

Aarhus Univesity, Molecular Biology and Genetics, Aarhus, Denmark

Primates include more than 500 species that vary vastly in many aspects of their behavior, including mating patterns. Mating systems characterised by a single female mating with multiple males in a short period of time, namely polygynandry and polyandry, are associated with high levels of sperm competition. Sperm competition, in turn, is thought to result in strong selective pressures on traits that maximize sperm production, the classical example being the relative testes size. However, less is understood about how sperm competition affects molecular evolution.
Here, we address this question using data on genetic diversity within species generated by the Primate Genome Diversity Project. For this, we focused on ~520 genes associated with the 'spermatogenesis' GO term. We then calculated the ratio of nucleotide diversity at nonsynonymous and synonymous sites (piN/piS) in the concatenated sequence of those genes in each species. We adjusted it for the piN/piS in a concatenated sequence of ~19,000 genes. Next, we tested if the interspecies differences in this value can be explained by a) the mating system as a binary variable ('polygynandrous' vs 'other'; 155 species) or b) the relative testes mass (52 species) as predictors. The effects of the polygynandrous mating system in model a and of relative testes size in model b are -0.015 (95%CI: -0.023; -0.007) and -0.007 (95%CI: -0.014; -0.0014), respectively, suggesting lower piN/piS and hence stronger purifying selection on spermatogenesis-relevant genes in species with more intense sperm competition.

CHARACTERIZING CELL-TYPE-SPECIFIC ISOFORMS USING LONG-READ TRANSCRIPTOMICS TO ENHANCE RARE DISEASE VARIANT INTERPRETATION

Katherine L Pardo, David R Adams, May C Malicdan

National Institutes of Health, Undiagnosed Disease Program, Bethesda, MD

Prioritizing and interpreting non-coding variants remains a challenge in rare disease genomics. While genome sequencing is essential for identifying candidate variants, it is often limited in the prioritization of functional non-coding or coding variants. Transcriptomic analysis has emerged as a complementary approach to genomic analysis by capturing expression outliers, aberrant splicing isoforms, allele-specific expression, and transcriptomic structural variants. Transcriptomics can also be used to classify variants of unknown significance (VUS), reveal insights into disease mechanisms, and identify the second allele in a recessive disorder when genomic sequencing only detected one.

Current clinical transcriptomics studies use short-read RNA-sequencing, limiting the accuracy of computational reconstruction of full-length transcripts. Long-read sequencing offers key advantages by improving resolution of transcript isoforms, identifying novel splicing events and gene fusions, and stratifying allele-specific expression using haplotype phasing. Despite these advantages, one of the primary challenges in transcriptomic interpretation is the variability of gene expression and isoform variation across different tissues and conditions.

To address these challenges, we performed long-read RNA-sequencing using PacBio's HiFi real-time sequencing technology on cell lines commonly used in rare disease research, including fibroblasts, PBMCs, melanocytes, iPSCs, and iPSC-derived neurons. HeLa cells were also included as a widely used reference cell line. This study aims to address the key research questions: 1) what are the isoform profiles of these cell types, and 2) how are these isoform profiles affected by varying conditions?

We plan to characterize cell-type-specific isoform profiles using bioinformatics tools using orthogonal data to detect novel, lowly expressed, and rare isoforms with high sensitivity and precision. By generating a detailed catalog of cell-type-specific isoforms, including previously unannotated or rare transcripts, we aim to create a reference dataset for the scientific community. This dataset will provide isoform profiles used to compare patient-derived transcriptomic data to matched healthy controls, ultimately enhancing the interpretation and prioritization of non-coding variants in diagnostics. Integrating transcriptomics into genomic analysis has the potential to reveal pathogenic variation undetected by solely using genomic analysis, which can be transformative in variant reprioritization and classification, disease mechanism elucidation, and enhanced diagnostics.

ORIGINS AND IMPLICATIONS OF INTRON RETENTION QUANTITATIVE TRAIT LOCI IN HUMAN TISSUES

Eddie Park[1], Yi Xing[1,2]

[1]The Children's Hospital of Philadelphia, Center for Computational and Genomic Medicine, Philadelphia, PA, [2]The Children's Hospital of Philadelphia, Department of Pathology and Laboratory Medicine, Philadelphia, PA

Intron retention is a type of alternative splicing in which introns remain unspliced in mature RNA transcripts. In order to explore the landscape and consequences of genetically regulated intron retention, we perform an intron retention quantitative trait loci (irQTL) analysis in 49 human tissues across 838 individuals. We identify 8,624 unique intron retention events associated with genetic polymorphisms. 1,369 irQTLs (16%) are also associated with genome-wide association study (GWAS) traits. 1,999 irQTLs (23%) colocalize with eQTLs to their respective gene. We demonstrate that irQTLs are sufficient to generate eQTLs when one of the alternatively spliced transcripts is preferentially targeted by the nonsense mediated decay (NMD) pathway. Surprisingly, for intron retention events whose potential NMD effects can be confidently predicted based on their positions within known gene annotations, we find that 58.8% (923/1570) of the colocalized irQTL and eQTL pairs show effect size directions that are discordant with the NMD model. Moreover, we find that irQTLs are significantly more likely to occur in the same gene with the same effect size direction as compared to exon skipping QTLs. Through mathematical modeling and analysis of experimental perturbation data, we provide evidence that eQTLs are able to generate irQTLs by altering the steady state ratios of spliced and unspliced transcripts, and we postulate that this mechanism may partially underlie the widespread intron retention observed previously in various biological conditions. Taken together, these results show that intron retention and steady state gene expression levels are closely intertwined to regulate phenotypic traits.

DYNAMICS OF RPS24 ALTERNATIVE SPLICING IN BREAST CANCER AND THERAPEUTIC IMPLICATIONS

Jiyeon Park[1], Da Hae Nam[2], Seung-Hyun Jung[3,4], Yeun-Jun Chung[1,2,4]

[1]The Catholic University of Korea, College of Medicine, Precision Medicine Research Center, Seoul, South Korea, [2]The Catholic University of Korea, College of Medicine, Department of Microbiology, Seoul, South Korea, [3]The Catholic University of Korea, College of Medicine, Department of Biochemistry, Seoul, South Korea, [4]The Catholic University of Korea, College of Medicine, Integrated Research Center for Genomic Polymorphism, Seoul, South Korea

Alternative splicing (AS) is a fundamental mechanism that enriches the diversity of gene expression patterns. Through our AS database analysis, we identified significant variations in the AS of ribosomal protein S24 (RPS24) across breast cancer subtypes. While comprehensive analysis has been challenging due to a complex region containing three consecutive microexons (3bp, 18bp, and 22bp), we developed a specialized approach combining splice junction read analysis with fragment analysis to accurately quantify these isoforms. Our investigation revealed three key aspects of RPS24 AS regulation in breast cancer: First, we found distinct isoform compositions across cell lines, with the 3bp exon-containing isoform showing specifically high expression in estrogen receptor-positive cells and strong association with estrogen receptor signaling. Second, this isoform demonstrated dynamic regulation in response to therapeutic interventions, being upregulated following mTOR or CDK4/6 inhibition but significantly reduced in drug-resistant cells. Third, analysis of patient data showed that decreased expression of this isoform correlated with poor differentiation and metastatic progression. These findings establish RPS24 AS as a potential molecular marker in breast cancer with multiple clinical applications: monitoring pathway activity, predicting drug response, and assessing disease progression.

CHARACTERIZING COVERAGE BIASES IN LONG-READ DIRECT RNA SEQUENCING FOR IMPROVED ISOFORM QUANTIFICATION

Sowmya Parthiban[1], Casey Keuthan[3], Sheridan Cavalier[4], Winston Timp[2], Donald J Zack[3], Stephanie C Hicks[1,2]

[1]Johns Hopkins School of Public Health, Biostatistics, Baltimore, MD, [2]Johns Hopkins University, Biomedical Engineering, Baltimore, MD, [3]Johns Hopkins School of Medicine, Dept of ophthalmology, Wilmer Eye Institute, Baltimore, MD, [4]Johns Hopkins School of Medicine, Biochemistry, Cellular and Molecular Biology, Baltimore, MD

Short-read sequencing has been the predominant RNA-sequencing method for over a decade, but its reliance on short fragment lengths imposes inherent limitations. This constraint has hindered our understanding of key biological processes such as alternative splicing (AS), which affects over 95% of protein-coding genes and plays a crucial role in various cellular mechanisms. Dysregulation of AS is linked to complex diseases, including cancer and neurodegenerative disorders.

Long-read direct RNA sequencing, particularly using Oxford Nanopore Technology (ONT), theoretically enables full-length mRNA sequencing from the 3′ to 5′ end, providing more accurate gene isoform quantification. However, RNA fragmentation—occurring within the cell, during extraction, library preparation, or sequencing—often results in truncated reads. Consequently, coverage biases in long-read sequencing remain poorly characterized, leading to distorted read counts and inaccurate inferences about differential isoform expression and usage.

To address this, we systematically characterize these biases across both in-house and publicly available datasets. Our findings can guide downstream decisions on read filtering and retention, ultimately improving isoform quantification accuracy.

EVOLUTION OF TOXIN RESISTANCE IN THE GRASSHOPPER MOUSE

Claudia Perez-Calles[1,2], Ashlee Rowe[3], David Thybert[4], Jingtao Lilue[5], Elisabeth Anderson[6], David Adams[6], Thomas Keane[1]

[1]EMBL-EBI, Wellcome Genome Campus, Hinxton, United Kingdom, [2]University of Cambridge, Cambridge, United Kingdom, [3]University of Oklahoma, Department of Biology, Oklahoma, OK, [4]The Jackson Laboratory, Bar Harbor, ME, [5]Oujiang Laboratory, Wenzhou, China, [6]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Novel traits enable many rodents to thrive in extreme environmental niches. For example, predatory grasshopper mice have co-evolved resistance to painful and lethal neurotoxins produced by their scorpion prey. Interestingly, toxin resistance in the three species of grasshopper mice (*Onychomys torridus*, *Onychomys leucogaster*, and Onychomys arenicola) varies; *Onychomys torridus* exhibits the highest resistance and *Onychomys leucogaster* the least. Grasshopper mice also feed on pinacate beetles, whose toxic sprays irritate the eyes and nasal tissues of predators. It has been reported that the grasshopper mice have structural and functional modifications in the sodium channel Nav1.8 that decrease the effect of the toxins. This raises interesting questions: What are the molecular adaptations underlying toxin resistance in *Onychomys*? What genes contribute to pain and toxin response? Which sodium and potassium channels are involved?

To address these questions, we followed a comparative genomics approach. We produced the first high-quality reference genomes and annotations for *Onychomys* species and *Peromyscus eremicus* (a closely related outgroup). We implemented a comprehensive pipeline to detect positive selection across genome-scale datasets and identified a promising candidate gene for pain resistance in *Onychomys*: Nav1.3 (*Scn3a*). This sodium channel gene is expressed in the central nervous system and plays an important role in nociceptive signalling. Additionally, we detected an *Onychomys*-specific tandem gene duplication of the *Cblif* gene, which encodes a glycoprotein crucial for vitamin B12 absorption. This adaptation likely supports the species' dietary specialisation and modified stomach morphology, where parietal cells expressing *Cblif* are especially numerous. To better understand the molecular response to toxin exposure, we generated RNA-Seq data from the dorsal root ganglion and the trigeminal ganglion from mice that have been exposed to toxic sprays from pinacate beetles. We observed the upregulation of genes associated with transcriptional processes that can be linked to cellular repair mechanisms, and also identified the regulation of genes involved in diverse functions.

In conclusion, our findings provide a key step for establishing the *Onychomys* species as a model system for studying toxin resistance, pain response, and behavioural traits.

EARLY-LIFE ADVERSITY PREDICTS CROSS-TISSUE DNA METHYLATION PATTERNS ASSOCIATED WITH AGE IN RHESUS MACAQUES

Rachel M Petersen[1], Baptiste Sadoughi[2], Sam K Patterson[3], Angelina V Ruiz-Lambides[4], Michael J Montague[5], Michael L Platt[5,6], James P Higham[3], Lauren J Brent[7], Noah Snyder-Mackler[2], Amanda J Lea[1]

[1]Vanderbilt University, Department of Biological Sciences, Nashville, TN, [2]Arizona State University, Center for Evolution and Medicine, Tempe, AZ, [3]New York University, Department of Anthropology, New York, NY, [4]University of Puerto Rico, Caribbean Primate Research Center, Punta Santiago, PR, [5]University of Pennsylvania, Department of Neuroscience, Philadelphia, PA, [6]University of Pennsylvania, Department of Psychology, Philadelphia, PA, [7]University of Exeter, Centre for Research in Animal Behaviour, Exeter, United Kingdom

Adversity during early life is associated with poor adult health and increased mortality risk in humans and non-human species. The mechanisms linking early life adversity (ELA) to adult health remain unclear, but epigenetic modifications, including DNA methylation (DNAm), are thought to play a role. While previous studies have shown how ELA shapes DNAm in blood, its impact on other organ systems is poorly understood despite its broad implications for all-cause mortality. Aging is also linked to DNAm changes, which serve as predictors of aging-related diseases and mortality. We hypothesized that ELA would lead to widespread DNAm variation across tissues and that its influence on mortality may stem from its tendency to mirror or amplify DNAm changes seen with aging. To test this hypothesis, we use a free-ranging non-human primate model, the rhesus macaques (Macaca mulatta) of Cayo Santiago, to map ELA-associated DNAm across 12 tissues from 190 individuals aged 3-26 years (n=2,003 samples). We tested 1.15M CpG sites across 7 established sources of ELA and identified 411,458 unique CpG sites associated with adversity (285,773- 386,383 per tissue). Most ELA-associated sites were linked to a single type of adversity (69.1%) and exhibited similar methylation changes (in both magnitude and direction) across multiple tissues. ELA-linked sites shared across all 12 tissues were enriched for less active genomic regions, such as heterochromatin (OR= 1.1-1.3, $p<0.001$), while those shared across subsets of tissues were enriched for more active regions, including promoters (OR=1.7-3.2, $p<0.001$) and enhancers (OR=1.1-1.3, $p<0.001$), suggesting that ELA may have functional impacts in specific tissue subsets. For example, we found greater ELA susceptibility (larger effect sizes) in tissues with long-lived cell types and in immune and endocrine tissues ($p< 0.001$ for all). Lastly, we found that ELA exposure generally recapitulated age-related DNAm patterns. Sites that showed higher methylation in older individuals were also more methylated in those exposed to adversity. Interestingly, this trend was reversed in immune tissues, suggesting that ELA may have tissue-specific effects on biological aging, potentially influencing immune function in distinct ways. This study is one of the most comprehensive investigations into how ELA shapes the epigenome across tissues and organ systems, providing new insights into the mechanisms that contribute to early-life environmental sensitivities and their long-term health impacts.

PER-NUCLEOTIDE SOMATIC MUTATION MODELLING REVEALS STRONG PATIENT-SPECIFIC SEQUENCE PREFERENCE OF MUTAGENESIS

Mario Aguilar-Herrador, Yana Vassileva, Jessica do Amaral Andrade, Melissa Sanabria, Anna R Poetsch

TU Dresden, Biotechnology Center, Dresden, Germany

DNA sequence preferences of somatic mutagenesis are dependent on the underlying epigenome and sequence content. Also, transcription and replication, which themselves have a strong relationship with the DNA sequence, determine sequence susceptibility to damage, repair preferences, and mutation. Together, these processes create a tissue-specific and individual balance between functionality and stability for somatic genomes. We aim to understand this balance through DNA-sequence-based deep learning, so called DNA language models.
EAGLE-MUT (Efficient Analysis with a Genome-wide LSTM to Evaluate per-nucleotide MUTation susceptibility) learns sequence context of single base substitutions in individual tumor samples. Applied to 423 samples of esophagus adenocarcinoma, the model shows that the mutation probability distribution over the genome is very heterogenous and individually different. EAGLE-MUT performs with up to 90-fold prediction over a random model. We derive sequence patterns for mutation cold- and hotspots, which reveal asymmetric motifs of up to 200 bp. Interestingly, the hotspot motifs from stomach-acid related mutagenesis resemble motifs of nucleosome occupancy and indeed show association with nucleosome binding. Although revealed via a question on mutagenesis, these observations may have substantial general impact for genome biology. They suggest i) a surprisingly specific sequence preference for nucleosome binding in the human genome, ii) that these locations have an inherent deficiency to resolve mismatches with gastric acid-associated oxidative DNA damage; iii) that these mutations, while currently being considered of little consequence, could lead to a disruption of chromatin organisation. Mutation cold-spots associate with promoters, CTCF binding sites, potential origins of replication, intron-exon boundaries, and nucleosome linker DNA. Interestingly, also cold-spots are associated with clear sequence motifs spreading several 100 bp which seem to serve the purpose of increasing stability of sequence with above-mentioned functionality. We also applied EAGLE-MUT to locations of cancer driver genes. Their recurrences can partially be explained by local differences of mutagenesis probabilities that span several orders of magnitude. Therefore, in addition to processes of evolutionary selection, also local genome instabilities affect the emergence of drivers for somatic genome evolution.
Models like EAGLE-MUT overcome the sparse nature of somatic genome instability data and can be used in an interpretable way to discover relationships between DNA sequence, genome function, and stability.

CELL-TYPE-RESOLVED CHROMATIN ACCESSIBILITY IN THE HUMAN INTESTINE IDENTIFIES COMPLEX REGULATORY PROGRAMS AND CLARIFIES GENETIC ASSOCIATIONS IN CROHN'S DISEASE

Yu Zhao*[1], Ran Zhou*[2], Zepeng Mu[3], Peter Carbonetto[4], Xiaoyuan Zhong[4], Bingqing Xie[2], Kaixuan Luo[4], Candace M Cham[2], Jason Koval[2], Xin He[4], Andrew W Dahl[2], Xuanyao Liu[2], Eugene B Chang[2], Anindita Basu[2], Sebastian Pott[2]

[1]University of Chicago, Pritzker School of Molecular Engineering, Chicago, IL, [2]University of Chicago, Department of Medicine, Chicago, IL, [3]Brigham and Women's Hospital, Harvard Medical School, Center for Data Sciences, Boston, MA, [4]University of Chicago, Department of Human Genetics, Chicago, IL

*Autors contributed equally

Crohn's disease (CD) is a complex inflammatory bowel disease resulting from an interplay of genetic, microbial, and environmental factors. Cell-type-specific contributions to CD etiology and genetic risk are incompletely understood. Here we built a comprehensive atlas of cell-type-resolved chromatin accessibility including biopsies from terminal ileum or ascending colon from 23 patients with active and inactive CD and 16 healthy controls. This atlas comprises a total of 178,030 cells, capturing 29 cell types and identifying 557,310 candidate cis-regulatory elements (cCREs). Using these data, we identified cell-type-, anatomic location-, and context-specific cCREs and characterized the regulatory programs underlying inflammatory responses in the intestinal mucosa of CD patients. Genetic variants that disrupt binding motifs of cell-type-specific transcription factors significantly affected chromatin accessibility in specific mucosal cell types. We found that CD heritability is primarily enriched in immune cell types. However, using fine-mapped non-coding CD variants we identified 29 variants located within cCREs several of which were accessible in epithelial and stromal cells implicating cell types from additional lineages in mediating CD risk in some loci. Our atlas provides a comprehensive resource to study gene regulatory effects in CD and health, and highlights the cellular complexity underlying CD risk.

REFERENCE GENOMES AND CONSERVATION APPLICATIONS FOR EMBLEMATIC AND ENDANGERED ECUADORIAN SPECIES

Gabriela Pozo[1,2], Martina Albuja-Quintana[1], Maria de Lourdes Torres[1,2]

[1]Laboratorio de Biotecnología Vegetal, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito (USFQ), Quito, Ecuador, [2]Instituto Nacional de Biodiversidad, (INABIO), Quito, Ecuador

Genome sequencing has become increasingly common, providing valuable insights into species' genetic makeup and fitness. The advent of next-generation sequencing has made this technology more accessible in terms of costs and speed. However, the majority of the world's biodiversity is found in developing countries, where historically access to genomic technologies has been limited.

Our research aims to obtain high-quality reference genomes of emblematic and endangered species in Ecuador, one of the most biodiverse countries in the world. As part of this initiative, we have successfully sequenced the genomes of three critically endangered or endangered primates: the brown-headed spider monkey (*Ateles fusciceps fusciceps*), the white-bellied spider monkey (*Ateles belzebuth*), and the Ecuadorian capuchin (*Cebus aequatorialis*). These reference genomes provide crucial insights into the evolutionary history and genetic makeup of these species.

Ongoing research includes a comparative analysis of the *AMY1* gene, which is linked to saliva amylase production, across the three primates. This study will specifically focus on variations in *AMY1* gene copy number and explore how these differences may be associated with the primates' habitats and the impacts of habitat fragmentation and loss. For the brown-headed spider monkey, we are conducting a comprehensive population genomics study to assess its genetic health. This study investigates genetic diversity, population structure, and potential indicators of inbreeding, offering critical data for conservation strategies.

Additionally, we have expanded our genomic efforts to include other endangered species, such as Orcés' Blue Whiptail (*Holcosus orcesi*) and an endemic plant from the Galapagos Islands (*Scalesia gordilloi*). By providing high-quality genomic resources for these species, we aim to empower researchers and conservationists with the tools needed to better understand population genetics, inform conservation management, and mitigate biodiversity loss in Ecuador.

HARNESSING DRUG-INDUCED GENE EXPRESSION CHANGES FOR IMPROVED DRUG RESPONSE PREDICTION

Henry W Raeder[1], Hae Kyung Im[2]

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

Each year, billions are spent on clinical trials that ultimately fail. The ability to predict drug response *a priori* would be transformative, as it could allow for streamlining of the early stages of the clinical trial process. Recent advances in genomic and computational tools provide new opportunities for accurately predicting drug response.

Previous studies have demonstrated that gene expression is a powerful predictor of drug response. As such, many existing methods rely on baseline gene expression profiles. However, an untapped opportunity exists in leveraging differential expression (DE) following drug exposure. These measurements integrate the drug's perturbation effect directly, and thus will likely have more predictive power.

Several large-scale repositories now provide extensive datasets of drug-induced DE. However, these do not generally have matched phenotype data on which further predictions can be run (i.e. measures of cell survival). To bridge this gap, we propose a framework that employs AI tools trained on the aforementioned DE databases. These models can take chemical structure as input, and use it to predict drug-induced DE which we can then apply to other phenotype-based drug response datasets.

Initial results indicate that one can make informative predictions of cytotoxicity metrics (specifically EC50 and area under the dose-respone curve (AUC)) using predicted DE. After training a predictive tool (ChemCPA) on the CMap L1000 DE dataset, we generated post-perturbation gene expression predictions for over 18,000 drug/cancer cell line combinations from the CTRPv2 database. Elastic Net linear regression, as well as nonlinear modeling through XGBoost, indicates that altered gene expression can explain 32% and 51% of the variance in AUC respectively.

By incorporating predicted drug-induced DE, our framework could offer a more powerful approach to drug response prediction. This strategy could also be expanded into patient-derived samples, enhancing its translatability for real-world therapeutic applications. Ultimately, if proven out, this approach has the potential for broad applications in preclinical drug prioritization, precision oncology, and personalized medicine.

LIVING FOSSILS: LEVERAGING SINGLE-MOLECULE SEQUENCING TO DECODE THE COMPLEX GENOMES OF ANCIENT PLANT LINEAGES

Srividya Ramakrishnan[1], Dennis Stevenson[2], Cristiane de Santis Alves[3], Veronica M Sondervan[2,6], Melissa Kramer[3], Sara Goodwin[3], Shujun Ou[12], Cecilia Zumajo-Cardona[2], Laís Araujo Coelho[5], Samantha Frangos[3], Katherine Jenike[1], Olivia Mendevil Ramos[3], Gil Eshel[6], Xiaojin Wang[7], Maurizio Rossetto[8], Hannah McPherson[9], Sebastiano Nigris[10], Silvia Moschin[10], Damon P Little[2], Manpreet S Katara[6], Kranthi Varala[7], Sergios-Orestis Kolokotronis[5], Barbara Ambrose[2], Larry J Croft[11], Gloria M Coruzzi[3], Michael C Schatz[1], Robert A Martienssen[3,4], Richard McCombie[3]

[1]Johns Hopkins University, Baltimore, MD, [2]The New York Botanical Garden, Bronx, NY, [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [4]Howard Hughes Medical Institute, Cold Spring Harbor, NY, [5]SUNY Downstate Health Sciences University, Brooklyn, NY, [6]New York University, New York, NY, [7]Purdue University, West Lafayette, IN, [8]Royal Botanic Garden Sydney, Sydney, Australia, [9]Australian Botanic Garden, Mount Annan, Australia, [10]Università degli studi di Padova, Padova, Italy, [11]Deakin University, Victoria, Australia, [12]The Ohio State University, Columbus, OH

Living fossils are ancient species that have endured millions of years of climate change with little to no morphological change, offering insights into genomic resilience and evolution. Gymnosperm lineages include a mix of living fossils such as Wollemia nobilis and recently radiated species like Gnetum gnemon, Araucaria angustifolia, and Juniperus communis, many exhibiting distinctive traits shaped by their evolutionary histories and adaptations to their respective ecological niches. Emerging as early as the Devonian, these species have withstood diverse environmental challenges, yet their massive genomes—up to 20 times larger than those of angiosperms—remain poorly understood due to transposable elements (TEs), which constitute over 70% of their content and have hindered short-read sequencing efforts.

Long-read sequencing now enables us to resolve these complex genomes, providing opportunities to explore gymnosperm evolution, adaptation, and genome architecture. Using long-read sequencing technologies, we generate high-quality assemblies for comparative analyses and identification of resilience signatures. Functional genomics and population resequencing reveal gene gains, losses, and expression shifts underlying resilience and adaptation, including the silencing and activation of defense genes in response to environmental stresses. We also investigate TE-driven genome dynamics, uncovering small RNA-mediated silencing and DNA methylation mechanisms that regulate TE activity. Notably, retrotransposon bursts coinciding with population declines suggest an adaptive role in enhancing epigenetic diversity. This study provides the first comprehensive genomic analysis of gymnosperm living fossils, shedding light on the genetic basis of evolutionary stasis.

PHYLOGENETIC PATTERNS OF CONTEXT-SPECIFIC MUTATION SPECTRA ACROSS 113 EUKARYOTES

Fabian Ramos-Almodovar[1], Ziyue Gao[1], Benjamin F Voight[1,2], Iain Mathieson[1]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA, [2]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA

Mutation spectra vary across genetic and environmental contexts, leading to differences between and within species. Most research has focused on the trinucleotide mutation spectrum in mammals, limiting the breadth and depth of the variation surveyed. Here, we leverage whole-genome population resequencing data from 113 eukaryotic species–including mammals, fish, plants, and invertebrates–to investigate variation in context-specific mutation spectra and underlying mechanisms in a wide phylogenetic context. We apply Baymer, a Bayesian hierarchical tree approach with context regularization, to non-coding polymorphisms in each species to generate 5-mer context-specific mutation spectra. Despite technical variation in variant calling, our inferred mutation spectra cluster by phylogenetic clade and biological factors instead of technical factors, attesting to the reliability of the results.

We find that variation in cytosine mutability at CpG and CHG sites accounts for 89.29% of differences in 5-mer mutation spectra across eukaryotes. Moreover, the inferred CpG transition rate strongly correlates with genomic CpG depletion. This observation provides support for our inferred mutation spectra, suggesting that the mutation landscape determines genomic CpG content. While methylation is necessary for higher CpG mutability, genome-wide CpG methylation levels surprisingly do not explain differences in CpG mutability across species. We find a similar result in plants, where mutation rates at CHG sites are elevated, due to cytosine methylation in those contexts, but CHG methylation levels do not predict the relative rate of CHG transitions across species. Although variation in mutation spectrum across species is strongly associated with the presence and sequence targets of DNA methylation, these results suggest that relative mutation rates at methylated cytosines are not determined by genome-wide methylation levels. Instead, we posit that they depend on unknown genetic or environmental factors that modify deamination or repair rates.

Finally, in contrast to previous reports, we find that variation in mutation spectra across species can be modeled as a function of three processes: cytosine methylation in CpG context, in CHG contexts, and a relatively constant background mutational process. Together, our analyses demonstrate a relatively simple architecture for variation in the context-specific mutation spectrum across eukaryotes but highlight a gap in our knowledge about what factors drive variation in methyl-cytosine mutation rates.

DISSECTING THE MULTI-OMIC RISK FACTORS FOR DELIRIUM

Vasilis Raptis[1,2], Youngjune Bhak[1], Tim Cannings[4], Alasdair Maclullich [5], Albert Tenesa[1,3]

[1]University of Edinburgh, Roslin Institute, Edinburgh, United Kingdom, [2]University of Edinburgh, Advanced Care Research Centre (ACRC) Academy, Edinburgh, United Kingdom, [3]University of Edinburgh, MRC Human Genetics Unit, Edinburgh, United Kingdom, [4]University of Edinburgh, School of Mathematics, Edinburgh, United Kingdom, [5]University of Edinburgh, Usher Institute, Edinburgh, United Kingdom

Delirium is a complex neurocognitive condition, characterised by an acute, but usually reversible, deterioration of the patient's cognitive ability, attention and awareness. Being a common complication in hospitalised older adults, delirium has been associated with high healthcare and human cost worldwide. In this work we shed light into the currently poorly understood multi-omic background of delirium, focusing on genomic, proteomic and metabolomic risk factors.

We conducted the largest to date multi-ancestry analysis of genetic variants associated with delirium (1,059,130 individuals, 11,931 cases), yielding the *Apolipoprotein E* (*APOE*) gene as a strong risk factor with possible population and age-varying effects. A multi-trait analysis of delirium with Alzheimer disease identified 5 delirium genetic risk loci. Investigation of plasma proteins associated with up to 16-years incident delirium (32,652 individuals, 541 cases) revealed known and novel protein biomarkers, implicating brain vulnerability, inflammation and immune response processes. Integrating proteins and *APOE* genetic risk with demographics significantly improved incident delirium prediction compared to demographics alone. Finally, delirium risk was associated with plasma metabolomic signatures of several disease-relevant gut microbiota.

Our results pave the way to better understanding delirium's complex aetiology and guiding further research on clinically relevant biomarkers.

GENOME-WIDE PERTURBATIONS LINK AUTOIMMUNE GENETIC RISK TO PRIMARY T CELL EXPRESSION AND FUNCTION

Ching-Huang Ho[1,2], Maxwell A Dippel[1], Meghan S McQuade[1], LeAnn P Nguyen[1], Arpit Mishra[2], Stephan Pribitzer[1], Samantha Hardy[3], Harshpreet Chandok[4], Florence Chardon[5,6], Troy A McDiarmid[5], Hannah A DeBerg[1], Jane H Buckner[2], Jay Shendure[5,6,7], Carl G de Boer[8], Michael H Guo[9], Ryan Tewhey[4], John P Ray[1,5,10]

[1]Benaroya Research Institute, Systems Immunology, Seattle, WA, [2]Benaroya Research Institute, Translational Immunology, Seattle, WA, [3]University of Washington, Molecular and Cellular Biology program, Seattle, WA, [4]The Jackson Laboratory, -, Bar Harbor, ME, [5]University of Washington, Genome Sciences, Seattle, WA, [6]Seattle Hub for Synthetic Biology, -, Seattle, WA, [7]Howard Hughes Medical Institute, -, Seattle, WA, [8]University of British Columbia, Biomedical Engineering, Seattle, WA, [9]University of Pennsylvania, Neurology, Seattle, WA, [10]University of Washington, Immunology, Seattle, WA

~8 percent of the US population have an autoimmune disease, many of which are debilitating and cause early mortality and put an enormous financial burden on the US health system. Most treatments for autoimmune diseases target major inflammatory pathways or entire immune cell populations, which leads to many undesired side-effects. If we better understood the mechanisms that drive disease, we could create therapeutics that are more targeted to desired pathways and with less side effects. Genome-wide association studies have identified hundreds of autoimmune disease-associated genomic regions, but defining the precise genetic variants that cause disease has been exceptionally difficult due to tight linkage disequilibrium between causal and non-causal variants and because >90% are found in non-coding regions where their effect on gene expression and disease-relevant cellular function is difficult to define. To address this, we recently overcame technological hurdles in assaying non-coding variants in primary human T cells. Through testing ~18,000 autoimmune GWAS variants for allele-specific effects on cis-regulatory element (CRE) activities with massively parallel reporter assays (MPRAs), we identified 545 expression-modulating variants (emVars) according to differences in reporter expression between variant alleles. emVars in T cell accessible chromatin enriched 122-fold for statistically fine-mapped variants with high posterior inclusion probabilities. emVars had putative target genes that enriched within gene networks linked to lymphocyte activation, translation, mRNA processing and splicing, and transcriptional regulation. We linked 41 emVars in cis-regulatory elements to the genes they regulate using single-cell CRISPR-interference screens, identifying genes that participate in T cell signaling, activation, and development. Using bulk CRISPR screens, we find 14 emVar CREs within the CD28, IL2RA, OX40, and other loci that significantly positively regulate T cell proliferation. Furthermore, we identified PPP5C as a novel regulator of effector T cell signaling and metabolic programs. Thus, using MPRA and CRISPRi screens in primary T cells, we for the first time systematically identified likely autoimmune-causal non-coding variants and their effects on human primary T cell expression networks and function, allowing us to assign plausible mechanisms of genetic risk to loci.

LONG-READ TRANSCRIPTOMICS OF A DIVERSE HUMAN COHORT REVEALS WIDESPREAD ANCESTRY BIAS IN GENE ANNOTATIONS.

Fairlie Reese*[1], Pau Clavell-Revelles*[1,2], Sílvia Carbonell-Sala[3], Fabien Degalez[3], Winona Oliveros[1,2], Carme Arnan[3], Roderic Guigó[3,4], Marta Melé[1]

[1]Barcelona Supercomputing Center, Life Sciences, Barcelona, Spain, [2]Universitat de Barcelona, Barcelona, Spain, [3]The Barcelona Institute of Science and Technology, Centre for Genomic Regulation, Barcelona, Spain, [4]Universitat Pompeu Fabra, Departament de Ciències i de la Salut, Barcelona, Spain

Background: Gene annotations are essential for interpreting biological findings in genetics and genomics. Current human genome annotations are primarily based on transcriptomic data from European-descent individuals. The extent to which these annotations are representative of all human populations at the full-length transcript level is unknown. Recent human annotations are built on GRCh38, which might further bias gene annotation efforts due to its inability to represent haplotypes or genetic diversity.
Methods: We present the first population-diverse long-read RNA-seq (LR-RNA-seq) dataset with lymphoblastoid cells from 43 individuals from 8 genetically-distinct different human populations using >600 million full-length reads, which we used to build a population-diverse gene annotation with >40k novel transcripts.
Results: Current gene annotations better represent transcriptomes from Europeans than non-Europeans. Novel transcripts are more often discovered in non-European populations. Transcripts exclusively discovered in non-Europeans are more commonly novel. Using our population-diverse gene annotation significantly increases the potential for discovery of allele-specific transcript usage, especially in non-Europeans, despite containing half as many transcripts as recent annotations. We quantify the effects of genome assembly choice on LR-RNA-seq transcript discovery using personal genomes and find that large-scale genomic variation harbors little novel transcription while local variation marginally impacts our ability to call putative splicing events.
Conclusion: Overall, current gene annotations are European-biased and negatively impact downstream analyses. Personal genomes in isolation offer few benefits over GRCh38 for transcript discovery compared to the relative gain in transcript diversity afforded by sequencing more diverse populations. Our study offers new insights into how ancestry-specific genetic variation shapes transcriptome diversity and demonstrates the importance of inclusive reference annotations for advancing our understanding of interindividual variation and health. We emphasize the pressing need for a gene annotation that better represents the diversity of the human transcriptome and for the improvement of pangenome tools that allow for analysis and interpretation of LR-RNA-seq.

SWEEPS IN SPACE: LEVERAGING GEOGRAPHIC DATA TO IDENTIFY BENEFICIAL ALLELES IN *ANOPHELES GAMBIAe*

Clara T Rehmann[1,2], Scott T Small[1,2], Peter L Ralph[1,3], Andrew D Kern[1,2]

[1]University of Oregon, Institute of Ecology and Evolution, Eugene, OR, [2]University of Oregon, Department of Biology, Eugene, OR, [3]University of Oregon, Department of Data Science, Eugene, OR

As organisms adapt to environmental changes, natural selection modifies the frequency of non-neutral alleles. For beneficial mutations, the outcome of this process may be a selective sweep, in which an allele rapidly increases in frequency and perhaps reaches fixation within a population. Selective sweeps have well-studied effects on patterns of local genetic variation in panmictic populations, but much less is known about the dynamics of sweeps in continuous space. In particular, because limited movement across a landscape leads to unique patterns of population structure, spatial dynamics may influence the trajectory of selected mutations. Here, we use forward-in-time, individual-based simulations in continuous space to study the impact of space on beneficial mutations as they sweep through a population. In particular, we show that selection changes the joint distribution of allele frequency and geographic range occupied by a focal allele and demonstrate that this signal can be used to identify selective sweeps. We then leverage this signal to identify in-progress selective sweeps within the malaria vector *Anopheles gambiae*, a species under strong selection pressure from vector control measures. By considering space, we identify multiple previously undescribed variants with potential phenotypic consequences, including mutations impacting known insecticide resistance-associated genes and altering protein structure and properties. Our results demonstrate a novel signal for detecting selection in spatial population genetic data that may have implications for genomic surveillance and understanding geographic patterns of genetic variation.

WORKFLOW FOR POLYGENIC SCORE ANALYSIS AND VISUALIZATION FROM SINGLE-SAMPLE WHOLE GENOME SEQUENCING VCF DATA

Raimonds Reščenko-Krūms

University of Latvia, Department of Biology, Riga, Latvia

Currently, over 29 million people globally have access to their raw genetic data, and as the sequencing costs continue to decline, this number is expected to increase. Accurate computation and visualization of this data is a crucial intermediate step in the practical application of genetic information to determine individual polygenic risk scores (PGS). However, existing tools are focused on the computation of multi-sample aggregates, lack understanding of model limits and do not address common whole genome sequencing (WGS) derived data formats such as GVCF. Additionally, result visualization in an user friendly way has been explored but has not been broadly implemented in a web compatible manner. Here we use the nextflow workflow management system and nf-core community framework for open-source development and build a modular workflow to compute and visualize PGS from single-sample whole genome sequencing VCF data using PGS models from PGS-catalog. Additionally, we investigate the internally validated PGS model and characterize its limitations for correct interpretation by the user. Finally, we use chartJS javascript library to visualize individual PGS percentile in a geographically matched population.

UNIFIED META REGRESSION MODEL FOR RARE VARIANT ASSOCIATION STUDIES (RVAS): MISSENSE PATHOGENICITY, CONSTRAINT, AND LOSS-OF-FUNCTION

Manuel A Rivas[1], Larissa Lauer[2]

[1]Stanford University, Biomedical Data Science, Stanford, CA, [2]Stanford University, Statistics, Stanford, CA

Background: Rare variant association studies (RVAS) of complex traits have emerged as a powerful approach for advancing drug discovery and diagnostics. Genebass, a publicly available online resource of exome-based association statistics gathered from 394,841 exome samples from UK Biobank provides valuable data for an RVAS study.
Missense pathogenicity predictions from AI AlphaMissense model based on protein language and structural context enable the differentiation between benign and deleterious variants. Constraint metrics allow researchers to identify regions of the genome under selective pressure, which are more likely to harbor mutations that affect genome function.
Loss-of-function (LoF) variants, which result in the complete or partial loss of protein function, are particularly informative is it is more straightforward to assess their downstream functional consequences.
Methods: We present a unified meta regression model approach that incorporates the probability of pathogenicity, probability of constraint, and indicators for whether a variant is a predicted loss-of-function or a missense variant as features to predict the observed effect size and uncertainty of effect size obtained from single variant genetic analysis. We applied the model to 1,144 continuous phenotypes from UK Biobank using single variant summary statistics obtained from Genebass. We replicated discoveries using the AllofUS cohort. For each gene discovery we make available a characterization of whether constrained sites are associated with the phenotype, along with a diagram of constraint in the gene, whether pathogenic sites determined by structural based predictions are associated with phenotype.
Results: When applied to the phenotype for epilepsy, our unified meta regression model identified a set of ($p < 1\times10^{-4}$) genes with stronger signals than previously published studies including *KDM5B*, *KCNQ2*, *CACNA1A*, *CACNA1B*, *RYR2*, and *ATP2B2*. When applied to the phenotype for pulse rate, we identified a gene *CASZ1* ($p < 1 \times 10^{-12}$) which had not been identified in previous studies. We will also present findings across the thousands of phenotypes in UK Biobank where we can partition variance explained of rare variant association signals into: i) loss-of-function, ii) structure based predictions, iii) constraint, and iv) broad missense category.
Results: We showed that through a meta regression model which unifies missense pathogenicity, constraint, loss-of-function predictions, and missense variants, we can observe new rare variant associations between genes and phenotypes, using publicly available UK Biobank data.

ADMIXTURE DYNAMICS OF A HYBRID BABOON POPULATION REVEALED BY NEAR-T2T ASSEMBLIES

Iker Rivas-González[1], Moisès Coll Macià[2], Mikkel H Schierup[2], Asger Hobolth[3], Susan C Alberts[4], Elizabeth A Archie[5], Jeffrey Rogers[6], Karen Miga[7], Jenny Tung[1,4]

[1]Max Planck Institute for Evolutionary Anthropology, Department of Primate Behavior and Evolution, Leipzig, Germany, [2]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark, [3]Aarhus University, Department of Mathematics, Aarhus, Denmark, [4]Duke University, Departments of Evolutionary Anthropology and Biology, Durham, NC, [5]University of Notre Dame, Department of Biological Sciences, Notre Dame, IN, [6]Baylor College of Medicine, Human Genome Sequencing Center and Department of Molecular and Human Genetics, Houston, TX, [7]University of California, Santa Cruz, UC Santa Cruz Genomics Institute and Department of Biomolecular Engineering, Santa Cruz, CA

Speciation is a complex process that often involves gene flow after the initial split, coupled with natural selection that influences whether introgressed genetic material is retained or purged. Natural hybrid zones provide excellent opportunities to understand these selective dynamics in actively mixing populations. However, computing accurate introgression maps requires high-quality, ideally haplotype-resolved genomes, as well as methods to differentiate introgression from other evolutionary processes, especially incomplete lineage sorting (ILS). To overcome these limitations, here we report a novel algorithm, iTRAILS, that jointly estimates key demographic parameters and locus-specific multi-species ancestral recombination graphs to distinguish ILS from introgression at base-pair resolution. I apply it to four newly-generated near-telomere-to-telomere (T2T) assemblies for baboons, a radiation often presented as a living model for admixture in our own lineage. These new assemblies include two yellow baboons (*Papio cynocephalus*), one anubis baboon (*P. anubis*), and one natural hybrid from a well-studied hybrid zone in Amboseli, Kenya, where previous work supports selection against gene flow. Our results offer novel estimates of migration rates, migration times, speciation times, and ancestral effective population sizes for the Amboseli hybrid zone, providing an unprecedentedly detailed reconstruction of the evolutionary history of this population. We then demonstrate how the inferred ILS/introgression map can help identify regions subject to natural selection during the hybridization process. Finally, we investigate how such analyses can be extended to include resequencing data from hundreds of individuals in the Amboseli population to explore hybridization and selection at the population level. Combined, these approaches hold substantial promise for enriching our understanding of hybridization in natural populations, including in our closest living relatives.

ENABLING LARGE-SCALE INTERPRETATION OF GENOMIC FOUNDATION MODELS THROUGH KNOWLEDGE DISTILLATION

Kaeli Rizzo, Jessica Zhou, Peter Koo

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

Deep learning has given rise to a new era of biological modeling, allowing researchers to make significant strides in predicting how genetic information translates to gene expression. Current state-of-the-art models like Enformer and Borzoi employ complex attention mechanisms to model regulatory interactions, demonstrating the potential of deep learning to capture biological patterns from sequence data. While these models offer impressive predictive power, using them for downstream applications – such as variant effect prediction, sequence design, and in silico experiments – often demands millions of predictions and creates a significant computational bottleneck. This is pivotal to address as understanding how genetic variation influences gene regulation is crucial for linking genotype to phenotype and advancing disease research. Further, these models provide predictions without indicating their reliability – a critical limitation when using these predictions to guide experimental studies. To address these challenges, we investigate knowledge distillation as a means to capture the learned representations of large models in more compact, efficient architectures. First, we demonstrate this method to predict chromatin accessibility profile data, leveraging a combination of knowledge distillation with ensemble learning. Our distilled models achieve comparable performance while showing more consistent interpretations of regulatory sequences and providing estimates of experimental variation and model uncertainty - crucial information for assessing prediction reliability. Building on these results, we extend our framework to current state-of-the-art genomic models, developing a systematic approach for creating more practical versions that maintain their biological insights. Our approach leverages optimizations like flash attention to dramatically reduce GPU memory requirements and inference time, all while providing uncertainty estimates and robust attribution analyses. This work makes powerful genomic deep learning models both more accessible and more reliable for the broader research community, providing a foundation for systematically investigating cis-regulatory mechanisms genome-wide.

DECODING A COMPLETE GENOMIC REPOSITORY OF NORTH AMERICAN CAPTIVE MARMOSETS: INSIGHTS INTO RECENT POPULATION HISTORY AND BIOLOGY

Murillo F Rodrigues[1], Philberta Leung[1], Alexandra Stendahl[1], Jenna Castro[1], Ricardo del Rosario[2], Joanna Malukiewicz[4], Jamie A Ivy[3], Jeff D Wall[1], Don F Conrad[1]

[1]Oregon National Primate Center, OHSU, Genetics, Beaverton, OR, [2]Broad Institute of Harvard and MIT, Genetics, Cambridge, MA, [3]Independent Consultant, LLC, Erie, CO, [4]University of Hamburg, Institute of Animal Cell and Systems Biology, Hamburg, Germany

The common marmoset (*Callithrix jacchus*) is an emerging non-human primate model for biomedical research. There are currently about 2,500 captive marmosets in the United States, but little is known about their genomic diversity, and population structure and history. As part of the Marmoset Coordinating Center, we have identified, registered, and sampled most marmosets in the U.S. research population. Here, we present a repository of over 800 genomes, generated from hair follicle DNA. Using 140 high coverage genomes, we show that individuals with low coverage (<5x) can be accurately imputed. We found there is moderate population structure across colonies, with the highest genetic differentiation (F_{ST}) reaching ~0.2. We observe a mean heterozygosity rate of 0.15%, which is comparable to other primates such as rhesus macaques. However, there is considerable variation across colonies likely due to inbreeding, and we found that many captive animals had more than 5% autozygosity. We estimate a sharp decline in population size over the last few generations, though long term population sizes were quite high (~50,000). Leveraging sequenced families, we provide the first pedigree-based recombination map and expand our understanding of de-novo mutations in marmosets. Our results regarding genetic diversity, structure and inbreeding will inform breeding strategies and management practices to maintain healthy, genetically diverse colonies. Further, our genomic repository is bound to facilitate research into the budding marmoset model.

INCREASED POWER IN eQTL STUDIES HELPS CLOSE COLOCALIZATION GAP WITH GWAS SIGNALS

Jonathan D Rosen[1], Sarah M Brotman[1], K A Broadaway[1], Karen L Mohlke*[1], Michael I Love*[1,2]

[1]University of North Carolina, Department of Genetics, Chapel Hill, NC,
[2]University of North Carolina, Department of Biostatistics, Chapel Hill, NC

Expression quantitative trait locus (eQTL) studies have supported gene expression as a mediator of genome-wide association study (GWAS) signals through colocalization, yet several recent reports have illustrated that many GWAS loci without colocalizing eQTLs remain. Insufficient power for eQTL detection is a contributor to this colocalization gap. We used effect sizes observed in recent eQTL studies to estimate the statistical power of eQTL detection at common sample sizes.

We specifically address the following questions: How much does increasing eQTL sample size increase the number of eQTL signals identified at various signal strengths? How do the characteristics of eQTL signals of various signal strengths compare to those of GWAS signals? Would detection of additional eQTL help close the colocalization gap? We used data from publicly available eQTL studies to address these questions. We estimated power via a standard linear model framework using a single metric (r) to describe signal strength that simultaneously captures the effect of magnitude of the linear coefficient (beta) and minor allele frequency (MAF). For example, r = 0.2 corresponds to a strong eQTL signal (e.g. beta = 0.33, MAF = 0.25), while r = 0.05 corresponds to a weak signal (e.g. beta = 0.08, MAF = 0.25). We fit parametric models to empirical effect size distributions to estimate the proportion of eQTL signals detected and compared colocalization results at different sample sizes.

We observed that the power to detect even strong eQTL signals is modest (<60%) at a sample size of 500. Even at sample sizes of 5,000, the estimated power to detect signals at r = 0.05 is very low (<25%). Analogously, the sample size required to achieve 80% power at such signal strength exceeds 10,000.

We next compared the genomic characteristics of eQTL signals by signal strength. Lower strength eQTL signals were located further from the transcription start site of their cognate gene and showed less overlap with promoter regions and more overlap with distal enhancer regions. The genes associated with lower strength eQTL signals were more likely to have a high probability of loss of function intolerance (pLI) and more likely to encode transcription factors than genes associated with stronger signals.

Finally, we compared GWAS colocalizations between adipose studies of two sample sizes, 2,200 and 420, to determine the effect of increased eQTL detection power on the number of colocalizations. We observed that the weakest signals were just as likely to colocalize as the strongest, strongly suggesting that additional colocalizations would be identified by expanding the range of detectable signal strength.

MODELLING GENE DOSAGE RESPONSE ACROSS MODALITIES AT SINGLE CELL RESOLUTION

Leah U Rosen[1], Jasper Panten[1], Tuuli Lappalainen[1,2]

[1]Science for Life Laboratory, KTH Royal Institute of Technology, Department of Gene Technology, Solna, Sweden, [2]New York Genome Center, New York, NY

As our understanding of the human genome has exploded since the early 2000s, over 90% of the human trait and disease-linked genetic variants have been found in the non-coding genome. While we are now able to link many of these variants to quantitative changes in gene expression, as well as to traits and disease, how quantitative changes in gene expression ("cis gene dosage") affect other genes within the same cell (so-called "trans genes"), as well as the cell state and tissue composition remains poorly understood.

Recent studies have leveraged technologies that produce variable levels of cis gene down- or upregulation to study the effects of gene dosage. While most of these studies have used single cell sequencing technologies to allow for scalable pooled screens, due to the noise and technical artifacts in single cell data, most of the analysis has either pooled cells that have a given perturbation, and/or been performed at a transcriptome-wide level. To leverage the full power of single-cell data, we here present a Bayesian model for studying Dosage Response Effects Across Modalities (bayesDREAM) at single cell resolution.

BayesDREAM models technical confounders (e.g. cell line effect) and true underlying cis gene expression, before fitting the trans effect to the inferred cis gene expression. Both the modelled trans effect modality as well as the function used to model it are flexible. We show this by modelling both gene expression and splicing. This approach allows for the study of dosage effects without a large number of guides per gene and enables hitherto elusive biological questions to be answered. Firstly, an open question in the field has been whether there indeed is a continuous response to dosage, or rather a more switch-like on/off response. Pooled analyses were not able to resolve whether a given quantitative change is at the cell level, or rather reflects the proportion of pooled cells in a given state. Secondly, the field has long modelled dosage responses using a sigmoid-like function, frequently a Hill equation. However, this functional form implies that either a trans gene can only be either up- or down-regulated in response to changes in cis gene expression, or it is, in the unperturbed state, very sensitive to small changes in cis gene expression. By carefully modelling the data at single cell resolution, bayesDREAM is able to provide insights into the cellular sensitivity to small changes in gene expression. Finally, bayesDREAM reveals bimodality in underlying cis gene expression, revealing clear constraints on gene dosage.

AN ATLAS OF ALLELE-SPECIFIC DNA METHYLATION IN THE HUMAN BODY

Jonathan Rosenski[1], Ayelet Peretz[2], Judith Magenheim[2], Netanel Loyfer[1], Ruth Shemer[2], Benjamin Glaser[3], Yuval Dor[2,4], Tommy Kaplan[1,2,4]

[1]The Hebrew University of Jerusalem, School of Computer Science and Engineering, Jerusalem, Israel, [2]The Hebrew University of Jerusalem, Dept. of Developmental Biology and Cancer Research, Jerusalem, Israel, [3]The Hebrew University of Jerusalem, Dept. of Endocrinology and Metabolism, Jerusalem, Israel, [4]The Hebrew University of Jerusalem, Center for Computational Medicine, Jerusalem, Israel

While hundreds of thousands of genomic loci have been associated with phenotypic variation, the epigenetic mechanisms underlying most of these phenotypes remain largely unknown. To address this gap, we generated a whole-genome human DNA methylation atlas, profiling >200 purified WGBS samples from ~40 cell types. We then developed computational algorithms to jointly analyze genetic variation, DNA methylation, and gene expression.

Using this integrative approach, we identified 325,000 genomic regions with a bimodal distribution of methylation. Among these, 40,000 regions exhibited genetic variation correlated with methylation patterns, and 460 regions displayed parental allele-specific methylation. By incorporating allele-specific expression data, we uncovered mechanisms driving tissue-specific escape of imprinted expression, such as for *IGF2* in the liver, and characterized tissue-specific imprinting of various disease-associated genes. Notably, we validated novel tissue-specific, maternal allele-specific methylation of *CHD7*, offering insights into the paternal bias in CHARGE syndrome inheritance.

Our analysis of cis-meQTLs revealed a cell-type-specific map of genomic variants that interact with DNA methylation through altered transcription factor binding affinity. By integrating these data with eQTLs, we uncovered networks of epigenetic regulators that influence phenotypes and disease.

This comprehensive approach provides unprecedented insights into the regulatory switches that drive cell-type specificity and the role of DNA methylation in gene expression. The resulting atlas serves as a critical resource for studying differentiation regulators, understanding how methylation controls gene activity, and elucidating its impact on heritable disease.

CELL-TYPE- AND CONTEXT-SPECIFIC EFFECTS OF ARCHAIC INTROGRESSION ON MODERN HUMAN IMMUNE RESPONSES

Zhi Li[1], Gaspard Kerner[1], Javier Mendoza-Revilla[1], Fumitaka Inoue[2], David Gokhman[3], Lluis Quintana-Murci[1], Maxime Rotival[1]

[1]Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, CNRS, Paris, France, [2]ASHBi, Kyoto, Japan, [3]Weizmann Institute of Science, Department of Molecular Genetics, Rehovot, Israel

Admixture between modern and archaic humans has resulted in contemporary populations outside Africa carrying ~1-4% of Neanderthal or Denisovan ancestry in their genomes. While most introgressed archaic variants are rare, adaptive events at genes involved in innate immunity genes (IIG) or encoding virus-interacting proteins (VIPs) can occasionally drive introgressed haplotypes to higher frequencies. Yet, the regulatory basis of such adaptive events remain unclear. Here, we investigated the cis-regulatory effects of 4,628 introgressed variants, located proximal to IIG or VIPs, among the 5% most frequent across human populations. Leveraging a lentivirus-based massively parallel reporter assay, we examined the regulatory effect of these archaic variants in three cell lines — HepG2 (hepatocytes), A549 (lung epithelial cells), and K562 (hematopoietic progenitors) — exposed to various immune or infectious stimuli (IFNα, Dexamethasone, TNFα, IAV and SARS-CoV-2). Of the tested loci, 58% (2,171/4,628) exhibited significant regulatory activity (FDR<1%) in at least one cell type, and ~8% (358/4,628) displayed differential regulatory activity between modern and archaic alleles (FDR<5%). For example, we find that the Neandertal rs121913171-A allele, located 133bp upstream of *IFNGR1* (33% and 70% frequency in Pakistanis and Papuans, respectively), is associated with decreased transcription in hepatocytes. We further show that immune stimulation alters activity of enhancers bound by specific transcription factors (e.g., IRFs for IFN stimulation), and identify 51 variants whose effect on expression is either revealed or amplified by stimulation (FDR_{GxE}<5%). Among these, we unveil a lung-specific regulatory variant at the *LZTFL1* COVID-19 locus, whose archaic allele rs17713054-A is associated with both increased basal activity and increased induction by TNFα. We also reveal the effect of the Denisova rs17066192-C allele, reaching 63% frequency in Papuans, which disrupts a TNF-inducible enhancer of *TNFAIP3* in HepG2 cells. Finally, we report the disruption in IFN-stimulated K562 cells of a *POLR3H* enhancer by the Neanderthal rs75784-G allele associated with altered white blood cell counts and increased risk of asthma in Europeans. Altogether, our study sheds light on the molecular and regulatory consequences of archaic introgression and highlights the need to expose cells to external stimuli to enable the detection of context-specific effects.

PREDICTING ALLELE-SPECIFIC EFFECTS USING A LOCAL SEQUENCE BASED TRANSFORMER MODEL

Joel Rozowsky, Jacqueline Wang, Andrei Onut, Tianxiao Li, Mark Gerstein

Yale University, Molecular Biophysics & Biochemsistry, New Haven, CT

We present an approach for modeling allele-specific behavior (such as allele-specific transcription factor binding or allele-specific gene expression) using a deep-learning transformer model. Traditionally allele-specific behavior is measured by mapping sequenced functional genomic data to a personalized diploid genome sequence and using statistical tests to distinguish balanced behavior from allele-specific behavior by counting the functional genomic reads that map to each haplotype using heterozygous SNVs that differentiate between them. Using the EN-TEx resource of matched functional genomic data and personalized diploid genomes for four individuals we train a transformer model using DNABERT to predict which heterozygous SNVs exhibit allele-specific activity using a 200 base pair window of the DNA sequence centered on the SNV.

We demonstrate that this transformer model approach can accurately (cross-validated AUROC ~ 0.75) predict allele-specific expression and binding for a number of different transcription factors (such as CTCF and POL 2) and histone modifications (such as H3K4me3 and H3K27ac). We train the model using data from one individual and assess the performance on hetSNVs from other individuals. We also investigate which DNA sequence features are important for the transformer model performance. By observing the attention score patterns around allele-specific hetSNVs we find that the transformer model focuses on distinct sequence features in the local neighborhood of the hetSNV. For example, for allele-specific CTCF binding the transformer model identifies not only CTCF motifs a set of known TF motifs corresponding to associated TF cofactors that bind in the vicinity of CTCF that are used as additional features for the performance of the transformer model.

EXPLORING RESIDUAL HETEROZYGOSITY IN INBRED RAT STRAINS: HOW MUCH, WHERE, AND WHY?

Farnaz Salehi[1], Andrea Guarracino[1], Denghui Chen[3], Flavia Villani[1], David G Ashbrook[1], Vincenza Colonna[1], Abraham Palmer[3], Robert W Williams[1], Hao Chen[2], Erik Garrison[1]

[1]University of Tennessee Health Science Center, Department of Genetics, Genomics and Informatics, Memphis, TN, [2]University of Tennessee Health Science Center, Department of Pharmacology, Addiction Science, and Toxicology, Memphis, TN, [3]University of California San Diego, Department of Psychiatry, San Diego, CA

Inbred strains are expected to achieve 98.7% homozygosity after 20 filial (F) generations of matings and 99.98% after 40, yet genome assemblies of inbred rats even after F100 reveal heterozygous regions, raising concerns about genetic stability. We are interested in resolving sources of residual heterozygosity and possible mechanisms of its maintenance.
We generated PacBio HiFi sequences (40–52X) for strains from three groups of male samples: common inbred strains (n = 16, mean F>100), the HXB/BXH RI family (n = 9, average F98), and the FXLE/LEXF RI family (n = 1, F27). Initial HiFi assemblies contained primary and alternate contigs totaling ~3.27 Gbp per strain. We hypothesized that the additional 0.5 Gbp per assembly relative to the reference length comes from alternate haplotypes and regions of residual heterozygosity.
To investigate this, we combined primary and alternate assemblies with GRCr8 to create strain-specific pangenome graphs with the PanGenome Graph Builder. We detected ~2.3 million heterozygous SNPs per assembly (~0.07%), ranging from 0.06% in BN to 0.11% in WKY/NCrl. The F27 FXLE12 sample showed 0.08% heterozygosity with a heterozygous locus on Chr 9. Regions within 5 Mb of centromeres showed unusually high heterozygosity on Chrs 2, 10-12, and X. Heterozygosity was intermediate on Chrs 1, 3, 4, 5, 7, and 8, while notably low on shorter chromosomes (Chrs 9, 13-19).
Next, we combined genome assembly and comparative analysis methods to identify heterozygous regions while removing false signals from sequencing errors and difficulty genotyping in duplicated sequences. We carried out an analysis for comparing both raw reads and assembled sequences independently to cross validate the heterozygous sites and restricted our study to regions which are not segmental duplications in the GRCr8 reference or any strain assembly.
The high levels of heterozygosity near centromeres, particularly on chromosomes 2 and 10, 11, and 12, suggest these regions may be under selection to maintain genetic diversity, which could affect chromosome stability in laboratory strains. Our ongoing work will show the significant differences before and after eliminating technical confounders.

NANOPORE DUPLEX SEQUENCING REVEALS PATTERNS OF ASYMMETRIC STATES OF 5HMC AND 5MC IN THE MEDAKA BRAIN GENOME

Walter Santana Garcia[1], Tomas Fitzgerald[1], Joachim Wittbrodt[2], Felix Loosli[3], Ewan Birney[1]

[1]European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, [2]Centre for Organismal Studies, Heidelberg University, Heidelberg, Germany, [3]Institute of Biological and Chemical Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany

Nucleotides in DNA can be covalently modified, and they can contribute to the regulation of biological processes across the tree of life. The methylation of the fifth carbon of the cytosine ring, 5-methylcytosine (5mC), has been extensively characterised and is associated with a gene-repressive role. However, the functional contributions of its oxidised state, 5-hydroxymethylcytosine (5hmC), and its interplay with 5mC remain to be further characterised. In addition, the emerging view in the DNA modification field on the importance of hemi-modified states, such as hemi-methylation and hemi-hydroxymethylation of CpG dinucleotides in the genome, also remains largely unexplored.

Although useful, traditional assays for the detection of 5mC and 5hmC in DNA are unable to provide enough resolution to address these research questions, as the simultaneous profiling of 5mC and 5hmC in the same DNA molecule is not possible. Recent advances in Oxford Nanopore Technologies (ONT) not only enable the sequencing and simultaneous profiling of different DNA modifications in the same molecule, but also allow to identify and pair modifications in the two self-complementary strands of a duplex DNA molecule, termed duplex sequencing.

Here, we explore the co-occurrence of 5hmC with 5mC in genomic CpGs of medaka (japanese rice paddy fish) brain tissues, and we characterise asymmetrical patterns of hemi-methylated and hemi-hydroxymethylated states in CpGs using ONT duplex sequencing which we can achieve at high coverage given the small genome size. We show at duplex resolution that the association of 5hmC with 5mC (5hmC/5mC) is the most prevalent state of 5hmC in CpGs, and that in this asymmetrical state, 5hmC is preferentially found in a strand specific pattern around splice sites with a clear strand bias shift between introns and exons. Furthermore, in the 5hmC/5mC state, 5hmC preferentially locates in one strand of the DNA duplex in certain families of transposable elements (TEs) in medaka. We are extending this work to other organisms, in particular Mouse and Human, and initial results show a similar pattern. Altogether these results provide a deeper understanding of the genomic function of 5hmC and its interplay with 5mC in their double stranded context.

ADAPTIVE INCREASE OF AMYLASE GENE COPY NUMBER IN PERUVIANS DRIVEN BY POTATO-RICH DIETS

Kendra Scheer[1], Luane J.B. Landau[1], Kelsey Jorgensen[2,3], Charikleia Karageorgiou[1], Lindsey Siao[1], Can Alkan[4], Angelis M Morales-Rivera[5], Christopher Osborne[1], Obed Garcia[3], Laurel Pearson[6], Melisa Kiyamu[7], Fabiola Leon-Velarde[7], Frank Lee[8], Tom Brutsaert[9], Abigail Bigham[2], Omer Gokcumen[1]

[1]University at Buffalo, Department of Biological Sciences, Buffalo, NY, [2]University of California, Los Angeles, Department of Anthropology, Los Angeles, CA, [3]University of Kansas, Department of Anthropology, Lawrence, KS, [4]Bilkent University, Department of Computer Engineering, Ankara, Turkey, [5]Universidad de Puerto Rico en Cayey, Departamento de Biología dentro de las Ciencias Naturales, Cayey, PR, [6]The Pennsylvania State University, Department of Anthropology, State College, PA, [7]Universidad Peruana Cayetano Heredia, Laboratorio de Fisiología del Transporte de Oxígeno y Adaptación a la Altura, Laboratorios de Investigación y Desarrollo, Lima, Peru, [8]University of Pennsylvania Perelman School of Medicine, Department of Pathology and Laboratory Medicine, Philadelphia, PA, [9]Syracuse University, Department of Exercise Science, Syracuse, NY

The salivary amylase gene (AMY1) exhibits remarkable copy number variation linked to dietary shifts in human evolution. While global studies highlight its structural complexity and association with starch-rich diets, localized selection patterns remain under explored. Here, we analyzed AMY1 copy number in 3,723 individuals from 84 populations, revealing that Indigenous Peruvian Andean populations possess the highest AMY1 copy number globally. A genome-wide analysis showed significantly higher amylase copy numbers in Peruvian Andean genomes compared to closely related populations. Further, we identified positive selection at the nucleotide level on a haplotype harboring at least five haploid AMY1 copies, with a Peruvian Andean-specific expansion dated shortly after potato domestication (~6–10 kya). Using ultra-long-read sequencing, we demonstrated that recombination-based mutational mechanisms drive the formation of high-copy AMY1 haplotypes. Our study provides a framework for investigating structurally complex loci and their role in human dietary adaptation.

SCALING DEEP LEARNING-BASED CANCER DRUG RESPONSE PREDICTION MODELS FOR PRECISION ONCOLOGY APPLICATIONS

Casey Sederman, Gabor Marth

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

The availability of large-scale pharmaco-omic screens in cancer cell lines has facilitated the development of deep learning (DL)-based cancer drug response prediction (CDRP) models with the potential to transform precision oncology practice. However, pharmaco-omic associations learned from cell lines do not always generalize to patients in a straightforward manner, limiting the clinical utility of models trained on cell lines alone. Data generation in more clinically relevant tumor models such as patient-derived organoids will be crucial for the development of DL-based CDRP models with clinical utility. However, the interplay between data generation and model performance is poorly understood. To enhance the synergy between resource-intensive data generation efforts and model development, here, we study the scaling of generalization error in CDRP models with increasing training dataset size. Using a field-standard neural architecture, we trained DL models to predict drug efficacy—measured by the area under the dose-response curve—for tumor-drug pairs based on the tumor's transcriptomic profile and the drug's chemical structure. We examined the impact of training data size on model performance by progressively increasing the number of tumor-drug pairs seen during training. Specifically, we gradually enriched the training dataset by sampling additional tumor-drug pairs from a pool of over 300,000 drug responses in the Genomics of Drug Sensitivity in Cancer database. As expected, generalization error scaled as a power law with increasing dataset size, appearing to converge towards an irreducible error region with sufficient training data. However, when decomposing generalization error using a hierarchical additive model, we found that this apparent plateau in model performance was driven by a saturation in the learning of tumor- and drug-specific mean effects corresponding to a tumor's average sensitivity across drugs and a drug's average response across tumors, respectively. In comparison, the DL model's ability to predict the effect of higher-order tumor-drug interactions remained data-dependent and continued to improve with additional training data. Critically, we found that including more tumors, but not more drugs, in the training dataset enhanced the model's ability to capture these higher-order tumor-drug interactions. As the capacity of CDRP models to predict these interactions represents the precision oncology-relevant component of performance, this insight suggests that resource allocation should favor the generation and pharmaco-omic profiling of additional tumor models rather than the screening of more drugs on existing samples. As the field shifts towards pharmaco-omic data generation in more clinically relevant patient-derived tumor models, this work will guide the optimal allocation of experimental budgets during resource development.

DECODING THE MECHANISTIC IMPACT OF GENETIC VARIATION ON REGULATORY SEQUENCES WITH DEEP LEARNING

Evan Seitz, David McCandlish, Justin Kinney, Peter Koo

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

Deciphering how DNA sequences encode cis-regulatory mechanisms is a central challenge in biology. Cis-regulatory elements integrate signals—such as specific transcription factor binding sites combined with broader sequence context—to orchestrate gene expression. Although deep learning has advanced our ability to predict regulatory activity from DNA, these models have yet to systematically reveal how genetic variation reshapes the underlying regulatory mechanisms. Here, we introduce SEAM, a computational framework that uses deep learning and explainable AI to uncover how small changes in DNA can reconfigure cis-regulatory logic. SEAM generates sets of variant sequences and applies model interpretation methods to pinpoint the key nucleotide positions that together form a regulatory "mechanism"—a composite signature reflecting both transcription factor binding sites and their surrounding features. By mapping sequences into a learned "mechanism space" (a data-driven representation of regulatory signatures) and clustering those with similar profiles, SEAM reveals how individual mutations remodel regulatory DNA and drive functional diversity. Applied to human and fly regulatory elements, SEAM not only highlights the remarkable evolvability of cis-regulatory sequences but also disentangles the specific contributions of transcription factor binding from broader sequence context. In doing so, it offers critical insights into the regulatory grammar underlying gene expression and provides a robust framework for guiding the rational design of synthetic DNA with tailored functions.

INVESTIGATING THE ROLE OF MITOCHONDRIAL DNA IN SPERM IMMOTILITY

Isabel Serrano[1], Emma James[2], Jason Kunisaki[1], Xiaoxu Yang[1], Kenneth I Aston[2], Aaron Quinlan[1]

[1]University of Utah, Human Genetics, Salt Lake City, UT, [2]University of Utah, Surgery, Salt Lake City, UT

Mitochondria have an average of four genome (mtDNA) copies per mitochondrion, with each genome encoding thirteen proteins that are essential for energy production. While mtDNA is maternally inherited in humans, mitochondria maintain an important role in sperm as their energy production is vital for sperm motility. However, multiple studies have made the counterintuitive observation that, despite this reliance on mitochondria as an energy source, an *increase* in mtDNA copy number coincides with a *decrease* in motility. Furthermore, studies uncovering the molecular mechanisms underlying the uniparental inheritance of mtDNA have shown that a key process in the development of mature sperm cells is the degradation of mtDNA. In humans, 0.58 to 1.24 mtDNA copies exist per motile spermatozoon, while immotile sperm have at least a 30-fold higher mtDNA copy number abundance. Recent work in *Drosophila* has shown that infertility phenotypes can be modulated by mtDNA abundance, with fertility being restored by mtDNA degradation.

Thus, while mitochondria are essential for sperm motility, mtDNA are purposefully removed from mature sperm, though the reason for mtDNA degradation during spermatogenesis remains unclear. Studies have shown an increase in mtDNA mutation frequency and nuclear DNA fragmentation with higher mtDNA content, suggesting that the presence of mtDNA may impact the genomic stability of both genomes in mature sperm. Altogether, these findings suggest that the presence of mtDNA may contribute to a mechanism impairing sperm motility and/or leading to male infertility.

In our ongoing work, we are exploring protocols to sequence the nuclear and mitochondrial genomes of sperm across different motility ranges to investigate the role mtDNA plays in sperm motility. We will present initial findings that characterize mtDNA abundance across motility gradients. We quantify mtDNA content via digital droplet PCR and sequencing data. The former is the most sensitive technique for quantifying low copy number molecules, allowing us to identify possible mtDNA copy number ranges across motility conditions. Secondly, we demonstrate our ability to recapitulate mtDNA copy number measurements through sequencing data, accounting for contamination and technical artifacts that more heavily impact low mtDNA copy number samples. Ultimately, our work begins to unravel the molecular consequences of retaining mtDNA in mature sperm.

A-TO-I EDITING GENERATES UNPARALLELED COMPLEXITY IN THE NEURAL PROTEOME OF CEPHALOPODS

Kobi Shapira[1,2], Ruti Balter[3], Joshua J Rosenthal[4], Erez Y Levanon[1,2], Eli Eisenberg[3]

[1]Bar-Ilan University, Mina and Everard Goodman Faculty of Life Sciences, Ramat Gan, Israel, [2]Bar-Ilan University, The Institute of Nanotechnology and Advanced Materials, Ramat Gan, Israel, [3]Tel Aviv University, Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv, Israel, [4]The Marine Biological Laboratory, The Eugene Bell Center for Regenerative Biology and Tissue Engineering, Woods Hole, MA

Post-transcriptional and post-translational modifications lead to the generation of diverse protein products from a single gene and make a crucial contribution to cells' complexity. Among these modifications, adenosine-to-inosine RNA editing (A-to-I editing) is a post-transcriptional mechanism capable of reprogramming protein-coding sequences ("recoding"). While prevalent across various metazoan taxa, it attains unparalleled proportions in coleoid cephalopods, where tens of thousands of sites undergo recoding. Numerous messages contain multiple editing sites, each presenting a binary option, resulting in an exponentially large number of potential protein isoforms. However, the extent to which this complexity is realized in the cephalopod nervous system remains unknown. Here, we employ deep-sequencing complemented by a graph-theoretic computational analysis to unravel the extent of this phenomenon in the Longfin Inshore Squid (*Doryteuthis pealeii*) and the common octopus (*Octopus vulgaris*). Targeted sequencing, utilizing both short- and long-read approaches, reveals an unprecedented abundance of encoded isoforms in highly edited squid transcripts, up to 67,000 from a single gene. These numbers even surpass the number of splice variants encoded by *Drosophila melanogaster*'s *dscam* gene (Down Syndrome Cell Adhesion Molecule), a renowned example of extreme diversity due to post-transcriptional RNA processing. Analysis of whole-transcriptome sequencing data from the common octopus further underscores the widespread diversification across genes in coleoid cephalopods. Remarkably, at least 21% of well-covered genes manifest at least 50 distinct editing-isoforms. The distribution of expression levels per isoform exhibits a broad spectrum with no discernible dominance of a small subset of isoforms, suggesting a functional role for numerous distinct isoforms.

CELL-TYPE-SPECIFIC AGE AND SEX EFFECTS ON GENE REGULATION IN IMMUNE RESPONSES TO VIRUSES

Marwan Sharawy, Aurelie Bisiaux, Jan Madacki, Yann Aquino, Milena Hasan, Etienne Patin, Darragh Duffy, Maxime Rotival, Lluis Quintana-Murci

Institut Pasteur, Genetics, Paris, France

Human immune variation is influenced by diverse factors, including age, sex, lifestyle, and genetics. To assess their relative contributions, we collected peripheral blood mononuclear cells from 415 donors, aged 30 to 80, with a balanced sex ratio, and stimulated them with SARS-CoV-2 (BA.5), influenza A virus (pdm09), or a mock control. We then profiled the transcriptomes of >1.5 million cells across 42 immune cell types using single-cell RNA-sequencing. Focusing on the effects of age on gene regulation across cell types, we identified 1,703 genes whose expression or 3′UTR isoform usage change with age, with CD4+ naive T cells exhibiting the most significant changes in gene expression and CD8+ naive T cells displaying the largest shifts in 3′UTR isoform usage. In contrast, the effects of sex were less cell-type-specific, with 1,073 genes differentially expressed between males and females in at least one cell type, including 62 located on sex chromosomes. Interestingly, 71% of age-related effects on gene regulation were only observed after viral exposure, while 65% of sex-related effects were already present at the basal state, suggesting that the impact of age is more context-specific. These effects translated into functional differences between sexes; for example, females exhibited stronger antiviral responses mediated by type I interferon signaling (GO:0034340) and inflammatory regulation (GO:0050729) in CD14+ monocytes. This prompted us to investigate how sex-specific effects could arise from age-associated mosaic loss of the Y chromosome (mLoY). We found that mLoY is lineage- and cell-type-specific, primarily affecting NK and myeloid cells. Furthermore, genes overexpressed in mLoY cells after viral stimulation, compared to non-mLoY cells from the same individuals, were enriched in innate immune pathways (GO:0045087). These findings suggest that Y chromosome loss could contribute to inter-individual differences in antiviral responses through stronger innate immunity . Finally, we mapped 18,532 independent expression quantitative trait loci jointly regulating expression of 7,703 eGenes (FDR < 5%) in at least one cell type and stimulation condition. While we found no evidence of G×Sex interactions, we identified eight eGenes with differential genetic regulation in younger and older donors (G×Age, FDR < 5%), four of which were observed after viral stimulation. Overall, our study provides a comprehensive view of how age and sex shape immune responses, linking immune aging to viral responses and uncovering novel regulatory mechanisms underlying immune senescence.

MUMEMTO: EFFICIENT MAXIMAL MATCHING ACROSS PANGENOMES

Vikram Shivakumar, Ben Langmead

Johns Hopkins University, Computer Science, Baltimore, MD

Aligning genomes into common coordinates is central to pangenome analysis and construction, though computationally expensive. Recent pangenomes comprise hundreds of complete or near-complete assemblies, necessitating efficient methods for pangenome-wide comparison and alignment. Multi-sequence maximal unique matches (multi-MUMs) are guideposts that help frame core and multiple genome alignments. Collinear multi-MUMs represent syntenic anchors that can span large, conserved regions in the multiple alignment. However, current methods for computing multi-MUMs do not scale efficiently beyond small bacterial collections.
We introduce Mumemto, a tool that computes multi-MUMs and other match types across large pangenomes. Mumemto is the only method able to compute multi-MUM synteny across all 474 human assemblies from the Human Pangenome Reference Consortium (HPRC). Mumemto can compute matches across the 474 human genomes (~2.8TB) in under 2 days across 8 nodes using ~566 GB each, and can merge in new assemblies without recomputing over the full dataset. Mumemto (on a single thread) is up to 15X faster than existing multi-MUM finders (using 48 threads), with a 50% smaller memory footprint for human genomes.
Mumemto enables scalable visualization of pangenome synteny, highlighting sequence conservation and structural variation. It can reveal aberrant assembly artifacts and errors in reference-guided contig scaffolding using pangenome-wide syntenic information. It also accelerates existing pipelines for pangenome graph construction and core genome alignment. We show that collinear multi-MUM-based pangenome graphs can be constructed ~25% faster, yielding comparable graphs for downstream read alignment. Mumemto can also slot into existing pipelines for core genome alignment, running up to 12X faster for human assemblies.
Mumemto can compute multi-MUMs across all subsets of sequences in a collection, revealing information about assembly sub-clusters and genomic distance. In a dataset of *A. thaliana* assemblies from diverse geographical regions, an aggregated metric of partially-shared multi-MUMs reveals subgroups corresponding to geographically isolated regions. Similarly, in a collection of Cyanobacteria genomes, aggregated shared MUM length across sequence pairs correlated highly with true phylogenetic tree distance, outperforming standard pairwise comparison methods, highlighting the potential for genomic distance estimation using multi-MUMs.
Mumemto serves as a core tool for pangenome construction, visualization, and interpretation. It scales with the growing size of newly released pangenome collections, and is the only method capable of running on hundreds of human genomes. Mumemto is implemented in C++ and Python and available open-source at https://github.com/vikshiv/mumemto.

OVERLAPPING READING FRAMES WITHIN THE MTDNA ARE DEEPLY CONSERVED AND ASSOCIATE WITH A PROGRAMMED FRAME SHIFT MECHANISM

Noam Shtolz[1], Michele Brischigliaro[2], Dan Mishmar[1], Antoni Barrientos[2,3]

[1]Ben-Gurion University of the Negev, Life Sciences, Beer-Sheva, Israel, [2]University of Miami Miller School of Medicine, Neurology, Miami, FL, [3]University of Miami Miller School of Medicine, Biochemistry and Molecular Biology, Miami, FL

Throughout evolution the mitochondrial DNA (mtDNA) encodes essential subunits of the oxidative phosphorylation system, which is crucial for cellular energy production. Unlike nuclear DNA-encoded genes, human mtDNA is transcribed from strand-specific promoters into polycistronic RNA molecules that in turn are cleaved into mature transcripts. While most mature mtDNA transcripts are monogenic, two (ND4L-ND4 and ATP8-ATP6) remain bicistronic, and contain overlapping open reading frames. Previously, we found that both gene pairs remained adjacent across metazoan evolution far more than any other mtDNA gene pair, suggesting that their proximity is subjected to purifying selection. Recently, a coordinated translation mechanism of the ATP8-ATP6 gene pair was discovered in human mtDNA via programmed mitochondrial ribosome frameshift (PRF), which enables independent expression of both proteins from a single transcript. We hypothesized that this mechanism served as a selective constraint to conserve the ATP8-ATP6 gene pair during metazoan evolution. To test this hypothesis, we sought to characterize these overlapping ORFs across 17,530 eukaryotic mtDNA sequences, and to determine the conservation of the PRF mechanism. We found that the length of sequence overlaps of both ATP8-ATP6 and ND4L-ND4 gene pairs is highly conserved per-phylum (appearing in 99% of Chordata species) and is not correlated with mtDNA size. Using motif and RNA structure analyses, we identified conservation of elements associated with the ATP8-ATP6 PRF mechanism across metazoans. This study not only sheds light on the evolutionary dynamics of mtDNA translation, but also provides a mechanistic explanation for the selective constraints underlying conservation of mtDNA organization across metazoan evolution, with implications for deeper understanding of mitochondrial gene expression regulation.

ORAL MICROBIOME DIVERSITY ACROSS DIFFERENT ETHNICITIES IN THAILAND

Faith Chin Yee Sim [1], Hie Lim Kim [1,2]

[1]Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, Singapore, [2]The Asian School of the Environment, Nanyang Technological University, Singapore

The diversity of the human oral microbiome plays a crucial role in maintaining oral and systemic health, influencing processes such as immune response, nutrient metabolism, and disease susceptibility. While lifestyle factors such as diet, dental hygiene and smoking habits are known to influence the oral microbiome, the impact of the host's genetic background on oral microbiome diversity remains largely unknown. In this study, we aim to investigate the interaction between host genomic variation and oral microbiome diversity and composition across six different indigenous ethnic groups in Thailand. We sequenced and analyzed the oral metagenome data and host genomes of 172 individuals, identifying 3185 species, of which 1079 (33.9%) were shared across all ethnicities. By comparing genetic distances among ethnic groups based on whole-genome data with the alpha diversity and composition of their oral microbiomes, We found significant differences in alpha diversity across ethnicities. The pairwise PERMANOVA analysis revealed that the microbial composition varied significantly between genetically distant ethnic groups. These findings suggest that host genetic differences contribute to shaping the oral microbiome diversity in the Thai population. Future studies will integrate immune variation and lifestyle factors to better understand their combined effects on oral microbiome diversity.

EFFECTIVE SINGLE CELL COUNTS ANALYSIS: FEATURE SELECTION IN THE ORIGINAL GENES SPACE AND CHOICE OF NUMBER OF COORDINATES IN REDUCED DIMENSIONS PRINCIPAL COMPONENTS SPACE HOLD THE KEY

Amartya Singh, Mona Arabzadeh, Daniel Herranz

Rutgers Cancer Institute of New Jersey, Center for Systems and Computational Biology, New Brunswick, NJ

The application of single-cell RNA-sequencing (scRNA-seq) technologies to uncover novel biological signatures at the level of individual cells has seen a tremendous growth over the past few years. Concomitantly, significant amount of effort has also been devoted to develop normalization/transformation approaches to transform the raw counts data prior to application of principal components analysis (PCA) for dimensionality reduction. It is widely acknowledged that normalization plays a vital role in determining the outcomes of clustering analyses performed downstream. Thus, it came as a surprise to many when in a recent comparative study of transformation methods, Ahlmann-Eltze and Huber arrived at the surprising conclusion that the simple shifted log-transformation based transformation approach performed just as well or better than more conceptually sophisticated approaches. Curiously, while examining the effect of post-processing steps they shortlisted only 1000 genes as highly variable genes (HVGs) and based on the results concluded that shortlisting HVGs did not lead to improvements for identifying nearest neighbors. Additionally, they did not systematically examine how changing the number of PCs affects the kNN graph and the inferred clusters.

Recently, we proposed a simple bin-based feature selection method that relies on estimation of quasi-Poisson dispersion coefficients based on the observed counts for each gene to help identify HVGs. This method has been implemented in our R package for scRNA-seq analyses called Piccolo. We used this feature selection method and found that in fact HVG selection and the choice of number of PCs prior to generating the kNN graph for downstream clustering analyses are key to ensuring robust and informative clustering outcomes. Normalization is of course a vital step prior to application of PCA but the clustering outcomes were not found to be significantly different between the various transformation methods provided each of them relied on informative sets of HVGs identified by our feature selection method and choice of number of PCs that best resulted in most pronounced separation between distinct cell groups corresponding to various cell-types/states. We further propose a simple unsupervised kNN graph-based framework to inform the choice of appropriate number of PCs. In summary, we highlight the crucial role of feature selection in scRNA-seq analyses supplemented by a conceptually simple approach to inform the choice of an appropriate number of PCs for downstream clustering analyses.

NON-CANONICAL DNA IN HUMAN AND OTHER APE TELOMERE-TO-TELOMERE GENOMES

Linnéa Smeds[1], Kaivan Kamali[1], Iva Kejnovská[2], Eduard Kejnovský[3], Francesca Chiaromonte[4,5,6], Kateryna D Makova[1,5]

[1]Penn State University, Department of Biology, University Park, PA, [2]Institute of Biophysics of the Czech Academy of Sciences, Department of Biophysics of Nuclleic Acids, Brno, Czech Republic, [3]Institute of Biophysics of the Czech Academy of Sciences, Department of Plant Developmental Genetics, Brno, Czech Republic, [4]Penn State University, Department of Statistics, University Park, PA, [5]Penn State University, Center for Medical Genomics, University Park, PA, [6]Sant'Anna School of Advanced Studies, L'EMbeDS, Pisa, Italy

Non-canonical (non-B) DNA structures—e.g., bent DNA, hairpins, G-quadruplexes, Z-DNA, etc.—which form at certain sequence motifs (e.g., A-phased repeats, inverted repeats, etc.), have emerged as important regulators of cellular processes and drivers of genome evolution. Yet, they have been understudied due to their repetitive nature and potentially inaccurate sequences generated with short-read technologies. Here we comprehensively characterize such motifs in the long-read telomere-to-telomere (T2T) genomes of human, bonobo, chimpanzee, gorilla, Bornean orangutan, Sumatran orangutan, and siamang. Non-B DNA motifs are enriched at the genomic regions added to T2T assemblies, and occupy 9-15%, 9-11%, and 12-38% of autosomes, and chromosomes X and Y, respectively. Functional regions (e.g., promoters and enhancers) and repetitive sequences are enriched in non-B DNA motifs. Non-B DNA motifs concentrate at short arms of acrocentric chromosomes in a pattern reflecting their satellite repeat content and might contribute to satellite dynamics in these regions. Most centromeres and/or their flanking regions are enriched in at least one non-B DNA motif type, consistent with a potential role of non-B structures in determining centromeres. Our results highlight the uneven distribution of predicted non-B DNA structures across ape genomes and suggest their novel functions in previously inaccessible genomic regions.

ENHANCER GRAMMAR OF DEVELOPMENTAL ENHANCERS

Joe J Solvason*[1,2], Fabian Lim*[1,2], Benjamin P Song[1,2], Jessica L Grudzien[1,2], Sophia H Le[1,2], Katrina M Olson[1,2], Granton A Jindal[1,2], Krissie Tellez[1,2], Emma K Farley[1,2]

[1]University of California, San Diego, Department of Biology, La Jolla, CA, [2]University of California, San Diego, Department of Medicine, La Jolla, CA

Enhancers direct precise patterns of gene expression during development by interactions with transcription factors (TFs). To explore the role of transcription factor binding site (TFBS) grammar within enhancers, we conducted a high-throughput screen testing 460,800 different organizations TFBSs for activity within developing embryos. Using statistical and ML approaches to mine our dataset we discover grammatical and syntax rules governing enhancer tissue-specificity. With this knowledge, we design synthetic tissue-specific enhancers with diverse sequence information but conserved grammatical rules.

SPCORR MODELS SPATIALLY VARIABLE GENE CO-EXPRESSION PATTERNS IN SPATIAL TRANSCRIPTOMICS

Chenxin Jiang[1], James Y. H. Li[2], Jingy Jessica Li[1], Dongyuan Song[2]

[1]University of California, Los Angeles, Statistics & Data Sciences, Los Angeles, CA, [2]University of Connecticut Health Center, Genetics & Genome Sciences, Farmington, CT

Spatial transcriptomics has transformed our ability to explore gene expression within its native tissue context, enabling us to dissect subtle yet biologically significant variations *in situ*. While numerous methods have been proposed for per-gene modeling of *Spatially Variable Gene* (SVG), few methods aim to model the spatial organization of gene-gene co-expression patterns. To fill this gap, we introduce spCorr --- a flexible and scalable semi-parametric regression framework for modeling and testing *Spatially Variable Co-expression* (SVC). Unlike methods focusing solely on single-gene spatial expression, spCorr distinguishes between changes in mean or variance and alterations in gene-gene correlation across space, enabling the detection of regulatory changes. Our method not only overcomes the computational burdens and interpretability issues in non-parametric methods but also provides rigorous testing for detecting spatial changes in correlation strength. Through comprehensive simulation studies, we show that spCorr archives the highest power while strictly controlling the False Discovery Rate (FDR), and is at least 20 times faster than existing methods.

We further demonstrate the utility of spCorr through two real data applications. In the first application, spCorr is applied to 10x Genomics Visium data of HPV-negative oral squamous cell carcinoma (OSCC). Focusing on the correlation of ligand-receptor pairs, our analysis reveals pronounced spatial co-expression heterogeneity within the tumor region, highlighting distinct zones of intercellular communication that may underlie tumor progression and microenvironment dynamics. In the second application, spCorr is employed on high-resolution spatial transcriptomics data of the mouse brain hippocampus generated by the 10x Genomics Visium HD technology. We concentrate on the between-domain SVC among three spatial regions: CA1, CA2, and CA3. By constructing differential correlation networks, we identify gene modules in the CA2 region significantly associated with reduced synaptic plasticity, offering novel insights into the molecular underpinnings of hippocampal function. Overall, spCorr offers a robust and interpretable framework for analyzing SVC, a critical second-order property that reflects the context-specific gene regulatory relationships. This approach enables the detection of complex molecular interactions driving biological processes and disease mechanisms, providing novel insights into the interdependencies of gene regulation that are often overlooked by conventional analyses.

WORLDWIDE PATTERNS OF DIVERSITY AT THE 17Q21.31 LOCUS IN MODERN AND ANCIENT HUMAN GENOMES

Samvardhini Sridharan[1,2], Runyang N Lou[3], Victor Borda[4], Santiago G Medina-Munoz[5], Simon Gravel[6], Brenna Henn[7], Peter H Sudmant[1,2,3]

[1]University of California, Department of Molecular and Cell Biology, Berkeley, CA, [2]University of California, Center for Computational Biology, Berkeley, CA, [3]University of California, Department of Integrative Biology, Berkeley, CA, [4]University of Maryland, School of Medicine, Baltimore, MD, [5]CINVESTAV, LANGEBIO, Mexico City, Mexico, [6]McGill University, Department of Human Genetics, Montreal, Canada, [7]University of California, Department of Anthropology, Davis, CA

The 17q21.31 locus in humans contains an inversion spanning over 900 Kbp that encompasses several medically relevant genes, notably *MAPT* and *KANSL1*, which are associated with neurodegenerative disease. Additionally, the locus harbors a complex duplication architecture that results in variation in both the orientation and copy number of distinct segments of the locus. Different 17q21.31 haplotypes are associated with variation in recombination rate, fecundity, and disease prevalence, however the full extent of diversity at this locus has not been characterized. By combining long-read and short-read sequencing data, we characterize the structural diversity and trace the recent evolution of different 17q21.31 structural haplotypes in humans. We analyzed 212 recently published long-read haplotype-resolved assemblies using a pangenome-graph-based method and identified 29 unique structural haplotypes at this locus. The diversity among these haplotypes primarily arises from their inversion orientation and their copy numbers of the *NSF* and *KANSL1* genes. Based on this knowledge, we use a read-depth based approach to characterize the haplotype diversity worldwide in ~4500 short-read-sequenced genomes. These haplotypes are grouped based on their orientation and duplication status (e.g. H2 for inverted orientation with a single copy of *KANSL1*, H1D for direct orientation with two copies of *KANSL1*). We uncovered unexpected patterns of variation in previously understudied superpopulations such as East Asians, South Asians, and Africans. Expanding on this further, we analyzed 626 ancient human genomes and revealed that H1 haplotype which was almost ubiquitous in Eurasia ~12 thousand years ago has declined in frequency ~40% concomitant with an increase in both H1D and H2D haplotypes. Finally, we identified several likely double-recombination, or long-patch gene conversion events, occurring mostly in African populations, that impact haplotype diversity. Altogether, our findings offer new perspectives on the evolutionary forces acting on this complex and medically significant locus, contributing to a broader understanding of human genetic diversity and structural variation

COMPARATIVE DEMOGRAPHIC ANALYSIS OF CARDAMINE HIRSUTA AND ARABIDOPSIS THALIANA

Rachita Srivastava[1], Bjorn Pieper[1], Sileshi Nemomissa[2], Donavan Bailey[3], Christian Brochmann[4], Sebsebe Demissew[5], Angela Hancock[6], Carlos Alonso-Blanco[7], Stefan Laurent[8], Miltos Tsiantis[1]

[1]Max Planck Institute for Plant Breeding Research, Comparative Development and Genetics, Cologne, Germany, [2]Addis Ababa University, Plant Biology & Biodiversity Management, Addis Ababa, Ethiopia, [3]New Mexico State University, Department of Biology, Las Cruces, NM, [4]University of Oslo, Natural History Museum, Oslo, Norway, [5]Addis Ababa University, Plant Biology & Biodiversity Management, Addis Ababa, Ethiopia, [6]Purdue University, Botany and Plant Pathology, West Lafayette, IN, [7]Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Plant Molecular Genetics, Madrid, Spain, [8]BioNTech SE, Tailored Omics Technologies, Mainz, Germany

Understanding the dynamics of populations and their genetic divergence in relation to historical ecological and climatic events is a major goal of evolutionary biology. Comparing the population dynamics and distribution of closely related species can illuminate the roles of common versus divergent ecological parameters in shaping trait and sequence diversity patterns, and identifying convergent genetic responses to global ecological changes. Cardamine hirsuta (Cardamine) is a small crucifer related to Arabidopsis thaliana (Arabidopsis), and shares features with Arabidopsis, making it an attractive experimental system. Population genetics analyses of Pan-European Cardamine samples comprising 746 natural accessions revealed that Cardamine from the Iberian Peninsula and the Macaronesian Islands
retain considerable genetic diversity, thus resembling Iberian relict populations of Arabidopsis. However, Cardamine strains extend further into mainland Europe than Arabidopsis relicts, suggesting a greater ability to establish beyond glacial refugia. In Arabidopsis, the deepest population splits have been reported to be those between the African populations. The absence of information on Cardamine African diversity leaves many open questions on the comparative understanding of the population dynamics of the two species. In this study, we developed a standardized bioinformatics workflow to generate variation data for two related Brassicaceae species sampled from Europe, North Africa, and East Africa. Next, we joint-analyzed the demographic history of Cardamine from Europe and Africa and compared it to that of Arabidopsis. Our findings shed light on the complex interplay of historical events, geographic influences, and ecological factors in shaping these closely related plant species population history and genetic
divergence.

BAYESIAN INFERENCE OF THE METASTASIS GRAPH FROM CANCER CELL LINEAGE TRACING DATA

Stephen J Staklinski, Adam Siepel

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

Inferring tissue-migration dynamics from cell-lineage tracing data in cancer metastasis presents a phylogenetic challenge similar to biogeographic inference. Migration graphs must be reconstructed from evolutionary trees which are themselves inferred from DNA mutations that accumulate as cells divide and move between tissues. Existing migration-graph inference methods rely on a single pre-estimated phylogeny as a fixed template for combinatorial optimization of migrations under parsimony constraints. This approach fails to account for uncertainty in the phylogeny and can lead to a bias toward overly parsimonious migration histories. Here, we present Bayesian Evolutionary Analysis of Metastasis (BEAM), a fully probabilistic model that jointly describes CRISPR barcode evolution and tissue migration using two conditionally independent continuous-time Markov chains. Implemented in the BEAST framework, BEAM uses MCMC inference to obtain an approximate posterior distribution over migration graphs while marginalizing out phylogenetic uncertainty. Barcode mutations are modeled by adapting a custom rate-matrix for the discrete irreversible state space of CRISPR mutations and tissue migrations are modeled with a flexible tissue transition rate matrix that enables formal hypothesis testing of migration routes and topological features through Bayes-factor model selection. A comprehensive benchmarking evaluation on simulated data demonstrates that BEAM outperforms available methods for migration-graph inference based on CRISPR cell-lineage tracing data, and that parsimony-based methods often oversimplify migration histories. Applications to real prostate-cancer and lung-cancer datasets showcase BEAM's ability to quantify uncertainty, infer complex metastatic scenarios validated by hypothesis testing, and assess the timing of metastatic events across tree and tissue spaces. Work is in progress on an extension to variational phylogenetic inference to improve scalability to larger datasets based on single-cell genomics. By integrating phylogenetic and spatial dynamics, BEAM provides a comprehensive framework for lineage tracing and evolutionary hypothesis testing in cancer metastasis.

LINKPREP™: A RAPID HIGH-RESOLUTION METHOD THAT IMPROVES CHROMATIN CONFORMATION DATA

Ericca Stamper, Cory Padilla, Jonathon Torchia, Daniel Hwang, Mital Bhakta, Lisa Munding

Dovetail Genomics, Part of Cantata Bio, LLC, Scotts Valley, CA

Epigenetic dysregulation, including enhancer hijacking and changes in 3D genome organization, is increasingly recognized as a key driver of disease. Hi-C technology, and its derivatives, have historically been used to query 3D chromatin states. Although significant advances have been made to improve data quality, Hi-C remains challenging due to assay complexity and high project costs. We developed LinkPrep™, a high-resolution, unbiased linked-read NGS assay, enabling sensitive chromatin topology detection with a streamlined, single-day workflow—less than half the time required by Hi-C methods.

We benchmarked the use of LinkPrep for chromatin topology studies using the well-characterized GM12878 cell line. LinkPrep libraries more efficiently captured 50% more topologically associated domains (TADs) and chromatin loops at a 5 kb resolution compared to Hi-C at the same sequencing depth. The unbiased genomic coverage provided more precise feature positioning, demonstrating 98% accuracy, while Hi-C showed only 73% accuracy, often resulting in incorrect loop anchors and misidentification of regulatory elements linked to promoters.

We then applied LinkPrep to the chronic myeloid leukemia (CML) cell line K562 to explore enhancer hijacking. LinkPrep data captured alterations in a complex structural variant event (four breakpoints across three chromosomes) linked to the BCR-ABL fusion and described a novel enhancer network bringing the known leukemogenic gene NUP214 (chr9) into contact with enhancers on chr13. Furthermore, LinkPrep identified a neoTAD event disrupting the therapy resistance gene ARMH1 (chr6) into a new regulatory cluster on chr16. These data provide mechanistic insights on how re-wired regulatory networks can lead to oncogenic expression.

LinkPrep represents a major advancement in 3D genomics technology, overcoming the limitations of Hi-C by offering a simplified and faster workflow, while delivering precise and sensitive detection of 3D genome features.

3,023 HUMAN GENOMES FROM MAINLAND SOUTHEAST ASIA DISCLOSE HIDDEN GENETIC DIVERSITY AND SIGNATURES OF TROPICAL ADAPTATION

Yaoxi He*, Xiaoming Zhang*, Min-sheng Peng*, Yuchun Li*, Kai Liu*, Qingpeng Kong, Yaping Zhang, Bing Su

State Key Laboratory of Genetic Evolution and Animal Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

Mainland Southeast Asia (MSEA) harbors rich ethnic and cultural diversity with a population of nearly 300 million. However, MSEA people are underrepresented in the current genomic database of global populations. Here we present the SEA3K genome dataset (Phase-I), consisting of the deep whole-genome sequencing (WGS) data of 3,023 individuals from 30 MSEA populations, together with the long-read WGS data of 40 representative MSEA individuals. We identified 71.91 million short variants and 43,462 structural variants, nearly half of which are novel. Using the long-read genome assemblies, we generated the first pangenome of MSEA populations, adding 339.36 Mb novel sequences to the current human reference genomes. We observed a high level of genetic heterogeneity across MSEA populations, reflected by the varied combinations of genetic components inherited from ancestral hunter-gatherers and Neolithic farmers. Remarkably, we discovered strong signatures of Darwinian positive selection, suggesting a comprehensive adaptation to tropical environment that explains the distinct phenotypic traits in MSEA populations, including short stature, dark skin pigmentation, wide and snub nose, curly hair and prevalence of thalassemia. Furthermore, we observed different patterns of archaic Denisovans introgression in MSEA populations, supporting the proposal of at least two distinct instances of Denisovan admixture into modern humans in Asia. Importantly, we detected the genomic regions indicating “borrowed” fitness in MSEA populations, due to either Neanderthal or Denisovan adaptive introgressions. The reported multi-million novel genomic variants in MSEA highlight the necessity of studying regional populations that can help answer the key questions on prehistory, genetic adaptation and complex diseases.

GrgPhenoSim - A PHENOTYPE SIMULATOR FOR GENOTYPE REPRESENTATION GRAPHS

Aditya Syam[1], Xinzhu Wei[2]

[1]Cornell University, Computing and Information Science, Ithaca, NY,
[2]Cornell University, Department of Computational Biology, Ithaca, NY

The Genotype Representation Graph (GRG) is a graph representation of whole genome polymorphisms, designed to encode the variant hard-call information in phased whole genomes. It uses a fraction of the space required by present alternatives and can be traversed efficiently, enabling dynamic programming-style algorithms on applications such as GWAS (genome-wide association studies) that run faster on biobank-scale data. To facilitate the adoption of the GRG in statistical genetics, we present GrgPhenoSim, a phenotype simulator for GRGs, suitable for simulating phenotypes on biobank-scale datasets. The simulator contains all the primary functionalities of a phenotype simulator (simulating continuous as well as binary phenotypes), uses standardized output formats, and provides functionalities supporting customized simulations. GrgPhenoSim's accuracy has been verified against external benchmarks like the state-of-the-art graph-based simulator tstrait. Further, its runtime scales linearly with a rise in the number of causal sites and outperforms tstrait in approximately ~99% of cases. The simulator has been released as an open-source library with a GPL license (easily available via the Python Package Index) and can be used for efficient biobank-scale simulations.

CAP-TRAP FULL-LENGTH cDNA SEQUENCING UNCOVERS NOVEL CELL TYPE-SPECIFIC CAPPED TRANSCRIPTS AND DIVERSE CODING AND NON-CODING RNA ISOFORMS

Hazuki Takahashi[1], Hiromi Nishiyori-Sueki[1], Diane Delobel[1], The FANTOM6 Consortium[1], Chi Wai Yip[1], Piero Carninci[1,2]

[1]RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan, [2]Human Technopole, Research Center for Genomics, Milan, Italy

One of the major challenges of transcriptomics consists in correctly identifying and quantifying the diverse set of cell type-specific coding and non-coding (nc)RNAs and regulatory elements present in the genome. The FANTOM consortium previously annotated the cell type-specific promoters and enhancers in mammals, and provided a comprehensive atlas of the 5' end transcription starting sites (TSSs) using the Cap Analysis of Gene Expression (CAGE) method. However, this approach relies on short-read sequencing, which drastically limits the precision at which RNAs can be characterized. Long-read sequencing, particularly in the context of transcriptomics, offers significant advantages in understanding the complexity and diversity of RNA molecules, as it can help identify previously unknown or poorly annotated transcripts by capturing their full-length sequences, as well as find novel alternative splicing events. Here, we harnessed the power of the cap-trap method to develop Cap-trap full length cDNA sequencing (CFC-seq). This protocol enables precise identification of TSSs and full-length transcript models, providing a comprehensive view of the transcriptome. To assemble the long-read data with TSS confidence, we developed the transcript Start-site Aware Long-read Assembler (SALA) and robustly identified ~40,000 novel genes and ~20,000 novel isoforms using 236 million reads derived from 5 cell types. By experimenting with sequencing depth, we found that around 70 million mapped reads in a single bulk RNA sample are sufficient to describe both known and novel capped RNAs, including enhancer (e)RNAs, long ncRNAs, and novel isoforms. Notably, the 24,000 eRNA transcript models original to our CFC-seq datasets reveal that the splicing activity, polyadenylation status and transcript length of eRNAs correlate with enhancer structures such as CpG islands and TATA boxes, an observation which would have been challenging to make using only short-read sequencing. Overall, CFC-seq is a cutting-edge method that offers new opportunities to explore the intricacy and variety of RNAs. The transcripts and gene catalogues generated by this protocol will greatly help us identify the functional molecules in both healthy and diseased individuals, in addition to better sampling the transcriptional diversity present within the human population.

MOUSE AND HUMAN CENTROMERIC AND PERICENTRIC SATELLITES SHARE A COMMON EVOLUTIONARY TRAJECTORY.

Jitendra Thakur, Gitika Chaudhry, Jingyue Chen, Lucy Snipes, Smriti Bahl, Xuan Lin

Emory University, Department of Biology, Atlanta, GA

Satellite DNA makes up ~11% of the mouse genome and is primarily located in centromeric and pericentric regions, which are crucial for chromosome segregation. While comprehensive genomic and epigenomic maps of these regions have been established in the human genome, they are still lacking in the mouse genome. Previous studies suggested that mouse satellite regions are highly homogeneous, unlike human satellites, which form complex higher-order repeat structures and consist of several satellite sub-families. In this study, we used PacBio long-read sequencing, CUT&RUN sequencing, DNA methylation analysis, and RNA sequencing to generate genomic and epigenomic maps of these regions. We find that centromeric core regions are primarily occupied by 120-mer Minor satellites, with other Minor Satellite length variants, 112-mers and 112-64-dimers, localized at centromere-pericentric junctions. Pericentromeres are mainly composed of homogeneous Major satellites, while pericentric-chromosomal junctions contain a higher density of divergent satellites. Additionally, the density of non-satellite repeats increases progressively from centromeres to pericentromeres, and further toward chromosomal arm junctions. Our results indicate that mouse satellites have developed significant variations and begun to form complex patterns and sub-families similar to those observed in humans. Next, we found that DNA methylation levels are lower in centromeres compared to pericentric regions. Interestingly, only a small subset of satellites is transcribed into RNA, particularly regions exhibiting lower DNA methylation density. Furthermore, we found that 120-mer Minor satellites in the core centromere are highly enriched with CENP-A, while the 112-mers and 112-64-dimers show lower CENP-A enrichment. Homogeneous Major satellites are more enriched with H3K9me3 heterochromatin, whereas divergent Major satellites are preferentially associated with H3K27me3. Our key findings and characterization of the genomic and epigenomic landscape of mouse centromeric and pericentric regions have major implications for satellite biology and completing the mouse telomere-to-telomere (T2T) assembly annotations.

ELECTRONIC GENOME MAPPING FOR VERIFYING SOMATIC STRUCTURAL VARIANTS

John F Thompson[1], Lindsay Schneider[1], Reger Mikaeel[1], William Jastromb[1], Kaylee Mathews[1], Xu Tan[2], Michael Kaiser[2]

[1]Nabsys LLC, Applications, Providence, RI, [2]Nabsys LLC, Informatics, Providence, RI

Structural variants (SVs) play a crucial role in both the onset and progression of cancer, as well as in the development of drug resistance over time. Despite their importance, identifying and validating SVs remains a significant challenge due to the limitations of current sequencing and genomic technologies. Even when multiple platforms and analysis algorithms are utilized, the accuracy and reliability of SV detection often remain uncertain. While variants smaller than 50 base pairs (bp) are usually identified with confidence, larger SVs—those exceeding 300 bp—become increasingly difficult to characterize as their size grows. The ability to detect these larger SVs depends on their size, type, and genomic context. Cancer genomes add further complexity, often exhibiting abnormal chromosome counts, numerous somatic variants with varying allele frequencies, and intricate rearrangements caused by events like chromothripsis. These challenges underscore the urgent need for novel methods to reliably detect and validate SVs, which are critical for advancing our understanding of cancer biology.

We examined high-confidence structural variants (hcSVs) in tumor cells using electronic genome mapping (EGM). Previous work by Talsania et al. (2022) identified 1,788 hcSVs through a multi-platform approach. They further analyzed these variants using multiple validation methods, yet fewer than 25% were confirmed, highlighting persistent challenges in SV detection. Given the complexity of structural variants in cancer genomes, where abnormal chromosomal structures and varying allele frequencies complicate analysis, we leveraged EGM to independently verify hcSVs across a wide size range. Our findings demonstrate EGM's utility in confirming SVs that were difficult to validate with conventional methods, reinforcing its role as a complementary technology for SV analysis.

EGM demonstrates its capacity to verify somatic SVs from tumor-derived cell lines across a wide size range, spanning from 300 base pairs (bp) to several megabases (Mb). EGM serves as a valuable complementary technology for confirming SVs across diverse size ranges, verifying insertions and deletions larger than 4 kb at a rate similar to long-read sequencing technologies. Its straightforward workflow and cost-effectiveness make it an appealing option today. Future advancements in algorithms and data quality are expected to enhance EGM's performance further, solidifying its role as a crucial tool in cancer genomics research.

FUNCTIONAL AND EPIGENETIC CHARACTERIZATION OF AFRICAN PAN-GENOME CONTIGS: IMPLICATIONS FOR REFERENCE GENOME BIAS AND HUMAN GENOMIC DIVERSITY

Rachel Martini[1], Abdulfatai Tijjani[2], Kyriaki Founta[3], Daniel Cha[4], Sebastian Maurice[5], Jason White[1], Melissa Davies[1], Nyasha Chambwe[2]

[1]Institute of Translational Genomic Medicine, Morehouse School of Medicine, Atlanta, GA, [2]Feinstein Institutes for Medical Research, Northwell Health, New York, NY, [3]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Northwell Health, New York, NY, [4]Carnegie Mellon University, Pittsburg, PA, [5]City College of New York, New York, NY

The African Pan-Genome (APG) Project has characterized African-specific sequences missing from widely utilized reference genomes, such as GRCh38. This has raised concerns about reference genome bias in interpreting sequencing results, particularly from diverse populations such as those of African descent. We perform a comprehensive analysis of APG contigs by aligning them to recent long-read reference assemblies and evaluating their functional implications.

We aligned 124,240 APG contigs to the Telomere-to-Telomere (T2T-CHM13) genome and 47 linear assemblies from the Human Pangenome Reference Consortium (HPRC) using BWA-MEM. Approximately 80% of the APG contigs mapped successfully to the T2T genome, mainly in repetitive and complex regions. When aligned to HPRC assemblies, 83% showed nearly perfect alignment, with the highest number of population-specific contigs (8,131) linked to African genetic ancestry.

Additionally, to investigate the functional relevance of the APG-aligned sequences, we examined their epigenetic potential by predicting novel CpG islands using EMBOSS CpGPlot. This analysis identified 8,353 CpG islands across the contigs, roughly 5% of which (6,352) contain putative regulatory elements. Further annotation revealed that 3,061 contigs overlap with known genes, including 161 protein-coding genes and 343 non-coding genes. Many of these genes are implicated in biological processes related to cellular immunity, synaptic function, and intracellular signaling.

These results highlight the limitations of GRCh38 and T2T, which primarily represent Eurocentric haplotypes, and emphasize the importance of incorporating diverse populations into reference genome frameworks. Our findings demonstrate the value of African pan-genome sequences in improving genomic representation, particularly in regions previously unresolved in linear references. As part of ongoing efforts to enhance genome equity, we are investigating the human graph genome as a more inclusive alternative to capture human genetic diversity better. This study contributes to the broader initiative of refining reference genomes, ultimately promoting equity in genomic studies and precision medicine.

DISSECTING FUNCTIONAL ELEMENTS IN GIANT GENOMES USING THE FIRST GENERATION OF HIGH-QUALITY SALAMANDER ASSEMBLIES

Nataliya Timoshevskaya*[1,2], S. Randal Voss*[2,3], Jeramiah J Smith*[1]

[1]University of Kentucky, Biology, Lexington, KY, [2]University of Kentucky, Department of Neuroscience, Spinal Cord and Brain Injury Research Center, Lexington, KY, [3]University of Kentucky, Ambystoma Genetic Stock Center, Lexington, KY

Salamanders serve as important models for studying vertebrate evolution (from the perspective of amphibians), genome size evolution (owing to their exceptionally large genomes), and large-scale epigenetic reprogramming events (during limb regeneration and metamorphosis). Advances in sequencing and assembly methods have dramatically improved our ability to work with these large genomes and dissect functional elements that are interspersed across large intergenic and intronic regions. Here we report a highly accurate single haplotype assembly for the 32 Gb axolotl (Ambystoma mexicanum) genome. This assembly was generated using an F1 hybrid between axolotls and a divergent relative, the tiger salamander. The resulting assembly yielded highly contiguous reference genomes for both species and provides a platform for discovering regulatory elements that define unique and conserved aspects of salamander biology. Comparisons with other recent salamander species (including three from the Darwin Tree of Life Project) reveal strong conservation of large-scale karyotypic structure over the last 150 million years of evolution and at finer scales multispecies alignments permit the identification of distinct highly conserved elements within massively expanded noncoding regions. Cross referencing these conservation tracks with other marks of active chromatin identifies subclasses of conserved elements that correspond to classically defined regulatory classes (enhancers and proximate regulatory elements) and classes that are not predicted by other existing chromatin datasets. These analyses suggest that comparative evolutionary studies among these large genomes are likely to identify distinct functional elements that have thus far escaped detection via chromatin-based analyses.

INSIGHTS INTO THE GENETIC DIVERSITY AND ADAPTATION MECHANISMS OF THE ANDEAN BLUEBERRY TO EXTREME ENVIRONMENTS USING GENOMIC APPROACHES.

Maria de Lourdes Torres[1], Chelsea Specht[2], Milton Gordillo[1], Jacob Landis[2], Martina Albuja[1]

[1]Laboratorio de Biotecnología Vegetal, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito (USFQ), Quito, Ecuador, [2]School of Integrative Plant Science, Plant Biology Section, Cornell University, Ithaca, NY

The Andean blueberry (*Vaccinium floribundum* Kunth) is a perennial shrub native to the Andean region, thriving across a broad altitudinal range (1.600–4.500 masl). In Ecuador, this species has successfully adapted to the extreme conditions of the Andean paramo, a high-altitude ecosystem (>4.000 masl) characterized by low temperatures, intense UV radiation, and high humidity. Understanding the genetic mechanisms underlying these adaptations can provide insights into plant resilience to harsh environments. The objective of this study was to investigate the genomic adaptations of *V. floribundum t*o high-altitude conditions. To this end, we genotyped 40 individuals from the northern and southern Ecuadorian highlands, collected from high-altitude (>4.000 masl) and low-altitude (<3.000 masl) locations, using short-read sequencing with AVITI Technology (Element Biosciences). Genetic diversity analyses based on SNP data revealed overall low heterozygosity values, ranging from 0.001 to 0.009. Notably, high-altitude sites (Napalé and Quilotoa) exhibited the lowest heterozygosity (~0.001), while low-altitude sites (Yangana and Sevilla de Oro) showed higher heterozygosity (0.0057 and 0.0068, respectively), suggesting elevation influences genetic diversity and potentially adaptive capacity. Principal component analysis (PCA) and STRUCTURE analysis indicated that genetic diversity is shaped by both elevation and geographic location. High-altitude populations were genetically distinct from low-altitude ones, while within low-altitude populations, northern and southern groups exhibited differentiation. The STRUCTURE analysis (K = 3) suggested limited gene flow between elevations, likely due to adaptive pressures or geographic barriers, while some admixture hinted at historical gene exchange. The identified SNPs not only provided insights into genetic diversity but also serve as a foundation for detecting candidate loci associated with high-altitude adaptation. Unravelling the genomic responses of plants to extreme environments can help identify genes related to climate resilience, elucidate the genetic basis of complex traits, and predict plant responses to environmental changes.

Keywords: Andean blueberry, population genomics, molecular ecology, environmental adaptations

ASSESSING CELLULAR CONTEXTS OF TYPE 2 DIABETES-ASSOCIATED VARIANTS AT SCALE

Adelaide Tovar[1], Amy Etheridge[2], Romy Kursawe[3], Kirsten Nishino[1], Jonathan D Rosen[2], Ziwei Chen[4], Daniel Dicorpo[5], James Meigs[6], Alisa Manning[7], Anshul Kundaje[4], Kimberly Lorenz[8], Benjamin F Voight[8], Sarah Schoenrock[2], Ryan Tewhey[3], Michael Stitzel[3], Karen Mohlke[2], Jacob O Kitzman[1], Stephen C Parker[1]

[1]University of Michigan, Ann Arbor, MI, [2]The University of North Carolina at Chapel Hill, Chapel Hill, NC, [3]The Jackson Laboratory, Bar Harbor, ME, [4]Stanford University, Stanford, CA, [5]Boston University, Boston, MA, [6]Massachusetts General Hospital, Boston, MA, [7]Broad Institute of Harvard and MIT, Cambridge, MA, [8]University of Pennsylvania, Philadelphia, PA

Type 2 diabetes (T2D) is a common metabolic disorder characterized by dysregulation of glucose metabolism. Genome-wide association studies have identified >660 loci associated with T2D. While much of this genetic risk is predicted to act through insulin-producing pancreatic islets, heritability is also distributed across regulatory regions active in other important metabolic tissues including adipose, liver, and skeletal muscle. Beyond specific tissues of action, there is mounting evidence that genetic effects on gene regulation are influenced by environmental context. High-throughput variant characterization assays such as massively parallel reporter assays (MPRAs) represent a tractable method to survey context-dependent regulatory activity of disease-associated variants at scale. Previously, we demonstrated widespread condition-specific allelic bias in a small MPRA library delivered to the LHCN-M2 human skeletal muscle myoblast cell line across four relevant states: (1) undifferentiated, or differentiated with (2) basal media, (3) AICAR to mimic exercise or (4) palmitate to induce insulin resistance. Specifically, 41.5% of tested variants (122/295) displayed allelic bias in only a single condition (*FDR* < 0.05) compared to just 7 variants across all conditions. We have since constructed an MPRA library to assess regulatory activities of >23k common variants in high linkage with 667 independent T2D association signals ($R^2 > 0.7$) and a set of ~1.5k TOPMed-contributed rare variants, comprising one of the largest single disease-associated MPRA libraries to date. Given previous evidence of enhancer-promoter regulatory specificity, we generated parallel versions of this library with several housekeeping and tissue-specific promoters. We delivered this library paired with either the potent synthetic promoter SCP1 or the skeletal muscle-specific *MYBPC2* promoter to differentiated LHCN-M2 skeletal muscle myoblasts (n = 4 for both promoter contexts). As an example, the multi-ancestry T2D GWAS signal variant rs146716733 displayed stronger allelic bias when paired with the ^{MYBPC2} promoter compared to SCP1. This variant is located in the third intron of *ARID5B*, an important global regulator of metabolism. The C risk allele displays weaker enhancer activity compared to the T non-risk allele, and the risk allele is highly prevalent across diverse ancestry groups (~85-95%). We have since delivered this library to three other cell types (hWAT adipocytes, EndoC-βH3 beta cells, and HepG2 hepatocytes) in basal state and are currently performing integration with other genomic datasets to annotate all ~25k common and rare variants and uncover novel disease mechanisms.

SCiMS: SEX CALLING IN METAGENOMIC SEQUENCES

Hanh N Tran[1,2], Kobie J Kirven[2,3], Emily R Davenport[1,2]

[1]Department of Biology, Pennsylvania State University, University Park, PA, [2]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, [3]Department of Chemistry, Pennsylvania State University, University Park, PA

Background: In microbiome studies, the diversity and abundance of microbial communities are closely associated with host characteristics, such as sex. Differences in sex hormones, other sex-related physiology, and sex-stratified behaviors can lead to different microbial profiles between males and females. Thus, host sex is an important factor to consider in microbiome analysis. However, sex information is sometimes not available or compromised by sample processing errors. Existing metagenomic tools lack the ability to determine sex of the host from sequencing data alone. Recognizing this gap, we aimed to create a bioinformatic tool that leverages the unbiased approach of metagenomic sequencing—the process of capturing all DNA present in a sample, including both microbial and host-derived genetic material—to infer host sex from metagenomic reads.

Results: Here, we introduce SCiMS: Sex Calling in Metagenomic Sequences. SCiMS combines a robust supervised learning algorithm with Bayesian probabilistic methods to accurately determine host sex from metagenomic sequencing data. It works by analyzing the reads that align with host's sex chromosomes—such as the X and Y chromosomes in humans—and calculating two key ratios: (1) the coverage of X chromosome relative to autosomes and (2) the coverage of the Y chromosome relative to the total sex chromosomes. These ratios are then compared against sex-specific multivariate Gaussian distributions. By assessing how closely each sample's ratios fit the male or female distribution, SCiMS confidently predicts host's sex, even when there is limited host reads present in the metagenome. We validated the performance of SCiMS using both simulated and real datasets. In simulations, we generated 24,000 samples with various host read depths ranging from 150 to 1,000,000 reads. Among samples with definitive predictions, SCiMS consistently achieved over 85% accuracy at a depth of 150 host reads, exceeded 94% accuracy at 250 – 450 host reads, and reached 100% at 1000 host reads and above. The proportion of uncertain calls dropped from 50% at 150 reads to 10% at 1,000 reads and reached zero at 5,000 reads or higher. Next, we applied SCiMS to real metagenomic datasets, including 1450 human samples from the Human Microbiome Project (HMP)s, 111 mouse cecal samples, and 94 chicken cecal samples. Among samples with definitive predictions, SCiMS accurately determined the sex of 98.7% of HMP samples, 100% of mouse fecal samples, and 77.4% of the chicken cecal samples. These results confirm SCiMS' accuracy and highlight its applicability across diverse host organisms and sampling sites.

Conclusions: We provide SCiMS as a user-friendly and effective tool to call host sex from metagenomic data in organisms with a heterogametic sex system. By accurately predicting host sex even under low host DNA coverage, SCiMS provides a reliable approach for data quality in metagenomic studies. The software is freely available at github.com/davenport-lab/SCiMS.

DETECTING GERMLINE MUTATIONS IN LOW-COVERAGE SEQUENCE DATA USING PEDIGREES

Georgia Tsambos[1], Daniel Seidman[2], Kelley Harris[1], Nancy Chen[2]

[1]University of Washington, Department of Genome Sciences, Seattle, WA,
[2]University of Rochester, Department of Biology, Rochester, NY

De novo mutations in the germline (DNMs) can have significant consequences for the individuals, populations and species that they arise in because of their contribution to genetic diversity, as well as their links with disease, cancer and ageing. It is therefore important to understand the factors that influence their prevalence. Despite their origin in repair and replication processes shared by all cellular organisms, DNM rates differ on both long and short evolutionary timescales. Understanding how mutation rates vary in response to inbreeding, mutator alleles, parental age, population size and other factors requires methods that can sample mutation rates broadly within many different species.

However, conventional DNM calling methods are limited by their dependence on high-coverage sequencing. A high sequencing depth is needed to confidently determine that a detected variant is indeed in the individual's germline and not in the soma, or simply an error. This requirement is prohibitively costly for large samples, especially in non-model organisms that lack a well-annotated reference genome. However, a growing number of datasets provide information from each sample individual's extended pedigree, which can grow over time as new generations are recruited into the study. In this work, we show that this added information can improve our power to detect DNMs by compensating for the low sequencing coverage of any single individual. The structure of the pedigree gives an informative constraint about which individuals might conceivably share a DNM, a constraint that can be further refined with inferred tracts of identity-by-descent (IBD).

Our preliminary findings suggest that even with very low sequencing coverage (3x) and flow-on errors in variant calling and phasing, conventional IBD inference methods are robust enough to enable this analysis. To show the power of this method, we compare DNMs detected in this setting to a set of "gold standard" DNMs called with conventional methods on high coverage trios from the CEPH family panel. Additional benchmarking illustrates how the statistical power varies according to sequencing depth and pedigree size, shape, and missingness. By facilitating the study of DNMs in a wider range of datasets, we hope our method will reveal new insights about the complex causes and consequences of germline mutation.

EXTENSIVE ADAR-MEDIATED RNA EDITING SHAPES KRAB-ZFP DIVERSITY THROUGH DYNAMIC MODIFICATION OF DNA-BINDING DOMAINS

Itamar Twersky, Erez Y Levanon

Bar-Ilan University, Mina and Everard Goodman Faculty of Life Sciences, Ramat Gan, Israel

Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) comprise the largest family of transcriptional regulators in mammals, with their DNA-binding specificity determined by arrays of zinc finger domains. Here, through comprehensive analysis of human transcriptomes, we reveal an unexpected and innovative finding: while coding sequence editing is generally rare in humans, KRAB-ZFPs emerge as a striking exception, showing extensive adenosine-to-inosine (A-to-I) RNA editing catalyzed by ADAR enzymes. Using transcriptome-wide analysis of human tissues, we discovered widespread RNA editing within KRAB-ZFP coding sequences, specifically enriched in their tandem DNA-binding zinc-finger domains. This editing is particularly pronounced in neural tissues, suggesting tissue-specific regulation of KRAB-ZFP function. Notably, human KRAB-ZFP genes show a striking bias toward inverted genomic arrangements (4:1 ratio over tandem orientations), facilitating RNA duplex formation and subsequent ADAR-mediated editing.
The combinatorial nature of multiple editing sites within individual KRAB-ZFP transcripts enables the generation of up to 2^n distinct protein variants, each potentially harboring unique DNA-binding specificities. As a possible functional outcome, we propose a model where this expanded repertoire of binding specificities could enhance recognition of diverse DNA sequences, including mutated endogenous retroelements (EREs), potentially contributing to genomic defense mechanisms. Our findings illuminate a novel regulatory mechanism where RNA editing expands the KRAB-ZFP functional repertoire and establishes RNA editing as an unexpected major contributor to transcription factor diversity in humans.

A NOVEL FRAMEWORK FOR BUILDING CELL-SPECIFIC GENE REGULATORY NETWORKS WITH SINGLE-CELL MULTI-OMICS

Yasin Uzun[1,2,3], Eric Moeller[1], Karamveer Karamveer[1], Hannah Valensi[1]

[1]Penn State College of Medicine, Department of Pediatrics, Hershey, PA, [2]Penn State College of Medicine, Department of Molecular and Precision Medicine, Hershey, PA, [3]Penn State, Cancer Institute, Hershey, PA

Gene regulatory networks (GRNs) define the gene expression programs that drive cellular functions and differentiation. They are essential for modeling and understanding biological processes related to development and disease. The advent of **single-cell multi-omics sequencing** technologies, which enable the simultaneous profiling of transcriptomic and epigenomic features within the same cells, has made it possible to construct GRNs with greater modeling power and accuracy.

Recently, multiple methods have been introduced for constructing GRNs using single-cell multi-omics data, leveraging various inference approaches. However, these studies utilize diverse benchmarking datasets and performance metrics, making it difficult to objectively compare available methods. Consequently, an unbiased benchmarking of GRN inference methods that use single-cell multi-omics data remains lacking.

To address this gap, we first built a **publicly accessible, comprehensive single-cell GRN repository**. This repository includes reference networks built using transcription factor (TF)-DNA interaction datasets and functional perturbation studies in addition to a diverse set of single-cell multi-omics datasets covering various cell types that match those of the reference networks.

Using these curated datasets, we conducted an unbiased benchmarking of state-of-the-art single-cell multi-omics GRN inference methods. To assess accuracy, we compared the inferred networks against the reference networks using established metrics. We also evaluated the stability of each algorithm by inferring networks with subsampled cell sets. Furthermore, we assessed each method's scalability in terms of its ability to infer networks from large datasets. Our benchmarking analysis highlighted key limitations of existing GRN inference methods for single-cell multi-omics data.

To overcome these limitations, we developed a novel machine learning-based algorithm for inference of GRNs from single-cell multi-omics data. Existing methodologies typically approach GRN inference as a regression problem, where the expression of a target gene is treated as a dependent variable predicted by the expression levels of its regulators. In contrast, we formulated the task of network inference as a binary classification problem, in which the goal is to determine whether an interaction (edge) exists between regulator-target gene pairs, using both gene expression and chromatin accessibility data.

We tested our novel methodology using existing single-cell multi-omics datasets and reference networks with matching cell types. Compared to existing methods, our approach provided **substantial improvement in the accuracy of inferred networks**. This novel framework provides a robust and reliable approach for accurately inferring gene regulatory networks from single-cell multi-omics data.

NON-CODING MUTATIONS IN DIFFUSE LARGE B-CELL LYMPHOMA: A CROSS-SPECIES STUDY

Anna D van der Heiden[1,2], Suvi Mäkeläinen[1,2], Raphaela Pensch[1,2], Sergey V Kozyrev[1,2], Sophie Agger[3], Cheryl London[4], Jaime F Modiano[5], Karin Forsberg Nilsson[6], Maja L Arendt[1,3], Kerstin Lindblad-Toh[1,2,7]

[1]Uppsala University, Department of Medical Biochemistry and Microbiology, Uppsala, Sweden, [2]Uppsala University, SciLifeLab, Uppsala, Sweden, [3]University of Copenhagen, Department of Veterinary Clinical Sciences, Copenhagen, Denmark, [4]Tufts University, Cummings School of Veterinary Medicine, North Grafton, MA, [5]University of Minnesota, Masonic Cancer Center, Minneapolis, MN, [6]Uppsala University, Department of Immunology, Genetics and Pathology, Uppsala, Sweden, [7]Broad Institute of MIT and Harvard, Cambridge, MA

Diffuse large B-cell lymphoma (DLBCL) is an aggressive lymphoma subtype affecting both dogs and humans. While research has advanced our understanding of this disease, most studies have focused on the protein-coding genome. However, emerging evidence suggests that non-coding mutations also play a crucial role in cancer. In this study, we aim to explore the non-coding genome, identifying novel driver mutations, genes, and pathways linked to DLBCL. Dogs were chosen as model due to their biological closeness to humans and clinical parallels between human and canine DLBCL. We used tumor-normal WGS data from 72 canine and 39 human patients and applied phyloP scores–a measure of evolutionary constraint–to prioritize non-coding constraint mutations (NCCMs). This approach hypothesizes that highly constrained regions are likely functionally important, and mutations in such regions may disrupt gene expression, contributing to oncogenesis and cancer progression. Our analysis identified 85 and 219 genes enriched with NCCMs in dogs and humans, respectively. Of these, 27 were shared, including four (*BCL6, BCL7A, POU2AF1, RUNX1T1*) listed in COSMIC as cancer genes linked to hematologic neoplasms. In this shared gene set, coding mutations were rare, with NCCMs predominating in 55.6% of genes. Notably, 15 genes showed NCCMs clustering in regions of high transcriptional activity and putative super-enhancers. Among these, *BACH2* emerged as an intriguing candidate, given its critical role in B-cell differentiation and its numerous mutational hotspots within intronic and upstream regions. Further investigation through *in-silico* transcription factor binding affinity analysis revealed a hotspot that significantly reduced transcription factor TFAP4 binding. These findings demonstrate that leveraging evolutionary constraint and cross-species analysis can uncover novel candidate genes affected by mutations in regulatory regions. Moving forward, we aim to investigate these candidate NCCMs through wet-lab validation, with the ultimate goal of identifying potential biomarkers and therapeutic targets.

INDIGENE (GENEtics of INDIviduality): RNA-SEQ ANALYSIS OF SELECTED TISSUES AND DIFFERENT ENVIRONMENTS IN MEDAKA FISH

Christina Vasilopoulou[1], Tomas Fitzgerald[1], Ian Brettell[1], Adrien Leger[1], Nadeshda Wolf[2], Natalja Kusminski[2], Jack Monahan[1], Carl Barton[1], Cathrin Herder[2], Narendar Aadepu[3], Jakob Gierten[3], Clara Becker[3], Omar T Hammouda[3], Eva Hasel[3], Colin Lischik[3], Katharina Lust[3], Natalia Sokolova[3], Risa Suzuki[3], Erika Tsingos[3], Tinatini Tavhelidse[3], Thomas Thumberger[3], Philip Watson[3], Bettina Welz[3], Nadia Khouja[2], Kiyoshi Naruse[4], Ewan Birney[1], Joachim Wittbrodt[3], Felix Loosli[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, [2]Institute of Biological and Chemical Systems, Biological Information Processing (IBCS-BIP), Karlsruhe Institute of Technology, Karlsruhe, Germany, [3]Centre for Organismal Studies, Heidelberg University, Campus Im Neuenheimer Feld, Heidelberg, Germany, [4]National Institute for Basic Biology, Laboratory of Bioresources, Okazaki, Japan

One main challenge in genetics is to decipher the relationship between natural genetic variation and phenotypic traits while accounting for environmental factors. The Japanese rice-paddy fish, Medaka (*Oryzias latipes*) is an established genetic model system with high tolerance to inbreeding from the wild. In the past decade, the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel was established, comprising 80 unique near-isogenic inbred lines from the original wild population. Medaka fish is an ideal model organism to investigate gene-by-environment interactions (GxE), enabling experiments where both the environment and the genome are controlled.

One main aim of this project is to investigate GxE effects in different tissues and between environmental contrasts of a subset of the MIKK panel at the transcriptome level. The dataset includes 1317 Medaka fish samples derived from six tissues (brain, eyes, gills, heart, liver and gonads), both sexes (male and female), and exposed to two laboratory environments, summer and winter.

We carried out extensive quality assurance steps to the RNA-seq dataset, including the thorough inspection of the nf-core/rnaseq pipeline multiQC metrics, and the development of a refined and comprehensive SNP check strategy to identify potential sample swaps. We explored tissue-, sex- and environment-specific patterns using the nf-core/differentialabundance pipeline and MOFA (Multi-Omics Factor Analysis), for comprehensive exploratory analyses and interpretable low-dimensional representation of complex data in terms of a few latent factors, respectively. Lastly, we show the genotypic patterns associated with expression profiles across multiple tissues and different environments in Medaka fish, using a newly developed scalable and flexible eQTL (expression-based quantitative trait) Nextflow pipeline based on the recently published birneylab/flexlmm software.

EXTENSIVE STRUCTURAL VARIATION AND LONGEVITY-ASSOCIATED ADAPTATIONS IN NEARCTIC *MYOTIS* BATS

Juan M Vazquez[1], Mary E Lauterbur[2,3], David Bahry[4], Meaghan Birkemeier[4], Eric Chen[5], Petar Pajic[4], Sarah Kassem[4], Omer Gokcumen[4], Michael Singer[1], Sarah Villa[1], Saba Mottaghinia[6], Carine Rey[6], Sarah Maesen[6], Michael Buchalski[7], Lucie Etienne[6], David Enard[2], Vincent J Lynch[4], Peter Sudmant[1]

[1]University of California, Berkeley, Integrative Biology, Berkeley, CA, [2]University of Arizona, Ecology & Evolutionary Biology, Tucson, AZ, [3]University of Vermont, Biology, Burlington, VT, [4]University of Buffalo, Biological Sciences, Buffalo, NY, [5]Harvard University, Molecular and Cell Biology, Cambridge, MA, [6]École Normale Supérieure de Lyon, Centre International de Recherche en Infectiologie, Lyon, France, [7]California Department of Fish and Wildlife, Sacramento, CA

The 100-fold difference in lifespans across mammalian species provides a rich trove of novel pathways and mechanisms underlying exceptional differences in longevity-associated traits such as DNA damage repair. Among mammals, bast of the genus Myotis exhibit some of the most extreme variations seen in lifespans. To study the evolution of longevity-associated traits, we generated primary cell lines for over 28 species of bats, and assembled near-complete diploid genome assemblies for 8 Myotis species. Using genome-wide screens of positive selection, analyses of structural variation, and functional experiments in primary cell lines, we identify new patterns of adaptation contributing to longevity, cancer resistance, and viral interactions in bats. We find that Myotis bats have some of the most significant variation in cancer risk across mammals and demonstrate a unique DNA damage response in primary cells of the long-lived M. lucifugus. We also find evidence of extensive structural variation both within and between species, including in genes central to stress response and viral immunity. Together, our results demonstrate how genomics and primary cells derived from diverse taxa uncover the molecular bases of extreme adaptations in non-model organisms.

IMMUNE PLEIOTROPY AND EVOLUTION IN THE RESPONSE TO *YERSINIA PESTIS*

Tauras Vilgalys, Anne Dumaine, Mari Shiratori, Luis Barreiro

University of Chicago, Genetic Medicine, Chicago, IL

The human immune system is under strong selective pressure from both infectious and non-communicable diseases, each of which requires a unique protective response. Competing selective pressures (antagonistic pleiotropy) may explain why natural selection has maintained risk alleles in modern populations. However, immunogenic pleiotropy is poorly understood. Here, we use *Yersinia pestis*, the causative agent of plague, to investigate the pleiotropic effects of genetic variants that affect the immune response to a historic pathogen.

We stimulated peripheral blood mononuclear cells (PBMCs) from 90 individuals with live, fully virulent *Y. pestis*, and characterized the response to infection using single-cell RNA sequencing. We mapped expression quantitative trait loci (eQTL) and detect 2,388 genes with at least one eQTL. 32% of eQTL had different effects in the stimulated and unstimulated conditions (i.e., immune response eQTL), including 383 variants that only affect gene expression levels in *Y. pestis* infected cells. Immune response eQTL are systematically different than other eQTL, with greater cell-type specificity and affecting genes under greater selective constraint. Immune response eQTL also have pleiotropic consequences for immune function. They colocalize with disease variants for inflammatory and autoimmune disorders and are up to 4x as likely to affect the expression response to other infectious diseases. Finally, we show that immune eQTL are more likely to be under recent positive selection than other variants.

Together, our results suggest that genetic variation shaping the immune response to *Y. pestis* also impacts the response to other infectious agents and disease risk in modern individuals. These findings are consistent with widespread pleiotropy in the immune response, shaping the risk for infectious and non-communicable diseases.

INVESTIGATING HOW POISON EXONS MODULATE ALTERNATIVE SPLICING TO SHAPE TRANSCRIPTOMES IN PLURIPOTENCY AND DIFFERENTIATION.

Isha A Walawalkar[1,2], Nathan Leclair[1,2], Mattia Brugiolo[1], Olga Anczukow[1,2]

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

Poison exons (*PE*s) are non-coding alternative exons that introduce a premature termination codon when spliced in, targeting the transcript for nonsense-mediated decay and thereby reducing protein levels. *PE*s that are genomically ultraconserved throughout evolution are enriched within RNA-binding proteins, many of which function as splicing factors (SFs). *PE*s in splicing factors (*SF-PE*s) have been shown to regulate expression of their own SF and in some cases, cross-regulate expression of other SFs within a finely tuned network. Since many SFs are critical for organismal development and *SF-PE*s are key regulators of SF levels, *SF-PE*s likely play a key role in shaping cellular and tissue identity. Indeed, work from us and others demonstrates that several *SF-PE*s are required for pluripotent and cancer cell survival. However, the regulation of *SF-PE*s through alternative splicing in pluripotency remains poorly understood. Here, we are characterizing pluripotency-associated *SF-PE* networks by leveraging publicly available long-read RNA-sequencing data from pre-implantation human and mouse embryos. To further investigate the role of *SF-PE*s in safeguarding pluripotency and driving differentiation, we are developing mouse models with embryo-wide as well as lineage-specific conditional *SF-PE* knockouts (*SF-PE* cKO) to examine developmental phenotypes. In parallel, we are assessing the molecular and functional phenotypes of *SF-PE* cKO upon differentiation into mature cell types. Our findings will elucidate how *SF-PE*s modulate transcriptomes, safeguard cell pluripotency, and drive differentiation. Furthermore, we aim to identify RNA isoforms that could be clinically targeted to drive cells towards stemness or differentiation.

COMPRESSIVE PANGENOMICS USING MUTATION-ANNOTATED NETWORKS

Sumit Walia[1], Harsh Motwani[2], Kyle Smith[3], Yu-Hsiang Tseng[1], Russell Corbett-Detig[4], Yatish Turakhia[1]

[1]University of California San Diego, Department of Electrical and Computer Engineeringl, San Diego, CA, [2]University of California San Diego, Department of Computer Science and Engineering, San Diego, CA, [3]University of California San Diego, Department of Biological Sciences, San Diego, CA, [4]University of California San Diego, Department of Electrical and Computer Engineering, San Diego, CA, [5]University of California Santa Cruz, Department of Biomolecular Engineering, Santa Cruz, CA, [6]University of California San Diego, Department of Electrical and Computer Engineering, San Diego, CA

Pangenomics is an emerging field that focuses on studying a collection of genomes of a same species (i.e., the pangenome) rather than a single reference genome, to overcome reference bias and enable the exploration of genetic diversity within species. Fueled by advancements in sequencing technologies, pangenomics has garnered significant attention for its potential in diverse fields, including epidemiology, metagenomics, evolutionary biology, and medicine. For instance, during the COVID-19 pandemic, over 16 million SARS-CoV-2 genomes were sequenced globally, providing invaluable data for tracking variants, monitoring their spread, assessing their fitness, and guiding vaccine development.

As the scope of pangenomics expands, the future pangenomics applications will require analyzing large and ever-growing pangenomes, containing hundreds of thousands to millions of genome sequences of the same species. Therefore, the choice of data representation is a key determinant of the scope, as well as the computational and memory performance of pangenomic analyses. Current pangenome formats, while capable of storing genetic variations across multiple genomes, fail to capture the shared evolutionary and mutational histories among them, thereby limiting their applications. They are also inefficient for storage and therefore face significant scaling challenges.

We propose PanMAN (Pangenome Mutation-Annotated Network), a novel data structure that is information-wise richer than all existing pangenome formats – in addition to representing the alignment and genetic variation in a collection of genomes, PanMAN represents the shared mutational and evolutionary histories inferred between those genomes. Through the use of “evolutionary compression”, PanMAN achieves 3.5 to 1391-fold compression over other pangenomic formats on various microbial datasets. PanMAN's relative performance generally improves with larger datasets and it is compatible with any method for inferring phylogenies and ancestral nucleotide states. Using SARS-CoV-2 as a case study, we show that PanMAN offers a detailed and accurate portrayal of the pathogen's evolutionary and mutational history, facilitating the discovery of new biological insights. To demonstrate scalability, we constructed a PanMAN containing 8 million SARS-CoV-2 sequences, the largest pangenome in terms of number of sequences, with a file size of just 39MB. We also propose panmanUtils, a software toolkit that supports common pangenomic analyses and makes PanMANs interoperable with existing tools and formats. PanMANs are poised to enhance the scale, speed, resolution, and overall scope of pangenomic analyses and data sharing.

ESTIMATING RECENT POPULATION SPLIT TIMES IN NON-PANMICTIC POPULATIONS

Jeff Wall

Oregon Health and Science University, Division of Genetics, Beaverton, OR

Standard models for estimating divergence times between populations suffer from several modeling shortfalls, including an inability to distinguish between genetic drift caused by small population sizes or longer divergence times, and the reliance on an unrealistic assumption of random mating within a population. Here we present a new method for estimating recent divergence times that utilizes information in shared identity-by-descent (IBD) segments within and between populations, and which is insensitive to assumptions about random mating or population size. We apply our approach to whole-genome sequence data from Indian caste groups sampled from the Birbhum district in West Bengal (as part of the GenomeAsia 100K project), and estimate average divergence times of <50 generations between different self-identified Hindu caste groups. Comparisons between Hindu caste groups and other (e.g., Austroasiatic language-speaking but geographically co-incident) groups yield slightly older divergence time estimates, and highlight the complex structure of population assimilation/replacement that highlights most of our species' history.

AIRQTL DISSECTS CELL STATE-SPECIFIC CAUSAL GENE REGULATORY NETWORKS WITH EFFICIENT SINGLE-CELL eQTL MAPPING

Lingfei Wang

University of Massachusetts Chan Medical School, Department of Genomics and Computational Biology, Worcester, MA

Single-cell expression quantitative trait loci (sceQTL) mapping offers a powerful approach for understanding gene regulation and its heterogeneity across cell types and states. It has profound applications in genetics and genomics, particularly causal gene regulatory network (cGRN) inference to unravel the molecular circuits governing cell identity and function. However, computational scalability remains a critical bottleneck for sceQTL mapping, prohibiting thorough benchmarking and optimization of statistical accuracy. We present airqtl, a novel method to overcome these challenges through algorithmic advances and efficient implementations of linear mixed models. Airqtl achieves superior time complexity and over eight orders of magnitude of acceleration, enabling objective method benchmarking and optimization. Airqtl offers *de novo* inference of robust, experimentally validated cell state-specific cGRNs that reflect perturbation outcomes. Our results dissect the drivers of cGRN heterogeneity and underscore the value of natural genetic variations in primary human cell types for biologically relevant single-cell cGRN inference.

SINGLE-MOLECULE SEQUENCE MODELS TO DECODE THE REGULATORY GENOME

Ruoyu Wang, Junru Jin, Jian Zhou

University of Texas Southwestern Medical Center, Lyda Hill Department of Bioinformatics, Dallas, TX

Sequence-to-function deep learning models have revolutionized our understanding of the regulatory genome in health and disease. However, these models are primarily designed to predict average activity across molecules, limiting insights into dynamic regulatory genomic events. Recent advances in single-molecule regulatory genomics have opened new avenues for observing dynamic regulatory chromatin events. Here, we sought to develop a new framework of sequence models capable of learning the regulatory genome at the single-molecule level. These next-generation models leverage state-of-the-art probabilistic generative frameworks and can be trained on single-molecule regulatory genomics datasets. Through interpretation techniques, these models can elucidate the sequence determinants of single-molecule chromatin events. Importantly, these single-molecule sequence models can be applied to reveal fundamental mechanisms of chromatin regulation and evaluate the impact of genetic variants on gene regulation. By decoding the regulatory genome at the single-molecule level, this new sequence model framework will serve as an *in silico* genome observatory for future genome regulation research in both health and disease.

REVISIT GLOBAL EXPRESSION CHANGE IN SINGLE-CELL PERTURBATION DATA

Shuyue Wang[1,2], Han Xu[1,2]

[1]MD Anderson Cancer Center, Department of Epigenetics and Molecular Carcinogenesis, Houston, TX, [2]MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, GSBS, Houston, TX

Global expression change (GEC), defined as the up- or down- regulation of a vast majority of genes in a single cell, provides a quantitative framework to investigate the genome-wide transcriptional behavior. An elevated expression of MYC has been observed across multiple cancer types and is associated with increased risk of aggressive tumor progression and poor survival outcomes. However, the broader landscape of regulators capable of driving global transcriptional amplification or repression remains poorly characterized. By leveraging total UMI counts as an indicator of global transcriptional activity, perturbations that trigger GEC can be identified through comparisons between non-targeting control cells and perturbed cells in single-cell RNA sequencing data. In this study, we analyzed five genetic perturbation datasets across four distinct cell lines and identified hundreds of global regulators capable of inducing either increased or decreased GEC. Notably, perturbation of MYC consistently resulted in decreased GEC across all datasets. Moreover, we observed that perturbation of KDM1A – a therapeutic target in leukemia – led to decreased GEC exclusively in leukemia-derived cell models, and pharmacological inhibition of KDM1A in single-cell leukemia patient data recapitulated this GEC reduction pattern. Together, these finding suggests that global expression analysis holds the promise for facilitating the identification of novel drug targets and advancing therapeutic strategies for various diseases, including cancer.

COUTURE: FACILITATING INTERPRETATION FROM GENOTYPE TO MOLECULAR AND FUNCTIONAL PHENOTYPE IN SINGLE-CELL CRISPR SCREENING

Jun Cao, Xiaoyue Wang

Institute of Clinical Medicine & Peking Union Medical College Hospital, National Infrastructures for Translational Medicine, Beijing, China

Single-cell CRISPR screening has revolutionized high-throughput functional genomics by enabling the simultaneous analysis of multiple perturbations and cellular responses at single-cell resolution. Nonetheless, several gaps remain in the analysis linking genotype to molecular or functional phenotypes, impeding a comprehensive understanding and effective application of the substantial data. Establishing robust genotype-molecular phenotype relationships is challenging due to technical complexities, such as variable cell counts, inconsistent perturbation efficiency, and both biological and technical noise that obscure cellular responses. Conventional analytical methods typically rely on threshold-based classification of individual features, failing to account for the interconnected nature of cellular responses. COUTURE proposes that the propagation of perturbation effects through molecular modularity leads to the formation of responsive patterns and further functional phenotype. Instead of evaluating features in isolation, COUTURE identifies responsive features through their contribution to cellular neighborhood structure and response patterns. In contrast, noise features negatively impact the cellular neighborhood, show random patterns, and lack reliability across sub-samples. Rather than excluding unperturbed cells through cell refinement, COUTURE implements a multi-sampled closed-loop iterative procedure that simultaneously refines both the cell neighborhood and feature patterns. COUTURE maintains robustness across various data settings and superiority over traditional strategies incorporating cell refinement. We validated COUTURE's effective recognition of modular features with several diverse data. Given abundant targets per dataset, functional phenotype queries similar to those of the Connectivity Map (CMap) present potential applications. However, traditional CMap relies on feature ranking-based enrichment analyses, overlooking relationships among features in single-cell data. In contrast, COUTURE focuses on the independent and joint queries of external gene sets using feature graphs. We will develop a functional phenotype query platform based on existing databases, expecting that the functional phenotypic pairwise relationships will yield deeper insights. Overall, COUTURE, as an innovative approach, assists single-cell CRISPR screening address various challenges, thereby enabling a better interpretation of genotype-phenotype relationships.

COMPREHENSIVE FUNCTIONAL ASSESSMENT OF *NF1* AND *NF2* VARIANTS WITH HIGH-RESOLUTION BASE EDITING SCREENS

Jiayu Wu[1], Guangyu Li[2], Liheng Luo[1], Chenyu Ma[1], Shangqi Zhao[1], Zhuang Du[1], Xiaoyue Wang[1]

[1]Institute of Clinical Medicine & Peking Union Medical College Hospital, Center for Bioinformatics, Beijing, China, [2]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, State Key Lab of Molecular Oncology, Beijing, China

Variants of uncertain significance (VUS) in the tumor suppressor genes *NF1* and *NF2* impede clinical decision-making in neurofibromatosis and cancer. Here, we developed an optimized near-PAM-less base editing screening approach combined with a dynamic Bayesian framework (EDGE-BE) to functionally assess 15,876 *NF1* and 7,386 *NF2* variants. By integrating temporal data and editing efficiency into a unified model, EDGE-BE robustly quantified variant effects, identifying 505 and 166 loss-of-function (LOF) variants for *NF1* and *NF2*, respectively. These LOF variants were significantly enriched in neurofibromatosis patients and various cancer cohorts, particularly melanoma. Through structural analysis and experimental validation, we uncovered diverse mechanisms of variant pathogenicity, including destabilization of dimerization interfaces, disruption of splicing regulation, and alterions in transcriptional control. Single-cell profiling of drug-resistant variants demonstrated NF1-LOF reactivates MAPK signaling through enhanced RAS signaling. This comprehensive functional map of NF1/NF2 variants advances our understanding of disease mechanisms and improves VUS interpretation in clinical settings.

ALLELE SPECIFIC EXPRESSION IN ALZHEIMER'S DISEASE

Zishan Wang[1], Delowar Hossain[2], Varun Subramaniam [1], Bin Zhang[1], Minghui Wang[1], Kuan-lin Huang[1]

[1]Icahn School of Medicine at Mount Sinai, Genetics and genomic sciences, New York, NY, [2]McGill University, Division of Experimental Medicine, Montréal, Canada

Allele-specific expression (ASE), preferential RNA expression of one allele over its counterpart, has been implicated in various diseases. However, its systematic role in Alzheimer's Disease (AD) remains unclear. We systematically analyzed ASE across five brain regions using RNA-sequencing data from two independent cohorts, identifying 42,025 unique variants showing ASE across 9,345 genes. ASE events were rare, occurring in 1.6%–3.9% of heterozygous variants, and enriched in imprinted genomic regions (e.g., chr6, chr14q32, chr15q11). ASE frequency negatively correlated with age of death and showed correlation trends with APOE genotype. Exonic ASE variants were identified in several AD-associated genes (e.g., APOE, HLA-DRB1, CLU), suggesting regulatory mechanisms. Lastly, we identified 18 variants exhibiting AD-associated ASE and integration with single-cell RNA-seq identified downregulated ASE genes in AD neurons, including VPS13C, MICU3, and LMO7. These findings highlight dysregulated ASE events that may link genetic variations to downstream expression and functional consequences, and potentially contribute to AD pathogenesis.

INDUSTRIALIZATION INFLUENCES BIOLOGICAL AND MOLECULAR MECHANISMS OF AGING IN IMMUNE CELLS IN THREE NON-INDUSTRIAL POPULATIONS

Marina M Watowich[1], Amy Longtin[1], Julien F Ayroles[2], Kenneth Buetow[3], Hillard Kaplan[4], Yvonne Lim[5], Dino Martins[6], Kee-Seong Ng[5], Sospeter Njeru[7], Jonathan Stieglitz[8], Benjamin Trumble[3], Vivek V Venkataraman[9], Ian J Wallace[10], Michael Gurven[11], Thomas S Kraft[12], Alexander G Bick[1], Amanda J Lea[1]

[1]Vanderbilt University, Department of Biological Sciences, Nashville, TN, [2]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [3]Arizona State University, Center for Evolution and Medicine, Tempe, AZ, [4]Chapman University, Economic Science Institute, Orange, CA, [5]Universiti Malaya, Department of Parasitology, Kuala Lumpur, Malaysia, [6]Stony Brook University, Turkana Basin Institute, Stony Brook, NY, [7]Kenya Medical Research Institute, Centre for Community Driven Research, Nairobi, Kenya, [8]Institute for Advanced Study in Toulouse, Toulouse School of Economics, Toulouse, France, [9]University of Calgary, Department of Anthropology and ArchaeologyCalgary, Canada, [10]University of New Mexico, Department of Anthropology, Albuquerque, NM, [11]University of California Santa Barbara, Department of Anthropology, Santa Barbara, CA, [12]University of Utah, Department of Anthropology, Salt Lake City, UT

Age-associated chronic inflammation is prevalent in Western, high-income societies and linked to a myriad of non-communicable diseases. In contrast, many subsistence-level societies experience low rates of non-communicable diseases and minimal evidence of chronic systemic inflammation during aging, pointing to effects of lifestyle. We hypothesize that lifestyle may impact immunological aging, and investigate these effects via two salient molecular mechanisms: 1) DNA methylation (DNAm), an epigenetic gene regulatory modification, and 2) clonal hematopoiesis of indeterminate potential (CHIP), the mutation and clonal expansion of blood cells. We quantified genome-wide DNAm (n=3974) and CHIP (n=1225) in whole blood samples from three societies that vary in their levels of market integration and industrialization: the Turkana of Kenya, the Tsimane of Bolivia, and the Orang Asli of Peninsular Malaysia. Leveraging long-term research studies in these populations, we combined detailed questionnaires, biomedical data, and genomic assays to (1) assess how age and lifestyle influence CHIP (including occurrence, variant allele frequency, and mutation type) and DNAm patterns across these populations. We find that DNAm is strongly perturbed by industrialized lifestyles and age in the three populations (~55% CpGs; n CpGs > 400,000; FDR<5%), in both shared and population-specific ways. We find that CHIP risk increases with greater industrialization (p=0.037) and with age (p=7.07 x 10-9), and predicts variation in DNAm. Finally, using an epigenetic clock trained on the DNAm data, we find that CHIP accelerates biological aging (p = 8.16 x 10-6) and industrialization has similar effects, but in population-specific ways dependent on how the industrialization transition is occurring in each country. Together, these findings provide important information about the degree to which immunological aging is universal versus environmentally-influenced across diverse human lifestyle contexts.

NETWORK-LEVEL CONVERGENCE OF RARE AND COMMON VARIANTS UNDERLYING COMPLEX TRAITS.

Sarah N Wright, Trey Ideker

University of California, San Diego, Medicine, La Jolla, CA

Understanding the contribution of rare and common genetic variants to complex traits is crucial for elucidating disease mechanisms, predicting disease risk, and prioritizing drug candidates. While variants across the frequency spectrum are known to contribute to genetic etiology, their shared and distinct influences remain poorly understood. Here, we systematically analyze rare and common variant associations across the human phenome by integrating association data for 330 traits within protein knowledge networks, identifying significant network-level convergence in 68% of traits. We further examine how genomic features such as heritability and gene conservation influence network convergence and drive differences in gene discovery across rare and common variant studies. Representative examples illustrate how networks derived from rare and common variants lead to an improved understanding of disease mechanisms and prioritization of disease genes. These findings highlight the importance of integrating variants across the frequency spectrum and establish a foundation for network-based investigation of variants across diverse traits.

HIGH-THROUGHPUT *IN SILICO* SCREEN DISCOVERED NOVEL REGULATORS OF 3D GENOME ORGANIZATION

Jiangshan Bai[1], Qingji Lyu[1], Jimin Tan[1,2], Bailey Tischer[1], Xinyu Ling[1], Viraat Goel[3], Aristotelis Tsirigos[4], Bradley E Bernstein[1], Anders S Hansen[3], Bo Xia[1,5]

[1]Broad Institute of MIT and Harvard, Gene Regulation Observatory, Cambridge, MA, [2]NYU Grossman School of Medicine, Institute for Systems Genetics, New York, NY, [3]MIT, Department of Biological Engineering, Cambridge, MA, [4]NYU Grossman School of Medicine, Department of Pathology, New York, NY, [5]Harvard University, Society of Fellows, Cambridge, MA

High-throughput screen approaches play a pivotal role in identifying candidate regulators that guide follow-up mechanistic investigations. Traditional experimental screen typically relies on high-throughput perturbation-to-measurement setup, which is limited by the available resources and feasibility. The rise of deep machine learning / artificial intelligence (AI) approaches has enabled accurate and context-specific predictions through learning unprecedented latent features from the data. These high-performance predictive AI models start to replace cumbersome experiments, enabling *in silico* experimenting to generate high-conviction hypotheses that inform the mechanistic investigation.

The vertebrate genome is hierarchically organized into three-dimensional (3D) conformations, which further shape the gene expression landscape. Yet the technological challenges to measuring chromatin interactions have limited the implementation of screening approaches to discover novel regulators that can resolve several important questions in the 3D genome field.

To address this challenge, we recently developed a high-throughput *in silico* screen approach, leveraging our C.Origami model that accurately predicts cell-type-specific chromatin interactions. C.Origami model adopts an encoder-decoder design, with two encoders that separately process DNA sequence and cell-type-specific chromatin feature, a transformer module that enables long-range information exchange, and a task-specific decoder that outputs the chromatin interaction heatmap, similar to experimental data. The high performance of C.Origami prediction can largely replace experimental measurements and thus enables in silico experimentation and screening with high conviction.

Applying this *in silico* screen approach, we identified several previously uncharacterized proteins that uniquely bind at specific 3D genome features, including chromatin domain boundaries, stripe anchors, and loop anchors. Subsequent experimental validation and mechanistic investigation confirmed their roles in demarcating chromatin domains, thus providing critical new insights into the fundamental mechanism of vertebrate 3D genome organization. Together, these results highlight the potential of applying high-performance predictive AI tools for *in silico* experimentation to accelerate fundamental discoveries in biology.

CHARACTERIZATION OF CODON AND AMINO ACID FREQUENCY VARIATION IN THE HUMAN GENOME

Zhuorui Xie[1], Ziyue Gao[2]

[1]University of Pennsylvania, Genomics and Computational Biology Graduate Program, Philadelphia, PA, [2]University of Pennsylvania, Department of Genetics, Philadelphia, PA

The genetic code defines how nucleotide sequences are translated into amino acids and is shared across functionally all domains of life; however, individual codons occur at unequal frequencies across genes and species. Most hypotheses for this variation are either mutation- or selection-driven, but the extent to which both forces act on the human genome is not fully understood. To address this question, we examined nucleotide trimer frequencies across three compartments of the human genome – coding, intergenic, and intronic – and compared against the expected frequencies based on GC%. We identified consistent differences in trimer frequencies in the coding compartment compared to the intergenic and intronic compartments across all chromosomes, suggesting that codon frequency variation is driven by selection acting on coding sequences. As expected, hypermutable CpG sites are highly depleted in all three compartments; however, we found CpG sites to be less depleted in coding than in non-coding compartments. Bisulfite sequencing data from human sperm cells showed similar CpG methylation rates across compartments, implying stronger preservation of CpG sites rather than lower mutation rate is responsible for the weaker depletion of CpG sites in coding regions. We also observed variation in synonymous trimer frequencies within the coding compartment, suggesting factors other than amino acid function are under selection. Overall, our results point to selection in coding sequences as the major driver of observed variation in codon frequency, although the exact targets are still unclear. Further investigation into the mechanisms of selection will shed more light on human genome evolution.

DEFINING THE LANDSCAPE OF POISON EXONS AND THEIR INVOLVEMENT IN HUMAN DISEASES

Huilin Xu[1,2,3], Paolo Pigini[1,2], Yan Ji[1], Hannah Lindmeier[1], Maria Catarina Lima Da Silva[1,2], Dadi Gao[1,2,3], Elisabetta Morini[1,2]

[1]Massachusetts General Hospital Research Institute, Center for Genomic Medicine, Boston, MA, [2]Massachusetts General Hospital Research Institute and Harvard Medical School, Department of Neurology, Boston, MA, [3]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Cambridge, MA

Poison exons (PEs) play critical roles in diversifying gene regulation mechanisms by introducing premature termination codons into transcripts and in turn triggering nonsense-mediated decay (NMD) to repress transcription. Growing evidence suggests that mutations affecting PE splicing can be pathogenic in diseases. Despite their importance, PEs remain poorly annotated and studied. To address this, we first scrutinized all the annotated human exons by harmonizing the genetic criteria of PE previously reported. This resulted in cataloging 12,014 PEs genome wide. By leveraging RNASeq from healthy individuals, we further investigated PEs' tissue-specific or developmental stage-specific regulations. To unravel tissue-specificity, we measured PE usage in percent-spliced-in (PSI) values from 17,382 GTEx RNASeq libraries and modeled the PSI changes across tissues via a generalized linear model. In the human brain for example, we found 117 PEs uniquely expressed and never used in non-brain tissues. Gene ontology (GO) analysis revealed that they were enriched for ion channel functionality. To investigate PEs' developmental trajectories, we analyzed 607 BrainSpan RNASseq libraries by applying generalized additive models to PSI values across 12 age intervals from fetal to 70 years old. In the primary auditory cortex for example, we identified 211 PEs with distinct time dependency in two linear and four non-linear trajectories. GO analysis of these PEs highlighted synaptic maturation and neuron-to-neuron synapse formation, agreeing with neurodevelopment. With normative PE dynamics established, we interrogated the association between PE mis-splicing and human diseases. By computationally incorporating pathogenic ClinVar mutations into the human genome and extracting mutations leading to PE mis-splicing predicted by SpliceAI, we found 217 pathogenic variants affecting the splicing of 233 brain PEs. Eight (out of 10) candidates were confirmed by CRISPR prime editing and with or without CHX treatment that blocks NMD. Overall, we depicted a comprehensive landscape of human PEs with insights into their differential regulations and potential pathogenicity when mis-spliced. Our results add a useful angle for variant-to-function interpretation and could inspire innovative therapeutic design to target PE dysregulation.

CHRONODE: A FRAMEWORK TO INTEGRATE TIME-SERIES MULTI-OMICS DATA BASED ON ORDINARY DIFFERENTIAL EQUATIONS COMBINED WITH MACHINE LEARNING

Beatrice Borsari*[1,2], Mor Frank*[1,2], Eve S Wattenberg#[1,2], Ke Xu#[1,2,3,4], Susanna X Liu#[1,2], Xuezhu Yu[1,2], Mark Gerstein[1,2,5,6]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, [3]Department of Biostatistics, Yale University, New Haven, CT, [4]Department of Computer Science, Yale University, New Haven, CT, [5]Department of Biomedical Informatics and Data Science, Yale University, New Haven, CT, [6]Department of Statistics and Data Science, Yale University, New Haven, CT

* Equally contributing authors
#Equally contributing authors

Many genome-wide studies capture isolated moments in cell differentiation or organismal development. Conversely, longitudinal studies provide a more direct way to study these kinetic processes. Here, we present an approach for modeling gene-expression and chromatin kinetics from such studies: chronODE, an interpretable framework based on ordinary differential equations. ChronODE incorporates two parameters that capture biophysical constraints governing the initial cooperativity and later saturation in gene expression. These parameters group genes into three major kinetic patterns: accelerators, switchers, and decelerators. Applying chronODE to bulk and single-cell time-series data from mouse brain development revealed that most genes (～87%) follow simple logistic kinetics. Among them, genes with rapid acceleration and high saturation values are rare, highlighting biochemical limitations that prevent cells from attaining both simultaneously. Early- and late-emerging cell types display distinct kinetic patterns, with essential genes ramping up early. Extending chronODE to chromatin, we found that genes regulated by both enhancer and silencer cis-regulatory elements are enriched in brain-specific functions. Finally, we developed a bidirectional recurrent neural network to predict changes in gene expression from corresponding chromatin changes, successfully capturing the cumulative effect of multiple regulatory elements. Overall, our framework allows investigation of the kinetics of gene regulation in diverse biological systems.

ON THE ACCURATE IMPUTATION OF COMMON INVERSIONS IN THE HUMAN GENOME

Illya Yakymenko[1,2], Adrià Mompart[1], Mario Cáceres[1,2,3]

[1]Hospital del Mar Research Institute, Research Program on Biomedical Informatics (GRIB), Barcelona, Spain, [2]Universitat Autònoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra, Barcelona, Spain, [3]ICREA, Barcelona, Spain

Genomic inversions are structural variants where a DNA segment gets reversed, usually without gain or loss of DNA, which have been associated to phenotypic traits and adaptation in both humans and other organisms. Inversions originated by homologous mechanisms are especially challenging to characterize due to the presence of highly identical inverted repeats at the breakpoints and the fact that most of them are recurrent. Thus, their functional effects have been typically understudied. Imputation is widely used as an alternative to infer genotypes of missing variants with great success. However, it has been mainly limited to simple variants and little is known about the imputation accuracy of human inversions. Here, we benchmarked the performance of five imputation software – Beagle5.4, Impute2, Impute5, Minimac4 and scoreInvHap – in a set of 55 inversions experimentally genotyped in multiple samples of the 1000 Genome Project and that lacked variants in perfect linkage disequilibrium. We have examined the overall and per-inversion imputation accuracy of each tool using whole genome sequencing (WGS) data and simulated microarrays with different SNP density. The results suggest an overall good inversion imputation performance for all methods, with up to ~65% of inversions being accurately imputed in different human populations from WGS data and slightly less from SNP arrays. In particular, Minimac4 and Impute5 show higher imputation accuracy and lower loss of poorly imputed individuals with respect to the other methods. Also, we found that sample size and the application of posterior genotype probability filtering are key factors for inversion imputation accuracy. Therefore, this work provides insights into the optimal conditions for inversion imputation and highlights the potential of these methods to enhance inversion genotype prediction, contributing ultimately to a better understanding of the functional impact of these variants.

RECONSTRUCT HUMAN CLONAL DEVELOPMENT WITH MOSAIC VARIANTS

Xiaoxu Yang

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

Lineage reconstructions have been accomplished by artificial dye and viral labeling, or recent crispr barcodes, which are extremely helpful in non-human organisms. Clonal lineage reconstruction could be achieved by using mosaic variants. Cells in our bodies continuously accumulate variants due to DNA replication, recombination, damage, or environmental exposure. While most of these variants are repaired, a small portion remains in the genome and passed to daughter cells—this distribution of cells carrying different variants results in genomic mosaicism. Fraction differences of mosaic variants reflect developmental relationships between different developed cell types and organs, revealing potential lineage patterns.
We have established a combined computational and experimental methodology to reconstruct clonal relationships with naturally occurring mosaic variants, termed mosaic variant barcode analysis (MVBA). MVBA successfully reconstructs human developmental patterns. MVBA utilizes deep sequencing at genome or exome scale as well as machine-learning-based variant calling pipelines on DNA extracted from bulk tissue to profile candidate mosaic SNV/INDELs with high accuracy. Mosaic variants from bulk samples were further quantified using massive parallel amplicon sequencing (MPAS) and single nuclei MPAS (snMPAS). Lineages were orthogonally validated by an integrated transcriptomic and genomic analysis from the same single nuclei.
Lineages are deconvolved based on the unequal representation of fractions of mosaic variants in different tissue samples and cell populations, reflecting the actual proportion of progenitors at the time the variants occurred. Our single-cell lineage reconstruction validates the hypotheses generated from bulk mosaic variant data.
MVBA detected approximately 50% of cortical inhibitory neurons are dorsally derived in the central nerve system and found early developmental fate determinations between dorsal root ganglia and sympathetic ganglia, which have not been thoroughly elaborated in human development due to technical and ethical limitations. MVBA as a top-down method ensures sufficient marker sharing between samples and reduces false positive detections compared to single-cell-originated bottom-up methods. Additionally, MVBA can be further applied to understanding cell-type-specific developmental events due to sample limitations.

SPATIAL DOMAIN DETECTION USING CONTRASTIVE SELF-SUPERVISED LEARNING FOR SPATIAL MULTI-OMICS TECHNOLOGIES

Jianing Yao[1], Jinglun Yu[2], Brian Caffo[1], Stephanie C Page[3], Keri Martinowich[3,4,5], Stephanie C Hicks[1,6,7,8]

[1]Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, MD, [2]Johns Hopkins University, Department of Electrical and Computer Engineering, Baltimore, MD, [3]Johns Hopkins Medical Campus, Lieber Institute for Brain Development, Baltimore, MD, [4]Johns Hopkins School of Medicine, The Solomon H. Snyder Department of Neuroscience, Baltimore, MD, [5]Johns Hopkins School of Medicine, Department of Psychiatry and Behavioral Sciences, Baltimore, MD, [6]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, [7]Johns Hopkins University, Center for Computational Biology, Baltimore, MD, [8]Johns Hopkins University, Malone Center for Engineering in Healthcare, Baltimore, MD

Recent advances in spatially-resolved single-omics and multi-omics technologies have led to the emergence of computational tools to detect or predict spatial domains. Additionally, histological images and immunofluorescence (IF) staining of proteins and cell types provide multiple perspectives and a more complete understanding of tissue architecture. Here, we introduce Proust, a scalable tool to predict discrete domains using spatial multi-omics data by combining the low-dimensional representation of biological profiles based on graph-based contrastive self-supervised learning. Our scalable method integrates multiple data modalities, such as RNA, protein, and H&E images, and predicts spatial domains within tissue samples. Through the integration of multiple modalities, Proust consistently demonstrates enhanced accuracy in detecting spatial domains, as evidenced across various benchmark datasets and technological platforms.

Expanding the Readable Genome: A Novel Approach for Analyzing Mononucleotide C Repeats

Zhezhen Yu[1,2], Inessa Hakker[1], Antoine Gruet[1], Anya 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—short, repetitive sequences scattered throughout the genome—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 report 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 obtain an accurate measure of somatic variation. We find 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.

ASSESSING THE IMPACT OF GENETIC VARIATION ON CHROMATIN INTERACTION DURING BRAIN DEVELOPMENT.

Samantha Zarnick[1,2], Lydia Adams[1,2], Jingying Wang[1,2], Tatiana Ulloa Avila[1,2], Ellen Hu[1,2], Jordan Valone[1,2], Brandon Le[1,2], Jason Stein[1,2], Hyejung Won[1,2]

[1]UNC, Neuroscience Center, Chapel Hill, NC, [2]UNC, Department of Genetics, Chapel Hill, NC

Many common variants have been identified that are associated with risk for neuropsychiatric disorders, but are often in non-coding regions of the genome, making their functional impact difficult to interpret. These variants could influence the regulation of gene expression through changes in chromatin folding that alter interactions between enhancers and promoters. While the role of genetic variation in gene expression has been extensively documented through expression quantitative trait loci studies, its influence on chromatin architecture, particularly in the developing human cortex, remains poorly understood. To address this gap, we have generated Promoter Capture Micro-C data from over 76 post-mortem developing human cortical tissue samples during the period of neurogenesis. To ensure data quality, we have sequenced 28 samples with an average read-depth of ~140 million paired-end reads. We assessed on-target rates, coverage depth, PCR duplication rate, and cis-ratio metrics. Our libraries exceed the expected on-target rate of 40%, with promoter region coverage 23 times higher than non-target regions. Additionally, all samples are within the expected PCR duplication range of 10-35% and exceed the 60% cis-ratio benchmark, confirming high-quality data. These samples have also undergone genome-wide genotyping and transcriptomic profiling, enabling us to integrate chromatin conformation with genetic and expression data. By mapping chromatin interactions at high resolution, we aim to identify genetic variants that alter chromatin folding and to assess their colocalization with known neuropsychiatric disorder risk loci from genome-wide association studies (GWAS). This work will provide new insights into how genetic variation shapes three-dimensional genome organization during human cortical development, ultimately improving our understanding of the regulatory mechanisms that govern brain development and contribute to disease risk.

WHEN ARCHAIC GENES BOOST GROWTH: THE NEANDERTHAL GROWTH HORMONE RECEPTOR

Philipp Kanis[1], Miriam Berreiter[2], Daniel Sieme[3], Xiang-Chun Ju[4], Nicholas E Holzwart[5], David H Ziliang[2], Shu Tadaka[6], Makiko Taira[6], Kengo Kinoshita[6], Richard Ågren[2], Johan G Olsen[3], Tomislav Maricic[1], Birthe Kragelund[3], Svante Pääbo[1,4], Andrew J Brooks[5], 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, [3]University of Copenhagen, Department of Biology, Copenhagen, Denmark, [4]Okinawa Institute of Science and Technology, Human Evolutionary Genomics Unit, Onna, Japan, [5]The University of Queensland, Frazer Institute, Brisbane, Australia, [6]Tohoko Medical MegaBank Organization, Department of Integrative Genomics, Sendai, Japan

Neanderthals had robust physique and exhibited several distinct skeletal characteristics. The hypothalamic-pituitary–somatotropic (HPS) axis, with growth hormone (GH) serving as a key signaling molecule, plays a central role in controlling skeletal growth. To investigate whether Neanderthal-specific genetic variants within this axis contributed to their robust build, we investigated genes responsible for encoding the relevant receptors and hormones. We discovered two amino acid substitutions and a deletion in the Neanderthal growth hormone receptor (GHR). In a GH-dependent cell line, stimulation with pituitary GH elicited a more rapid cell proliferation in cells expressing the Neanderthal GHR compared to those expressing the modern human GHR; no difference was observed under placental GH stimulation. Furthermore, we found that certain individuals today carry the Neanderthal version of the *GHR* gene, and these individuals tend to be taller, exhibit greater muscle mass, and display craniofacial features reminiscent of Neanderthals, such as short-rooted teeth and reduced mandibular height. As a result, traces of Neanderthal anatomy persist in some present-day humans.

GENOME-WIDE INFERENCE OF POSITION-SPECIFIC ELONGATION RATES USING TIME-COURSE NASCENT RNA-SEQ DATA

Xin Zeng, Rebecca Hassett, Adam Siepel

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

Deciphering transcriptional regulation dynamics is essential for understanding how cells adapt to external signals. Recent advances in nascent RNA-seq techniques enable precise capture of RNA polymerase distribution along the gene body at specific time points. In our previous study, we introduced a unified probabilistic model to describe RNA Pol II dynamics (Siepel, 2021). Based on this model, we used PRO-seq data to infer genome-wide patterns of promoter-proximal pausing and local elongation rates under steady-state conditions (Liu et al., n.d.; Zhao et al., 2023).

Here, we extend the model to analyze time-course transcriptional elongation within the first 2000 bp downstream of the transcription start site (TSS), making it applicable to both run-on-based methods (e.g., PRO-seq) and metabolic labeling-based approaches (e.g., 4sU-seq). We evaluated our model using simulated data and benchmarked it against existing methods, demonstrating consistency while providing higher-resolution estimates of position-specific elongation rates across thousands of genes. We further analyzed elongation patterns, including the duration of proximal pausing and the acceleration following pausing release. Additionally, we examined how these patterns correlate with DNA sequence features near the TSS and steady-state epigenomic marks. Our method provides a unified framework for analyzing time-course nascent RNA data, enabling higher-resolution estimation of elongation rates. This approach may offer new insights into gene regulation under different environmental conditions.

EXPLORING SELECTIVE SCANNING WITH DZ STATISTIC: SIMULATION AND EMPIRICAL STUDIES

Alouette Zhang[1,2], Aaron Ragsdale[3], Kevin R Thornton[4], Simon Gravel[1]

[1]McGill University, Department of Human Genetics, Montreal, Canada, [2]Kyoto University, Faculty of Medicine, Kyoto, Japan, [3]University of Wisconsin-Madison, Department of Integrative Biology, Madison, WI, [4]University of California, School of Biological Sciences, Irvine, CA

Selective sweeps have been widely studied due to their evolutionary importance. A sweep occurs when a beneficial variant increases rapidly in frequency and reaches fixation in the population. This process reduces diversity and elevates linkage disequilibrium (LD) in the surrounding region. Different methods incorporate a combination of these patterns to detect sweeps. Here, we consider the effect of hard sweeps on Dz, a measure of LD between rare variants which was recently shown to be informative about deeply coalescing human demography.

We demonstrate via theory and simulation that Dz can utilize both the reduced diversity and the excess LD around the sweep site for sweep detection, and is more informative about recent and ancient sweeps compared to other commonly used LD statistics. Dz maintains adequate detection power for sweep detection up to 150 thousand years ago in simulations with human-like parameters. Through applying a window scan that incorporates the expectation of Dz, we successfully identify well-characterized selective sweeps in different populations from the 1000 Genomes Project data. We also identify elevated Dz regions that were shared across different populations. Furthermore, we discuss how various annotated regions, including long inversions and archaic introgressions, influence our observed signal. Our findings highlight Dz as a promising statistic for sweep detection, capturing patterns that are not fully considered by most existing methods.

THE ROLE OF RARE NON-CODING VARIANTS IN BICUSPID AORTIC VALVE PATHOLOGY

Artemy Zhigulev[1], Madeleine Petersson Sjögren[1], Andrey Buyan[2], Vladimir Nozdrin[3], Karin Lång[4], Rapolas Spalinskas[1], Raphaël Mauron[1], Enikő Lázár[1], Sailendra Pradhananga[1], Anders Franco-Cereceda[5], Joakim Lundeberg[1], Ivan V Kulakovskiy[2], Per Eriksson[4], Hanna M Björck[4], Pelin Sahlén[1]

[1]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Division of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden, [2]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia, [3]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [4]Cardiology Unit, Department of Medicine Solna, Karolinska Institutet, Karolinska university hospital, Stockholm, Sweden, [5]Department of Molecular Medicine and Surgery, Karolinska Institutet, Karolinska university hospital, Stockholm, Sweden

The bicuspid aortic valve (BAV), a congenital condition affecting ~1% of the population, exhibits high heritability. Previous genetic studies identified few coding or common GWAS variants associated with BAV, explaining only a minority of cases. Recognizing the gap in understanding the disease etiology, we hypothesized that less common regulatory variants in promoters and enhancers might play a critical role in BAV pathogenesis. Ascending aorta tissue samples were collected from 16 patients undergoing open-heart surgery, evenly distributed between BAV and normal tricuspid aortic valve (TAV) cases. We generated high-resolution transcriptome and promoter-enhancer interaction maps for endothelial cells derived from the biopsies. Whole-genome sequencing was performed to determine sequence variants.

Notably, none of the BAV patients possessed protein-coding mutations in BAV-related genes. Bulk RNA-seq analysis revealed differentially expressed genes, but enriched gene sets reflected consequences of BAV pathology. To delve into the causal mechanisms, we isolated BAV and TAV-specific promoter-enhancer interactions that occurred or were disrupted only in individuals with the rare mutation in the interacting regulatory regions (RMiRRs). We retained only the cases where the RMiRR was predicted to change a TF's binding affinity.

Using human embryonic heart scRNA-seq datasets spanning aortic valve developmental stages, we found that RMiRRs with minor allele frequencies 0–3% predominantly affected differentially expressed genes in spatially relevant fetal mesenchymal cells and fibroblasts in BAV patients. These genes were uniquely enriched for pathways involved in BAV pathogenesis, aortic valve development, and TGFBR signaling. Gene network analyses at the single-patient level revealed genetic heterogeneity within the BAV cohort. Our analysis significantly expanded the BAV pathway gene set, uncovering new potential molecular targets.

IMPROVING GENETIC SCORE PORTABILITY AND CAUSAL VARIANT DETECTION THROUGH A NOVEL JOINT BAYESIAN FRAMEWORK

Helyaneh Ziaei Jam

Department of Computer Science and Engineering, La Jolla, CA,
Department of Medicine, La Jolla, CA

Genetic score prediction accuracy remains limited in individuals, specifically in those from non-European populations, primarily due to smaller sample sizes. A major challenge is that current methods often identify tagging variants rather than the causal variant, due to strong correlations among nearby variants. Since linkage disequilibrium (LD) patterns vary across populations, findings from one population often fail to generalize well to others.

To address these challenges, we introduce a novel multi-ancestry approach that (1) jointly models the posterior distribution of effect sizes across multiple populations and (2) leverages differences in LD patterns in different populations to achieve better resolution for capturing causal effects and therefore make more portable scores. Unlike other methods such as PRS-CSx, our method simultaneously estimates posterior effect sizes for all populations in a single Bayesian shrinkage framework. Notably, by introducing a parameter modeling the correlation of effect sizes across populations, our method maintains a consistent interpretation of shared genetic signals while allowing for population-specific effect sizes.

Our results on simulated African and European datasets demonstrate that our approach (1) improves effect size estimation by yielding more consistent and accurate posterior distributions across populations and (2) enhances causal variant detection compared to models that analyze each population separately or do not jointly model effect sizes. Moving forward, we will assess the impact of variant-specific correlation parameters, explore the method’s performance on diverse real-world datasets, and benchmark it against leading polygenic risk score prediction tools to evaluate its effectiveness in multi-ancestry genetic studies.

STUDYING THE EMERGENCE OF DE NOVO COPY NUMBER VARIATIONS IN MALARIA PARASITE *PLASMODIUM FALCIPARUM* WITH LONG-READ SEQUENCING

Julia Zulawinska[1], Noah Brown[1], Aleksander Luniewski[1], Shiwei Liu[1,2], Jennifer Guler[1]

[1]University of Virginia, Department of Biology, Charlottesville, VA, [2]Indiana University, School of Medicine, Indianapolis, IN

Plasmodium falciparum is a protozoan parasite that is largely responsible for human malaria deaths. Despite continuing attempts at finding a reliable antimalarial treatment, the parasite easily develops resistance to newly introduced drugs. Changes in gene copy number are an important adaptive strategy for these parasites, contributing directly to drug resistance and fitness. Apart from "common" CNVs that occur widely in parasite populations under drug selection, our group has also observed distinct CNVs at various locations across individual parasite genomes utilizing two complementary approaches: (1) a low-cell short-read sequencing approach combined with specialized CNV analysis tools, and (2) ONT long-read sequencing with novel single-read visualization technique. We hypothesize that these relatively rare "*de novo*" CNVs prepare the parasite for future stressors and may be an important source of genetic diversity. Here, we present results from single long-read visualizations, which allow direct quantification of gene copy number and structural assessment (i.e. deletions, tandem or inverted duplications, etc). Because the long reads often cover entire duplicated regions (>50kb and tens of genes), we can assess CNV boundary sequences to uncover mechanisms of generation and change over time. This work has proven challenging, as studying the emergence of *de novo* CNVs requires high-coverage sequencing data with a high N50. However, through these investigations, we have uncovered a positive relationship between CNV frequency and replication stress, heterogeneity of CNVs in clonal lines, and a pattern of AT-richness at CNV boundaries. This provides invaluable information as we seek to understand how the *Plasmodium* genome adapts and evolves to confer new phenotypes. The single-long-read visualization method can be employed to any organism, including cancers and other microorganisms, to facilitate investigations of CNV emergence and analysis of complex structural variations.

GENOME-WIDE CRISPR SCREENING IDENTIFIES A NOVEL MEMBRANE PROTEIN GOVERNING *SALMONELLA* REPLICATION AND PERSISTER FORMATION

Sehee Yun, Seoyeon Kim, Seonggyu Kim, Hunsang Lee, Eunjin Lee

Korea University, Department of Life Sciences, Seoul, South Korea

Salmonella spp. are major foodborne pathogens that pose significant global health and economic challenges, particularly due to increasing antibiotic resistance. Although extensive research identified a few host factors, there still exist certain gaps. Key host factors exploited by *Salmonella* to form the persisters that survive antibiotic treatment remain elusive. To address this, we conducted a genome-wide CRISPR screening to identify host factors that govern infection efficiency. We challenged genome-wide knockout HAP1 cells with GFP-expressing *Salmonella* and allowed *Salmonella* to replicate in the host cells for 21 hours. Subsequently, we sorted for GFP-negative populations aim to identify factors that govern *Salmonella* replication inside the host. Upon analyzing the population, we identified a novel regulator governing *Salmonella* infection and replication. Using a candidate knockout cell line, we have verified the protective effect against *Salmonella* infection and confirmed their potential as therapeutic targets. We are currently conducting additional experiments to investigate whether this is involved in forming persisters. Our findings provide new insights into host-pathogen interactions exploited by *Salmonella* and may reveal novel targets for combating antibiotic-resistant infections.

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Bethlehem Abebe
UConn Health Center
abebe@uchc.edu

Mr. Temidayo Adeluwa
The University of Chicago
temi@uchicago.edu

Dr. Nadav Ahituv
University of California, San Francisco
nadav.ahituv@ucsf.edu

Ms. Clara Albinana
University of Oxford
clara.albinana@psych.ox.ac.uk

Mr. Abdalla Alkhawaja
University of North Carolina at Chapel Hill
abdalla.alkhawaja@unc.edu

Dr. Luay Almassalha
Northwestern Memorial Hospital
luay-almassalha@northwestern.edu

Ms. Atia Amin
McGill University
atia.amin@mail.mcgill.ca

Ana Beatriz Silva Amorim
Max Planck Institute for Evolutionary Anthropology
beatriz_amorim@eva.mpg.de

Ms. Ernestine Amos-Abanyie
University of Tennessee health Science
ekubi@uthsc.edu

Ms. Sambina Islam Aninta
University of British Columbia
sambina.islam@gmail.com

Mr. Alber Aqil
University at Buffalo
alberaqil@gmail.com

Dr. Mona Arabzadeh
Rutgers Biomedical and Health Science
mona.arabzadeh@rutgers.edu

Dr. Peter Arndt
Max Planck Institute for Molecular Genetics
arndt@molgen.mpg.de

Ms. Audrey Arner
Vanderbilt University
audrey.m.arner@vanderbilt.edu

Mahsa Askary Hemmat
Iowa State University
mahsa@iastate.edu

Thomas Atkins
Princeton University
thomas.atkins@princeton.edu

Peter Audano
The Jackson Laboratory
peter.audano@jax.org

Mr. Betselot Ayano
Ethiopian Public Health Institute
betselotzerihun@ephi.gov.et

Ms. Sara Azidane
Universitat Autonoma de Barcelona
sara.azidane@gmail.com

Ms. Ayesha Bajwa
UC Berkeley
arbajwa@berkeley.edu

Parithi Balachandran
The Jackson Laboratory
parithi.balachandran@jax.org

Zhigui Bao
Max Planck Institute for Biology Tübingen
zhigui.bao@tuebingen.mpg.de

Dr. Vanessa de Araujo Barbosa
Livestock Improvement Corporation
vanessa.barbosa@lic.co.nz

Dr. Alyson Barnes
American Society of Human Genetics
abarnes@ashg.org

Thomas Bataillon
Aarhus University
tbata@birc.au.dk

Dr. Alexis Battle
Johns Hopkins University
ajbattle@jhu.edu

Dr. Christine Beck
UConn Health and the Jackson Laboratory
Christine.Beck@jax.org

Dr. Pascal Belleau
Cold Spring Harbor Laboratory
belleau@cshl.edu

Dr. Jonathan Belyeu
PacBio
jrbelyeu@gmail.com

Dr. Kynon Benjamin
Northwestern University
kynon.benjamin@northwestern.edu

Mr. Thomas Bertino
Stony Brook University
thomas.bertino@stonybrook.edu

Prof. Minou Bina
Purdue Unversity
bina@purdue.edu

Ms. Naomi Boldon
University of Wyoming
nboldon@uwyo.edu

Prof. Hidemasa Bono
Hiroshima University
bonohu@hiroshima-u.ac.jp

Layla Brassington
Vanderbilt University
layla.brassington@vanderbilt.edu

Sean Bresnahan
The University of Texas MD Anderson Cancer Center
stbresnahan@mdanderson.org

Tylor Brewster
UConn Health and the Jackson Laboratory
tylor.brewster@Jax.org

Eva Brill
EpiCypher, Inc.
ebrill@epicypher.com

Ms. Mia Broad
Northwestern University
mia.broad@northwestern.edu

Dr. Lawrence Brody
National Human Genome Research Institute/NIH
lbrody@mail.nih.gov

Mr. Noah Brown
University of Virginia
njb8sg@virginia.edu

Prof. Brielin Brown
University of Pennsylvania
brielin.brown@pennmedicine.upenn.edu

Seyoun Byun
UNC at Chapel Hill
sbyun@unc.edu

Prof. Mauro Calabrese
University of North Carolina at Chapel Hill
jmcalabr@med.unc.edu

Yuwei Cao
University of California, San Diego
yuc408@ucsd.edu

Dr. Clayton Carey
University of Utah
clay.carey@utah.edu

Maria Carilli
Caltech
mcarilli@caltech.edu

Sara Carioscia
Johns Hopkins University
scarios1@jhu.edu

Dr. Piero Carninci
Human Technopole
piero.carninci@fht.org

Mr. Marc Carrillo Perez
Helsinki Institute of Life Science
marccarprz@gmail.com

Dr. Ava Carter
Harvard Medical School
Ava_Carter@hms.harvard.edu

Dr. Gemma Carvill
Northwestern University
gemma.carvill@northwestern.edu

Dr. Nyasha Chambwe
Feinstein Institutes for Medical Research
nchambwe@northwell.edu

Nikol Chantzi
Penn State University College of Medicine
nmc6088@psu.edu

Margaret Chapman
University of Michigan
marcha@umich.edu

Mr. Bide Chen
Duke University
bide.chen@duke.edu

Dr. Nancy Chen
University of Rochester
nancy.chen@rochester.edu

Dr. Haoyu Cheng
Yale University
haoyu.cheng@yale.edu

Dr. Jia Cheng
Cell Press
jcheng@cell.com

Ashwin Chetty
University of Chicago
achetty@uchicago.edu

Maria Chikina
University of Pittsburgh
mchikina@gmail.com

Eunice Choi
University of California, San Diego
ejc043@ucsd.edu

Elysia Chou
University of Michigan
elysian@umich.edu

Dr. Francesca Ciccarelli
The Francis Crick Institute / QMUL
francesca.ciccarelli@crick.ac.uk

Prof. Nathan Clark
University of Pittsburgh
nclark@pitt.edu

Mr. Shahein Clay
Stony Brook University
shahein.clay@stonybrook.edu

Dr. Alanna Cohen
Rutgers University
alanna.b.cohen@rutgers.edu

Dr. Alessio Colantoni
Sapienza University of Rome
alessio.colantoni@uniroma1.it

Dr. Laura Colbran
University of Pennsylvania
colbranl@pennmedicine.upenn.edu

Dr. Charles Cole
Juniper Genomics
chkcole@gmail.com

Dr. Vincenza Colonna
University of Tennessee
vcolonna@uthsc.edu

Isabelle Cooperstein
University of Utah
isabelle.cooperstein@genetics.utah.edu

Mr. Miguel Cortes
Centre for Genomic Regulation/UPF
miguel.cortes@crg.eu

Ms. Christina Costa
New York University
cec701@nyu.edu

Mr. Arun Das
Johns Hopkins University
arun.das@jhu.edu

Dr. Jishnu Das
University of Pittsburgh
jishnu@pitt.edu

Dr. Carl de Boer
University of British Columbia
carl.deboer@ubc.ca

Dr. Carolina de Lima Adam
University of Oregon
carolid@uoregon.edu

Prof. Jared Decker
University of Missouri
DeckerJE@missouri.edu

Mr. William DeGroat
Rutgers, The State University of New Jersey
williambdegroat@gmail.com

Ms. Mrunal Kishor Dehankar
University of Minnesota
dehan002@umn.edu

Ms. Allyson Dekovich
University of Tennessee, Knoxville
adekovic@vols.utk.edu

Dr. Jennifer DeLeon
Genome Research, Senior Assistant Editor
deleon@cshl.edu

Mr. Weixia Deng
Iowa State University
wdeng@iastate.edu

Dr. Ahmet Denli
Genome Research, Associate Editor
denli@cshl.edu

Ms. Astrid Deschenes
Cold Spring Harbor Laboratory
deschene@cshl.edu

Mr. Alex Diaz-Papkovich
Brown University
alex_diaz-papkovich@brown.edu

Tristram Dodge
Stanford University
tododge@stanford.edu

Dr. Anders Dohlman
Dana Farber Cancer Institute
andersb_dohlman@dfci.harvard.edu

Dr. Alexander Downie
Max Planck Institute for Evolutionary Anthropology
adownie14@gmail.com

Max Dudek
University of Pennsylvania
maxdudek@upenn.edu

Mr. Muhammed Rasit Durak
Max Planck Institute for Evolutionary Biology
durak@evolbio.mpg.de

Dr. Gokcen Eraslan
Genentech
eraslan.gokcen@gene.com

Mr. Kaan Ihsan Eskut
University of Kentucky
eskutkaan@gmail.com

Dr. Gilad Evrony
NYU
Gilad.Evrony@nyulangone.org

Ms. Sonia Eynard
Inrae Genphyse
sonia.eynard@inrae.fr

Dr. Khalid Fakhro
Sidra Medicine / Weill Cornell Medicine
kfakhro@sidra.org

Dr. Lingzhao Fang
Aarhus University
lingzhao.fang@qgg.au.dk

Ms. Lynn Fellman
Fellman Studio Inc
lynn@fellmanstudio.com

Dr. Scott Ferguson
UC Berkeley
scott.ferguson@berkeley.edu

Prof. Matt Field
James Cook University
matt.field@jcu.edu.au

Mr. Kwesi Forson
University of Virginia
bty6kj@virginia.edu

Ms. Kimberleigh Foster
Saint Louis Community College
fosterkimberleigh@gmail.com

Ms. Kyriaki Founta
Zucker School of Medicine at Hofstra/Northwell
kfounta@northwell.edu

Eden Francoeur
The Jackson Laboratory/ UConn Health
eden.francoeur@jax.org

Connor Frasier
Epicypher
cfrasier@epicypher.com

Dr. Megan Frayer
Yale University
megan.frayer@yale.edu

Dr. Donald Freed
Sentieon Inc
don.freed@sentieon.com

Dr. Juan Fuxman Bass
Boston University
fuxman@bu.edu

Mr. Nicolas Gaitan
Barcelona Supercomputing Center
nicolas.gaitan@bsc.es

Dr. Pedro Galante
Hospital Sirio-Libanes
pgalante@mochsl.org.br

Dr. Judit Garcia Gonzalez
Icahn School of Medicine at Mount Sinai
judit.garciagonzalez@mssm.edu

Dr. Erik Garrison
University of Tennessee Health Science Center
erik.garrison@gmail.com

Dr. Kristina Garske
Princeton University
kg8086@princeton.edu

Devin Gee
Feinstein Institutes
dgee@northwell.edu

Ilias Georgakopoulos-Soares
Penn State University College of Medicine
igeorgakopoulossoare@pennstatehealth.psu.edu

Stephanie Georges
University of Utah
stephanie.georges@genetics.utah.edu

Mr. Samuel Ghatan
New York Genome Center
sghatan@nygenome.org

Mr. Alejandro Gil Gomez
Stony Brook University
alejandro.gilgomez@stonybrook.edu

Dr. Richard Gill
UC Davis
rtgill@ucdavis.edu

David Gokhman
Weizmann Institute of Science
david.gokhman@weizmann.ac.il

Huanfa Gong
Zhejiang University
gonghuanfa@zju.edu.cn

Ms. Silvia Gonzalez-Lopez
Centre for Genomic Regulation (CRG)
silvia.gonzalez@crg.eu

Dr. M Grace Gordon
Genentech
gordon.gracie@gene.com

Dr. Alon Goren
University of California, San Diego
agoren@ucsd.edu

Simon Gravel
McGill University
simon.gravel@mcgill.ca

Dr. Olivia Gray
Variant Bio
olivia@variantbio.com

Dr. Marta Gronska-Peski
NYU Grossman School of Medicine
gronskapeski@gmail.com

Dr. Rodrigo Gularte Merida
Memorial Sloan Kettering Cancer Center
gularter@mskcc.org

Dr. Jennifer Guler
University of Virginia
jlg5fw@virginia.edu

Mr. Sakuntha Gunarathna
University of North Dakota School of Medicine
sakuntha.gunarathna@und.edu

Dr. Laura Gunsalus
Genentech
gunsalus.laura@gene.com

Ms. Ridhi Gutta
Curabitrix LLC
ridhigutta@gmail.com

Ms. Jessica Hacheney
Karolinska Institutet
jessica.hacheney@ki.se

Mr. Yoav Hadas
Icahn School of Medicine at Mount Sinai
yoav.hadas@mssm.edu

Dr. Maximilian Haeussler
University of California, Santa Cruz
maxh@ucsc.edu

Nadia Haghani
Stanford University
nhaghani@stanford.edu

Mr. Charles Hale
Cornell University
coh22@cornell.edu

Dr. Ira Hall
Yale University
ira.hall@yale.edu

Mr. Xikun Han
Peking University
hanxikun2017@gmail.com

Mr. Soeren Hansen
University of Copenhagen
soren.blikdal.hansen@sund.ku.dk

Dr. Hannah Happ
University of Uath
hannah.happ@genetics.utah.edu

Dr. Taslima Haque
University of Michigan
tahaque@umich.edu

Prof. Shinichi Hashimoto
Wakayama Medical University
hashimot@wakayama-med.ac.jp

Dr. Yaoxi He
Kunming Institute of Zoology. CAS
heyaoxi@mail.kiz.ac.cn

Mr. Jakob Heinz
Harvard University
jheinz@g.harvard.edu

Dr. Johannes Hellmuth
University of Munich (LMU)
jch2004@med.cornell.edu

Mx. Laurel Hiatt
University of Utah
laurel.hiatt@hsc.utah.edu

Dr. Stephanie Hicks
Johns Hopkins University
shicks19@jhu.edu

Carla Hoge
University of Chicago
choge@uchicago.edu

Dr. Aaron Holleman
Alnylam Pharmaceuticals
aholleman@alnylam.com

Rezwan Hosseini
University of Pittsburgh
SEH197@pitt.edu

Genevieve Housman
Max Planck Institute for Evolutionary Anthropology
genevieve_housman@eva.mpg.de

Mr. David Ziliang Hu
Karolinska Institutet
david.hu@ki.se

Hongru Hu
University of California-Davis
hrhu@ucdavis.edu

Mengying Hu
University of Pittsburgh
meh251@pitt.edu

Dr. Emilia Huerta-Sanchez
Brown University
emilia_huerta-sanchez@brown.edu

Naomi Huntley
Pennsylvania State University
nhuntley@umich.edu

Prof. Matthew Hurles
The Wellcome Sanger Institute
meh@sanger.ac.uk

Dr. Tadashi Imafuku
Wakayama Medical University
imafuku@wakayama-med.ac.jp

Dr. Mariko Isshiki
Albert Einstein College of Medicine
mariko.segawa@einsteinmed.edu

Dr. Nada Jabado
McGill University
nada.jabado@mcgill.ca

Dr. Kishore Jaganathan
Illumina Inc
kjaganathan@illumina.com

Chris Jakobson
Stanford University School of Medicine
cjakobso@stanford.edu

Mr. Zoeb Jamal
University of California, Davis
znjamal@ucdavis.edu

Mr. Benjamin James
Massachusetts Institute of Technology
benjames@mit.edu

Ms. Haerin Jang
Wellcome Sanger Institute
hj10@sanger.ac.uk

Dr. Peilin Jia
Beijing Institute of Genomes
pjia@big.ac.cn

Dr. Sheethal Jose
NIH/ NHGRI
sheethal.jose@nih.gov

Christina Jurotich
University of Wisconsin - Madison
jurotich@wisc.edu

Jorge Kageyama
Genentech
kageyamj@gene.com

Cynthia Kalita
University of Chicago
cakalita@gmail.com

Dr. Morten Kallberg
Novo Nordisk
qmok@novonordisk.com

Reshma Kalyan Sundaram
University of Pennsylvania
reshmaks@seas.upenn.edu

Mr. Nolan Kamitaki
Brigham and Women's Hospital
nolan_kamitaki@hms.harvard.edu

Bard Karlsen
Nordlandssykehuset, Forskningslab
baard.ove.karlsen@gmail.com

Dr. Rebecca Keener
Johns Hopkins University
rkeener@jhmi.edu

Dr. Andrew Kern
University of Oregon
adkern@uoregon.edu

Dr. Gaspard Kerner
Harvard School of Public Health
gkerner@hsph.harvard.edu

Devishi Kesar
DKFZ
devishi.kesar@kitz-heidelberg.de

April Kim
Johns Hopkins University
aprilkim@jhu.edu

Prof. Hie Lim Kim
Nanyang Technological University
hlkim@ntu.edu.sg

Dr. Doyeon Kim
Illumina
dkim5@illumina.com

Mr. Taeho Kim
University of Utah
taeho.kim@utah.edu

Dr. Sarah Kim-Hellmuth
LMU & Helmholtz Munich
sarah.kimhellmuth@med.uni-muenchen.de

Dr. Hamish King
Walter and Eliza Hall Institute
king.h@wehi.edu.au

Magda Kmiecik
The Jackson Laboratory
magda.kmiecik@jax.org

Ms. Sachiko Kobayashi
Shimadzu Corporation
kobayashi.sachiko.6jy@shimadzu.co.jp

Dr. Linda Koch
Springer Nature
l.koch@nature.com

Dr. Vamsi Kodali
National Library of Medicine, NIH
kodalivk@nih.gov

Mr. Samuel Koehler
Los Alamos National Laboratory
sikoehler@lanl.gov

Mr. Justin Koesterich
Rutgers University
jhk148@dls.rutgers.edu

Dr. Peter Koo
Cold Spring Harbor Laboratory
koo@cshl.edu

Dr. Amnon Koren
Roswell Park Comprehensive Cancer Center
amnon.koren@roswellpark.org

Dr. Diane Korngiebel
The Hastings Center
dmkorngiebel@gmail.com

Dr. Kanako Koyanagi
Hokkaido University
kkoyanag@ist.hokudai.ac.jp

Dr. Anat Kreimer
Rutgers University
kreimer@cabm.rutgers.edu

Mr. Arjun Sai Krishnan
Przeworski/Andolfatto Labs, Columbia University
ak4890@columbia.edu

Dr. Vivek Kumar
Cold Spring Harbor Laboratory
vkumar@cshl.edu

Dr. Nurdan Kuru
Cold Spring Harbor Laboratory
kuru@cshl.edu

Alexander Kwakye
Stony Brook University
alexander.kwakye@stonybrook.edu

Dr. Avantika Lal
Genentech
avantikalal02@gmail.com

Dr. Alice Lambolez
RIKEN
alice.lambolez@riken.jp

Mr. Mikel Lana Alberro
Max Planck Institute for Evolutionary Anthropology
mikel_lana_alberro@eva.mpg.de

Dr. Eric Lander
Broad Institute of MIT and Harvard
esloffice@broadinstitute.org

Prof. Tuuli Lappalainen
New York Genome Center & SciLifeLab
tlappalainen@nygenome.org

Ms. Larissa Lauer
Stanford School of Medicine
larissal@stanford.edu

Dr. Mara Lawniczak
Wellcome Sanger Institute
mara@sanger.ac.uk

Dr. Amanda Lea
Vanderbilt University
amanda.j.lea@vanderbilt.edu

Christopher Lee
UCSC
chmalee@ucsc.edu

Dr. Marcela Legue Cordero
NIMH/NIH
augusk@nih.gov

Marta Lemanczyk
Cold Spring Harbor Laboratory
lemanczyk@cshl.edu

Mr. Samuel Leppiniemi
University of Helsinki
samuel.leppiniemi@helsinki.fi

Steven Lewis
CSHL
steven.lewis@stonybrookmedicine.edu

Tony Li
University of Washington
tli824@uw.edu

Mr. Taibo Li
Johns Hopkins School of Medicine
taiboli@jhu.edu

Stacy Li
Sudmant Labk, University of California Berkeley
stacy-l@berkeley.edu

Dr. Qiuhui Li
Johns Hopkins University
qli111@jh.edu

Xintong Li
Cold Spring Harbor Lab
xili@cshl.edu

Mr. Jun Li
University of Michigan
junzli@med.umich.edu

Dr. Jiadong Lin
University of Washington
jdlin@uw.edu

Ms. Jiayi Liu
Rutgers University
jl2791@scarletmail.rutgers.edu

Prof. Boxiang Liu
National University of Singapore
boxiangliu@nus.edu.sg

Dr. Lingjie Liu
BioMarin Pharmaceutical Inc.
ling.liu@bmrn.com

Dr. George Liu
USDA-ARS
george.liu@usda.gov

Dr. Cong Liu
UCSD
col003@ucsd.edu

Dr. Kaiser Loell
Cold Spring Harbor Laboratory
loell@cshl.edu

Amy Longtin
Vanderbilt University
amy.l.longtin@vanderbilt.edu

Hailey Loucks
University of California, Santa Cruz
hloucks@ucsc.edu

Shuangjia Lu
Yale University
shuangjia.lu@yale.edu

Suhasini Lulla
Baylor College of Medicine
slulla@bcm.edu

Ms. Ava Mackay-Smith
Duke University
apm58@duke.edu

Dr. Kyle MacQuarrie
Stanley Manne Children's Research Institute
kmacquarrie@luriechildrens.org

Nicholas Mancuso
University of Southern California
nicholas.mancuso@med.usc.edu

Dr. Riley Mangan
MIT/Broad Institute
rimangan@mit.edu

Dr. Tomislav Maricic
Max Planck Institute for Evolutionary Anthropology
tomislav.maricic@gmail.com

Dr. Maximillian Marin
Harvard Medical School
mgmarin@g.harvard.edu

Dr. Franco Marsico
University of Tennessee Health Science Center
franco.lmarsico@gmail.com

Prof. Gabor Marth
University of Utah
gabor.marth@gmail.com

Ms. Dionne Martin
UGA
dm25256@uga.edu

Dr. Cristina Martin Linares
Johns Hopkins School of Medicine
cristina.martinlinares@jhu.edu

Dr. Nicole Martinez-Martin
Stanford Center for Biomedical Ethics
nicolemz@stanford.edu

Dr. Rachel Martini
Morehouse School of Medicine
rmartini@msm.edu

Arya Massarat
UC San Diego
amassara@ucsd.edu

Dr. Kaia Mattioli
Brigham and Women's Hosp. / Harvard Med. School
kaia.mattioli@gmail.com

Prof. Steven McCarroll
Harvard Medical School + Broad Institute
smccarro@broadinstitute.org

Dr. William McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Ms. Hunter McConnell
University of Missouri
hlmkh9@umsystem.edu

Cecilia McCormick
New York Genome Center
cmccormick@nygenome.org

Dr. Rajiv McCoy
Johns Hopkins University
rajiv.mccoy@jhu.edu

Dr. Matthew Meyerson
Dana-Farber Cancer Institute
matthew_meyerson@dfci.harvard.edu

Dr. Stephen Meyn
University of Wisconsin - Madison
stephen.meyn@wisc.edu

Mr. Nikhil Milind
Stanford University
nmilind@stanford.edu

Mr. John Miraszek
University of Missouri
jlmmpk@missouri.edu

Tara Mirmira
University of California, San Diego
tmirmira@ucsd.edu

Dr. Dan Mishmar
Ben-Gurion University of the Negev
dmishmar@bgu.ac.il

Dr. Dewi Moonen
Embl
dewi.moonen@embl.de

Prof. Jill Moore
University of Massachusetts Chan Medical School
Jill.Moore@umassmed.edu

Mr. Ricardo Moreira
Universitat Autònoma de Barcelona
r.moreira.pinhal@gmail.com

Dr. Stephen Mosher
Johns Hopkins University
smosher3@jhu.edu

Dr. Sara Mostafavi
University of Washington
saramos@cs.washington.edu

Ioannis Mouratdis
Penn State College of Medicine
ipm5219@psu.edu

Mr. Surag Nair
Genentech Inc.
nair.surag@gene.com

Dr. RK Narayanan
Cold Spring Harbor Laboratory
narayan@cshl.edu

Dr. Nasna Nassir
Mohammed Bin Rashid University of Medicine and Health Sciences
nasna.nassir@dubaihealth.ae

Nicholas Navin
MD Anderson Cancer Center
nnavin@mdanderson.org

Akshatha Nayak
Penn state College of Medicine
abn5461@psu.edu

Mr. Bohan Ni
Johns Hopkins University
bni1@jhu.edu

John Novembre
University of Chicago
jnovembre@gmail.com

Dr. Ryo Nozu
Hiroshima University
rnozu22@hiroshima-u.ac.jp

Mr. Chosen Obih
The University of Arizona
chosenobih@arizona.edu

Dr. Dong-Ha Oh
National Center for Biotechnology Information
dongha.oh@nih.gov

Dr. Sungryong Oh
Children's Hospital of Philadelphia
ohs5@chop.edu

Dr. Naoki Osato
Institute of Science Tokyo
naokiosato11@gmail.com

Prof. Svante Paabo
Max Planck Institute for Evolutionary Anthropology
paabo@eva.mpg.de

Dr. Petar Pajic
University at Buffalo
petarpaj@buffalo.edu

Dr. Vasili Pankratov
Aarhus University
vasilipankratov@gmail.com

Ms. Katherine Pardo
National Institutes of Health
maychristine.malicdan@nih.gov

Dr. Jiyeon Park
The Catholic University of Korea
parkji7@gmail.com

Eddie Park
Children's Hospital of Philadelphia
parke2@chop.edu

Ms. Sowmya Parthiban
Johns Hopkins University
sowmyaparthiban@gmail.com

Ms. Raphaela Pensch
Uppsala University
raphaela.pensch@imbim.uu.se

Ms. Claudia Perez-Calles
EMBL-EBI & University of Cambridge
cperez@ebi.ac.uk

Dr. Rachel Petersen
Vanderbilt University
rpetersen42@gmail.com

Cheng Fei Phung
NA
phungchengfei@gmail.com

Dr. Luca Pinello
Harvard Medical School/ MGH/BROAD
lpinello@mgh.harvard.edu

Dr. Anna Poetsch
TU Dresden
anna.poetsch@tu-dresden.de

Sebastian Pott
University of Chicago
spott@uchicago.edu

Gabriela Pozo
Universidad San Francisco de Quito
gpozo@usfq.edu.ec

Dr. Aaron Quinlan
University of Utah
aquinlan@genetics.utah.edu

Mr. Henry Raeder
The University of Chicago
hraeder@uchicago.edu

Dr. Srilakshmi Raj
Albert Einstein College of Medicine
srilakshmi.raj@einsteinmed.edu

Prof. Sohini Ramachandran
Brown University
sramachandran@brown.edu

Ms. Srividya Ramakrishnan
Johns Hopkins University
sramakr4@jh.edu

Fabian Ramos-Almodovar
University of Pennsylvania
ramosalf@pennmedicine.upenn.edu

Mr. Vasilios Raptis
University of Edinburgh
V.Raptis@sms.ed.ac.uk

Dr. John Ray
Benaroya Research Institute
jray@benaroyaresearch.org

Dr. Soumya Raychaudhuri
Broad Institute of MIT and Harvard
soumya@broadinstitute.org

Dr. Fairlie Reese
Barcelona Supercomputing Center
fairlie.reese@gmail.com

Clara Rehmann
University of Oregon
crehmann@uoregon.edu

Mr. Raimonds Rescenko-Krums
University of Latvia
raimonds.rescenko@gmail.com

Mr. Manuel Rivas
Stanford University
mrivas@stanford.edu

Iker Rivas-González
Max Planck Institute for Evolutionary Anthropology
iker_rivas_gonzalez@eva.mpg.de

Kaeli Rizzo
Cold Spring Harbor Laboratory
rizzo@cshl.edu

Murillo Rodrigues
Oregon National Primate Center
rodrigmu@ohsu.edu

Dr. Sara Rohban
Cell Press
srohban@cell.com

Dr. Leah Rosen
Science for Life Laboratory/KTH (CBH)
leah.rosen@scilifelab.se

Dr. Jonathan Rosen
University of North Carolina at Chapel Hill
jdrosen@live.unc.edu

Jonathan Rosenski
The Hebrew University of Jerusalem
jonathan.rosenski@mail.huji.ac.il

Dr. Maxime Rotival
Institut Pasteur
maxime.rotival@pasteur.fr

Dr. Joel Rozowsky
Yale University
joel.rozowsky@yale.edu

Mr. Zubairu Saidu Rayyanu
Gombe State University
zubairbinauwam@gmail.com

Farnaz Salehi
The University of Tennessee Health Science Center
fsalehi1@uthsc.edu

Dr. Mark Sanda
Stony Brook University
mark.sanda@stonybrook.edu

Mr. Walter Santana Garcia
European Molecular Biology Laboratory
wsantana@ebi.ac.uk

Thomas Sasani
University of Utah
thomas.a.sasani@gmail.com

Dr. Michael Schatz
Johns Hopkins University
mschatz@cs.jhu.edu

Ms. Kendra Scheer
University at Buffalo
kendrasc@buffalo.edu

Dr. Robert Schnabel
University of Missouri
schnabelr@missouri.edu

Casey Sederman
University of Utah
casey.sederman@hsc.utah.edu

Evan Seitz
Cold Spring Harbor Laboratory
seitz@cshl.edu

Dr. Isabel Serrano
University of California, Berkeley
isabel.serrano@genetics.utah.edu

Dr. Mithun Shah
Mayo Clinic
shah.mithun@mayo.edu

Mr. Kobi Shapira
Bar-Ilan University
shapirakobi@gmail.com

Mr. Marwan Sharawy
Pasteur institue
marwan.sharawy@pasteur.fr

Ruhollah Shemirani
Icahn School of Medicine at Mount Sinai
ruhollah.shemirani@mssm.edu

Gloria Sheynkman
University of Virginia School of Medicine
gs9yr@virginia.edu

Mr. Vikram Shivakumar
Johns Hopkins University
vshivak1@jhu.edu

Mr. Noam Shtolz
Ben-Gurion University of the Negev
shtolz@post.bgu.ac.il

Ms. Kinga Sidzinska
Merritt College
ksidzinska@peralta.edu

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Ms. Faith Sim Chin Yee
Nanyang Technological University
simc0028@e.ntu.edu.sg

Dr. Amartya Singh
Rutgers Cancer Institute of New Jersey
as2197@scarletmail.rutgers.edu

Dr. Linnea Smeds
Penn State University
lbs5874@psu.edu

Dr. Jeramiah Smith
University of Kentucky
jjsmit3@uky.edu

Ms. Leslie Solorzano
Uppsala University
leslie.solorzano@igp.uu.se

Dr. Joseph Solvason
UCSD
solvason@ucsd.edu

Dr. Dongyuan Song
UConn Health Center
dosong@uchc.edu

Janet Song
Boston Children's Hospital
janet.song@childrens.harvard.edu

Dr. Cynthia Soto
Lieber Institute for Brain Development
cyntsc10@gmail.com

Dr. Pieter Spealman
Broad Institute of MIT and Harvard
pspealma@broadinstitute.org

Samvardhini Sridharan
University of California, Berkeley
sridharan@berkeley.edu

Mr. Itamar Twersky
Bar-Ilan University
Itamarty@gmail.com

Dr. Yasin Uzun
Penn State University College of Medicine
yuzun@pennstatehealth.psu.edu

Sarah Vaccaro
Stony Brook University
sarah.vaccaro@stonybrook.edu

Anna Darlene Van der Heiden
Uppsala University
anna.vd.heiden@imbim.uu.se

Dr. Christina Vasilopoulou
European Molecular Biology Laboratory
christina@ebi.ac.uk

Ms. Eduarda Vaz
Johns Hopkins
evaz2@jh.edu

Dr. Juan Vazquez
University of California, Berkeley
aging@berkeley.edu

Dr. Krishna Veeramah
Stony Brook
krishna.veeramah@stonybrook.edu

Dr. Tauras Vilgalys
University of Chicago
taur.vil@gmail.com

Dr. Deven Vyas
Stony Brook University
deven.vyas@stonybrook.edu

Ms. Isha Walawalkar
UConn Health Center
walawalkar@uchc.edu

Mr. Sumit Walia
UC San Diego
swalia@ucsd.edu

Prof. Jeff Wall
Oregon Health and Science University
walljef@ohsu.edu

Ms. Jiahui Wang
Division of Cancer Epidemiology and Genetics, NCI
jiahui.wang2@nih.gov

Dr. Chengqi Wang
University of South Florida
chengqi@usf.edu

Zishan Wang
Icahn School of Medicine at Mount Sinai
zishan.wang@mssm.edu

Ruoyu Wang
UT Southwestern
ruoyu.wang@utsouthwestern.edu

Lingfei Wang
UMass Chan Medical School
Lingfei.Wang.CN@gmail.com

Prof. Xiaoyue Wang
Chinese Academy of Medical Sciences
wxy@ibms.pumc.edu.cn

Dr. Jinhua Wang
University of Minnesota-Twin Cities, Masonic Compr
wangjh@umn.edu

Shuyue Wang
MD Anderson Cancer Center
swang21@mdanderson.org

Dr. Alistair Ward
University of Utah
AlistairNWard@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory
ware@cshl.edu

Dr. Doreen Ware
Cold Spring Harbor Laboratory
ware@cshl.edu

Dr. Marina Watowich
Vanderbilt University
marina.watowich@vanderbilt.edu

Tanye Wen
University of California Irvine
tanyew@uci.edu

Ian Whaling
Stanford University
iwhaling@stanford.edu

Mr. Johannes Wibisana
Okinawa Institute of Science and Technology
johannes.nicolaus@oist.jp

Dr. Patricia Wittkopp
University of Michigan
wittkopp@umich.edu

Ms. Sarah Wright
University of California, San Diego
snwright@ucsd.edu

Peipei Wu
Cold Spring Harbor Lab
pwu@cshl.edu

Dr. Bo Xia
Broad Institute of MIT and Harvard
xiabo@broadinstitute.org

Zhuorui Xie
University of Pennsylvania
sherry.xie@pennmedicine.upenn.edu

Dr. Jiawei Xing
Cold Spring Harbor Laboratory
xing@cshl.edu

Mr. Ke Xu
Yale University
k.xu@yale.edu

Dr. Huilin Xu
Mass General Hospital
huxu1@mgh.harvard.edu

Tianyao Xu
UC San Diego
tix034@ucsd.edu

Mr. Illya Yakymenko
Universitat Autònoma de Barcelona
illya.yakymenko@uab.cat

Dr. Xiaoxu Yang
University of Utah
xiaoxu.yang@genetics.utah.edu

Jianing Yao
Johns Hopkins Bloomberg School of Public Health
jyao37@jhmi.edu

Ms. Bin Ye
Regeneron Pharmaceuticals Inc.
bin.ye@regeneron.com

Zhezhen Yu
Cold Spring Harbor Laboratory
zhezhen@cshl.edu

Ms. Xinrui Yu
Michigan State University
yuxinrui@msu.edu

Ms. Sehee Yun
Korea University
seheeyun@korea.ac.kr

Samantha Zarnick
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
neuroadmin@unc.edu

Dr. Hugo Zeberg
Karolinska Institutet
hugo.zeberg@ki.se

Dr. Xin Zeng
Cold Spring Harbor Laboratory
zeng@cshl.edu

Mr. Shilong Zhang
Shanghai Jiao Tong University
shilong.zhang@sjtu.edu.cn

Dr. Xuan Zhang
UCSD
xuz043@ucsd.edu

Alouette Zhang
McGill University
zhiwei.zhang@mail.mcgill.ca

Mr. Artemy Zhigulev
KTH Royal Institute of Technology / SciLifeLab
artemy.zhigulev@scilifelab.se

Jian Zhou
UT SOUTHWESTERN MEDICAL CENTER
jian.zhou@utsouthwestern.edu

Helyaneh Ziaei Jam
University of California San Diego
hziaeija@ucsd.edu

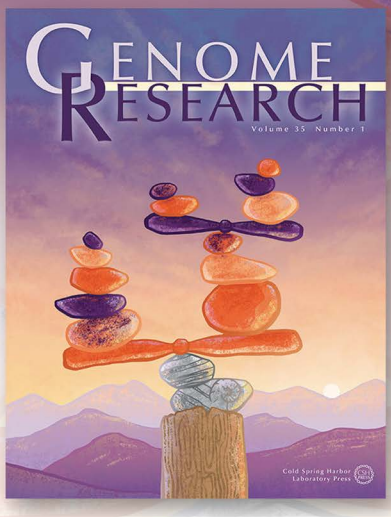
Ms. Julia Zulawinska
University of Virginia
egr3we@virginia.edu

Melissa Zwaig
Cold Spring Harbor Laboratory
zwaig@cshl.edu

Meet the Editors

Jennifer DeLeon, Ahmet Denli,
Hillary Sussman





Wednesday

7:00–9:00 PM

and

Friday

1:30–3:30 PM

Bush Auditorium Lobby

CSH PRESS

www.genome.org

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**
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400**

GENERAL INFORMATION

Meetings & Courses Main Office

Hours during meetings: M-F 9am – 9pm, Sat 8:30am – 1pm
After hours – See information on front desk counter
For assistance, call Security at 516-367-8870
(x8870 from house phone)

Dining, Bar

Blackford Dining Hall (main level):
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Blackford Bar (lower level): 5:00 p.m. until late

House Phones

Grace Auditorium, upper / lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

Books, Gifts, Snacks, Clothing

CSHL Bookstore and Gift Shop
516-367-8837 (hours posted on door)
Grace Auditorium, lower level.

Computers, E-mail, Internet access

Grace Auditorium
Upper level: E-mail and printing in the business center area
WiFi Access: GUEST (no password)

Announcements, Message Board Mail, ATM, Travel info

Grace Auditorium, lower level

Russell Fitness Center
Dolan Hall, east wing, lower level
PIN#: **(On your registration envelope)**

Laundry Machines
Dolan Hall, lower level

Photocopiers, Journals, Periodicals, Books
CSHL Main Library
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 Campus Map

CSHL Map