ProbGen 2026 Abstract Booklet

March 25–28, 2026

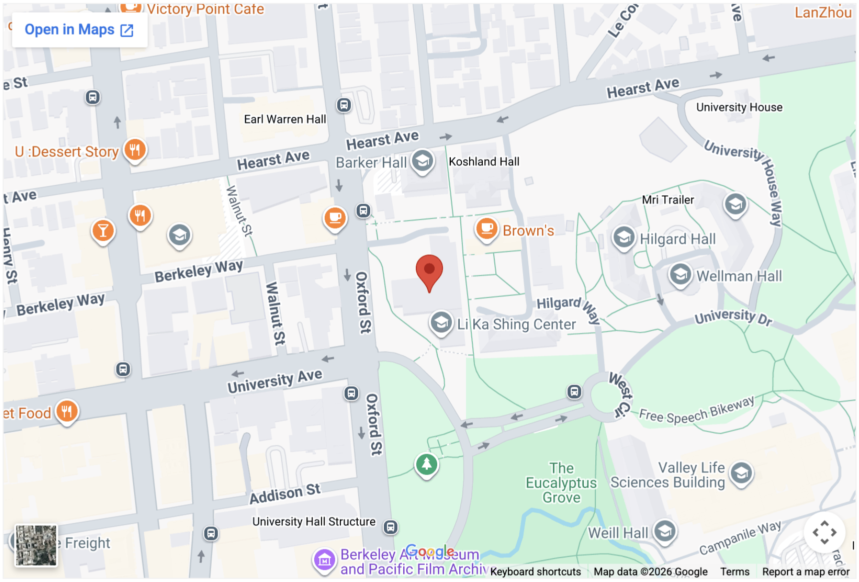
University of California, Berkeley

Venue

Li Ka Shing Center
University of California, Berkeley
1951 Oxford Street
Berkeley, CA 94720

View on Google Maps

Located here: Center for Biomedical and Health Sciences



Supported by

Center for Computational Biology, UC Berkeley
Center for Theoretical and Evolutionary Genomics

Industry Gold Sponsor

Genentech
A Member of the Roche Group

Industry Silver Sponsors:

Human Population Genetics and Genomics

ancestry

Program

For each talk: 20-minute presentation followed by a 5-minute Q&A.
Suggested poster size: 30" × 40".

Wednesday, March 25

8:00–8:45 Registration, Coffee, and Light Breakfast

8:45–9:00 Opening Remarks

9:00–10:15 Session: Machine Learning in Genomics I
Chairs: Sara Mostafavi and Sriram Sankararaman

- 9:00 AM — Sriram Sankararaman, *GPC: An expressive, tractable, and privacy-preserving deep generative model for genetic variation data*
- 9:25 AM — Matteo Fumagalli, *Generative models for clinical and population genetic data*
- 9:50 AM — Jan Engelmann, *Scalable and robust mapping of eQTL interactions reveals immune cell contexts of regulatory variant activity*

10:15–10:45 Break

10:45–12:00 Session: Machine Learning in Genomics II

- 10:45 AM — Sara Mostafavi, *Scaling Sequence-to-Function Models for Regulatory Variant Interpretation*
- 11:10 AM — Kuan-Hao Chao, *Predicting dynamic expression patterns in budding yeast with a fungal DNA language model*
- 11:35 AM — Chengzhong Ye, *Predicting functional constraints with phylogeny-informed genomic language models*

12:00–1:00 Lunch

1:00–2:00 GHIST Workshop (Optional)

2:00–4:00 Poster Session I

4:00–8:00 Wine & Cheese Reception

Thursday, March 26

8:00–9:00 Coffee and Light Breakfast

9:00–10:15 Session: Quantitative Genetics and Association Mapping I

Chairs: Sasha Gusev and Sohini Ramachandran

- 9:00 AM — Sohini Ramachandran, *It's All Relative: Turning Biobank 'Nuisance' Kinship into New Genetic Insights*
- 9:25 AM — Lin Poyraz, *Leveraging relatedness in genomic datasets to study variation in germline mutation rates*
- 9:50 AM — Roshni Patel, *The ubiquity of evolutionary theory: phylogenetic approaches to genetic confounding in epidemiology and psychology*

10:15–10:45 Break

10:45–12:00 Session: Quantitative Genetics and Association Mapping II

- 10:45 AM — Sasha Gusev, *Learning context specific disease mechanisms from single cell data*
- 11:10 AM — Scott Sauers, *Calibration of polygenic scores across ancestries*
- 11:35 AM — Max Dudek, *A meta-analysis of chromatin accessibility QTLs explains differences in QTL-GWAS colocalization*

12:00–1:30 Lunch

1:30–2:30 Keynote I: Jonathan Pritchard, *From genes to programs to traits: Building causal models for human genetics with GWAS and perturb-seq*

2:30–3:00 Break

3:00–4:15 Session: Systems Biology I

Chairs: Christina Leslie and Heng Li

- 3:00 PM — Christina Leslie, *Learning multiomic cell dynamics with neural ODEs*
- 3:25 PM — Nicholas Ingolia, *Interpretable systems biology models from genetic interaction maps*
- 3:50 PM — Alexis Garretson, *Bayesian approaches for considering cell type composition when assessing the heritability and covariance of aging-related genomic traits*

4:15–4:45 Break

4:45–6:00 Session: Systems Biology II

- 4:45 PM — Heng Li, *Modeling phylogenetic trees with Stochastic Context-Free Grammar (SCFG)*
- 5:10 PM — Magdalena Russell, *Inferring signatures of time-varying mutational mechanisms during antibody affinity maturation*
- 5:35 PM — Jiawei Xing, *VOUS: Variational Ornstein-Uhlenbeck linking single-cell lineage tracing with stochastic gene expression*

6:00–8:00 Self-organized Dinner

Friday, March 27

8:00–9:00 Coffee and Light Breakfast

9:00–10:15 Session: Phylogenetics and Evolutionary Dynamics I
Chairs: Erick Matsen and Julia Palacios

- 9:00 AM — Julia Palacios, *Coalescent Inference for Epidemics with Latent Periods*
- 9:25 AM — Priscilla Lau, *A phylogenetic state-dependent Ornstein-Uhlenbeck model to jointly infer discrete and continuous trait evolution*
- 9:50 AM — Maximillian Newman, *There are no unlinked loci: How pedigrees couple neutral genealogies across the genome*

10:15–10:45 Break

10:45–12:00 Session: Phylogenetics and Evolutionary Dynamics II

- 10:45 AM — Erick Matsen, *Transformer-powered mutation-selection models for antibody evolution*
- 11:10 AM — Annabel Large, *Thousand-parameter TKF-based models with latent and hierarchical structure predict protein sequence evolution competitively with million-parameter neural models*
- 11:35 AM — Adam Siepel, *Variational Inference with Node Embeddings (VINE) for scalable Bayesian phylogenetics*

12:00–1:30 Lunch

1:30–2:30 Keynote II: Sally Otto, *Evolutionary Dynamics when Selection and Drift Occur in Haploid and Diploid Stages*

2:30–3:00 Break

3:00–4:15 Session: Population and Statistical Genetics I
Chairs: Rori Rohlfs and Pier Palamara

- 3:00 PM — Rori Rohlfs, *Working with wildly varying mutation rates in TRs to investigate their role in adaptation*
- 3:25 PM — Wenhan Lu, *Heritability and effect-size distribution of rare protein-coding variation*
- 3:50 PM — Xinyi Li, *Improving gene effect estimates with functionally informed hierarchical GeneBayes*

4:15–4:30 Break

4:30–6:30 Poster Session II

6:30–9:00 Conference Dinner

Saturday, March 28

8:00–9:00 Coffee and Light Breakfast

9:00–10:15 Session: Population and Statistical Genetics II

- 9:00 AM — Pier Palamara, *Scalable inference and analysis of ancestral recombination graphs*
- 9:25 AM — Simon Gravel, *Inference in large pedigrees*
- 9:50 AM — Nicole Kleman, *Increased genetic diversity and residual stratification contribute to polygenic score prediction accuracy*

10:15–10:45 Break

10:45–12:00 Session: Demographic Inference I

Chairs: Jonathan Terhorst and Anastasia Ignatieva

- 10:45 AM — Anastasia Ignatieva, *Analysing haplotype structure in ARGs: From recombination suppression to phantom epistasis*
- 11:10 AM — Yulin Zhang, *Recovering signatures of archaic introgression using ancestral recombination graphs*
- 11:35 AM — Romain Fournier, *LDate: An expectation-maximization approach for dating ancient individuals based on linkage disequilibrium*

12:00–1:00 Lunch

1:00–2:15 Session: Demographic Inference II

- 1:00 PM — Jonathan Terhorst, *Demographic inference using ARGs: theory, methods, and statistical foundations*
- 1:25 PM — James Kitchens, *Reconstructing the spatial histories of populations from gene trees*
- 1:50 PM — Yixin Zhu, *A new approach for estimating the fitness effects of missense and non-coding variants*

2:15–2:30 Closing Remarks

Poster Schedule

Poster Session I

Poster #	Presenter	Poster Title
1	Ruoyao Shi	High dimensional confounder adjustment for Multivariable Mendelian randomization using Genetic Factor Analysis
2	Ahmed Selim	A Coalescent Latent Space Model for flexible depiction of structure through time
3	Aaron Ragsdale	Inferring ancient introgression from single unphased genomes and a two-locus statistic
4	Jiatong Liang	A fast general framework for computing functionals of coalescent rates for complex demographies
5	Hao Shen	Fine-Resolution Asymmetric Migration Estimation
6	Chris Kyriazis	The genomic consequences of near extinction and recovery in ʻAlalā, the Hawaiian crow
7	Dmitry Biba	Inferring epistasis in evolutionary accumulation processes
8	Margarita Geleta	A Tsallis-Entropy Lens on Genetic Variation
9	Guy Sella	Limits on the number of traits maintained by stabilizing selection: revisiting Barton's paradoxes
10	Kiley Hudson	Discovery of Pacific Islander-Specific Segmental Duplications in 46 Haploid Assemblies
11	Huisheng Zhu	Genetic architectures of human complex traits reveal stronger selective constraint on variants for brain-related traits
12	Lucas Sort	Robust detection of selection from ancient DNA time series using mixed models

Poster #	Presenter	Poster Title
13	Magnus Nordborg	Towards an unbiased characterization of genetic polymorphism
14	Maike Morrison	A general FST framework reveals the variability of rare versus common alleles
15	Marc Henein	Sampling Alternative Gene Genealogies in an Ancestral Recombination Graph
16	Runyang Nicolas Lou	The evolution of structural and single nucleotide mutation across haplotype-resolved vertebrate genome assemblies
17	Peter Laurin	Adaptations spreading across human gut microbiomes arise from complex multisite adaptive architectures
18	Ryan Gutenkunst	Quantifying selection on the nonsynonymous human mutation spectrum
19	Sebastián Iturbe	Leveraging the Ancestral Recombination Graph to Infer Microsatellite Mutational Models
20	Matin Saeidi	Dynamics of recessive mutator alleles: model and data
21	Tami Gjorgjieva	How natural selection shapes gene regulatory architecture: the case for 5' UTRs
22	Thomas Bøggild	Utilizing haplotype cluster assignments for fast and accurate local ancestry estimation and deconvolution
23	Alouette Zhang	Exploring selective scanning by LD statistic Dz: simulations and empirical studies
24	Amy Goldberg	Rare variation in malaria parasites biases population-genetic inference
25	Benjamin Peter	Population Genetic Analysis of highly degraded ancient-DNA data
26	Nathan W. Anderson	A model of background selection in changing populations
27	Sarah Johnson	Modeling Local Ancestry Covariance to Infer the Timing of Denisovan Admixture Events

Poster #	Presenter	Poster Title
28	Sean Yetter	General Moment Closure for the Neutral Two-Locus Wright-Fisher Dynamics
29	Tommaso Stentella	Bayesian Inference of Demographic History and Structure from the Distribution of Heterozygous Sites Distances
30	Tianyi Wang	Detecting deep-time selection sweeps using coalescent-based inference
31	Amy Williams	The Poisson recombination model artificially increases the spatial dependence of coalescence and ignores sex-biased transmission
32	Ivan Specht	Joint Phylogenetic and Transmission Inference with JUNIPER
33	Giovanni Mazzeo	An SFS based method for detecting seasonal balancing selection
34	Pratik Katte	Biobank-scale visualization and interactive exploration of ancestral recombination graphs with LORAX
35	Remus Stana	Systematic estimation of the number of mutational origins in large population cohorts
36	Zhigui Bao	The Arabidopsis pangenome reveals highly divergent haplotypes persisting across species boundaries and populations
37	Ziyue Gao	Trinucleotide genome composition reflects context-dependent mutation spectra in mammals and plants
38	Antoine Aragon	Searching for signatures of selection in B-cell affinity maturation
39	Ali Osman Berk Sapci	Pseudo-likelihood kmer-based calculation of regional genomic distances and applications to phylogenetics
40	Frederick Matsen	Separating selection from mutation in antibody language models
41	Fabian Ramos-Almodovar	Understanding the evolution of mutation and recombination with large scale phylogenomics

Poster #	Presenter	Poster Title
42	Junyan Dai	Quartet-based species tree methods enable fast and consistent tree of blobs reconstruction under the network multispecies coalescent
43	Julia A. Palacios	Coalescent Inference for Epidemics with Latent Periods
44	Audrey Li-Wen Wang	Phylogenetic Deconvolution of Wastewater Sequencing Data to Detect Influenza Reassortment
45	Lorenzo Cappello	Test for association between responses and tree structures
46	Maya Lemmon-Kishi	ratePlacer: A Phylogenetic Framework for Molecular Dating of Ancient Environmental DNA
47	Vladimir Seplyarskiy	A Phylogeny-Based Framework to Uncover the Evolution of Mutational Processes in Primates
48	Michael Wasney	Uniform bacterial genetic diversity along the gut
49	Isaac Goldstein	Coalescent Inference for Pathogens with Latent Periods
50	Mara Baylis	Lineage specific selection in the mound-building mouse Mus spicilegus
51	Peter Gerlach	The Genomic Landscapes of Oceania and ISEA: Insights from 92 populations
52	Portalier	Detecting changes in fitness optimum from genetic and phenotypic data
53	Yan Wong	ARGs of multiple chromosomes

Poster Session II

Poster #	Presenter	Poster Title
1	Ezequiel Alejandro Galpern	Disentangling biophysical effects of rare missense variants on cis and trans plasma protein levels in the UK Biobank
2	Michelle Kim	Inferring effect-size and selection coupling in admixed populations using machine learning models
3	Liana Lareau	Protein language models reveal evolutionary constraints on synonymous codon choice
4	Liubov Shilova	REECAP: Contrastive learning of retinal aging reveals genetic loci linking morphology to eye disease
5	Miquel Anglada-Girotto	Genetic disruption of transcriptional specificity underlies rare disease risk
6	Shiron Drusinsky	Genomic Features Explaining Deep Learning Predictions of cis-Regulatory Variant Effects
7	Xinru Zhang	Identification of Early Exhaustion–Specific Regulatory Programs in T Cells
8	Annie Xie	Comparison of Methods for Non-Negative Covariance Matrix Decomposition
9	Yazheng Di	Learning Lifetime Disease Liability Reveals and Removes Genetic Confounding in Electronic Health Records
10	Cindy Gilda Santander	A deep learning approach to detecting negative frequency-dependent selection
11	Mark Chernyshev	Detecting Sapiens-Specific Selective Sweeps by Leveraging Deep Divergence and Machine Learning
12	Moisès Coll Macià	Enhancement of hidden Markov model analyses for improved inference of archaic introgression in modern humans
13	Prakruthi Burra	Reconstruction of splicing regulatory networks from single cells

Poster #	Presenter	Poster Title
14	Arun Durvasula	Exposure accumulation drives age-dependent disease architectures and polygenic risk scores
15	Mark Bitter	High-resolution mapping of a rapidly evolving complex trait reveals genotype-phenotype stability and a complex genetic architecture of adaptation
16	Jingyou Rao	Combinatorial Mutational Scanning Libraries for Probing Global Epistasis in Proteins
17	Albert Xue	keju: powerful and accurate enhancer effect estimation in Massively Parallel Reporter Assays
18	Federico Billeci	Rare-common variant interactions influence genetic disease and explain incomplete penetrance
19	William H. Majoros	Structured Modeling Improves Estimation of Allelic Effects in Gene Expression
20	Marius Weidmann	GPU-accelerated coordinate ascent variational inference for scalable prediction and association analyses
21	Alejandro Ochoa	An elegant linkage disequilibrium model for structured populations applied to polygenic risk scores for linear mixed-effects model summary statistics
22	Javier Maravall-López	M. tuberculosis and gastrointestinal pathogens drove immune adaptations in ancient West Eurasians
23	Jeremy Wang	Concentrating association power along specific biological pathways by controlling for heritable covariates in proxy GWAS
24	Walid Mawass	Theoretical bounds and empirical diagnostics for confounding control in polygenic prediction
25	Michael Goldberg	Heritability of germline mutagenesis in 40 large three- and four-generation pedigrees
26	Nikhil Milind	Buffering of gene dosage response curves for human complex traits

Poster #	Presenter	Poster Title
27	Shaila Musharoff	Social Factors of Health Covary with Population Stratification and Confound Heritability Estimates
28	Joshua Schraiber	Pleiotropic stabilizing selection shapes genetic architecture of complex traits
29	Shivam Gandhi	Effect Size Correlations of Proximal SNPs in a Complex Trait
30	Yining Fan	Scalable genealogical association testing for binary traits using ancestral recombination graphs
31	Yun Deng	Uncovering the Dynamics of Population Structure Through Time Using Genome-Wide Genealogies
32	Rowan Hart	Cophylogeny between individuals within a single host species and their symbionts
33	Natanael Spisak	Collateral mutagenesis funnels multiple sources of DNA damage into a ubiquitous mutational signature
34	Thomas Atkins	Quantifying Within-Individual Immune Variability
35	Alexis Edozie	Accelerated Phylodynamic Inference via Neural ODE Solvers
36	Alvina Adimoelja	Inferring the origins of precancerous lesions from spatially-informed genomic sampling
37	Bjarke Meyer Pedersen	Shared polymorphism investigation on X and autosomes in primates
38	Genona T. Maseras	Demographic and familial information dominates psychiatric risk prediction, while polygenic risk shows age-dependent associations
39	Gillian Meeks	MeQTL Discovery in Admixed Human Genomes to Estimate Epistasis
40	Grace Brophy	Uncovering structure in genetic interaction networks of yeast
41	Chester Henry Charlton	Classification of Single-Strand Methylation with PacBio SMRT Data

Poster #	**Presenter**	**Poster Title**
42	Kai Shimagaki	Inferring epistasis from temporal genetic sequence data and the limitations of inference
43	Matteo	Cross-species gene expression encodes signatures of gene essentiality
44	Monica Arniella	Error correction of SARS-CoV-2 genomic sequences using a phylogenetic prior
45	Noel McAllister	A Neural Network Investigation of Evidence for Ghost Introgression in Denisovans
46	Sophie K. Joseph	Neanderthal Ancestry in East Asia
47	Steven Sun	The Impact of Associative Overdominance on Genetic Diversity
48	Leandra Braeuninger	Towards continuous ancestry PRS via genetic distance kernel methods
49	Christophe Thomassin	Polygenic risk and association beyond linearity
50	Dat Do	Inconsistency of model selection method in admixture model using second-order changes of log-likelihood

GHIST Genomic History Inference Strategies Tournament

PROBGEN 2025 - Wednesday at 1:00 pm

GHIST Workshop

Community benchmarks for population genomic inference

Unbiased community-driven benchmarking for population genomics methods

Population genomics offers a dizzying array of approaches for inferring demographic history, natural selection, and other evolutionary processes. But which approaches work best in which circumstances? The Genomic History Inference Strategies Tournament is an annual open competition in which participants analyze simulated genomic datasets using any approaches they choose - with no prior knowledge of the true parameters. The result: a transparent, crowd-sourced evaluation of what works, when, and for whom.

120+ registered participants between GHIST 2024 & 2025

MBE GHIST 2024 results published in *Mol. Biol. Evol.*

Workshop Wednesday 1:00 pm

Why Participate?

- Test your methods against ground-truth simulated data in a no-stakes environment
- Learn how your tools compare to the field - without publication bias
- Top competitors become co-authors on the annual results paper and earn cash prizes
- Excellent training for students and early-career researchers new to inference methods
- Use any approach you like: no prescribed tools or frameworks

TIP

Hosted web apps for bottleneck demographic inference and sweep detection generate submission-ready output files, no coding required to try GHIST.

How It Works

Each year, GHIST releases simulated population genomic datasets and a set of inference challenges. Participants analyze the data independently, then submit inferences by the deadline.

- **Datasets:** VCF files from forward-time or coalescent simulations with known parameters
- **Submission:** Simple text file with inferred values - no code required
- **Challenges:** Varied difficulty, from single-population demography to archaic admixture and complex sweep detection
- **Evaluation:** Blind automated scoring against true parameters
- **Transparency:** Results shared openly with the community via Synapse

2025

GHIST 2025 challenges included demographic history inference and selective sweep detection, at multiple difficulty levels.

This Workshop

This session walks through GHIST from the ground up: structure, challenge design, how to analyze and submit, and what we have learned so far about inference method performance.

- Overview of competition design philosophy
- Walkthrough of a GHIST challenge dataset
- Discussion of GHIST 2024 and 2025 results
- Hands-on: download data, run an inference, submit
- Community input on GHIST 2026 challenge design

JOIN

Subscribe to the GHIST mailing list to be notified of future competitions and results. Challenge design proposals are welcome!

https://ghist.bio

ORGANIZED BY

Ryan Gutenkunst
University of Arizona
rgutenk@arizona.edu

SUPPORTED BY

NIH NIGMS R35 GM149235

Talks

Ordered alphabetically except keynote speakers

Keynote I

From genes to programs to traits: Building causal models for human genetics with GWAS and perturb-seq

Jonathan Pritchard

Stanford University

Genome-wide association studies (GWAS) provide a unique and powerful tool for identifying causal links from variants to genes to human traits and diseases. Although modern GWAS gives an information-rich readout of the relevant variants and genes, it remains very challenging to turn this into mechanistic models of disease and clinical applications. In this talk I will describe how new genome-wide CRISPR-based perturbations provide a critical interpretive key for human genetics data, including our recent proof-of-concept study inferring causal graphs for red blood cell-related traits such as hemoglobin levels, and our new genome-wide perturb-seq of primary T cells in multiple stimulation contexts. I will close with a broader discussion of the opportunities and open challenges in this field.

Keynote II

Evolutionary Dynamics when Selection and Drift Occur in Haploid and Diploid Stages

Sally Otto

University of British Columbia

Classical population genetic theory generally assumes either a fully haploid or fully diploid life cycle. However, many organisms exhibit more complex life cycles, with both free-living haploid and diploid stages. Here we expand population genetic theory to account for drift and selection in organisms with haploid-diploid life cycles. We develop models that consider the dynamics of a population using both the Moran model and Wright–Fisher models, allowing for sexual and asexual reproduction. For haploid-diploid species, we determine the appropriate measure of the variance effective population size and find the fixation probability of beneficial and deleterious mutations, using branching processes and a diffusion approximation. In many cases, particularly when one phase predominates, the fixation probability differs substantially for haploid-diploid organisms compared to either fully haploid or diploid species. We also show how ploidy structures populations, in a manner similar to spatial structure, and show how the extent of Fst with ploidy "patches" can be used to estimate the extent of sexual versus asexual reproduction. This theory provides a framework for understanding evolution in organisms with both haploid and diploid phases.

Predicting dynamic expression patterns in budding yeast with a fungal DNA language model

Kuan-Hao Chao

Illumina Inc

Predicting gene expression from DNA sequence remains challenging due to complex regulatory codes. We introduce a masked DNA language model pretrained on 165 fungal genomes closely related to budding yeast that captures conserved regulatory grammar. Fine-tuning the LM on yeast RNA-seq data—including high-resolution transcriptional regulator induction time courses generated in this study—yielded Shorkie, a model that substantially improves gene expression prediction compared to baselines trained without self-supervision. Shorkie identified canonical transcription factor (TF) binding motifs and tracked their usage across induction experiments. Furthermore, Shorkie accurately predicted variant effects, outperforming leading sequence-to-expression models in cis-eQTL classification and achieving high concordance with massively parallel reporter assays. Interpretability analyses revealed Shorkie's ability to resolve promoter dynamics, splicing signals, and temporal changes in regulatory motif usage. This framework demonstrates that evolutionary-scale pretraining combined with transfer learning substantially improves our ability to decode gene regulation from sequence, providing insights into noncoding variants and regulatory networks.

Preprint: https://doi.org/10.1101/2025.09.19.677475

A meta-analysis of chromatin accessibility QTLs explains differences in QTL-GWAS colocalization

Max Dudek

University of Pennsylvania

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. However, more than half of GWAS signals do not colocalize with any expression quantitative trait loci (eQTLs) discovered by GTEx. This "colocalization gap" represents a major challenge for understanding the mechanisms of genetic trait association. A study by Mostafavi et al. [PMID:37857933] showed that compared to eQTLs, GWAS loci were further from promoters, and enriched near "core" genes – those with strong selective constraint and complex regulatory landscapes. The authors proposed a model of discovery in which selection depletes these "core" genes of variants with sufficiently high effects on expression (detected eQTLs), suggesting that non-colocalized GWAS loci are further from genes and may regulate expression weakly through distal cis-regulatory elements (cREs).

Hypothesizing that the lack of eQTL discovery at GWAS loci is due to weaker perturbations on expression mediated through cREs, we sought to investigate the role of variants that modulate cREs and are detectable as chromatin accessibility QTLs (caQTLs). Towards this end, we aggregated caQTL data from 8 studies derived across different tissues, cell-types and lines, representing a total sample size of 3,737 and consisting of 62,823 lead caQTLs. By mirroring the assessment of variants by Mostafavi et al. with this new dataset, we observed that caQTLs occur at "core" genes more often than eQTLs. Specifically, properties of these "caGenes" lay between those of eGenes and GWAS-implicated 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). We show that these observations are consistent with a causal model in which many eQTLs and GWAS hits are mediated through genetic effects on cREs. This model suggests that for such GWAS hits, chromatin accessibility QTLs (caQTLs) are better powered than eQTLs to detect colocalizations due to a more direct effect. Specifically, caQTLs are expected to colocalize with GWAS signals more in distal regions and at “core” genes compared to eQTLs. Our model predicts that with a limited sample size for discovery, epigenetic association signals can provide complimentary information to eQTLs by implicating functional mechanisms of additional disease-associated loci.

Scalable and robust mapping of eQTL interactions reveals immune cell contexts of regulatory variant activity

Jan Engelmann

Helmholtz Munich & Stanford University

Identifying the immune cell contexts in which specific genetic variants exert regulatory effects is key to linking genetic risk to cellular mechanisms. As single-cell eQTL datasets scale to thousands of individuals and millions of cells, new opportunities arise to resolve where and how these effects manifest. However, existing approaches either rely on discrete cell-type labels—masking continuous or subtle state variation—or do not scale to genome-wide analysis on population-level cohorts.

To address these limitations, we introduce AttenQTLoc, a scalable framework for localizing regulatory variant effects across the continuous landscape of single-cell states. AttenQTLoc leverages an attention mechanism to identify subsets of cells where a variant exerts a regulatory effect, without requiring predefined cell-type annotations. It also incorporates a formal statistical test to assess whether the observed localization reflects significant context specificity.

We applied AttenQTLoc to 8,517 fine-mapped loci with prior regulatory evidence, using single-cell data from the novel UK Biobank “Cardinal cohort”. We employed a two-step procedure: Discovery in 2,391 donors and validation in the remaining 2,392. We further tested for replication in 972 OneK1K individuals. We detected robust state-dependent effects in 35% of pseudobulk eQTLs (FDR 5%), including 218 in B, 694 in CD4 T, 581 in CD8 T cells, 372 in monocytes, and 331 in NK cells. To our knowledge, this constitutes the first genome-wide effort to systematically map the continuous cellular contexts of regulatory variants in immune cells.

AttenQTLoc recapitulates known biology—e.g., a previously reported variant, rs2736336, with increasing effect on BLK along the differentiation trajectory from naive to memory B cells—and reveals novel regulatory programs in transcriptionally defined subpopulations not captured by canonical annotations. For example, for the eczema-associated locus PRKCQ-AS1 in CD8+ T cells (P < 1.6 x 10e-81 for heterogeneity), we identified that regulation is concentrated in naive cytotoxic T cells. The model prioritized cells enriched for "CD28 costimulation" and "PD-1 signaling" pathways, validating the causal context of inflammatory mechanisms in eczema.

Our results demonstrate that scalable, attention-guided localization of eQTL effects in continuous single-cell landscapes reveals cell-state-specific regulatory mechanisms and offers mechanistic insight beyond traditional QTL approaches.

LDate: An expectation-maximization approach for dating ancient individuals based on linkage disequilibrium.

Romain Fournier

Harvard University

Understanding the chronology of ancient individuals is crucial for interpreting their history. While radiocarbon dating by accelerator mass spectrometry is the gold standard for determining chronology, its application is often limited by cost and the preservation of collagen proteins required for analysis. Consequently, many samples are dated on the basis of their archaeological context, which can span centuries or be compromised by intrusive burials.

Recombination events can serve as a biological clock to address these limitations, allowing the dating of an individual relative to a reference panel with trusted radiocarbon dates. Specifically, the expected linkage disequilibrium (LD) between a pair of individuals is a function of both the effective population size and the time gap separating the individuals. Our method, LDate, uses an Expectation-Maximization approach to jointly infer the effective population size history, Ne(t), and the age of each individual. With radiocarbon or archaeological contexts acting as priors on individual ages, the algorithm starts by estimating Ne(t) using an updated version of HapNe-LD that accounts for time gaps between pairs of individuals. It then uses this demographic history to refine the age distribution of each individual, repeating the cycle until convergence.

To validate our approach, we applied LDate to a dataset of 122 ancient individuals from the Ceramic Age in the Dominican Republic (39 radiocarbon-dated, 83 undated), spanning over a millennium. For this dataset, existing archaeological and radiocarbon data already constrain most samples to a relatively narrow 200-year window. To test our method, we systematically discarded the prior age information for each individual one at a time. In 10% of cases (n=13), LDate produced 95% credible intervals narrower than the 200-year archaeological baseline. Among the four radiocarbon-dated individuals in this subset, the LDate intervals correctly encompassed the direct radiocarbon age. These results underline the potential of our approach for refining chronologies where radiocarbon dating is unavailable.

Generative models for clinical and population genetic data

Matteo Fumagalli

Queen Mary University of London

Deep learning algorithms have been increasingly used in genomics studies to infer complex patterns and relationships among large-scale data sets. Among these new data-driven approaches, generative models, in the form of generative adversarial networks (GANs), have been proven successful at producing high-fidelity genomic data. GANs consist of a generator that simulates data, and a discriminator that attempts to separate real from synthetic data. At the end of training, the generator will simulate data that is ideally indistinguishable from real samples.

In this talk I will introduce how we employed a variation of GANs to generate realistic synthetic data from an association study on diabetes susceptibility in London residents with South Asian heritage. I will illustrate how we assessed the veracity of our artificial data and will discuss benefits and pitfalls of this approach. In the second part, I will then showcase how GANs can be utilised for obtaining direct estimates of evolutionary parameters by replacing the generator with a coalescent simulator. I will present how trained GANs onto population genetic data were able to infer population size changes and migration rates between pairs of populations of Anopheles mosquitoes from sub-Saharan Africa as accurately as the state-of-the-art. The main advantage of GANs over empirical and model-based methods lies in their ability to be trained on genome alignments without any loss of information due to compression to a finite set of summary statistics.

I argue that this technology could be applicable to data from iconic and threatened species, and in general nonmodel species, to monitor population variation.

Bayesian approaches for considering cell type composition when assessing the heritability and covariance of aging-related genomic traits

Alexis C. Garretson

University of Utah

Many genomic and molecular traits can be estimated via whole-genome sequencing vary predictably with age and are linked to disease. Examples of such somatic traits include telomere lengths, mutation burden, larger mosaic chromosomal alterations, as well as non-nuclear traits like mitochondrial copy number (mtCN) and heteroplasmy. Within blood cells, there are also sex differences in these traits, not limited to but perhaps most dramatically in the mosaic loss of X (mLOX) and Y (mLOY), a common event amongst aging males. Studies using large biobanks have shown many of these traits to be heritable. However, they are also strongly linked to blood composition, which is also heritable. Therefore, accounting for the impacts of blood cell fractions may change the estimated heritability and associated genetic signals, but is not routinely considered in models.

The large, multigenerational CEPH/Utah pedigrees are an ideal framework to understand the covariation, heritability, and signatures of aging because of their three and four generation structure, and the wealth of available health, laboratory, and sequencing data under relatively homogenous environments. Using a Bayesian approach, we control for blood composition while estimating heritability and covariance. Initial estimates suggest significant heritability of blood traits amongst our population in both males and females (0.1 – 0.9) and uniformly higher heritability for the blood traits in males. Correcting for blood composition reduces the heritability of telomere and mtCN by approximately half; for example, telomere length heritability drops from 0.76 to 0.43 and mtCN from 0.98 to 0.5. We also find genetic covariation between mtCN and both telomere length (0.33) and mLOY (-0.44) in males, but only between mtCN and mLOX in females (-0.42).

These results underscore that both cell composition and sex are core determinants of how aging-related genomic traits are measured and interpreted. Moreover, this approach provides a framework for disentangling aging signals from hematopoietic variation in examining genetic architecture. Future extensions involve estimating heritability across specific age brackets to determine how the genetic architecture of these traits varies across the lifespan. Importantly, this framework is extensible to traits in other tissues that may vary with cell type composition, provided that cell composition is either inferrable or directly measurable alongside focal traits of interest.

Inference in large pedigrees

Simon Gravel

McGill University

Relatedness between individuals can be measured at the genealogical level (by describing shared ancestors in a pedigree) or at a genetic level (by describing shared haplotypes across the genome). The shared haplotype structure can be conveniently summarized as a sequence of trees along the genome, where each tree describes the last shared genetic ancestors between individuals at a locus. While many tools exist to infer tree sequences from genetic data, and many large pedigree datasets are available, few tools exist to identify the relationship between the two – finding which genetic ancestor corresponds to which pedigree ancestor, and conversely.

In this presentation, we propose an algorithm to solve this problem by providing, for each genetic tree, the list of all consistent ancestry paths within a genealogical tree. We also provide variants of the algorithm with moderate robustness to both tree sequence and pedigree errors. We demonstrate the scalability of our approach on the BALSAC genealogical dataset, which includes millions of individuals in Quebec, Canada. We find that 20 carriers usually provide enough information to reliably identify a common ancestor 15 generations ago or find the parent of origin of alleles among probands, but the number of possible ancestry paths within the pedigree can remain large. We apply the method to reconstruct the inheritance of a causal allele for type 1 myotonic dystrophy.

Learning context specific disease mechanisms from single cell data

Sasha Gusev

Harvard Medical School

Genome-Wide Association Studies have now identified hundreds of thousands of disease-associated loci, but most associations still cannot be linked to specific gene functions: a “missing mechanism" problem. Some missing mechanisms may reside in specific contexts, such as individual cell types, cell states, or environmental/cellular exposures. In this talk, I will describe two new methods for identifying context-specific mechanisms in single cell data and connecting them to complex disease. First, we propose a method for quantifying cell-type/state interaction heritability of gene expression using population scale sing-cell data. Applying the method to single-nucleus RNA-seq from the ROSMAP brain study, we find that cell-type and cell-state specific heritability is sizable and comparable to the heritability of main effects. Consistent with prior work, we find that genes with high expression heritability are less likely to be under evolutionary constraint. However, we find that genes with high cell-state specific expression heritability are more likely to be under evolutionary constraint, suggesting that common cell-state (but not cell-type) specific effects may be particularly relevant to disease mechanisms. Second, we propose a method for inferring cell-type/state interaction effects from genome-wide perturbational screens by leveraging sparse, low rank representations of the data. We apply this approach to perturb-multiome data across hematopoietic differentiation and show that our approach can recover substantially more cell-type specific interactions than conventional methods in both scATAC-seq and scRNA-seq modalities. We show that these cell-type specific mechanisms are enriched for disease heritability. Overall, our findings demonstrate that single-cell data provides important new insights into the missing mechanisms of disease.

Analysing haplotype structure in ARGs: From recombination suppression to phantom epistasis

Anastasia Ignatieva

Oxford University

Ancestral recombination graphs (ARGs) reconstructed from large-scale sequencing data naturally capture local haplotype structure, enabling the analysis of clade-specific recombination rate variation. Modelling the persistence of clades of samples along the genome can both illuminate evolutionary processes and help to resolve confounding in association tests. In this talk, I will present three applications united by this genealogy-based perspective. Firstly, by comparing the observed genomic spans of haplotype blocks to theoretical expectations under the sequentially Markovian coalescent (SMC'), we explicitly detect regions of localised recombination suppression. Applying this to human data, we identify many novel and known structural variants driving this signal, including large inversions. Secondly, we identify clades with longer-than-expected spans perfectly overlapping single genes, indicative of allele-dependent crossover suppression in genes expressed during meiosis. Finally, we demonstrate how local haplotype structure can drive phantom epistasis: the spurious inference of gene-by-gene interactions between variants in cis. We introduce a novel ARG-based method to quantify this genealogical confounding, and hence distinguish between genuine interactions and false positives arising due to shared ancestry.

Interpretable systems biology models from genetic interaction maps

Nicholas Ingolia

UC Berkeley

When two genetic perturbations affect the same process, their combined phenotype will often be stronger or weaker than expected, based on the individual phenotypes of each perturbation alone. These genetic interactions can provide powerful insights into the organization and operation of proteins, complexes, and pathways. Emerging technologies now enable large-scale surveys of genetic interactions in individual protein and cis-regulatory sequences and CRISPR-mediated perturbations across the genome. Current approaches to analyze these data identify non-additive phenotypic effects and infer local structure, such as curvature in the genotype-to-phenotype map. The interpretation of this local structure in molecular terms is ambiguous, however, and can vary across different contexts. We describe an analytical framework that infers global structure: an effective systems biology model, with parameters that vary based on genetic perturbations, that generates the observed data. This model identifies the latent features affected by genetic perturbations and the way they interact to produce the phenotype, making it highly interpretable in cellular and molecular terms.

Reconstructing the spatial histories of populations from gene trees

James Kitchens

University of California, Davis

Patterns of genetic variation in spatially distributed populations reflect both isolation by distance and barriers to gene flow, such as topography, climatic conditions, and social/cultural divisions. Identifying these barriers, whether in the past or present, is critical to understanding the genetic structure that we observe in populations. When migration is constrained to only neighboring subpopulations, we often visualize migration rate variation using a heat map, or "migration surface". We present a method for calculating migration surfaces from samples' associated ancestral recombination graph (ARG). These graphs store an immense amount of information about the genetic relationships between samples, including the timings of common ancestors at various positions in the genome. We measure the accuracy of migration rate estimates through simulations under different demographic scenarios. We apply this method to ARGs inferred from empirical datasets and show how migration surfaces can be used to trace the movements of ancestral lineages over time, highlighting the promise of this approach for reconstructing the spatial histories of these populations.

Increased genetic diversity and residual stratification contribute to polygenic score prediction accuracy

Nicole Kleman

University of Minnesota

Recent efforts in human genetics have increasingly focused on analyzing data from diverse cohorts, given their potential for variant discovery and improved polygenic score prediction (PGS). While this is a welcome shift, the complex genetic structure of diverse cohorts can bias GWAS effect sizes and polygenic score (PGS) prediction accuracy (i.e. residual stratification). However, systematic analyses of residual stratification are lacking. To address this, researchers typically include ancestry components as covariates during PGS prediction, yet the effect of this practice on PGS accuracy is unclear.

To assess this, we derived the analytical expectation of PGS prediction accuracy of quantitative traits in the presence of stratification, with and without ancestry correction, in admixed cohorts. As expected, PGS r2 can be inflated in the presence of residual stratification but it can also be inflated or deflated even with unbiased effect sizes if the true genetic value is correlated with ancestry and therefore, with the environment. Including ancestry as a covariate mitigates this bias, but it can also underestimate accuracy by removing genetic variation along the ancestry axes.

Empirically, we asked (i) does multi-ancestry GWAS improve PGS r2 in admixed cohorts? And (ii) is this increase driven by higher genetic diversity or residual stratification? To disentangle between the two, we computed PGS r2 in 48,586 self-identified African American individuals in All of Us (AoU) using summary statistics derived from multi-ancestry, European-ancestry, and sibling-based GWAS (which are robust to environmental stratification). All traits exhibited equal or higher r2 when we use multi-ancestry GWAS, but some traits (e.g. height and systolic blood pressure) showed inflated accuracy due to residual stratification evident from decreased accuracy with sibling-based

GWAS. Including ancestry components as covariates did not change the r2 for these traits, suggesting the bias due to stratification is orthogonal to these ancestry components. Sibling-based PGS r2 for other traits (e.g. BMI) remained comparable to the r2 based on multi-ancestry GWAS, suggesting that the improvement in PGS accuracy is not because of stratification but due to increased genetic diversity.

Even though we show Increased genetic diversity in GWAS can improve PGS prediction accuracy, rigorous testing is needed to distinguish this improvement from residual stratification.

Thousand-parameter TKF-based models with latent and hierarchical structure predict protein sequence evolution competitively with million-parameter neural models

Annabel Large

University of California, Berkeley

Large protein language models have shown promise in capturing evolutionary information. However, classical discrete-state Markov chains remain the preferred choice of evolutionary model in statistical phylogenetics due to their principled construction, interpretable parameterization, tractability, and explicit incorporation of evolutionary time as a parameter. As a result, the realism of phylogenetic analyses of protein evolution has been limited: the relatively simple models used in phylogenetics rely on restrictive and biologically unrealistic assumptions, which limit their ability to capture heterogeneous selection pressures or broader sequence context. To address these issues, we extend the TKF92 model - the canonical hierarchical model combining substitution and multi-residue insertion-deletion (indel) events - by introducing mixture distributions at various levels of the stochastic process. We compare these elaborated versions of TKF92 to two classes of seq2seq models that use neural sequence embeddings, either (i) enforcing a TKF92-like structure within the model, by predicting sample- and site-specific parameters of an F81-TKF92 model, or (ii) predicting descendant sequences and alignments directly, without enforcing any TKF92-like structure. We start by introducing a mathematical framework encompassing both classical mixture models and neural extensions. We define these collectively as “alignment-Markovian” models: autoregressive models for generating an aligned descendant sequence, conditioned on an ancestral sequence, that are Markovian with respect to the alignment component. We benchmark all approaches on a dataset curated from high-quality Pfam seed alignments. Across all sequence embedding architectures, we find that neural models explicitly incorporating TKF92 structure consistently achieve better model fit than unconstrained alternatives. Notably, a nested 10-component

TKF-based model with only 32,000 parameters is highly competitive with neural models containing tens of millions of parameters (outperforming all but two). Together, these results demonstrate that approaches grounded in molecular evolutionary theory, whether neural or classical, are more parameter-efficient and provide better fit to real alignments than unconstrained alternatives, making a strong case for the continued use and extension of classical models of molecular evolution.

A phylogenetic state-dependent Ornstein-Uhlenbeck model to jointly infer discrete and continuous trait evolution

Priscilla Lau

Ludwig-Maximilians-Universität München

Macroevolutionary studies of adaptation have predominantly focused on the impact of environmental conditions and ecological niche a species occupies on morphological traits. In contrast, the association between genomic features (e.g., ploidy, gene expression level) and ecological, morphological, or behavioral traits is understudied. We are especially interested in how the adaptation of a continuously distributed variable depends on the state of a discrete character over millions of years and across a large phylogeny with hundreds or thousands of species. For example, are diploid and tetraploid cotton species adapting towards different optimal fibre productivity? How different is the evolution of gene expression between firefly species that express neoteny and those that do not?

To answer these questions, we present a novel Bayesian approach to jointly model the evolution of a discrete character under a continuous-time Markov process and the evolution of a continuous trait under a state-dependent Ornstein-Uhlenbeck (OU) process. Under a state-dependent OU process, the evolution of a continuous trait has adaptive properties that depend on the state of a discrete character. We derived an efficient pruning algorithm for our state-dependent OU model and implemented the method in RevBayes, a flexible phylogenetic software based on probabilistic graphical models.

In this presentation, I will demonstrate the performance of our state-dependent OU model using a series of simulation studies. First, I demonstrate the speed of our pruning algorithm across different tree sizes. Next, I assess how the precision and accuracy of parameter estimates change with increasing amounts of data. Lastly, I characterize the model behavior under different model configurations, specifically, in terms of false positive rate and power in inferring state dependency. In conclusion, our model and implementation facilitate application to many large-scale genomic and trait datasets to test different hypotheses of macroevolutionary adaptation.

Learning multiomic cell dynamics with neural ODEs

Christina Leslie

Memorial Sloan Kettering Cancer Center

We will present a generative neural ordinary differential equation (ODE) model for single-cell multiome data called DynaVelo, which we use to learn a functional form of the dynamics — corresponding to learned RNA velocity and transcription factor (TF) motif velocity vector fields — of wildtype and mutant germinal center B cells. We show that Jacobian analysis or in silico perturbations can recover dynamic regulatory networks governing cell state transitions in the germinal center. We also show how to predict the impact of genetic loss-of-function mutations on cell dynamics, and how to predict TF perturbations that rescue loss-of-function phenotypes.

Modeling phylogenetic trees with Stochastic Context-Free Grammar (SCFG)

Heng Li

Broad Institute

Stochastic Context-Free Grammar (SCFG) is a statistical model that generalizes Hidden Markov Model (HMM). We note the link between SCFG parse trees and phylogenetic trees and propose to model phylogenetic trees with SCFG. The inside algorithm of SCFG is identical to the Felsenstein's algorithm for calculating the tree likelihood. The inside-outside algorithm infers the posterior distributions of ancestral sequences and branch-specific transition matrices. This gives us an expectation–maximization (EM) algorithm to optimize parameters with no assumption about time reversibility and homogeneity and it enables model comparison on individual branches. We can also use nearest neighbor interchanges (NNI) to explore the local tree space without re-estimating the parameters of the whole tree. SCFG provides a new direction to model unrestricted phylogenetic trees within the maximum-likelihood framework.

Improving gene effect estimates with functionally informed hierarchical GeneBayes

Xinyi Li

Stanford University

A central goal of human genetics is to identify the key genes and pathways that drive disease. While gene-based burden tests have proven effective for aggregating the effects of loss-of-function (LoF) variants, their statistical power is often limited, particularly for genes harboring few LoF variants. To address this limitation, we extend GeneBayes to refine gene-level effect estimates by sharing information across biologically similar genes within a hierarchical Bayesian framework.

In our framework, observed gene-level effect size estimates from burden tests are modeled as noisy observations of the latent true effect through a normal likelihood with known standard errors. We then introduce a sign–magnitude prior on gene effects in which the magnitude follows a Gamma distribution and the direction of effect follows a Bernoulli distribution. The prior parameters are learned from a machine-learning model trained on genome-scale gene features, including protein structure embeddings and measures of selective constraint.

We evaluated the proposed method using simulations in which gene effects were drawn from a constraint-dependent mixture of normal distributions. Across a range of genetic architectures, the extended GeneBayes model substantially improved effect size estimation and increased power to detect disease-associated genes relative to standard burden tests, while maintaining appropriate false-positive control.

We further applied our method to 21 quantitative traits using UK Biobank data and assessed replication in the All of Us cohort. Compared to the raw burden effect, the cross-cohort correlations for the refined GeneBayes estimates increased by a median of 24.9%. For example, the cross-cohort binned correlation for low-density lipoprotein cholesterol increased from 0.57 to 0.71 after

GeneBayes refinement. We further examined which gene-level features most strongly accounted for the observed performance gains. Across 1,253 candidate features, measures of selective constraint and missense burden emerged as the top contributors to improved effect size estimation and cross-cohort replication.

Together, these results demonstrate that integrating machine-learned functional priors into hierarchical Bayesian gene-based models improves both statistical power and cross-cohort reproducibility in rare variant association studies.

Heritability and effect-size distribution of rare protein-coding variation

Wenhan Lu

Broad Institute

Rare predicted loss-of-function (pLoF) variants are consistently deleterious and mechanistically interpretable, contributing to most of the statistical power and burden heritability in rare variant association studies (RVAS) for complex traits and diseases. However, the genetic architecture beyond their aggregated effects on phenotypic variance, including how pLoF heritability is distributed across genes with different effect sizes, and how these effects may vary by ancestry, remains unclear.

We developed burdenEM, a mixture-model framework to estimate the distribution of gene-level burden effect sizes for rare protein-coding variants and their contributions to burden heritability. Using summary-level simulations across a range of realistic genetic architectures, we show that burdenEM produces unbiased estimates of effect size distribution and burden heritability. We then applied burdenEM to RVAS summary statistics from the UK Biobank (UKB; N=394,841), All of Us v8 (AoU; N = 392,030) across diverse ancestry groups, and their cross-biobank meta-analysis, for 82 well-powered traits harmonized between the two cohorts.

Burden heritability estimates are consistent across ancestries and biobanks. Cross-biobank meta-analysis results yield a mean burden heritability of 0.9% (s.e. = 0.01%) from rare pLoF variants, v.s. 0.2% for synonymous variants. Using UKB for discovery and AoU for replication, burdenEM shows that UKB-significant associations replicate at 60% of their expected rate in AoU across most traits, accounting for the winner's curse and limited power, indicating genuine differences between the biobanks.

Unlike common variants, rare variants do not show symmetric positive and negative effects. For height, we find that over 70% of the burden heritability is from genes with negative effect sizes. Similar patterns are observed for LDL cholesterol, haemoglobin, and calcium, whereas other traits, including BMI, weight, and blood glucose, are dominated by genes with positive effects. Another difference between the effect size distributions of rare and common variants is that for some traits, pLoF heritability is highly concentrated within just a few genes. For example, explaining 50% of burden heritability requires approximately five genes for LDL, but more than 1,000 genes for more polygenic traits such as diastolic blood pressure. However, all traits exhibit a long polygenic tail, requiring thousands of genes to explain 90% of burden heritability.

Transformer-powered mutation-selection models for antibody evolution

Erick Matsen

Fred Hutchinson Cancer Center

Mutation-selection models have been central to molecular evolution research for decades, cleanly separating the rate at which mutations arise from the fitness effects that determine whether they persist. In a parallel universe, protein "language models" have appeared, which are transformer networks trained on massive sequence databases. Because these models are trained with a masked objective, they conflate mutation and selection. We show that this conflation has real costs: the models implicitly learn codon tables, site-specific mutation rates, and germline identity, all of which are irrelevant to protein function. We propose a synthesis: a Deep Amino acid Selection Model (DASM) that updates the mutation-selection framework with modern transformer architectures. A neural mutation model is fit to neutrally evolving sequences; a transformer-encoder then predicts the selective effect of every possible substitution at every site, trained on millions of parent-child pairs from reconstructed phylogenies. Because mutation and selection are factored by construction, the model exclusively quantifies functional constraint. On experimental benchmarks, the DASM substantially outperforms existing protein language models while being an order of magnitude smaller and faster. The learned selection landscape reveals conserved structural motifs, a previously underappreciated CDR3-proximal aspartate, and diversifying selection at framework sites where only purifying constraint was expected. In summary, evolution-first principles beat brute-force scaling.

Scaling Sequence-to-Function Models for Regulatory Variant Interpretation

Sara Mostafavi

Understanding how genetic variation influences gene regulation remains a central challenge in biology. Sequence-to-function (S2F) models trained on large-scale regulatory assays now enable prediction of regulatory activity directly from DNA sequence, offering new ways to interpret genome function across cell types and states. In this talk, I will present our work examining the scaling and generalization of S2F models along three complementary directions. First, I will describe a study using a curated collection of MPRA datasets spanning diverse regulatory environments, showing how dataset size, sequence diversity, and model complexity influence performance and transferability across assays and cellular contexts. Second, I will introduce SAGE-net, a scalable and compact S2F framework that enables training on large collections of personal genomes, increasing sequence diversity and improving model sensitivity. Finally, I will discuss our latest evaluation of AlphaGenome, assessing the current state of sequence-based models for interpreting personal genome variation. Together, these results highlight both the promise and current limitations of S2F models for regulatory variant interpretation and outline principles for improving models that link sequence to regulatory function.

There are no unlinked loci: How pedigrees couple neutral genealogies across the genome

Maximillian Newman

University of Chicago

Implicit in current population genetic methods is the assumption that genes far enough apart in the genome, such as on different chromosomes, have independent genealogies. These gene genealogies, however, are subject to the same pedigree, the random graph capturing the reproductive history of the population. I show how the pedigrees of both well-mixed and structured populations record macroscopic demographic events, and how these events couple genealogies across the genome. In particular, I will explain how neutral gene genealogies far apart on the genome are only independent in the absence of large migrations and uneven offspring distributions. These results suggest that genome-wide association statistics, which aggregate weak signals across many loci, may be sensitive to pedigree-induced correlations even between unlinked regions of the genome.

Coalescent Inference for Epidemics with Latent Periods

Julia Palacios

Stanford University

Structured epidemiological coalescent models are used to study the transmission dynamics of rapidly evolving pathogens from molecular sequence data, however inference under these models rely on modeling approximations and remain computationally challenging. In this work, we focus on exposed-infectious (EI) epidemics and make three contributions. First, we provide a novel derivation of the EI coalescent as the thinning of an approximate marked joint point process of epidemic process and sampled genealogy. Second, we introduce a Bayesian inference framework that uses a competing-risk phase-type augmentation to integrate over latent states, enabling exact likelihood evaluation on fixed genealogies with heterochronous sampling. Third, we develop a Markov chain Monte-Carlo algorithm for inferring time-varying effective reproduction numbers and EI population trajectories using Gaussian Markov random field priors. We validate the accuracy of the EI coalescent approximation and the inferential procedure in simulation, including comparisons to existing coalescent and birth–death-sampling approaches. We reanalyze Ebola virus genomes from the 2014 Liberia outbreak to estimate transmission dynamics over time. Joint work with Isaac Goldstein.

Scalable inference and analysis of ancestral recombination graphs

Pier Palamara

Oxford University

Ancestral recombination graphs (ARGs) provide a compact representation of the genealogical history of a set of genomes and enable a wide range of analyses. We developed Threads, a scalable algorithm for ARG inference that can be applied to hundreds of thousands of genotyped, imputed, or sequenced genomes while retaining high accuracy in simulations, particularly at recent time scales. We applied Threads to infer genome-wide genealogies for 487,409 individuals from the UK Biobank and developed several downstream analytical tools that leverage these inferred ARGs. Using the ARG for genotype imputation from a reference panel of up to ∼200,000 exome-sequenced samples, we observed improvements in imputation accuracy (r^2) of up to ∼10% for ultra-rare variants (MAC $\leq$ 10). We further developed scalable linear mixed model algorithms and applied them to ARG-based variance component association testing in the UK Biobank, detecting more gene–trait associations than approaches relying solely on genotype imputation from ∼65,000 sequenced haplotypes. Finally, we implemented efficient algorithms to characterize recent shared ancestry and analyze genealogical relationships between ancient and modern genomes within large biobank-scale datasets. Together, these results illustrate the practical value of scalable ARG inference for large genomic datasets, particularly for analyses of population-specific and rare variation at recent time scales.

The ubiquity of evolutionary theory: phylogenetic approaches to genetic confounding in epidemiology and psychology

Roshni Patel

University of Oregon

Observational studies are commonly used in psychology and epidemiology to identify risk factors correlated with health outcomes. However, these studies are vulnerable to confounding when shared genetic variation influences both the putative risk factor and outcome. Researchers have historically controlled for this type of genetic confounding using polygenic scores, but these scores are often noisy and biased estimators of a trait's genetic component. Here, we develop a method that leverages solutions to a similar problem in the field of phylogenetics. By translating models from phylogenetics into a statistical genetics framework, we show that the genetic relationship matrix (GRM) can be used to control genetic confounding when testing for non-genetic risk factors. In simulations, we find that our method outperforms existing approaches, particularly in the sample sizes characteristic of datasets in psychology and epidemiology. We also demonstrate that while existing methods are susceptible to poor GWAS portability, our method is inherently robust to such concerns. Finally, we apply our method to 6,104 Indian and 8,483 Black British individuals in the UK Biobank to re-analyze social risk factors for health outcomes in historically understudied cohorts.

Leveraging relatedness in genomic datasets to study variation in germline mutation rates

Lin Poyraz

Columbia University

Germline mutation rates vary among individuals, influenced by parental ages, genetic effects and potentially environmental exposures, and evolve over time. Yet beyond a handful of very rare modifiers, we know little about genetic and environmental effects, in part because human germline mutation rates are difficult and expensive to estimate precisely. To overcome this limitation, we developed a new approach that utilizes existing genomic data from sibling pairs. This method leverages the fact that on average, siblings inherit the same two chromosomes from their parents in a quarter of their genomes (i.e., are IBD2). In that IBD2 subset, any true difference between siblings is a de novo mutation (DNM) (or a gene conversion event). We applied this approach to the UK Biobank and All of Us, which includes 22,344 and 7,701 sibling pairs, respectively, as well as 1,040 and 1,295 trios. After extensive filtering based on duplicates and identical twins and the subset of quad families, we identified over 1 million DNMs, the spectrum and distribution of which very closely resembles what has been reported based on trio studies. We further estimated parental mutation rates and spectra for >32,000 parental pairs. Imputing parental genotypes from the sibling genomes, we identified several significant associations between the mutation spectra and loss of function and missense variants in genes known to be involved in DNA repair and replication. We also found significant differences in the mutation spectra across genetic ancestries labeled on the basis of genetic similarity to the 1000G and HGDP samples, but no evidence for differences in the fraction of TCC>TTC mutations at present-day. This work thus provides a large-scale characterization of germline mutation rates across ancestries and a first look at its genetic architecture in a population cohort.

It’s All Relative: Turning Biobank "Nuisance" Kinship into New Genetic Insights

Sohini Ramachandran

Brown University

As genomic biobanks scale to millions of participants, the prevalence of both close and distant relatives being sampled presents significant methodological opportunities. We demonstrate that traditional approaches ignoring relatives are increasingly counterproductive in the biobank era. First, we introduce HAPTIC (Haplotype TIling and Clustering), a novel inter-chromosomal phasing approach that uses identity-by-descent (IBD) segments shared among distant relatives (>10 cM) to discretely assign paternal and maternal variants genome-wide. By representing shared segments as nodes in a signed graph and employing spectral clustering, HAPTIC achieves a median phase accuracy of 99.8% in the UK Biobank. We further explore the impact of "purple nodes"—ancestors shared by both parents—revealing that these rates vary significantly by ancestry, from <5% in admixed individuals to >60% in South Asian cohorts, containing important spatial genetics signals and complicating the inference of parent-of-origin effects. Using Wright-Fisher simulations and BALSAC genealogical data, we show that the common practice of filtering close relatives (e.g., 2nd-degree) to estimate effective population size introduces significant upward bias and artifacts in recent population history. We find that half-siblings and first cousins contribute a non-negligible fraction of the IBD segments used to infer effective population size even 7–10 generations in the past. Our results suggest that in randomly ascertained biobanks, retaining all degrees of relatedness provides the most stable and least biased reconstruction of recent population trajectories. Together, these methods provide a framework for utilizing the complex web of kinship inherent in modern biobanks to improve the precision of a range of genetic problems, ranging from individual-level phasing to demographic inference.

Working with wildly varying mutation rates in TRs to investigate their role in adaptation

Rori Rohlfs

University of Oregon

Tandem Repeats (TRs) are abundant, mutate rapidly, and are structural variants that impact gene expression and other downstream traits, making them putative targets of natural selection. However, not only do TRs mutate rapidly, but their mutational processes and rates vary dramatically across loci, complicating detection of selection in TRs. For instance, TR divergence could be explained by either selection or a high mutation rate. Now that long-read DNA sequencing facilitates reliable TR genotyping, the field is motivated to create robust approaches to test TR evolutionary hypotheses. Our lab has developed TR ARG Mutation Analysis (TRAMA), to infer the mutational model and estimate parameters for a single TR by conditioning on the ARG. We show that TRAMA's inferences and estimates are accurate, particularly with higher mutation rates and larger sample sizes. We also developed an HKA-like approach that compares TR variance between versus within species to account for locus-specific mutation rates. Among humans and chimpanzees, we find signatures of widespread stabilizing selection on TR allele length, as well as TR candidates for positive and balancing selection. We find that TRs with functional variation in humans are expanded in humans, a pattern consistent with recent selection. The talk will conclude with a brief discussion of applications of quantitative approaches beyond biology.

Inferring signatures of time-varying mutational mechanisms during antibody affinity maturation

Magdalena Russell

University of Washington

During adaptive immune responses, B cells produce antibodies that recognize foreign antigens. B cells undergo affinity maturation, an iterative process of mutation and selection that generates lineages of increasingly effective antibody variants. Recent studies suggest that this mutation process is affinity-dependent rather than uniform. High-affinity B cells undergoing clonal expansion appear to acquire fewer mutations per division, possibly due to a shortened window of mutator activity that preserves beneficial genotypes (Pae et al. 2025; Merkenschlager et al. 2025) and may confer an evolutionary advantage (Pyo et al. 2025). If these dynamics reflect changes in underlying mutational mechanisms, they should produce detectable shifts in antibody mutational spectra over time. However, existing approaches to quantifying antibody mutation patterns assume temporally constant mutational processes (e.g. Yaari et al. 2013) and therefore cannot capture such shifts, leaving the temporal structure of antibody mutagenesis poorly understood. Here, we recover mutation rate and spectral histories from mutations mapped to branches of B cell lineage phylogenetic trees, modeling mutation as an inhomogeneous Poisson process analogous to approaches used to infer changing mutational processes in population genetics (DeWitt et al. 2021; Deng et al. 2025). With this framework, we explore temporal variation in mutation spectra and the presence of mutational signatures active during specific stages of affinity maturation, supporting dynamic regulation of mutagenesis as an evolutionary mechanism in adaptive immune responses.

GPC: An expressive, tractable, and privacy-preserving deep generative model for genetic variation data

Sriram Sankararaman

UCLA

Generative models play an increasingly important role in population genetics: being used to generate artificial genomes (AGs) that are used to benchmark methods, test evolutionary hypotheses, and construct reference panels for imputation while working around data-sharing restrictions. Existing generative models of genetic variation, however, struggle to faithfully express dependencies in the data while retaining tractability and preserving privacy. I will introduce Genetic Probabilistic Circuits (GPC), a deep generative model for genetic variation data based on hidden Chow-Liu trees represented as probabilistic circuits. GPC generalizes traditional hidden Markov models by allowing arbitrary tree structures over latent variables, enabling it to capture long-range dependencies among SNPs. GPC is tractable, supporting exact computation of marginal and conditional probabilities that enables both AG generation and direct genotype imputation (avoiding the need for simulating AGs). We show that GPC generates AGs that accurately reproduce population structure and linkage disequilibrium patterns across a range of length scales. Compared to other deep generative approaches, GPC consistently improves imputation accuracy, with particularly strong gains for low-frequency SNPs and in populations that are not well-represented in public reference panels. Finally, we show that GPC better preserves privacy of the training data thereby providing a practical framework for AG generation in settings with limited data access.

Calibration of polygenic scores across ancestries

Scott Sauers

University of Minnesota

The application of polygenic scores in clinical risk prediction is limited by performance and score distribution differences across diverse human populations. Standard strategies to apply a single score across multiple genetic ancestries involve adjustments of the raw polygenic score based on principal components. However, these adjustments are phenotype-agnostic, meaning they normalize based on the relationship between scores and principal components within reference panels without considering the health outcome itself. However, this approach cannot determine if the correlation between a score and ancestry should be removed, or if the correlation is a signal that is essential for accurate risk prediction.

We propose a phenotype-informed calibration method that uses observed health outcomes to learn how the relationship between a polygenic score and a phenotype is modulated by genetic ancestry. We allow for non-linear interactions between the polygenic score and ancestry principal components by using flexible basis functions which empirically discover the appropriate calibration curves across the ancestry spectrum. This allows the model to retain ancestry-associated risk signals when they are predictively relevant while correcting for neutral drift and bias when they are not. Overconfident risk estimates are particularly problematic for minority populations where polygenic scores are less accurate. The model shrinks extreme predictions toward the population mean in regions of the principal component space where uncertainty is high. Finally, we sample the posterior distribution of the model coefficients to compute the final risk assessment and integrating the risk function over the parameter uncertainty. The implementation is designed end-to-end for biobank scale with support for binary, continuous, and time-to-event data.

By grounding the calibration process in observed phenotype data and probabilistic modeling, this method offers a step along the path toward more equitable and accurate clinical applications of polygenic risk scores across the global spectrum of human diversity.

Variational Inference with Node Embeddings (VINE) for scalable Bayesian phylogenetics

Adam Siepel

Cold Spring Harbor Laboratory

Bayesian phylogenetic inference is now widely used but remains heavily reliant on Markov chain Monte Carlo (MCMC) sampling, which is computationally intensive and requires careful convergence monitoring. Variational inference (VI) is an appealing alternative that approximates posterior distributions without sampling, but existing variational approaches for phylogenetics have seen limited adoption owing to constraints in accuracy, speed, and scalability. Here we introduce Variational Inference with Node Embeddings (VINE), a variational phylogenetic inference method with striking improvements over prior work. VINE supports both standard DNA substitution models and CRISPR barcode-mutation models for cell-lineage phylogenies. Its key innovations are: embedding taxa in a high-dimensional Euclidean space; backpropagating gradients through fast distance-based phylogeny inference algorithms; introducing a sampling-free approximate estimator for the VI evidence lower bound; and enhancing posterior flexibility using normalizing flows. Across simulated and empirical datasets, VINE yields accurate posterior approximations for datasets with as many as 1000 taxa in a fraction of the time required for state-of-the-art MCMC-based methods.

Demographic inference using ARGs: theory, methods, and statistical foundations

Jonathan Terhorst

University of Michigan

Recent advances in genome-scale ancestral recombination graph (ARG) estimation have made it increasingly feasible to infer demographic history directly from reconstructed coalescent times along the genome. In this talk, I will describe recent work from my group on the statistical foundations of population-genetic inference in this setting. First, I will present demestats, a differentiable software library for computing generalized instantaneous coalescent and cross-coalescent rate curves for arbitrary deme-based demographic models, enabling improved inference of recent growth and migration. Second, I will discuss a specialization of this idea designed to infer migration surfaces directly from sequence data. Third, I will discuss recent theoretical work on ergodicity of the sequentially Markov coalescent, which quantifies how quickly genealogical dependence decays along the genome and therefore how far apart loci should be subsampled to behave approximately independently. Time permitting, I will also discuss recent work on model selection in ADMIXTURE, showing that standard model selection criteria can be inconsistent and may miss fine-scale structure when populations are closely related.

VOUS: Variational Ornstein-Uhlenbeck linking single-cell lineage tracing with stochastic gene expression

Jiawei Xing

Cold Spring Harbor Laboratory

As single-cell technologies increasingly capture multimodal data, including transcriptomics, epigenetics, spatial locations, and lineage tracing, the need grows for integrative models that can handle sparse, noisy data and infer temporal, dynamic relationships. Here we present a probabilistic machine learning framework, VOUS, that combines Ornstein-Uhlenbeck (OU) stochastic processes with Negative Binomial-distributed read counts. VOUS models the stochastic dynamics of gene expressions along paired cell lineages, efficiently estimates latent expression states using mean-field variational inference, and enables hypothesis testing for lineage-specific expression shifts under high sparsity and low coverage. We validated our approach with simulated datasets generated from lineage-based stochastic simulations, and applied it to scRNA-seq data from metastatic lung cancers. Our model accurately recovers both known and novel gene programs involved in tissue-specific metastasis, outperforming current lineage-based models that ignore the real data distribution. Beyond embedding single-cell lineages with gene expression inferences, we envision extending this framework with other modalities, moving towards a more comprehensive atlas of cellular dynamics.

Predicting functional constraints with phylogeny-informed genomic language models

Chengzhong Ye

University of California, Berkeley

Genomic language models (gLMs) have emerged as a powerful approach for learning genome-wide functional constraints directly from DNA sequence, yet NLP-style gLMs often demand substantial compute and still lag classical evolutionary models on key tasks. We present GPN-Star (Genomic Pretrained Network with Species Tree and Alignment Representation), a biologically grounded gLM with a phylogeny-aware architecture that integrates whole-genome alignments and species trees to model evolutionary relationships explicitly. Trained on alignments spanning vertebrate, mammalian, and primate timescales, GPN-Star attains state-of-the-art accuracy across diverse variant effect prediction tasks in both coding and noncoding regions of the human genome. Analyses across timescales reveal task-dependent advantages of modeling more recent versus deeper evolutionary signal. In human genetics applications, GPN-Star improves prioritization of fine-mapped GWAS variants, yields strong enrichments of complex trait heritability, and increases power in rare variant association testing. Extending beyond humans, we applied GPN-Star to mouse, chicken, fruit fly, C. elegans, and A. thaliana, demonstrating robustness and generalizability. Overall, GPN-Star provides a scalable and flexible framework for genome interpretation that leverages expanding comparative genomics resources.

Recovering signatures of archaic introgression using ancestral recombination graphs

Yulin Zhang

UC Berekeley

The sequencing of the Neanderthal and Denisovan genomes has reshaped our understanding of archaic gene flow into modern humans. However, the limited availability of archaic genomes from deeper timescales or other regions (especially outside Eurasia), together with a lack of methods that can reliably identify archaic ancestry without unadmixed outgroup populations, has left the evolutionary history and impact of past introgression events largely unknown. We introduce TRACE, a novel approach to identify archaic ancestry in humans, by leveraging features of ancestral recombination graphs (ARGs) constructed from contemporary genomes alone, without requiring an archaic reference or an unadmixed outgroup. By performing extensive simulations, we show that TRACE has high sensitivity and specificity, comparable to existing methods. Applying TRACE to 1000 Genomes Project data reproduces known signatures of Neanderthal and Denisovan introgression in non-Africans. Moreover, TRACE reveals novel signals of "ghost admixture"–archaic gene flow from an uncharacterized hominin lineage–in both African and non-African populations, pointing to an introgression event predating the out-of-Africa expansion. Ghost ancestry is significantly depleted in conserved and low-recombination regions, yet notably persists within many deserts of Neanderthal and Denisovan ancestry. In Oceanian genomes, TRACE identifies significant enrichment of deep lineages within Denisovan–but not Neanderthal–introgression tracts, supporting a model of super-archaic gene flow into Denisovans and modern humans. Our results demonstrate the power of ARG-based approaches to recover hidden episodes of gene flow in our past and offer a scalable path towards mapping archaic ancestry in modern humans, even in the absence of archaic genome sequences.

A new approach for estimating the fitness effects of missense and non-coding variants

Yixin Zhu

Cornell University

Understanding the fitness consequences of genetic variation is a central problem in evolutionary genomics and human disease genetics. A key quantity is the distribution of fitness effects (DFE), which quantifies the strength of selection acting on new mutations and is informative about pathogenicity and disease risk. Existing genome-wide measures of mutational constraint (e.g., pLI, LOEUF, GNOCCHI) successfully highlight regions under strong purifying selection. However, these metrics focus primarily on loss-of-function variants, which represent only a small fraction of observed genetic variation.

The remaining variation is dominated by missense and noncoding mutations, many of which segregate at low frequencies due to purifying selection acting against deleterious effects. As a result, rare variants are especially enriched for deleterious mutations and contain much of the information needed to infer the DFE. Despite this, many existing approaches rely on the full site-frequency spectrum (SFS) and are constrained by the infinite-sites assumption, which become increasingly inappropriate in large cohorts where recurrent mutation is common and is computationally costly at the biobank scale.

We present a scalable framework for DFE inference that centers rare variants as the primary source of information about selection. Building on recent theoretical advances that approximate rare variant frequencies using birth-death processes, we derive efficient likelihoods for the rare-variant SFS that explicitly accommodate recurrent mutation and avoid the infinite-sites assumption. Crucially, our framework explicitly models gene-specific mutational tolerance, recognizing that the fitness impact of variants depends heavily on the local genomic context. We capture this structure using an empirical Bayes framework that enables information sharing across genes. Gene-level priors on selection are learned from rich functional and evolutionary annotations using a flexible, data-adaptive Deep Learning approach, allowing us to extract maximal information from heterogeneous annotations when mutation data are sparse.

We apply our approach to real-world datasets, such as the UK Biobank and All of Us. This work is critical for filling gaps in our understanding of genetic variation and disease, offering a more comprehensive view of selection pressures on the genome.

Posters

://moorjanilab.org/riesOrdered alphabetically

Inferring the origins of precancerous lesions from spatially-informed genomic sampling

Alvina Adimoelja

Stanford University

A fundamental question in cancer evolution is how cancers arise and how early dynamics shape subsequent evolution. While cancers are widely thought to originate from a single transformed cell, the origins of precancerous stages remain largely unclear. Barrett's Esophagus (BE)—a precursor to esophageal adenocarcinoma characterized by replacement of esophageal epithelium with intestinal-like crypts—has long been hypothesized to arise from single-cell origins, though this remains disputed.

To infer founder number and timing from sequencing data, we developed a model that generates spatial patterns of genetic variation under different founder scenarios. Using SLiM, we simulate spatial expansion: crypts, each maintained by a small population of stem cells, are modeled as discrete demes that expand into neighboring regions. Genealogies are tracked using tskit, and different founder scenarios are simulated by modifying these trees. This allows a single spatial simulation to generate predictions across multiple founder scenarios.

We tested our model with whole-genome sequencing of densely-sampled microdissected crypts from BE patients. Initial results support multiple founders. The site frequency spectrum shows many mutations shared by intermediate numbers of crypts, rather than by most samples as expected from a recent single founder. Phylogenetic reconstruction reveals multiple clades with spatial stratification by esophageal depth. Mutations shared across many samples consistently reach fixation within crypts, while private mutations show variable frequencies, reflecting differences in mutation age. We are extending to additional patients and developing a Bayesian inference framework to quantify founder number and timing across this cohort.

A model of background selection in changing populations

Nathan W. Anderson

University of Wisconsin - Madison

Whole genome sequencing data has been collected from a large and ever-growing number of human populations. From these data, we gain insight into the relationship between populations as well as their historical size. Most methods used to infer population histories assume selection is absent. However, even neutral loci are subject to selection through their genomic linkage with sites under selective constraint (background selection; BGS). Failing to account for BGS biases inference. Here, we present a site frequency spectrum (SFS) based demographic inference tool that corrects for biases caused by BGS. Because the probability that any linked site carries a deleterious mutation decreases backwards in time, extant variation tends to descend from an increasingly small, high fitness subset of the gene pool as we look further back in time. By appropriately rescaling the ancestral population sizes at different times, we can approximate distortions in the SFS due to BGS and perform unbiased inference of demographic histories. Using our approach, we reanalyze historical datasets of both model organisms and human populations. We find a reduced signal of population growth compared to previous models that assumed neutrality, indicating that some of the growth in recent history seen in previous results is due to uncorrected bias.

Genetic disruption of transcriptional specificity underlies rare disease risk

Miquel Anglada-Girotto

Centre for Genomic Regulation (CRG)

Cells acquire diverse identities by expressing genes in a context-specific manner across tissues, developmental stages, and cell types. While many genetic variants have limited phenotypic impact, variants that disrupt transcriptional specificity, that is, when and where genes are transcribed, may represent an important genetic vulnerability that contributes to disease. However, determining how variants alter specificity across contexts remains challenging because comprehensive multi-tissue profiling is infeasible for every individual. Here, we deploy sequence-based machine learning models to predict how DNA variants change RNA-sequencing coverage across tissue-, stage-, and cell type-resolved contexts, enabling genome-wide prioritization of variants that perturb transcriptional specificity at multiple regulatory resolutions. Using experimentally measured activity of human enhancers carrying specificity-disrupting variants across mouse embryonic tissues, we show that embryo-trained models recapitulate in vivo effects, including polydactyly-associated variants. In paired genotype-transcriptome datasets spanning healthy adult tissues and cell types, variants associated with altered transcriptional specificity are depleted, consistent with negative selection, and adult-trained models best capture these shifts. Applying this framework to rare disease genomes reveals an enrichment of specificity-disrupting variants affecting clinically relevant genes. Together, our results establish transcriptional specificity as a regulatory axis of disease risk and provide a scalable approach to identify specificity-altering variants genome-wide across biological contexts.

Searching for signatures of selection in B-cell affinity maturation

Antoine Aragon

École Normale Supérieure

Cellular diversification in processes from development to cancer progression and affinity maturation is often linked to the appearance of new mutations, generating genetic heterogeneity. Describing the underlying coupled genetic and selective growth processes that result in the observed diversity in cell populations is informative about the timing, drivers and outcomes of cell fates. Current approaches based on phylogenetic methods do not cover the entire range of evolutionary rates, often making artificial assumptions about the timing of events. We introduce CBA, a probabilistic method that infers the division, degradation and mutation rates as well as selection strength from the observed genetic diversity in a population of cells. It uses a summarized backbone tree, intermediary between the true cell tree and the allelic tree, representing the ancestral relationships between types, which allows for efficient sampling of possible phylogenies consistent with the observed mutational signatures. We demonstrate the accuracy of our method on simulated data and compare its performance to standard phylogenetic approaches. We further apply CBA to B-cell lineage trees inferred from high-throughput antibody repertoire sequencing experiments, quantifying effective growth and mutation rates and providing estimates for the strength of selection during affinity maturation.

Error correction of SARS-CoV-2 genomic sequences using a phylogenetic prior

Monica Arniella

UC Berkeley

Over the course of the SARS-CoV-2 pandemic, there has been a fast-paced and widespread effort to sequence clinical and wastewater samples around the world. Analysis of these sequences has quickly revealed potentially problematic sites with high error rates, necessitating careful filtering or masking of regions in the genome. Additionally, the error rate of genomic sequencing can be highly variable between experiments, and areas of low sequencing depth make base calling even more error-prone due to ambiguity in the identity of nucleotides. In the case of viral sequences, we can use additional information from phylogenies in order to infer the most probable nucleotide at a position. Here, we use the global SARS-CoV-2 phylogeny (involving millions of sequences from around the world) and the phylogenetic method tronko by Pipes et al. to calculate prior probabilities for each nucleotide along randomly chosen genomes, and use the observation of real sequencing reads to choose the bases with the highest posterior probability. To test the performance of this method, we test varied sequencing conditions using real samples downloaded from the Sequencing Read Archive (SRA), simulated reads from genomes obtained from the GISAID repository, and compare the performance of our method with standard viral base calling methods. Using this approach, we evaluate the utility of using phylogenetic information for more confident base-calling at uncertain sites and increasing the accuracy of downstream analyses.

Quantifying within-individual immune variability

Thomas Atkins

University of California Berkeley

Variance is traditionally studied at the population level by measuring differences between individuals. However, advances in single-cell sequencing enable the quantification of variability among cells from the same individual, allowing us to investigate the drivers of within-individual heterogeneity. We model cell-cell variability by applying biophysical models of transcription and sequencing to large single-cell PBMC atlases. We find that known environmentally responsive genes exhibit not only greater variability between individuals, but also increased variability within individuals. Furthermore, we examine how cellular heterogeneity changes under immune stress using COVID-19 and stimulation atlases. Together, these results link population scale immune variability to within-individual immune cell variation, establishing a new framework for how an immune system coordinates its response to a dynamic environment.

The *Arabidopsis* pangenome reveals highly divergent haplotypes persisting across species boundaries and populations

Zhigui Bao

Max Planck Institute for Biology Tübingen

Much of population genetics represents genetic variation primarily in the form of independently segregating biallelic sites. While this abstraction has been highly productive, it is increasingly clear from complete genome assemblies that substantial fractions of the genome violate these assumptions. In particular, structurally complex regions with extreme sequence divergence are not only difficult to represent using linear references and remain poorly characterized at the population scale, but they can also create extended haplotypes with extreme linkage disequilibrium.

Here, we curated and assembled 622 long-read de novo assemblies of Arabidopsis thaliana, together with 58 haplotypes from 10 closely related Brassicaceae species, to trace the evolutionary dynamics of genetic variation across populations and species boundaries. We show that highly divergent regions (HDRs) are pervasive in the genome and are characterized by extreme variant density, frequent read-mapping failure, and tightly clustered structural variants. These regions are predominantly multiallelic, violating biallelic assumptions that underlie most population-genetic analyses.

To capture those complex variations, we developed GRASP (Graph Reconstruction Anchored by Subgraph Partitioning), a scalable framework for characterizing structural diversity and allelic architecture at orthologous loci across large populations. Applying GRASP reveals that HDR haplotypes exhibit extensive structural divergence at the sequence level, despite sharing much of their protein-coding gene content. Comparisons with outgroup haplotypes indicate heterogeneous evolutionary histories, with some haplotypes persisting across species boundaries and others reflecting more recent population-specific diversification.

Together, our results demonstrate that a substantial fraction of genetic variation in Arabidopsis resides in extended linkage blocks formed by multiallelic haplotypes that are poorly captured by mapping-based methods. These findings motivate a shift from site-centric to haplotype-centric population genetics and highlight the importance of pangenome-scale frameworks for understanding evolutionary dynamics, genotype–phenotype relationships, and adaptive potential.

Lineage specific selection in the mound-building mouse Mus spicilegus

Mara Baylis

UC Berkeley

The steppe mouse Mus spicilegus exhibits unique behaviors within its genus, including mound-building, pubertal delay, and dispersal behaviors correlated with day length, making it an ideal system for linking lineage-specific molecular evolution to the genetics underlying these traits. Toward this end, we first analyzed coding alignments spanning Rattus and diverse Mus lineages, fitting codon models in PAML to test for lineage specific acceleration of nonsynonymous substitution on the M. spicilegus branch. Likelihood ratio tests identified 51 significant genes as candidates of positive selection along the M. spicilegus lineage, an excess relative to other Mus, including several characterized in model mice with roles in pubertal development, circadian regulation, and synaptic processes. Separately, we used phyloP to identify cis-regulatory loci whose sequences were significantly accelerated on the M. spicilegus branch, and identified elements under accelerated evolution. These elements are identified as regulatory regions with target genes involved in neuronal development and metabolism. To further contextualize gene candidates, we intersected gene sets with cell-type-resolved expression from mouse brain resources. Together, our data establish that M. spicilegus has adapted under evolutionary pressures distinct from other house mice, and they open a first window onto the genetic architecture of photoperiod gated dispersal, risk behavior, and mound-building phenotypes.

Inferring epistasis in evolutionary accumulation processes

Dmitry Biba

Cold Spring Harbor Laboratory

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

Rare-common variant interactions influence genetic disease and explain incomplete penetrance

Federico Billeci

Center of Genomic Regulation (CRG - Barcelona)

Penetrance is defined as the proportion of variant carriers who are affected by the disease. The growing availability of population-based biobanks has revealed larger-than-expected numbers of healthy individuals harbouring rare variants considered pathogenic. This phenomenon is referred to as incomplete penetrance, and raised concerns about the correct use of genetic testing and whether additional factors should be considered in genetic counselling. Recent work of Fahed et al. 2020 showed the importance of including common variants' genetic liability to disease in estimating risk across carriers of rare variants for common complex traits. However, the joint contribution of rare and common variants has not been studied beyond additive models. We hypothesise that genetic background not only constitutes an additional genetic risk factor but also modulates pathogenic variants' effect. We propose a novel Bayesian modelling strategy to estimate disease risk by modelling the interactions between common and rare coding variants. We trained our model on complex dichotomous traits on the UK Biobank whole-exome sequencing cohort. Our approach shows a substantial improvement in disease risk prediction across rare variant carriers relative to additive models. Among the individuals at high predicted disease risk, we observe a significantly higher penetrance than the average. Our results suggest that common variants can act as disease risk modifiers of rare pathogenic variants and highlight the importance of including genetic background when estimating their effect on disease.

High-resolution mapping of a rapidly evolving complex trait reveals genotype-phenotype stability and a complex genetic architecture of adaptation

Mark Bitter

Stanford University

A key application of association mapping of complex traits is to predict adaptation. Such predictability can be demonstrated either through correctly inferring whether or not a trait has evolved, given patterns of genomic data alone; or, correctly predicting which loci showed signatures of selection given observation of phenotypic change alone. Systematically testing the efficacy this application of association data is challenged in most systems (e.g., humans), largely due to the timescale over which evolutionary dynamics proceed. Accordingly, to circumvent this limitation, we leveraged rapid evolution of a model complex trait in a model species, and high-powered longitudinal sampling design to test the efficacy of genomic prediction of complex trait variation and evolution in ecologically relevant settings. Specifically, we monitored genome-wide allele frequencies and pigmentation variation in genetically diverse populations of Drosophila melanogaster across seven generations of evolution in both field mesocosms exposed to natural environmental fluctuations, as well as mesocosms housed in a controlled, lab-based setting. At two time points throughout trait evolution, we conducted a high-powered, tail-based mapping of pigmentation, producing a well-resolved genotype-phenotype map that reaffirms canonical pigmentation genes and unveils novel loci. While we were able to use this map to correctly infer the direction of pigmentation evolution in both the field and lab mesocosms, the particular loci responding to selection, and thus architecture of adaptation itself, were largely unpredictable. We suggest this unpredictability to be a result of pleiotropic constraint, which was more pronounced in the field, relative to the lab-based, environment. Finally, we quantified a striking stability of the genotype-phenotype map across genetically diverged populations, demonstrating that shifting epistatic landscapes associated with the evolutionary process

itself do not alter trait architecture and preclude phenotypic prediction, provided the mapping is sufficiently powered. In concert our results highlight both the promise and limitations of genomic prediction, and exemplify the challenges of applying lab-based studies of complex traits to their evolutionary dynamics in the wild.

Towards continuous ancestry PRS via genetic distance kernel methods

Leandra Braeuninger

University College London, Alan Turing Insitute

Polygenic risk scores (PRS) are quantitative summaries of a person's genetic predisposition to a disease or trait. While PRS have shown considerable potential for clinical utility, for many traits their predictive performance remains starkly uneven across diverse genetic ancestries. To improve transferability, numerous ancestry-aware approaches have been proposed, many of which aggregate models trained within separate genetic ancestry groups or adjust for group-specific differences. However, these approaches typically rely on discretising genetic ancestry into categorical groups, often using continental labels. Such discretisation risks conflating biological variation with sociopolitical constructs such as race and perpetuating discriminatory belief systems.

Here, we explore efforts to move towards continuous characterisations of genetic ancestry for PRS. We then examine the use of kernel methods to this end and propose a kernel-based formulation of genetic distance, which naturally captures smooth genetic variation without imposing discrete ancestry boundaries. Such an approach would implement contemporary conceptualisations of genetic ancestry, and may improve generalisability of PRS across heterogeneous populations.

Uncovering structure in genetic interaction networks of yeast

Grace Brophy

Tufts University

Biological traits are shaped by genetic variation. In many settings, quantitative traits can be modeled as a linear and pairwise function of variation at genetic loci, where pairwise terms suggest genetic interactions. While traits differ in their genetic dependencies and patterns of interaction, they are not independent and often share structure. Prior approaches aim to decompose and interpret the shared structure of linear genetic effects, with the assumptions that each locus participates in a small number of “core processes,” and each core process influences a small number of traits. In our work, we extend this idea to explore structure in pairwise epistatic patterns using penalized matrix decomposition. Our method identifies sets of sparse, shared “basis” networks that are repeatedly observed across traits. We apply our method to a yeast fitness dataset, where growth rates in different environments define the traits of interest and pairwise epistatic coefficients are inferred from a large-scale F1 cross. The basis networks that our model finds exhibit biologically meaningful activation across traits. We also find that our results complement those from linear decomposition methods applied to the same dataset, while capturing additional structure specific to epistatic interactions. This is joint work with Nicholas Lawson, Caroline Holmes, Michael Desai, and Samantha Petti.

Reconstruction of splicing regulatory networks from single cells

Prakruthi Burra

University of California, Berkeley

Reconstruction of splicing regulatory networks from single cells

Prakruthi Burra1, Liana F. Lareau1,2,3

1 Center for Computational Biology, University of California, Berkeley

2 Department of Bioengineering, University of California, Berkeley

3 Chan Zuckerberg Biohub, San Francisco

Alternative mRNA splicing dramatically shapes the transcriptome, adding complexity to the output of the genome. Regulated alternative splicing produces mRNA isoforms with distinct roles in differentiation, disease, and cell function. Reconstructing the splicing code and predicting alternative splicing outcomes from sequence has been a major goal of computational biology, with new relevance in the era of personalized genome interpretation. Here, we leverage deep sequencing of single cells to reconstruct regulatory networks of alternative splicing in development. Our method, Istos, uses scRNA-seq to uncover position-dependent splicing regulation by modeling interactions between regulatory sequence features and splicing factors. Specifically, Istos maps subtle changes in the abundances of splicing factors to changes in splicing outcomes in an interpretable, biochemically motivated regression framework that incorporates the strength and position of splicing factor binding sites. We build on our earlier methods to address the sparsity of single cell splicing data, enabling biological discovery from noisy data. Our method captures determinants of regulated alternative splicing, showing how exon inclusion levels depend on splicing factor abundances over a developmental process.

Utilizing haplotype cluster assignments for fast and accurate local ancestry estimation and deconvolution

Thomas Bøggild

University of Copenhagen

Computational efficiency is crucial when working on large scale genetic data and finding ways to infer and handle admixture at a local level is not yet a solved problem. Haplotype clustering is a process that bins SNP data from multiple sampled haplotypes in windows along the genome and groups them by similarity. Using the PDC-DP-Medians algorithm implemented in the software suite hapla, this transformation is relatively inexpensive and can be seen as a way to utilize the linkage disequilibrium (LD) to reduce the dimension of the data which makes it less costly to process while retaining most of the information. This also greatly reduces the violation of the independence assumption in downstream analysis caused by LD between the markers. The haplotype cluster assignments can be used in downstream analyses like ancestry proportions, PCA or GRM estimation in the place of SNP data, though with a smaller number of loci, but usually a larger number of alleles per site/window. They can also be used for estimating ancestry at a local level on the genome with our model fatash, which uses the inferred parameters of admixture modelling on the haplotype cluster assignments to set the parameters of an HMM, which then only requires decoding to yield estimates for local ancestry. Based on simulated data the method is at least as accurate as RFmix and FLARE, while being much faster. We show that we can deconvolute the ancestry of the admixed individuals from the 1000 genomes project into their ancestral components. We also show that using probabilistic PCA which handles missingness is better at clustering the decomposed individuals with their unadmixed counterparts, indicating correct inference, as well as reflecting capture of substructure within estimated source populations. It also points to this deconvolution process potentially being useful for utilizing otherwise discarded information in admixed samples, in methods that work with missing data, but cannot account for admixture.

Test for association between responses and tree structures

Lorenzo Cappello

Universitat Pompeu Fabra

In many applications, data naturally come with a tree structure that encodes hierarchical relationships among nn samples, together with additional response variables observed for each sample. For example, a tree is inferred from molecular sequences of viruses collected from infected individuals, and the associated responses correspond to patient characteristics such as disease severity or vital parameters. A central question in this setting is whether the responses exhibit any association with the hierarchical partitioning induced by the tree. We propose a methodology with statistical guarantees to address this question that makes no assumptions on the form of the dependence, accommodates multivariate responses, and provides principled measures of uncertainty. Joint work with Davis Berlind and Oscar Madrid Padilla.

Classification of single-strand methylation with PacBio SMRT Data

Chester Henry Charlton

BiRC, Aarhus University

Detection of CpG methylation is fundamental to epigenetic analysis, yet current methods often rely on dual-strand consensus, obscuring hemimethylation events. We demonstrate that kinetic signals derived from single-strand PacBio SMRT sequencing are sufficient for robust methylation attribution. We present a deep convolutional neural network (CNN) trained on the PacBio Primrose 5mC validation dataset, which successfully identifies methylation status without complementary strand information, achieving 82% Top-1 accuracy. Furthermore, we address the challenge of labeled data scarcity in epigenetics by proposing a Contrastive Predictive Coding (CPC) foundation model. We describe the architecture of this self-supervised framework and its potential to leverage vast amounts of unlabeled sequencing data to improve generalization in future single-molecule modification classifiers.

Detecting sapiens-specific selective sweeps by leveraging deep divergence and machine learning

Mark Chernyshev

Uppsala Universitet

The identification of ancient genetic adaptations can shed light on the traits which distinguish us from archaic humans, revealing human-specific disease vulnerabilities and molecular mechanisms which were crucial to survival. The present work describes a random forest model trained on a modified population branch statistic (sPBS) calculated from coalescent simulations to identify sapiens-specific selection and classify its strength and timing. We leverage the unique phylogenetic position of the newly sequenced genomes of ancient Southern Africans (Jakobsson et al. 2025), a non-admixed population representing the deepest split among all homo sapiens (~300kya), as well as the recently expanded list of five high-coverage archaic genomes (Neanderthal and Denisovan) to provide the best resolution possible. To ensure clean signal from sapiens-specific evolution, we utilize Yorubans as a second homo sapiens population, chosen for their lack of archaic introgression, preventing introgression from confounding simulation-based training. Our method achieves an ROC AUC of 0.89 for classifying selected windows across all simulated selection coefficients (s=0.05 and s=0.01) and mutation timepoints (330kya, 450kya, 590kya). Notably, the random forest can also differentiate between selection strengths across all timepoints with an ROC AUC of 0.91 and can classify whether selection occurred 330kya with an ROC AUC of 0.88. Finally, we apply this method to the empirical data to extend beyond the fixed coding variant analysis in Jakobsson et al. 2025 to a global sweep search for varying selection strengths and timings.

LAML-Pro: joint maximum likelihood inference of cell genotypes and cell lineage trees

Gillian Chu

Princeton University

Motivation. Recent dynamic lineage tracing technologies use genome editing to induce heritable mutations, or edits, that accumulate across successive cell divisions. These edits are measured using single-cell sequencing or imaging, providing data to reconstruct cell lineages at single-cell resolution. Current computational approaches to infer cell lineage trees, or phylogenies, from these data perform two separate steps: (1) Identify each cell's edits (genotype) from the raw sequencing or imaging data; (2) Infer a cell lineage tree from the cell genotypes. However, genotyping cells is an inexact process and genotype errors can yield an inaccurate lineage tree. For example, using fluorescence based-imaging to measure edits results in a high fraction (≈25-50%) uncertain or genotype errors.

Results. We introduce Lineage Analysis via Maximum Likelihood with PRobabilistic Observations (LAML-Pro), an algorithm that jointly infers cell genotypes and a cell lineage tree. LAML-Pro is based on the Probabilistic Mixed-type Missing Observation (PMMO) model, which we derive to describe both the genome editing and genotype observation processes. LAML-Pro constructs lineage trees from thousands of cells in under an hour by leveraging the sparsity of transitions under the PMMO model. On simulated data, we demonstrate that LAML-Pro corrects genotype errors and infers substantially more accurate trees than existing methods which are vulnerable to genotype errors. Applied to data from two recent imaging-based lineage tracing systems, LAML-Pro reduces genotype errors by 5-fold and produces more spatially coherent lineage trees compared to existing methods.

Availability. LAML-Pro is implemented in C++ and is freely available at:

github.com/raphael-group/LAML-Pro.

Conditions favoring alleles that disturb the sex determination pathway

Andrew G. Clark

Cornell University

Sex determination requires the tightly coordinated initiation of early developmental processes, gametogenesis, and dosage compensation. In Drosophila, the primary signal that establishes male or female fate is the X-chromosome-to-autosome (X:A) ratio. Depending on the expression of X and autosomal factors, this ratio is interpreted as either 1.0 (XX:AA for females) or 0.5 (XY:AA for males). Reduced X-linked gene expression has no effect in XY individuals but activates dosage compensation (hypertranscription of the single X chromosome) and causes lethality in XX individuals. Conversely, elevated expression of X-linked factors is benign in XX individuals but suppresses dosage compensation and causes lethality in XY individuals. Surprisingly, natural populations harbor substantial variation in these regulatory errors. Using QTL mapping, we identified multiple genes underlying this variation. Here, we evaluate the selective forces that might maintain this polymorphism and consider whether it could ever be adaptive. Drosophila populations also harbor several sex-ratio drive systems, often mediated by factors that kill Y-bearing sperm. Here we develop an explicit genetic model to test whether sex determination modifiers, (like those above) can invade when initially rare in drive-distorted populations. Despite elevating embryonic mortality, these modifiers can invade under certain conditions, offering a mechanism to maintain what initially appears to be a maladaptive polymorphism.

Identifying germline mutation hotspots and their determinants in humans

Nathan Cramer

University of California, Berkeley

Germline mutations are the ultimate source of genetic diversity within and across species. Understanding germline mutation rates and mechanisms is essential for studies of medical genetics (to interpret the incidence of de novo and heritable diseases) and evolutionary biology (to date demographic and adaptive events). Germline mutation rates vary across the human genome at multiple resolutions, ranging from adjacent base pairs to whole chromosomes. However, the biological mechanisms underlying mutation rate variation are yet to be fully characterized. To understand the determinants of local germline mutation rates, we analyzed single nucleotide polymorphisms from 76,156 whole-genome sequences of Europeans, East Asians, and Africans in the Genome Aggregation Database (gnomAD). We focus on putatively neutral mutations by removing conserved regions and exons. We apply a wavelet-based approach to identify mutation hotspots across varying genomic resolutions. Simulations demonstrate that our method outperforms the naive windowing approach for detecting hotspots across multiple scales (100 KB - 1 MB). Application to gnomAD shows that many factors correlate with local mutation rate variation and are enriched in mutational hotspots, including replication timing, recombination rate, germline methylation, and GC content. We recover known C $>$ G maternal hotspots and identify novel associations between scale-specific mutation hotspots and mutational determinants, such as replication timing. We then apply this approach to study mutation rate differences across human populations. Notably, we identify population-specific enrichment of TCC $>$ TTC mutational hotspots in Europeans compared to Africans. By associating DNA motifs and genomic features enriched in the hotspots, we uncover candidate trans and cis-acting factors associated with mutational modifiers in Europeans. Together, our method provides a novel framework for identifying mutation hotspots and

the genomic features and biochemical processes impacting the mutation rate landscape in humans.

Quartet-based species tree methods enable fast and consistent tree of blobs reconstruction under the network multispecies coalescent

Junyan Dai

University of Maryland, College Park

Gene flow between species or populations is an important force in evolution, modeled by the network multispecies coalescent. Reconstructing evolutionary histories, called species networks, under this model is notoriously challenging, with the leading methods scaling to just tens of species. Divide-and-conquer is a promising path forward; however, methods with statistical consistency guarantees require the \{}textit{tree of blobs} (TOB), which displays only the tree-like parts of the network, to perform subset decomposition. TOB reconstruction under the NMSC is challenging in its own right, with the only available method TINNiK having time complexity $O(n{̂}5 + n{̂}4k)$, where k is the number of input gene trees and n is the number of species. Here, we present a framework for TOB reconstruction that operates by (1) seeking a refinement of the TOB and then (2) contracting edges in it. For step (1), we show that an optimal solution to Weighted Quartet Consensus is a TOB refinement almost surely, as the number of gene trees increases, motivating the use of fast quartet-based methods for species tree estimation such as ASTRAL or TREE-QMC. For step (2), we contract edges in the refinement tree based on the same hypothesis tests as TINNiK, which are applicable to subsets of four taxa. We show that sampling just $O(n)$ four-taxon subsets around each edge enables statistically consistent TOB estimation, with asymptotic runtime dominated by tree reconstruction. Leveraging TREE-QMC for this step gives our method a time complexity of $O(n{̂}3k)$ and its name TOB-QMC. On simulated data sets, TOB-QMC is at least as accurate and often more accurate than TINNiK. Moreover, TOB-QMC scales to larger data sets and enables fast and interpretable exploration of hyperparameters used in hypothesis testing. We demonstrate the importance of this feature on phylogenomic data sets. Lastly, our framework is related to ad

hoc analyses performed by biologists, as network methods do not scale. Our theoretical results provide justification for such approaches as well as context for interpreting species trees estimated with quartet-based methods in the presence of gene flow; this is critical given the recent result that tree-based network inference with ASTRAL can be positively misleading(Link to our preprint: https://www.biorxiv.org/content/10.1101/2025.11.05.686850v1.full.pdf).

Uncovering the dynamics of population structure through time using genome-wide genealogies

Yun Deng

Stanford University

Understanding population structure and the underlying demographic history is a central goal in population genetics. Popular ancestry decomposition methods, such as STRUCTURE and ADMIXTURE, provide a summary of individual ancestry proportions. But they are extremely simplified representations of the much richer evolutionary processes. In particular: (1) they do not resolve local ancestry patterns along the genome; (2) they provide no information about when ancestries diverged or how drift accumulated over time; and (3) when many components are specified, the interpretation of “latent ancestries” becomes unclear, when no present-day population is a pure representative of some components.

Here we introduce, argmixture, a framework that transforms ancestry decomposition into a comprehensive reconstruction of demographic history. Argmixture jointly calls local ancestry tracts and performs time-resolved PCA between these ancestries, enabling direct visualization of the temporal dynamics of population divergence and admixture. Crucially, argmixture can also infer the local ancestry and demographic history of "ghost ancestries"—ancestral groups lacking modern representatives—without requiring reference genomes. This is particularly important because many ancestries involved in human admixture events are extinct or only partially sampled in present-day populations.

We applied argmixture to the Middle Eastern populations of the Human Genome Diversity Project and reconstructed their demographic histories. Notably, argmixture identifies and characterizes the Levantine Farmer ancestry, which is represented in ancient DNA but absent as a pure modern population. argmixture not only recovers the corresponding local ancestry tracts in modern individuals, but also reconstructs its divergence and admixture history—without using any ancient DNA. This demonstrates argmixture's capability to enable a richer, temporally explicit understanding of population structure than traditional ancestry-decomposition methods.

Learning lifetime disease liability reveals and removes genetic confounding in electronic health records

Yazheng Di

ETH Zurich, D-BSSE (Department of Biosystems Science and Engineering)

Population-scale biobanks and electronic health records (EHRs) are expanding genetic research into new phenotyping modalities. Prior work has characterized and corrected EHR biases for short-term clinical prediction, yet genetic studies require phenotypes that reflect lifetime-scale disease liability rather than instantaneous risk. Here we learn such a liability construct by jointly modeling EHR codes and routinely available clinical measurements, supervised on deeply phenotyped subsets.

We evaluate this model on nine diseases in UKBiobank: generalized anxiety disorder, major depressive disorder, anaemia, chronic obstructive pulmonary disease, diabetes, hyper-LDL-cholesterolemia, hypertension, non-alcoholic fatty liver disease, and osteoporosis. The framework achieves a macro-AUC of 0.74 when using EHR codes alone, and 0.97 when adding disease-relevant information. As deep phenotyping labels are costly, we develop an active learning callback strategy that improves label efficiency by 2-fold, maintaining prediction accuracy while halving required clinical labeling.

The learned liability increases genetic correlation with high-specificity deep phenotypes and yields more etiologically relevant GWAS loci than raw EHR codes. We also identify a shared genetic confounder in EHR that inflates cross-disease correlations and drives most spurious links with behavioral and socio-economic traits. This confounder generalizes across biobanks, and its correction abolishes most non-etiological genetic correlations, demonstrating a promising direction for restoring true disease genetic architecture directly from GWAS summary statistics. Our results demonstrate that EHR GWAS can be meaningfully refined, advancing genetic research toward more biologically grounded discovery.

Inconsistency of model selection method in admixture model using second-order changes of log-likelihood

Dat Do

University of Chicago

Choosing the number of populations K in admixture analysis to investigate population structure is a challenging problem, both in practice and in theory. In the original Bayesian implementation of the model (STRUCTURE software), the \$\{}Delta K\$ method (Evanno et. al. 2005) is perhaps the most popular choice to choose the number of populations, where it picks the \$K\$ with the highest second-order change in log-likelihood. In this project, we show that applying this strategy to the Maximum Likelihood Estimator (for example, using the ADMIXTURE program) can lead to inconsistent estimates of the populations when the amount of data tends to infinity, under a simple three-population setting. We quantify this phenomenon under the Jensen-Shannon divergence between the three populations, and then connect to F_2 and F_ST statistics. Simulation studies and a real-data experiment are presented to illustrate our theory. This is a joint work with Professor Jonathan Terhorst (University of Michigan)

Genomic features explaining deep learning predictions of cis-regulatory variant effects

Shiron Drusinsky

UC San Francisco/Gladstone Institutes

Genomic deep learning (DL) models that predict gene expression from DNA inputs underperform when given personal genome sequences, even after fine-tuning on personal genomes and paired gene expression values, but the reasons are unclear. To better understand why, we used Enformer as a representative model and introduced C/G substitutions into the consensus STAT1 motif sequence within personal genome sequences around different genes, along with strong, consistent gene expression increases associated with the G allele. In this way, we tested how well Enformer could learn this clear and causal, yet artificial, pattern after observing it repeatedly throughout the human genome and across many personal genomes, which mirrors how we would otherwise expect it to learn to detect the more subtle effects of natural genetic variation when using unaltered personal genomes. We found that Enformer underperforms on personal genome sequences because it systematically underestimates causal variant effect magnitudes. This phenomenon persists even when the number of artificial causal variants seen during training far outnumbers that which we expect to observe naturally, suggesting personal genome datasets exhibit insufficient variation to maximize performance for this task. Further, we show that Enformer underestimates causal variants because of substantial locus-to-locus diversity in sequence contexts surrounding these variants, which obfuscates the learnable pattern. We find that placing artificial causal variants into motifs that are longer, more palindromic, and/or appear more conserved between different genomic loci leads to less sequence context diversity and better learning of the causal variant effect, and we find these observations also hold for naturally observed genetic variants. Overall, we introduce a framework for investigating genomic DL models' abilities to learn causal variant effects without being hindered by uncertainty about variant causality, and identify genomic features

responsible for their accuracy on variant effect predictions, including within personal genomes.

Exposure accumulation drives age-dependent disease architectures and polygenic risk scores

Arun Durvasula

USC

Our understanding of the dependence of the genetic and environmental architecture of common diseases on age is incomplete. Here, we use longitudinal data to separate age-dependent genetic and environmental variance across complex traits and diseases. When applying our approach to 16 UK Biobank quantitative traits (average N=180K), we found environmental variance accumulates with age, leading to decreasing heritability with age, which follows an exposure accumulation model that is distinct from gene-age interaction model. Heritability decreases with age for 5 traits including systolic blood pressure and lung function (FEV1/FVC), with an average change of -17.8% with age per 10-years. We demonstrated that majority of the decreasing heritability comes from exposure accumulation, instead of gene-age interaction, using longitudinal phenotypic correlations. For diseases, environmental variance in disease liability also accumulates with age for 5 of 9 diseases, with an average decrease in heritability of -18.1% per 10 years. A liability-threshold model with exposure accumulation explains 86% of decreasing PRS prediction R2 with age in 9 UK Biobank complex diseases. Finally, we show that both genetic and non-genetic predictors have decreasing prediction accuracy with age in 61 predictor-disease pairs. Taken together, our results demonstrate that the age-dependent liability-threshold model with exposure accumulation is a general disease model, suggesting ascertaining younger cohorts is more powerful for training both genetic and non-genetic risk prediction models.

Accelerated phylodynamic inference via neural ODE solvers

Alexis Edozie

University of Michigan

Multi-type birth-death processes such as the birth-death-exposed-infected (BDEI) process are a class of stochastic models that are used for studying the spread of an infectious disease. When used in combination with genetic data, they allow for accurate epidemiological surveillance without the need for costly measures such as contact tracing. Fitting these models to data requires solving a large number of simultaneous ordinary differential equations (ODEs). This limits their ability to analyze data sets containing more than a few hundred samples. We propose an attention-based neural ODE method inspired by transformer architectures to fit BDEI models efficiently, eliminating the need for repeated ODE solvers. We train a neural network to learn a time-stepping integration scheme, analogous to a corrected forward Euler method, to directly compute the BDEI dynamics. Our method significantly reduces computational overhead, enabling rapid likelihood estimation even for large samples. This approach offers a scalable, fast, and accurate solution for disease monitoring, making it particularly useful for epidemiological applications where timely decision-making is critical.

Scalable genealogical association testing for binary traits using ancestral recombination graphs

Yining Fan

University of Oxford

Genome-wide association studies have been instrumental in identifying genetic variants associated with complex traits. However, detecting associations with rare variants, which often have large effects, remains challenging, particularly in cohorts that lack large, ancestry-matched sequencing reference panels, as is common for case–control cohorts and disease traits. Ancestral recombination graphs (ARGs) have been shown to complement genotype imputation in these analyses by leveraging genealogical relationships to infer the presence of unobserved variation. However, current approaches that use ARGs for complex trait analyses are limited to quantitative traits and cannot be applied to study rare variation in disease phenotypes.

We introduce computationally efficient algorithms for genealogical association testing for binary phenotypes using ARGs. Our method uses fast ARG-based matrix operations to compute score-based test statistics, enabling sub-linear time genotype–matrix products and efficient multi-trait analysis. The approach also incorporates Firth regression and saddle point approximation to control false positive rates in the presence of case–control imbalance and rare variants. Simulations show that these ARG-based operations are highly scalable, enabling analysis of biobank-scale datasets, and that the method provides increased power for detecting ultra-rare associations and complementary signals to imputation-based approaches, particularly when imputation quality is poor or reference panels are mismatched.

Disentangling biophysical effects of rare missense variants on cis and trans plasma protein levels in the UK Biobank

Ezequiel Alejandro Galpern

CRG

Genome sequencing has grown exponentially, but determining the phenotypic consequences of missense variants remains a key challenge in human genetics. Single amino-acid changes may alter biological functions, driving disease through diverse molecular mechanisms. Deep-learning unsupervised methods that model the distribution of sequence variation across organisms have emerged as promising tools for scoring variant effects. However, evolutionary-based models often fall short of attributing effects to specific biophysical mechanisms. Large-scale plasma proteomics in the UK Biobank provides a resource to evaluate variant effect predictors, as pathogenic variants often associate with lower measured plasma protein abundance. Proteogenomics studies have also identified widespread cis and trans associations between genetic variation and protein levels, but linking these associations to biophysically interpretable causes remains difficult.

Here, we integrate protein language models with folding energetics to disentangle effects on stability and on functions beyond folding, such as ligand/partner binding and downstream interactions. We propose a general method to quantify and separate mutational effects on the energetics of protein folding versus natural selection, capturing functional constraints imposed by evolution. We compute the protein “Dark Energy” (Galpern et al., PNAS, in press) as the difference between physical folding free energies and evolutionary free energies derived from a protein language model.

We relate predicted stability and Dark Energy changes from rare missense variants to plasma protein measurements in UK Biobank, considering both the encoded protein (cis) and distal proteins (trans). This method supports mechanistic hypotheses for trans associations by separating cases driven primarily by destabilization of the encoded protein from those more consistent with functional perturbations beyond folding. More broadly, it provides a scalable framework for proteome-wide mapping of biophysically interpretable cis and trans effects of rare missense variation.

Effect size correlations of proximal SNPs in a complex trait

Shivam Gandhi

Harvard University - Sunyaev Lab

A common modeling assumption in population genetic models of selection is that genetic variants do not directly interact with one another to determine the evolution and fate of mutations. However, several recent analyses suggest this assumption is often violated between proximal variants. Biobank analyses of complex traits have inferred genome-wide effect size correlations between proximal SNPs in positive LD. Similarly, analyses of the site frequency spectrum of missense variants suggest the presence of compensatory effects or antagonistic epistasis within proteins. To investigate how population genetic forces generate effect size correlations, we build a general two-locus model of segregating alleles influencing a phenotype under stabilizing selection. We derive analytical approximations and implement Wright-Fisher simulations to compute the strength of SNP-pair effect correlations as a function of the distribution of mutational effects, recombination, and the degree of polygenicity of the trait under selection. While stabilizing selection consistently generates negative correlations for pairs in positive LD, consistent with the Bulmer effect, correlations for pairs in negative LD depend strictly on the degree of polygenicity of the trait. We predict that positive correlations can arise for highly polygenic traits while no correlation is expected for simple stabilizing selection within a confined genetic element. As an application, we utilize this framework to investigate the architecture of protein stability. We analyze coding variants in the same protein where both GWAS and SFS-based analyses have suggested stabilizing selection or compensatory interactions. We use high-throughput experimental mapping of ddG values, a measure of how destabilizing a variant is, from over 500 protein domains to annotate human haplotypes observed in the UK Biobank. Consistent with our theoretical predictions and earlier empirical results, we observed negative effect size correlations between nearby variants in positive LD (r=-0.09, p=1.34e-2) and no correlation between variant pairs in negative LD. This suggests stabilizing selection on protein folding stability manifests locally rather than through a highly polygenic background. We conclude by modeling pleiotropy and other protein phenotypes to extend our understanding of the presence of compensatory variation and the population genetic processes governing the creation of correlations within proteins.

Trinucleotide genome composition reflects context-dependent mutation spectra in mammals and plants

Ziyue Gao

University of Pennsylvania

Genome nucleotide composition is shaped by the interplay of mutation, recombination, drift, and selection. While GC-biased gene conversion (gBGC) is known to drive intra-genome variation in GC content, it cannot fully explain the variation in multi-nucleotide composition observed across genomic regions or among species. Building on our previous observation of a nearly perfect negative correlation between CpG mutation rates and CpG content across 108 eukaryotes, we hypothesize that the k-mer composition of neutral genomic regions largely reflects context-dependent mutation spectra.

To test this, we modeled the evolution of genomic trinucleotide composition using a discrete-time Markov chain framework. We found a strong, positive correlation between the ratio of reciprocal trinucleotide mutation rates and the corresponding trinucleotide abundance ratio at equilibrium. Empirically, this theoretical prediction is supported by the observed trinucleotide proportions in reference genomes and mutation spectra inferred from genome-wide polymorphisms in multiple species, although with a weaker correlation partially driven by gBGC. Crucially, this relationship holds not only across trinucleotide pairs within a genome but also across species for the same trinucleotide pair. This suggests that trinucleotide genome composition can serve as a powerful, scalable proxy for inferring context-dependent mutation patterns.

Leveraging this discovery, we analyzed the frequencies of 32 trinucleotides in the reference genomes of 30 mammals and 46 plants. The intergenic trinucleotide composition varies drastically across species. Notably, compared to other plants, the Poaceae family displays a unique composition characterized by high GC content and attenuated CpG depletion, indicating distinctive mutation spectra. Treating pairwise trinucleotide ratios as quantitative traits, we performed a phylogenetic regression analysis with genes involved in DNA repair, replication, and methylation, and identified candidate genes that modify mutation rates during plant evolution. Overall, this study demonstrates strong connections between genome composition and mutation rates, suggesting that a single reference genome can provide rich information on mutational patterns, even in species where polymorphism or direct mutation data are lacking.

A Tsallis-entropy lens on genetic variation

Margarita Geleta

UC Berkeley

We introduce an information-theoretic generalization of the fixation statistic, the Tsallis-order q F-statistic, Fq, which measures the fraction of Tsallis q-entropy lost within subpopulations relative to the pooled population. The family nests the classical variance-based fixation index FST at q=2 and a Shannon-entropy analogue at q=1, whose absolute form equals the mutual information between alleles and population labels. By varying q, Fq acts as a spectral differentiator that up-weights rare variants at low q, while q>1 increasingly emphasizes common variants, providing a more fine-grained view of differentiation than FST when allele-frequency spectra are skewed. On real data (865 Oceanian genomes with 1,823,000 sites) and controlled genealogical simulations (seeded from 1,432 founders from HGDP and 1000 Genomes panels, with 322,216 sites), we show that Fq in One-vs-Rest (OVR) and Leave-One-Out (LOO) modes provides clear attribution of which subpopulations drive regional structure, and sensitively timestamps isolation-migration events and founder effects. Fq serves as finer-resolution complement for simulation audits and population-structure summaries.

The genomic landscapes of Oceania and ISEA: insights from 92 populations

Peter Gerlach

Stanford University

Oceania and Island Southeast Asia have a rich, yet understudied, human genomic landscape. This region encompasses some of the first areas inhabited by humans following the out-of-Africa expansion, includes populations with the highest levels of archaic hominid introgression, and contains Pacific islands that are among the most remote continuously inhabited locations in the world. Here, we describe the first region-wide analysis of samples from population groups spanning Oceania and its broad perimeter. In total we generate and analyze genome-wide data from 92 different populations, 58 separate islands, and 30 countries, covering one third of the planet. Leveraging this diverse dataset, we resolve genetic connections among islands, providing a detailed view of regional population structure and identifying the island groups likely involved in the settlement of several Polynesian Outliers. Ancestry-specific analyses allow us to deconvolve different layers of history from tracing groups deriving their Austronesian ancestry via the Lapita expansion to quantifying variable archaic introgression across the basal Papuan component of Oceanians and Southeast Asians. Together, these findings refine models of oceanic settlement and admixture and establish a comprehensive reference that will enable subsequent work on genetic, historical, and health-related diversity while advancing global efforts to ensure broad and equitable representation in human genomics.

How natural selection shapes gene regulatory architecture: the case for 5' UTRs

Tami Gjorgjieva

Stanford

Selective constraint, or the extent to which a genetic sequence tolerates variation, is a valuable proxy of biological importance. In fact, constraint captures the total effect of gene expression changes on all traits, and can thus be construed as a fundamental measure of a gene's importance for traits. Existing estimators of gene-level constraint—such as pLI, LOEUF and shet—have been widely utilized, but are limited to coding regions and rely on loss of function variation. Our work extends these frameworks to model selection on a variant-level, in non-coding regions, and across a wider range of effects. We introduce gene dosage selection curves (GDSCs), which describe the mathematical relationship between variant effects on gene expression and their selection coefficients. In this project, we study selection in 5' untranslated regions (UTRs), which are important regulators of translation. We first measure the effects of 1 million 5' UTR variants on translation in HEK239T cells using NaP-TRAP, a novel MPRA. We implement a robust error modeling approach and obtain reliable effect size estimates. Analyzing variant effects with their frequencies, we identify unique signatures of selection in 5' UTRs. We develop a statistical framework to infer GDSCs, and find that translation-increasing variants are under weaker selection than translation-decreasing variants. Finally, we investigate 5' UTR variant effects on protein expression in the UKB, and develop a simple algorithm to predict the direction of 5' UTR variant effects on translation. This work provides meaningful insights into the functional and evolutionary landscape of 5' UTRs and the inference of GDSCs.

Rare variation in malaria parasites biases population-genetic inference

Amy Goldberg

UCLA

Understanding how pathogens evolve is fundamental to disease control and a basic question in evolutionary biology, yet pathogens with complex life cycles violate assumptions of classic evolutionary models. Genetic analyses of the malaria parasite \{}textit{Plasmodium falciparum} have shown multiple surprising patterns. In particular, effective population size estimates that vary by orders of magnitude depending on the method, and an excess of genes with elevated nonsynonymous variation (measured as $\{}pi_N/\{}pi_S$). Here, reanalyzing genomic data from 18 worldwide populations, I show that these observations are unified predictions of a multiple-merger coalescent, which reflects the parasite's skew in reproductive success. The impact is particularly strong in endemic regions. This genealogical tree shape generates a massive excess of rare variants compared to the standard Kingman coalescent, causing common summary statistics to be biased in predictable directions. In particular, the abundance of rare variants interacts with the mathematical properties of ratio statistics to systematically inflate gene-level $\{}pi_N/\{}pi_S$ estimates. Notably, filtering this rare variation reveals previously masked candidates for selection, including well-characterized antigens such as merozoite surface proteins. This framework provides a foundation for interpreting genomic data in pathogens with high reproductive variance.

Heritability of germline mutagenesis in 40 large three- and four-generation pedigrees

Michael Goldberg

University of Utah

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

Here, we present the first phase of the Gametes Through Generations (GTG) project, which comprises new whole genome sequencing of >1000 individuals from 22 four-generation and 17 three-generation CEPH/Utah pedigrees. Prior studies of mutation rate heritability in humans have been limited to single-nucleotide mutations observed in trios. The large GTG pedigrees allow us to measure germline mutation in hundreds of individuals, eliminate false positives, and assign mutations to a parent-of-origin.

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

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

Coalescent inference for pathogens with latent periods

Isaac Goldstein

Stanford University

The effective reproduction number is an important descriptor of an infectious disease epidemic. Coalescent models can be used to infer the effective reproduction number from genealogies constructed from viral genomes sampled from infected individuals during an epidemic. Most methods based on the coalescent model either assume individuals to be infectious as soon as they are infected, or that the transmission rate is constant over time, both are strong assumptions violated by many real world pathogens such as SARS-CoV-2. Inspired by recent work connecting coalescent models and phase-type distributions, we propose a variation of a coalescent model which tracks the counts of lineages representing infected but not infectious or infectious individuals where the probability of coalescence is the probability of reaching an absorbing state in a competing risk phase-type model. We use a variation of conditional uniformization to augment the unobserved counts of lineages in each infectious stage in a computationally tractable manner, and combine this model with a non-parametric Gaussian Markov Random Field prior on the time-varying effective reproduction number to conduct semiparametric Bayesian inference of the effective reproduction number. Using simulated genealogies, we demonstrate our model's robust performance under nontrivial model misspecification. We apply our new model to estimating the effective reproduction number of the 2014 Ebola epidemic in Liberia.

Quantifying selection on the nonsynonymous human mutation spectrum

Ryan Gutenkunst

University of Arizona

Mutation rates and fitness effects are often treated as independent, but mutation rates are variable and evolve under indirect selection. For example, human European populations experienced a transient increase in the 3-mer mutation rate TCC -> TTC in the past 20,000 years. To quantify indirect selection on mutation spectra, we developed an approach to estimate the distribution of fitness effects (DFE) of nonsynonymous 3-mer mutation types, by analyzing pairs of complementary mutation types to account for GC-biased gene conversion and ancestral state misidentification. We then applied this approach to all 96 possible 3-mer mutation types in humans, using data from the 1000 Genomes Project. We found widely varying DFEs among mutation types and that inferences from a few hundred genomes could predict pN/pS in samples of tens of thousands of genomes. Our DFE estimate for TCC -> TTC mutations is consistent with recent theoretical predictions by Milligan, Amster, and Sella of scenarios under which a moderate number of modifier loci could yield population-specific transient bursts of a specific mutation type.

Cophylogeny between individuals within a single host species and their symbionts

Rowan Hart

University of Chicago

Cophylogeny, the study of phylogenetic similarity between interacting organisms, provides insights into the shared evolutionary history of symbiosis. While cophylogenetic analyses have primarily investigated macroevolutionary relationships between species, increasing population genomic data of host-symbiont systems opens opportunities to explore cophylogeny between individuals within a single host species and their symbionts. While the ecological drivers of cophylogeny have been investigated at the macroevolutionary scale, their influence on microevolutionary processes are unclear. In part, this arises from the fact that the evolutionary history between individuals within a host species cannot be well represented with a single phylogenetic tree. Here, we propose to measure microevolutionary cophylogeny by comparing the ancestral recombination graph of the host individuals with the genealogy of the associated symbionts. This approach provides cophylogenetic measurements of genome-wide trends, as well as locus-specific specific signals in the host. Through simulations, we demonstrate how transmission mode, population structure, admixture, and allelic incompatibility influence cophylogeny. We find that genome-wide cophylogenetic signals do not arise from vertical transmission when the host population is large and unstructured, and investigate the differences of allopatric and sympatric population divergence. Further, we test the effects of allele incompatibility between hosts and symbionts, and discover transient locus-specific signals. We then measure mitonuclear cophylogeny within the 1000 Genomes Project—a system with strict maternal transmission and well-studied population structure—and observe substantial variation across human populations. Within putatively unstructured populations, we observe no genome-wide signals of mitonuclear cophylogeny, while structured and admixed populations show genome-wide signals. We explore locus-specific cophylogeny and evidence

of mitonuclear incompatibilities within and across human populations.

Sampling alternative gene genealogies in an ancestral recombination graph

Marc Henein

McGill University

Recently developed algorithms infer ancestral recombination graphs (ARGs) over the entire genome at biobank scale. Genealogies at specific loci are used for imputation, association studies and population genetic inferences. However, ARG inference algorithms that scale to large-scale datasets (e.g. ARG-Needle) employ heuristic methods without modeling uncertainty. Conversely, Bayesian inference approaches (e.g. ARGweaver) do not scale to large datasets. The uncertainty in the topology of inferred gene genealogies can be important. In simulations of clades with recent ancestry, we find that tree topologies are rarely reconstructed correctly, even for samples sharing long and informative haplotypes. In this work, we propose a strategy to rapidly resample ARGs consistent with the data with a focus on a locus of interest. This is based on two observations. First, a reasonable ARG must capture the structure of haplotype sharing among samples. Using simulations, we found that the haplotype sharing structure is accurately reconstructed by ARG inference methods even when the topology of the local genealogy is incorrect. Second, very dissimilar ARG topologies can produce the same haplotype sharing structure. We therefore developed an algorithm to enumerate all such compatible topologies and to sample from them using a maximum likelihood approach. We apply this strategy to genealogical inference (i.e., aligning gene trees to population-scale pedigrees) and rare variant imputation. We discuss its potential to improve ARG inference in general.

Discovery of Pacific Islander-specific segmental duplications in 46 haploid assemblies

Kiley Hudson

University of Minnesota

Segmental duplications (SDs) are a type of structural variation consisting of DNA sequences that are at least 1 kilobase long with 90% or greater similarity with another genome sequence. SDs shape gene and genome evolution and play an important role in both common and rare disease susceptibility. Pacific Islander populations remain underrepresented in genomic studies, limiting understanding of population-specific variation. Here, we call, characterize, and analyze segmental duplications in Pacific Islander genomes using 46 high-quality haploid assemblies, comparing them to 94 high-quality assemblies in the Human Pangenome Reference Consortium (HPRC).

Using Whole Genome Assembly Comparison following repeat masking and quality filtering, we identified segmental duplications across Pacific Islander assemblies alongside reference assemblies from African, Admixed American, East Asian, European, and South Asian populations in HPRC. Our quality control pipeline addressed assembly errors, highly complex regions, assembly fragmentation, and alignment confidence to generate a high-confidence set of SDs.

The results of our analysis showed significant population differences in SD content. We identified population-specific segmental duplications unique to Pacific Islander assemblies that were absent from other reference populations. Many of these Pacific Islander-specific SDs overlap annotated genes and exons, suggesting possible functional relevance that may elucidate the etiology of disease risk in this underrepresented population.

This work expands our knowledge of structural genomic variation in Pacific Islander populations. The identified population-specific segmental duplications present a promising target for further investigations into the genomic architecture underlying health disparities in Pacific Islander populations.

Leveraging the ancestral recombination graph to infer microsatellite mutational models

Sebastián Iturbe

University of Oregon, International Laboratory for Human Genome Research (LIIGH)

Short tandem repeats (STRs), or microsatellites, are repetitive sequences that make up approximately 3% of the human genome. More than 10,000 STR variants are known to influence gene expression and account for 10–15% of its heritability. Additionally, many STRs contribute to various clinical conditions. Modern long-read sequencing technologies now enable us to determine STR genotypes with higher accuracy and investigate their role in complex traits. However, understanding the evolutionary and mutational behavior of STRs is essential to characterize their impact on different phenotypes. To this end, inferring the mutation model that governs STR variation is critical. This task is particularly complicated by the fact that the mutational dynamics vary substantially across loci. To solve this problem we created Tandem Repeat Ancestral recombination graph Mutational Analysis (TRAMA), a tool that infers the mutational process of each STR locus individually. For each STR, TRAMA uses information from the local Ancestral Recombination Graph (ARG), which encodes the complete local genealogical history of the samples. TRAMA uses a novel maximum-likelihood approach conditioning on the local ARG information to infer the most probable STR mutation model and its parameters per locus from observed variation in a set of samples. We demonstrate the method's accuracy through extensive simulations showing that it can infer the known mutational model acting on each loci along with the parameters that explain its evolutionary history across a wide range of scenarios. We further apply TRAMA to data from the Human Pangenome Reference Consortium - Phase 2. This work offers a path toward modeling STR mutational processes more accurately.

Modeling local ancestry covariance to infer the timing of Denisovan admixture events

Sarah Johnson

University of California, Berkeley

Gene flow from our extinct hominin cousins, Denisovans, has shaped the landscape of the modern human genome and contributed to adaptive variation in present-day populations. Yet our understanding of the timing and geographic context of Denisovan admixture remains poorly resolved. Existing approaches for dating admixture events typically rely either on linkage disequilibrium (LD) among a limited set of ascertained archaic-informative markers or on local ancestry inference, which has reduced power to detect short introgressed fragments. These limitations are particularly acute for Denisovan ancestry, which is rare (˜0.1–4%) in many populations and unevenly distributed across the globe. Here we introduce a composite method that uses LD in local ancestry across the genome. Our method models the decay of covariance in local ancestry across genomic distances, leveraging the recombination clock. By using genome-wide local ancestry information rather than sparse Denisovan informative positions, our method accesses orders of magnitude more data with an effective gain in power of nearly 1,000-fold relative to SNP-based methods. Additionally, by computing pairwise covariance across multiple individuals rather than directly modelling inferred archaic fragments, our method robustly captures Denisovan ancestry decay without relying on precise local ancestry inference in any single individual. We perform simulations under a range of demographic scenarios and demonstrate that our method works reliably to date both Neanderthal and Denisovan admixture, even in cases of low admixture proportions ($< 0.5\%$). We apply this method to a dataset of over 4,700 published whole genome sequences from across Asia and Oceania to infer the timing and structure of Denisovan admixture events in modern humans. Together, our study provides a powerful new method for dating admixture, and sheds light into the complex history of modern human-Denisovan interactions and human dispersal following the

Out-of-Africa migration.

Neanderthal ancestry in East Asia

Sophie K. Joseph

University of California, Berkeley

Gene flow from Neanderthals and Denisovans has played a critical role in shaping variation in human populations. It is associated with a range of functional effects in Eurasians, including susceptibility to chronic and infectious diseases (e.g. SARS-CoV-2, Type II diabetes). In particular, East Asians have a complex history of admixture with Neanderthals and Denisovans, reflected in a unique set of population characteristics not seen outside of East Asians. For example, East Asians harbor around 20% higher Neanderthal ancestry compared to other global populations, and at least two divergent pulses of gene flow from Denisovans. Thus, East Asians are a critical missing piece in understanding the evolutionary underpinnings of present-day human variation and health globally. Using over 27,000 newly generated whole genome sequences from present-day East asian individuals, as well as ancient genomes from East Asia ranging in age from 40,000–1,500 years before present, this work investigates the complex nature of Neanderthal introgression into East Asians.

Biobank-scale visualization and interactive exploration of ancestral recombination graphs with LORAX

Pratik Katte

University of California, Santa Cruz

Ancestral Recombination Graphs (ARGs) provide a complete representation of genetic ancestry within a population and form the basis of modern demographic inference, selection scans, disease association, and large-scale genomic data sharing. As genome sequencing has scaled to biobank-sized cohorts, ARG-based methods have become increasingly powerful, yet their complexity and size have limited practical exploration and interpretation.

We present Lorax, a GPU-accelerated tool for real-time, interactive visualization of biobank-scale ARGs. Lorax unifies genome-wide navigation, temporal structure, local tree topology, and metadata-aware filtering within a single interactive environment, enabling users to explore recombination-aware ancestry across the genome at population scale. By making complex genealogical structure directly observable, Lorax lowers the barrier to ARG-based analysis and supports more intuitive reasoning about population history and disease-relevant variation.

Inferring effect-size and selection coupling in admixed populations using machine learning models

Michelle Kim

University of Michigan

The relationship between allelic effect sizes and negative selection plays a central role in shaping the genetic architecture of complex traits, yet it is not directly observable in empirical association data. Numerous evolutionary models have been proposed to describe this coupling, including frameworks linking effect sizes to fitness consequences through selection. In this study, we focus on the Eyre-Walker model, which parameterizes the strength of coupling via the τ parameter, where larger τ values correspond to stronger correlations between allelic effect sizes and deleterious fitness effects.

Here, we develop a simulation-trained machine learning framework that treats τ as a latent evolutionary parameter and infers it from GWAS-like summary statistics. Using forward simulations under realistic demographic histories, including multiple ancestral populations and recently admixed populations, we generate complex traits spanning a wide range of τ values. From these simulations, we derive trait-level genomic summaries, including LD decay profiles, allele frequency spectra, and effect-size distributions, and use them to train supervised machine-learning models for both categorical and continuous prediction of τ.

We evaluate model performance under stringent generalization settings, including holding out entire demographic models, and demonstrate that τ induces robust and distinguishable signatures in GWAS-like summary statistics data across diverse population histories, even in the presence of admixture-driven LD complexity. Notably, our framework classifies effect-size and selection coupling regimes across diverse demographic histories, including recently admixed populations. This inference task has received limited attention in previous studies. Together, these results establish machine learning-based inference as

a powerful and scalable approach for recovering evolutionary constraints on complex traits from association data, enabling comparative analyses of genetic architecture across populations.

The genomic consequences of near extinction and recovery in ‘Alalā, the Hawaiian crow

Chris Kyriazis

San Diego Zoo Wildlife Alliance

Hawaiian birds are among the most imperiled groups of animals on earth, with ~75% of native species already thought to be extinct. ’Alalā, the endemic Hawaiian crow, narrowly avoided extinction when a captive breeding population was founded from just nine individuals, which has since recovered to a population size of ~120, almost all of which remain in captivity. To investigate the genomic consequences of this severe population bottleneck, we generated a chromosome-level reference genome assembly and resequenced dataset of 174 individuals. Consistent with the severity of this bottleneck, we observe a high abundance of runs of homozygosity in ’alalā, with a mean FROH=0.32. Surprisingly, we do not detect associations between FROH and numerous measures of survival and reproduction, suggesting that many aspects of inbreeding depression have been purged in ’alalā. However, we identify two putative recessive lethal haplotypes that substantially contribute to elevated rates of egg hatching failure. Using a genomics-informed simulation model, we explore numerous future reintroduction scenarios, highlighting pathways towards a robust recovery of the wild population. Our findings emphasize the value of genomic variation datasets and computational simulation approaches for understanding the fitness consequences of near-extinction events.

Protein language models reveal evolutionary constraints on synonymous codon choice

Liana Lareau

University of California, Berkeley

Evolution has shaped the genetic code, with subtle pressures leading to preferences for some synonymous codons over others. Codons are translated at different speeds by the ribosome, imposing constraints on codon choice related to the process of translation. The structure and function of a protein may impose pressure to translate the associated mRNA at a particular speed in order to enable proper protein production, but the molecular basis and scope of these evolutionary constraints have remained elusive. Here, we show that information about codon constraints can be extracted from protein sequence alone. We leverage a protein language model to predict codon choice from amino acid sequence, combining implicit information about position and protein structure to learn subtle but generalizable constraints on codon choice in yeast. In parallel, we conduct a genome-wide screen of thousands of synonymous codon substitutions in endogenous loci in yeast, reliably identifying a small set of several hundred synonymous variants that increase or decrease fitness while showing that most positions have no measurable effect on growth. Our results suggest that cotranslational localization and translational accuracy, more than cotranslational protein folding, are major drivers of selective pressure on codon choice in eukaryotes. By considering both the small but wide-reaching effects of codon choice that can be learned from evolution and the strong but highly specific effects determined via experiment, we expose unappreciated biological constraints on codon choice.

Adaptations spreading across human gut microbiomes arise from complex multisite adaptive architectures

Peter Laurin

University of California, Los Angeles

Adaptation is pervasive in human gut microbiomes, giving rise to fundamental traits like the ability to digest foods and metabolize drugs. Given large mutational input per gut microbiome, adaptation to selective pressures should be rapid and repeatable, resulting in independent origins of adaptations in separate hosts. Despite this expectation, recent work has shown that broadly-beneficial adaptations can spread globally across human gut microbiomes via migration and subsequent recombination onto novel strain backgrounds rather than mutating de novo within each host. In this work, we test whether adaptations that have spread across hosts have multiple origins arising from rapid mutation. We find, surprisingly, that many adaptations have only one or a few origins. Using a stochastic model of an across-host selective sweep, we show that adaptations attaining detectable frequencies require adaptive mutation rates orders-of-magnitude lower than the mutation rate at a single base pair, and, consequently, that adaptations spreading across hosts bear structural or epistatic variants more complex than a single site mutation. In summary, we demonstrate that recombination across human gut microbiomes permits the spread of widely-beneficial adaptations with complex genetic architectures that otherwise would require a long waiting time to generate within an individual host.

Mapping disease critical spatially variable gene programs by integrating spatial transcriptomics with human genetics

Hanbyul Lee

Memorial Sloan Kettering Cancer Center

Spatial gene expression patterns underlie tissue organization, development, and disease, yet current methods for detecting spatially variable genes (SVGs) lack the flexibility to capture multi-scale structure, ensure robustness across platforms, and integrate with genetic data to assess disease relevance. We present Spacelink, a unified framework that models spatial variability of a gene at both whole-tissue and cell-type resolution using an adaptive mixture of data-driven spatial kernels and summarizes it using an Effective Spatial Variability (ESV) metric. Spacelink achieved up to 3.2x higher detection power over eight existing global SVG and cell-type SVG methods while showing consistently superior FDR control across 34 different simulation settings and also showed superior cross-platform concordance in matched tissue Visium and CosMx datasets. Applied to 3 healthy CosMx human tissues (brain cortex, lymph node, liver), Spacelink revealed that SVGs are highly informative for 113 complex traits and diseases (average N = 340,406). Spacelink showed up to 2.2x higher disease informativeness over competing methods in tissue-relevant complex diseases and traits, conditional on putative non-spatial expression-level confounders. Applied to a mouse organogenesis Stereo-seq atlas (8 developmental stages), Spacelink identified 145 genes with stage-associated ESV within brain independent of mean expression, that are enriched in pathways like Wnt signaling and Rap1 signaling characterizing early and late development, respectively. Integration with in vivo Perturb-seq targeting 35 de novo ASD risk genes revealed that perturbations in excitatory neurons and astrocytes preferentially altered spatially structured downstream gene programs (1.7–2.2x higher average ESV across stages than other cell types), many of which were enriched for polygenic autism GWAS loci. In neurodegeneration, analysis of 32 Visium dorsolateral prefrontal cortex samples spanning Alzheimer's disease (AD) pathology stages identified 334 genes with decreasing ESV along amyloid burden (enriched for glycolysis) and 216 genes with decreasing ESV along tau tangle accumulation (enriched for apoptotic pathways). Several AD risk genes (PKM, CLU, GPI) showed conserved reductions in spatial variability with AD pathology in both human and 5xFAD mouse, with PKM linking to a colocalized splicing QTL and amyloid burden QTL variant.

ratePlacer: A phylogenetic framework for molecular dating of ancient environmental DNA

Maya Lemmon-Kishi

University of California Berkeley

Ancient environmental DNA (aeDNA) from permafrost, lake and marine sediments provides a rich source of genetic data that captures broad perspectives of past biodiversity. Accurate dating is crucial for discovering ecologically relevant patterns from aeDNA, and increasingly samples are too old for C-14 dating. While molecular dating allows for sample ages to be estimated from the recovered genetic material itself, the fragmented and damaged nature of short-read ancient DNA from multiple taxonomic sources poses significant challenges. We have developed ratePlacer, a phylogeny-based method for analyzing aeDNA that can combine information from many short reads in a sample while accounting for DNA damage to provide maximum likelihood estimates of sample ages. Simulations demonstrate that ratePlacer accurately dates samples even under the fragmented, damaged conditions characteristic of aeDNA and outperforms tip dating methods like BEAST for taxonomically mixed samples commonly found in aeDNA. However, applying ratePlacer to exceptionally old sediment samples revealed previously uncharacterized substitution patterns that substantially complicate molecular dating. These patterns deviate from typical ancient deamination damage, suggesting that aeDNA should be carefully evaluated in future genomic and evolutionary analyses.

Estimating effective population size from ARGs: An EM-based approach incorporating coalescent time uncertainty

Jacky Kaiyuan Li

UC Berkeley

New methods for estimating Ancestral Recombination Graphs (ARGs) are promising to transform population genetics, and the inference of demographic parameters, by providing access to full likelihood functions for large whole-genome datasets. However, fully harnessing this potential requires the development of statistical methods that can effectively utilize ARGs while addressing the inherent uncertainty in coalescent time estimation- a critical factor for accurate population size inference. Simulation studies reveal that existing methods often are biased due to the misplacement of coalescence events in adjacent time intervals. To overcome these limitations, we introduce a novel method for inferring effective population sizes by integrating ARGs with an Expectation-Maximization (EM) algorithm, which explicitly incorporates coalescent time uncertainty into the estimation framework. Using msprime to simulate large sequencing datasets and SINGER and POLEGON to infer ARGs, we demonstrate that our new method provides accurate estimates of effective population size over time, effectively correcting biases observed in existing methods. This approach exhibits superior statistical properties in terms of Mean Square Error in the estimation of effective population size, establishing the method as a powerful and reliable tool.

A fast general framework for computing functionals of coalescent rates for complex demographies

Jiatong Liang

University of Michigan

We introduce demestats, a software library for computing generalized instantaneous coalescent rate (ICR) curves for arbitrary demes-formatted demographic models. While pairwise ICR curves are widely used for inferring population history, features such as recent growth and migration are poorly resolved due to the scarcity of very recent coalescent events. demestats improves resolution by computing the rate of first coalescence among any number of sampled lineages drawn from multiple demes in a structured population model. It also computes cross-coalescence curves–the rate at which lineages sampled from different demes first share a common ancestor–enabling inference of very recent migration. The library is implemented in a differentiable programming framework and integrates naturally into likelihood-based inference pipelines. Benchmarks on standard simulation models show accurate recovery of present-day effective population size, the rate of recent exponential growth, and migration history within the past ten generations. Finally, we apply demestats to whole-genome human data and show that it recovers fine-scale recent population structure consistent with prior genetic studies.

The evolution of structural and single nucleotide mutation across haplotype-resolved vertebrate genome assemblies

Runyang Nicolas Lou

University of California, Berkeley

Structural variants (SVs) contribute substantially to genetic variation and play vital roles in adaptation and disease. However, SVs are poorly captured by short read sequencing and thus are understudied, particularly in non-model organisms. Here, taking advantage of recently generated haplotype-resolved genome assemblies from >600 vertebrate species, we present the most comprehensive survey of the diversity of SVs and single nucleotide variants (SNVs) across the vertebrate tree of life to date. By identifying SVs and SNVs that segregate across two representative haplotypes in each genome assembly, we confirm patterns of reduced diversity of both SNVs and SNVs in endangered or threatened taxa. However, we find that while SNV and SV diversity are correlated across species, the proportion of these two forms of genetic diversity is fundamentally distinct across taxa, with fish and amphibians harboring 3 to 6-fold more segregating SVs than amniotes given the same number of segregating SNVs. We show that recent transposable elements (TE) activity is a significant source of SVs across vertebrates, with particularly rapid turnover observed in several mammalian lineages and higher diversity in TE composition in fish, amphibians, and reptiles. Using machine learning models we identify genomic features predictive of structural mutations across taxa with the top features broadly conserved across species, reflecting common bases underlying genomic instability in vertebrates. Lastly, we demonstrate that SVs are more likely to alter protein coding sequences than SNVs. Most of these variants are likely deleterious, and species harboring less genetic diversity tend to have a higher proportion of putatively deleterious variants. However, several genes, many of which are involved in olfactory and immune systems, are repeatedly impacted by SVs in multiple species, hinting at the adaptive roles SVs can play in evolution. Together, this study characterizes the diversity of SNVs and SVs

across the vertebrate tree of life and highlights that patterns of segregating genetic variation are distinct across taxa with broad implications for vertebrate genome evolution, selection, and biodiversity conservation.

The site-frequency spectrum under selection and time-varying demography

Anastasia Lyulina

Stanford University

Demographic history and natural selection both influence the site-frequency spectrum of new mutations, but their joint effects remain difficult to disentangle – especially in populations far from equilibrium. In this work, we derive an analytical expression for the frequency spectrum of rare alleles for arbitrary time-dependent population sizes and deleterious fitness effects. This forward-time approach allows us to trace the trajectories of mutations contributing to different parts of the spectrum. We find that rapid population growth can produce an abundance of older deleterious variants at intermediate frequencies relative to the neutral expectation, resulting in a non-monotonic ratio of nonsynonymous-to-synonymous mutations. By applying these results to recent demographic reconstructions of European human history, we show that these nonequilibrium effects are likely to play an important role in shaping the observed distribution of deleterious variants. These results provide a new theoretical framework for interpreting the site-frequency spectrum in populations with complex demographic histories, and highlight scenarios where selection and demography interact in non-intuitive ways.

Enhancement of hidden Markov model analyses for improved inference of archaic introgression in modern humans

Moisès Coll Macià

Institut de Biologia Evolutiva (IBE), Barcelona

Neanderthal and Denisova introgressed fragments in present-day human genomes are commonly inferred using Hidden Markov Models (HMMs), such as hmmix. However, existing HMM decoding methods have important limitations. Viterbi decoding searches for a globally optimal path which tends to underestimate transitions between human and archaic ancestry tracks, whereas Posterior decoding focuses on local signals and tends to overestimate transitions. Furthermore, both approaches yield deterministic outputs, failing to capture uncertainty in summary statistics of admixture estimates.

Here, we introduce two methodological enhancements to HMM-based inference and implement them in hmmix, applying the framework to simulated data and 1000 Genomes Project genomes. First, we implement a sampling method that generates hidden state sequences conditional on the observed data, enabling robust estimation of the full distribution of admixture summaries such as introgressed proportion and fragment length. We further show that the same results can be obtained analytically by re-implementing the finite Markov chain imbedding framework. Second, we apply a hybrid decoding method that interpolates between Viterbi and Posterior decoding, combining global and local information to achieve an improved sensitivity–specificity trade-off in the detection of archaic fragments.

We demonstrate that these two frameworks substantially improve the inference of archaic introgression and are broadly applicable to HMM-based methods in general. Together, they increase decoding accuracy and provide principled uncertainty estimates for summaries derived from genomic data.

Structured modeling improves estimation of allelic effects in gene expression

William H. Majoros

Duke University

Abnormalities in gene expression can serve as important diagnostic signals in cases of undiagnosed genetic disease. Of particular relevance to the search for gene regulatory defects is allelic imbalance in expression. We describe a family of probabilistic graphical models for dissecting allelic signals in both endogenous genes and massively parallel reporter assays. For endogenous genes, we show that detailed modeling of haplotype structure can improve estimation accuracy when short read data are used, and in the context of pedigrees we show that joint modeling of the sharing of haplotypes and expression imbalance between individuals leads to improved identification of inheritance patterns of genetic causes of imbalance, even when causal factors are unobserved. In the case of pooled, multi-sample reporter assays, we show that imposing structure on the sample pools leads to higher estimation accuracy, due to heterogeneity in allele frequencies and an induced Poisson-binomial structure in the data-generating process. We also identify a number of important future directions for both of these applications.

M. tuberculosis and gastrointestinal pathogens drove immune adaptations in ancient West Eurasians

Javier Maravall-López

Harvard University

Insights into human biology and disease can be obtained by studying genetic variants affected by directional natural selection. Ancient DNA now enables direct study of human adaptation over the last few thousand years, and recent work (Akbari et al. 2024) has identified hundreds of loci with genome-wide evidence of positive selection in West Eurasia. However, the underlying adaptive phenotypes and selective pressures particularly those imposed by rising infectious burden in this time transect, remain poorly understood. Here, we characterize the immune landscape of recent human evolution by systematically integrating the Akbari et al. selection statistics with diverse genome-wide association (GWAS), quantitative trait locus (QTL), and molecular and gene pathway data, to identify variants, genes, phenotypes and microbes targeted by selection.

Genome-wide, we find that positively selected alleles are associated with reduced susceptibility to many infections, including those caused by respiratory and gastrointestinal pathogens (meta-analyzed genetic correlation with selection across 12 approx. independent infection traits: -0.05 (SE 0.01), P<4e-5). We resolve several novel adaptive loci by colocalizing convergent signals from selection, infectious disease GWAS and immune-gene QTLs to likely causal variants (e.g., a SNP upstream of FUT6, fine-mapped with 98% confidence for a GWAS of intestinal infections and with >99% confidence for expression of several key gut immunity genes, including MUC2, GP2 and PIGR). Adaptive loci colocalize with QTLs for genes strongly enriched for host-defense pathways (P<e-20), and the direction of QTL effects suggests that selection has favored alleles that increase activation of these pathways (e.g., “foreign body rejection”, FDR<1.5e-10).

Tissue- and cell type-specific analyses reveal enrichments in immune cells within barrier tissues that interface with pathogens, such as the respiratory tract and gut mucosa. Multiple lines of evidence point to selection driven by Mycobacterium tuberculosis, widely regarded as one of the deadliest infectious agents in human history. Consistent with a long-standing hypothesis of antagonistic pleiotropy, we find that adaptive alleles at loci related to tuberculosis susceptibility are risk alleles for autoimmune inflammatory bowel disease (genetic correlation = 0.66 (SE 0.19), $P<2e\text{-}4$), highlighting a possible evolutionary trade-off between host defense and immune-mediated pathology.

Sex-specific associations between polygenic risk scores and age-specific risk of psychiatric disorders

Genona T. Maseras

Aarhus University

Polygenic scores (PGS) are increasingly used to quantify genetic liability to psychiatric disorders, yet they are typically constructed and interpreted under assumptions of time-invariant effects and homogeneous risk expression. Whether these assumptions hold in population-based settings, where risk unfolds across development and differs by sex, remains unclear. Using a population-based case–cohort design within the Danish iPSYCH2015 cohort, we examined age- and sex-specific associations between PGS and risk of psychiatric disorders.

The study included individuals born between 1981 and 2008 and followed for psychiatric diagnoses from 1994 to 2015 (N = 119,941). Disorder-specific PGS were derived from large genome-wide association studies, and sex- and age-specific hazard ratios (HRs) for first-time hospital-based diagnoses were estimated across predefined age windows. Across disorders, PGS were associated with increased diagnostic risk; however, the magnitude and timing of associations varied substantially by age and sex. For schizophrenia (SCZ), the strongest association between SCZ-PGS and diagnosis in males was observed at ages 20–23 (HR per one–standard deviation increase, 1.71; 95% CI, 1.54–1.90), whereas in females peak associations occurred earlier, at ages 17–20 (HR, 1.29; 95% CI, 1.16–1.44). Across age windows, SCZ-PGS associations were consistently stronger in males than in females, and these differences were not fully explained by sex-specific incidence patterns.

These findings demonstrate pronounced temporal and sex-specific heterogeneity in PGS associations, indicating that standard, time-invariant PGS models may be poorly aligned with how genetic liability manifests in real population settings. Incorporating developmental timing and sex-specific structure may therefore be necessary for improving the interpretability and applicability of

PGS in epidemiological and predictive contexts.

Demographic and familial information dominates psychiatric risk prediction, while polygenic risk shows age-dependent associations

Genona T. Maseras

Aarhus University

Polygenic risk scores (PGS) are widely proposed as tools for stratifying risk of psychiatric disorders, yet their incremental value relative to routinely collected non-genetic information remains low. We assessed the relative contributions of demographic characteristics, socioeconomic indicators, family psychiatric history, and PGS to psychiatric disorder prediction across the life course.

Using longitudinal registry data, we fitted cross-validated lasso regression models to predict psychiatric disorder onset across multiple age-defined prediction windows. Predictors were structured into information blocks, allowing assessment of incremental predictive performance and window-specific associations under consistent regularization.

Across all prediction windows, individual demographic characteristics were associated with psychiatric risk in a stable manner across the life course. In contrast, both family psychiatric history and polygenic risk showed age-dependent associations, contributing differentially across windows. Demographic information accounted for most of the predictive variation consistently over time, whereas the incremental contribution of PGS was limited overall but strongest in earlier windows and attenuated later in life, consistent with temporal heterogeneity in genetic liability.

These results suggest that, in penalized high-dimensional models, demographic and familial information dominates short-term psychiatric risk prediction, while PGS captures time-dependent genetic susceptibility that may be more informative for etiological stratification than for near-term prediction.

Theoretical bounds and empirical diagnostics for confounding control in polygenic prediction

Walid Mawass

University of Chicago

Principal component analysis (PCA) is the standard for correcting population stratification in GWAS, yet residual confounding often persists, biasing polygenic scores and masking signals of adaptation. We combine a theoretical analysis of PCA performance with a diagnostic framework to quantify the resulting risks to genetic prediction and evolutionary inference.

We first establish that residual bias arises from the imperfect estimation of ancestry-informative eigenvectors in finite samples. This estimation error is most severe in a transition regime where population structure is statistically detectable but too weakly captured to be precisely removed. Consequently, residual bias peaks near the detection threshold rather than decreasing monotonically with signal strength, demonstrating that a component's statistical significance is a poor proxy for its adequacy as a covariate. We further establish that systematic error in meta-analysis is constrained by the ancestry-estimation accuracy of its constituent cohorts. Because meta-analysis relies on cohort-specific PCA, it preserves these local residuals, which then fail to attenuate even as the aggregate sample size grows.

Complementing this, we develop a framework to quantify the susceptibility of PGS-ancestry association tests to stratification. We define this susceptibility as the proportion of genetic variance in a GWAS panel explained by an external ancestry gradient, which dictates the rate of spurious correlation in downstream tests. Applying this to the UK Biobank, we find that while PCA effectively flattens the susceptibility disparity between diverse and homogeneous panels, residual structure often remains near the theoretical limit of detection. Because even undetectable structure can bias highly polygenic scores, we introduce a diagnostic to calculate the critical magnitude of environmental confounding required to explain an observed signal. Using this approach, we find that polygenic divergence in height and blood pressure within Europe appears relatively robust to residual stratification. Together, these results define the theoretical bounds of confounding control and provide essential diagnostics for interpreting signals of polygenic adaptation in large-scale genetic studies.

An SFS based method for detecting seasonal balancing selection

Giovanni Mazzeo

University of Virginia

Temporally fluctuating conditions, like seasonal changes in temperature, influence the evolution of many species, yet we lack precise methods for distinguishing the footprint of fluctuating selection from that of overdominance, and previous work has shown many common summary statistics confuse the two. While time-series data could be used to make the distinction, informative data of that type can be hard to acquire and would not provide evidence towards the long-term maintenance of polymorphism. Recent theoretical work by Wittman et al. (2023) has shown that fluctuating selection has a distinct signature when the allele frequency trajectory is cyclic, which can be detected at a single time point. This assumption is unproblematic, as any long-term pattern of fluctuating selection (in the absence of other forces) should have roughly cyclic trajectories. Another benefit of Wittman's approach is that it is parametrized only by the arithmetic and harmonic means of the allele frequency trajectory, avoiding direct modelling of changing selection and dominance. These parameters capture the midpoint and oscillation strength of the trajectory, respectively. However, in their model, this difference is only apparent at large genetic distances, and it is unlikely in natural populations that other forces will not come into play far away from a focal locus. We extend the model presented by Wittman et al. into a method of inference that can be used at more practical genetic distances, by deriving the normalized expected site frequency spectrum from it, using the block counting state space of Holboth et al. As the state space of the Markov chain grows superpolynomially (but subexponentially) with the number of lineages sampled, performing inference on the full sample can be implausible. To counteract this, we use a subsampling approach to compute the normalized expected site frequency spectrum of a fixed sample size from the empirical data. We then use multinomial composite likelihood to fit models

of seasonal balancing selection, overdominance, and neutral evolution to the observed data. We can compare models by using the Composite Likelihood Information Criterion (Varin et al, 2011). We test our model on synthetic data produced by the population genetic simulation engine SLiM (Haller & Messer 2023). We hypothesize that our method will allow for statistically significant classifications to be made and circumvent the issues discussed above.

A neural network investigation of evidence for ghost introgression in Denisovans

Noel McAllister

University of Michigan

Denisovans, close relatives of Neanderthals, were identified through the nuclear genome of a finger bone found in Denisova Cave in the Altai Mountains. Sequencing of their mitochondrial genome has revealed lineages deeply diverged from those of both modern humans and Neanderthals, hinting at possible introgression from an unknown super-archaic hominin group, known as ghost introgression. Previous works have attempted to detect ghost introgression by leveraging summaries of allele frequency data such as the site frequency spectrum (SFS). While informative, constraining to only statistics derived from the SFS results in large data compression that risks losing important evidence for or against ghost introgression. Neural networks provide a flexible framework for detection of ghost introgression with available Denisovan sequence, allowing us to utilize SFS derived statistics in combination with other data summaries such as statistics on the distribution of pairwise sequence differences within and between study individuals which give information on the time to the most recent common ancestor. This allows us to more finely tailor our model to the detection of deep lineages, older than the Neanderthal-Denisovan common ancestor, that we expect if ghost introgression occurred. Detection of ghost introgression in Denisovans is complicated by our only having access to two full genomes from the same geographic area, limiting our power to identify ghost introgressed sequence. To offset this data limitation, we train our model to identify the presence of ghost introgressed tracts in a combination of two Denisovan haplotypes and modern human haplotypes with possible Denisovan introgression over windows of fixed size. To train our model we simulated 50 Kb windows over our sample set. Two sets of statistics were then calculated over the simulated sequence for each common interval, one set calculated over only the human haplotypes and another over the combined human and Deniso-

van sequence set. We trained our network over 1.25 million windows, half with at least one ghost tract and half without any. We apply our network to a sample consisting of seven native Oceanians and the high coverage Denisovan genome. Our work explores the practicability of using machine learning tools to investigate complicated demographic histories with large data limitations.

MeQTL discovery in admixed human genomes to estimate epistasis

Gillian Meeks

University

The extent to which epistatic interactions contribute to ancestry-specific genetic effect sizes—and thereby limit the transferability of genetic prediction—remains unresolved. Progress has been limited by two major challenges: comparisons across continental populations confound epistasis with gene-by-environment effects, and prior estimates of ancestry-specific effect correlations have relied on a small number of traits, limiting statistical precision. We address these limitations using genetic data from an admixed South African population ($n = 366$), which enables direct estimation of epistatic interactions by comparing genetic effects of the same variants carried on different local chromosomal ancestries within a shared environmental context. This population includes Khoe-San, non–Khoe-San African, and out-of-Africa ancestries, allowing estimation of three pairwise ancestry-specific genetic effect correlations spanning a range of genetic divergence. We integrate newly generated genome-wide CpG methylation data with paired DNA sequence data, treating methylation at individual CpG sites as molecular phenotypes. This yields hundreds of thousands of phenotypes, providing substantial power to estimate the genome-wide correlation of genetic effect sizes across ancestries. Using a simulation framework that incorporates ancestry-matched genomes and a cis-methylation–relevant genetic architecture, we demonstrate accurate recovery of simulated ancestry effect correlations. When meta-analyzing 50 CpGs per chromosome, the mean absolute difference between simulated and estimated correlations is only 6% across the full range of simulated values. At individual CpG sites, we achieve 62%, 55%, 47%, and 36% power to reject the null hypothesis of equal ancestry effect sizes for simulated correlations of 0, 0.25, 0.5, and 0.75, respectively, while maintaining a false positive rate of 3.7%. Finally, applying CARMA-X, a fine-mapping method designed for admixed populations, we achieve 98% power to recover

the true causal variant within credible sets across all simulated ancestry effect correlations.Together, these approaches enable robust estimation of genome-wide ancestry effect correlations, identification of loci with ancestry-specific deviations in genetic effects, and accurate fine-mapping of causal variants with heterogeneous effect sizes.

Kernel robust bayesian clustering in high dimensions

Jingcheng Meng

Duke University

Single-cell RNA sequencing analysis (scRNA-seq) derives cell types from the expression profiles of individual cells across a massive number of genes. After pre-processing, the standard workflow is to project the expression profiles onto a low-dimensional space, then cluster the projections. Generally, these approaches lacks the ability to express uncertainty in cluster assignment, motivating probabilistic high-dimensional clustering. However, computational expense, nonidentifiability in the latent space, and misspecification of the true clustering model are major practical issues for proabilistic methods. In this article, we propose a holistic, flexible, and efficient Bayesian framework for high-dimensional clustering. In contrast to standard models, we assume that low-dimensional factors are generated from an unknown latent density, which we approximate with a Dirichlet Process Gaussian mixture model (DPGMM). To ensure identifiability and robustness to misspecification, we propose a pretraining approach and use a component-merging loss to estimate clusters. We provide a thorough theoretical analysis of our framework, including posterior contraction of the latent subspace, consistency in estimating the latent density, and convergence of the clustering risk function to an oracle. Finally, we validate our method through simulations and an scRNA-seq application.

Buffering of gene dosage response curves for human complex traits

Nikhil Milind

Stanford University

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 for individual genes reducing expression should have the opposite effect on a phenotype as increasing expression. To understand this paradox, we use the gene dosage response curve (GDRC), which relates changes in gene expression to expected changes in phenotype. We show that, for many traits, GDRCs are systematically biased in one trait direction relative to the other, a phenomenon we call trait buffering. A consequence of trait buffering is that traits are more easily modified in one direction than the other by genetic variation. We develop a simple theoretical model that explains this bias in trait direction. Our results have broad implications for complex traits, drug discovery, and statistical genetics.

Direct species tree inference from whole-genome alignments

Siavash Mirarab

UCSD

Genomes contain mosaics of discordant evolutionary histories, challenging the accurate inference of the tree of life. While genome-wide data are routinely used for discordance-aware phylogenomic analyses, due to modeling and scalability limitations, the current practice leaves out large chunks of genomes. As more high-quality genomes become available, we urgently need discordance-aware methods to infer the tree directly from a multiple genome alignment. Here, we introduce CASTER, a theoretically justified site-based method that eliminates the need to predefine recombination-free loci. CASTER is scalable to hundreds of mammalian whole genomes. We demonstrate the accuracy and scalability of CASTER in simulations that include recombination and apply CASTER to several biological datasets, showing that its per-site scores can reveal both biological and artefactual patterns of discordance across the genome.

A general FST framework reveals the variability of rare versus common alleles

Maike Morrison

Santa Fe Institute

Genetic variation across populations reflects evolutionary phenomena such as population structure, migration, and local adaptation. However, summarizing the shared variation of populations is challenging and question dependent. The statistic FST is a standard measure of such variation, but its reliance on heterozygosity (which squares every allele frequency) can obscure evolutionary phenomena whose signatures lie in patterns of rare genetic variation, such as recent population structure and gene flow. Inspired by the Hill numbers, a family of ecological diversity statistics, we present a generalized FST framework, FST-alpha, with tunable emphasis on rare versus common genetic variation. FST-alpha replaces heterozygosity with the Tsallis entropies, which use the parameter alpha to synthesize diversity measures such as richness (alpha=0; maximum emphasis on rare alleles), Shannon information (alpha approaches 1), heterozygosity (alpha=2), and fixation (alpha approaches infinity; maximum emphasis on common alleles). When alpha equals 2, FST-2 simplifies to standard, heterozygosity-based FST. We demonstrate that analyses using a range of alpha values can reveal important patterns of genetic variation that are overlooked when only standard FST (FST-2) is considered. In application to a global human genetic dataset, we find that the well-established negative relationship between genetic diversity and FST-2 reverses when more emphasis is placed on rare variation. We also identify scenarios where FST-2 values are nearly identical, but FST-alpha values for alpha not equal to 2 are very different. For example, when comparing a group of high-diversity populations and a group of low-diversity populations with similar FST-2 values, FST-0 and FST-1 captured substantial variation of rare alleles present only in the high-diversity population. Finally, simulations demonstrate that FST-alpha is better powered than FST-2 alone for detecting phenomena such as recent population

structure. This framework expands the toolkit of FST-like statistics, enabling future studies to utilize the full range of genetic variation.

User-friendly, general inference on the coalescent and ARG using efficient computation of exact PDFs, moments, and joint probabilities

Kasper Munch

Aarhus University

Formulating generative models in the coalescent and ancestral recombination graph (ARG) is conceptually straightforward. However, inference on such models - computing distributions, moments, and joint probabilities - often proves computationally demanding or infeasible in practice. I present Phasic, a framework that makes exact inference on coalescent and ARG models both tractable and accessible. Phasic applies to any finite-state Markov model and represents the waiting time phase-type distributions as directed graphs and employs efficient graph algorithms to compute statistics exactly. The computational bottleneck—solving the state transition matrix—is computed by Gaussian elimination on the model graph. Symbolic recording and caching allow re-evaluation with different parameter values by retracing computations in linear time, enabling efficient parameter estimation and confidence interval construction via Stein variational gradient descent. Built-in support for sharing symbolic recordings in a research community means each model needs to be computed only once. I demonstrate its flexibility, and usability in two applications: (1) Inference of ghost populations from genealogical data, computing exact joint probabilities of SNPs at arbitrary genetic distances in the two-locus ARG, and (2) inference of positive selection coefficients using exact likelihoods under the coalescent with selection.

Towards an unbiased characterization of genetic polymorphism

Magnus Nordborg

Gregor Mendel Institute

Our current view of genetic polymorphism is extremely biased because it is based on aligning short sequences to a reference genome. This only works in well-behaved regions of the genome, and only for simple polymorphisms. It is also leads to bias w.r.t. the reference genome. The solution is obviously switching towards comparing independently assembled genomes. Thanks to advances in technology, generating such genomes is now a solved problem, but making sense of them is not, as appropriate alignment and interpretation is a difficult inference problem. I will present progress we have made on this problem, focusing in particular on 500 almost complete Arabidopsis thaliana genomes. I will cover what we have already about the unbiased polymorphism, focusing on particular on what we are learning about the epidemiology of actually mobile transposable elements, which are finally made visible by these methods.

Uncovering non-identifiable demographic histories using generative AI

Ekaterina Noskova

University of Edinburgh

Reconstructing demographic history from genetic data provides key insights into population size changes, splits, and migrations. However, demographic inference often faces the problem of model non-identifiability, where distinct histories generate similar genetic patterns. Generative AI is well suited for uncovering complex structures and subtle patterns in high-dimensional data, making it a promising approach to investigate this problem. I will present how I plan to use generative AI models to explore the space of indistinguishable demographic scenarios and how it can provide new perspectives on demographic inference.

An elegant linkage disequilibrium model for structured populations applied to polygenic risk scores for linear mixed-effects model summary statistics

Alejandro Ochoa

Duke University

Polygenic risk scores (PRS) are widely used to predict disease risk, but most applications exclude minor ancestries and admixed individuals. In particular, most PRS methods employing summary statistics assume homogeneous populations, and existing multi-ancestry methods (such as PRS-CSx and PROSPER) require multiple homogeneous ancestries analyzed separately. We propose to develop a PRS framework for cohorts with arbitrary relatedness analyzed jointly for both association and linkage disequilibrium (LD). We identified two important gaps: (1) most researchers estimate LD using the Pearson correlation, which is poorly suited for structured populations; and (2) although associations with linear mixed-effects models (LMMs) are already successfully model complex cohorts, existing PRS models are not designed for LMM data. Here, we first introduce LDkin, an LD estimator based on a new "LD-kinship" model of LD cross-covariance that explicitly models population kinship. We then develop a corresponding summary statistic model for LMMs under our LD-kinship model, which reveals a model analogous to that of linear models under no structure but with generalized quantities, including a new effective sample size formula. Thus, this framework suggests a drop-in replacement for existing summary statistic models for single populations, enabling LDpred2 and related approaches to operate correctly in structured populations with minor modifications. Using the 1000 Genomes Project and HAPGEN2 simulations, we show that LDkin eliminates long-range LD artifacts and substantially improves PRS accuracy in multi-ancestry and admixed cohorts. Our LMM summary statistic model also has direct applications for fine mapping and LD score regression. Overall, our work expands the set of genetic models applicable to arbitrary individuals regardless of their ancestry and relatedness.

Investigating the multiomic susceptibility of TB

Oshiomah Oyageshio

University of California Davis

Tuberculosis (TB) is the world's leading cause of death due to infectious disease, currently greater than COVID-19. The causative agent, Mycobacterium tuberculosis (M.tb), is an obligate intracellular pathogen mainly infecting the lungs, and sometimes other organs. Approximately 25% of the world's population is infected with M.tb and the annual death toll is similar to COVID-19 (~1.5 million deaths). South Africa is among the world's top eight tuberculosis (TB) burden countries, and despite a focus on HIV-TB co-infection, most of the population living with TB are not HIV co-infected. The disease is endemic across the country, with 80–90% exposure by adulthood.

TB is endemic in our study population, with an incidence of 643/100,000. We have sampled PBMCs from 75 LTBI controls and 75 cases. We propose a novel eQTL mapping approach combining single-cell RNA sequencing and ancestry-adjusted analyses on the repertoire of peripheral blood mononuclear cells (PBMCs) from TB cases (positive GeneXpert) and latent (LTBI) controls (positive interferon-gamma release assay and negative GeneXpert) to discover critical variants associated with TB.

Our sequencing approach combines gene expression, cell surface protein, T-cell receptor and B-cell receptor sequencing. This integrative approach will first allow us to accurately identify heterogeneous cell-types involved in TB disease

.

First, we will use CITE-seq for single-cell profiling of the entire PBMC repertoire and its cell surface proteins. This leverages multimodal clustering to integrate gene and protein expression profiles to accurately define cell types and states. Then, we will identify differentially expressed genes (DEGs) in these cell clusters between cases and controls.

Next, we will identify genomic variants (eQTLs) that modulate the TB immune response. To do this we will utilize whole genome sequences (WGS) from our cases and controls and estimate global and local ancestry. Then we will combine WGS data with significant DEGs. Finally, we will use eQTL mapping models that incorporate ancestry proportions to identify novel TB susceptibility variants.

Shared polymorphism investigation on X and autosomes in primates

Bjarke Meyer Pedersen

Aarhus university

Previous investigations into ancient shared polymorphisms (ASP), single nucleotide variation shared between distinct species, have mainly targeted ape lineages and potential sites under strong balancing selection in humans. Primate diversity data provides a rich resource to examine ASP across multiple species and understand fundamental evolutionary mechanisms throughout the primate phylogeny. Coalescent theory predicts the number of ASP based on divergence time, sample sizes, and population sizes. However, directional selection, balancing selection, population structure and ancient admixture also shape ASP patterns. Examining genomic distributions of ASP will reveal how common these processes are among primates.

Additionally, since the X chromosome is disproportionally important for reproductive barriers it might have less ASPs than expected from each relative effective population size due to less impact by introgression. This predicts significantly fewer ASP on the X relative to autosomes.

To investigate whether selection and ancient admixture is common across the primate phylogeny, we intersected SNPs from the autosomes and X chromosomes of ~1500 samples from 259 species. By stratifying the SNPs into mutational classes and CpG/Non-CpG sites we get the number of ASPs for the entire primate phylogeny. We find that the number of ASP are clustered along the genome unless species are very divergent. There are far fewer ASPs on the the X than on the autosomes, indicating that the X chromosome is less affected by ancient admixture

Population genetic analysis of highly degraded ancient-DNA data

Benjamin Peter

University of California, Los Angeles & Max Planck Institute for Evolutionary Anthropology

Ancient DNA studies are largely restricted by DNA preservation; samples from hot or humid climates, from old layers or from sediments frequently only have minuscule amounts of DNA preserved. For Neandertals, studies have focused on the few fossils with extraordinary DNA preservation, but we lack tools for nuclear DNA analyses of contaminated low-coverage DNA, which, as a result, often remain unpublished.

Here, we introduce admixslug, a method designed to analyze ancient DNA data at ultra-low (<0.01x) coverages. admixslug leverages characteristics of individual sequenced molecules, such as fragment length and deamination patterns, to compute contamination-aware genotype likelihoods. Using these likelihoods, we estimate the conditional site-frequency spectrum of a low-coverage sample relative to a reference panel of high-coverage genomes.

Applying admixslug to published Neandertal data, we show that contamination patterns, the conditional site-frequency spectrum, and F-statistics can be reliably estimated. For example, when we downsample the data from Scladina (Belgium) to 0.001x (5,000 reads overlapping informative sites), we correctly identify it as most-closely related to the high-coverage Vindija Neandertal. We further find that contamination rates vary substantially with read length; (25% contamination for 35bp molecules, >90% contamination for molecules 60bp or longer). Thus, admixslug substantially extends the scope of nuclear DNA analyses for old and highly degraded ancient-DNA samples.

Quantifying sex and asexuality using an ancestral recombination graph (ARG) approach

Tymoteusz Pieszko

University of Oxford

The distinction between sexual and asexual reproduction is fundamental to eukaryotic evolution. Testing theories about the evolution of reproductive modes first requires knowing whether sex is present in a population and, if so, how often it occurs. A major complicating factor is that asexuality is commonly accompanied by homologous recombination occurring independently of the sexual pathway, which is typically ignored in classical treatments. I present a new framework that leverages ancestral recombination graphs (ARGs) to address this challenge in detecting and quantifying reproductive modes from population genomic data. Based on realistic individual-based models of asexuality, I demonstrate that, where standard summaries of sequence diversity fail to distinguish between obligate asexual and sexual scenarios, features of the ARG can be explicitly linked to sexual and asexual forms of recombination. I apply the new approach to population genomic datasets for bdelloid rotifers, a group of micrometazoans famous for their asexuality, yet recently the subject of debate over presumed signatures of cryptic sex. The work demonstrates the potential of ARG thinking for improved quantitative inference of reproductive modes in a broad range of non-model eukaryotes.

Detecting changes in fitness optimum from genetic and phenotypic data

Swan Portalier

University of Cologne

Knowing which selection pressures act on phenotypes would allow us to better understand which traits are crucial for an ongoing adaptation. Techniques to detect selection acting on monogenic traits have already been developed (Ewens, 2004; Tajima, 1989). Selection acting on polygenic quantitative continuous traits is typically more challenging to detect because the selection signature is often diluted at many loci throughout the genome. Different modes of selection can shape such phenotypes. Under stabilising selection, individuals that display intermediate phenotypes are favoured. Those phenotypes are close to the fittest phenotype which is called the fitness optimum. On the other hand, under directional selection, the fittest individuals can either have the highest or the lowest phenotypic values.

Natural populations may undergo an environmental change that induces an optimum shift, resulting in a shift from stabilising to directional selection for a trait of interest. From sequencing and phenotyping, we can obtain the joint distribution of the effect sizes and frequencies of the alleles underlying the trait. We call this distribution the genetic architecture. We aim to develop a new method which relies of the genetic architecture of a trait, to assess if this trait is undergoing a shift in fitness optimum.

Inferring ancient introgression from single unphased genomes and a two-locus statistic

Aaron Ragsdale

University of Wisconsin-Madison

The emerging picture of the deep population histories of humans and our close relatives is one of complexity, population structure, and gene flow. We have been able to add resolution to this picture by inferring parameterized models based on two-locus statistics, which summarize the information contained in genealogical correlations between linked loci and are shaped by demographic history. These methods have relied on large sample sizes from present-day populations, though the inclusion of ancient DNA (aDNA) would provide increased power to distinguish competing models. However, it is challenging to apply existing demographic inference tools based on two-locus statistics to aDNA data, where samples are sparse, unphased, and time-stratified. Here, we introduce a two-locus statistic defined as the probability of observing heterozygosity at two loci in a two-haplotype sample. We can estimate this statistic from a single unphased diploid genome, which makes it well-suited for work with aDNA, and show how to extend it to a multi-population setting. We develop theory to describe its expected behavior and show that it captures both one-locus diversity and genealogical covariance, adding increased resolution over standard f-statistic based approaches. We relate the statistic to a tractable system of two-locus summaries whose expectations can be computed under arbitrarily complex demographic models. With this, we develop a model that describes the ancestry of seven ancient hominids, including three Neanderthals, a Denisovan, and early humans in Eurasia, and a contemporary human. In agreement with earlier work, we infer two episodes of gene flow from the ancestors of contemporary humans to Neanderthals and an introgression from a long-diverged 'Superarchaic' lineage to the ancestors of Denisovans, with reasonably narrow confidence intervals on estimated parameters. We also infer a parameterized model relating early modern human lineages in Europe, and we show that the

accurate inference of each of these features requires their joint inference due to their correlated effects on diversity.

Understanding the evolution of mutation and recombination with large scale phylogenomics

Fabian Ramos-Almodovar

University of Pennsylvania

Modeling genetic diversity requires understanding evolutionary forces that generate variation, including mutation and recombination. Here, we use whole-genome polymorphism data from 108 eukaryotic species spanning mammals, fish, invertebrates, and plants to study the evolutionary dynamics of mutation spectra and recombination landscapes, indicating widespread inter-species heterogeneity.

We characterized 5-mer mutation spectra using a regularized hierarchical Bayesian framework to identify mutational mechanisms driving variation in polymorphism patterns. We find that differences in cytosine transitions rates at CpG sites (and CHG sites in plants) are the primary driver of mutation spectrum variation across eukaryotes, capturing 54% of variance. Surprisingly, genome-wide average methylation levels do not predict CpG transition rates across species after phylogenetic correction ($p = 0.21$), suggesting that additional modifiers of deamination and repair influence mutation rates. We find an association between genomic composition and mutation rates ($p = 4\times10\hat{\{\}}$(-4)), indicating that mutational pressures shape CpG/CHG depletion across eukaryotes.

We inferred fine-scale recombination rates in 40 vertebrate species using pyrho. We examined the evolutionary dynamics of balance between PRDM9-directed and promoter-targeted recombination. We find that loss of PRDM9 is associated with more dispersed recombination maps ($p = 7.5\times10\hat{\{\}}$(-3)), and both increased recombination rate ($p = 1.6\times10\hat{\{\}}$(-4))) and more recombination hotspots ($p = 6.9\times10\hat{\{\}}$(-6)) at promoter-like features. However, extensive variation in these metrics show that most species exist along a spectrum of mechanism usage rather than relying exclusively on one pathway.

Together, these analyses provide a framework for understanding how mutation and recombination processes evolve and interact to generate the patterns of genetic diversity underlying comparative genomic inference. Our findings have direct implications for improving probabilistic models that assume uniformity in mutational and recombination processes across species.

Combinatorial mutational scanning libraries for probing global epistasis in proteins

Jingyou Rao

UCSF

Residue–residue interactions underlie nearly every aspect of protein behavior, from folding and stability to binding and allosteric regulation. Despite their fundamental importance, these molecular contacts are often difficult to characterize functionally because most experimental approaches measure only aggregate or indirect effects. As a result, our understanding of how specific residue pairs contribute to protein function remains limited and often relies on computational inference, slowing the development of therapeutics and engineering proteins with valuable properties.

Current high-throughput genotype–phenotype mapping approaches, including deep mutational scanning, primarily focus on single amino acid substitutions or combinations involving only a small number of sites. This limitation arises because the number of possible variants grows exponentially with protein length, rendering large-scale combinatorial in-vitro screens infeasible for most proteins. Consequently, epistatic interactions remain poorly characterized even in small proteins (˜50 residues), let alone those of typical size (˜300 residues) or larger membrane proteins (>500 residues).

Here, we introduce a barcoded combinatorial mutational scanning strategy that leverages oligonucleotide pools to pair single insertions, deletions, or missense variants with diverse genetic backgrounds, generating double-mutant libraries to quantify global epistasis. We use membrane proteins as the model system, enabling high-resolution mapping of epistatic interactions in challenging therapeutic targets previously inaccessible to such analysis.

Pleiotropic stabilizing selection obscures signatures of directional selection against complex disease

Kellen Riall

University of Chicago

Despite the presumed fitness costs of complex diseases such as schizophrenia or type 2 diabetes, genome-wide association studies (GWAS) have failed to uncover consistent genomic signatures of directional selection against disease-risk alleles. Motivated by this puzzle, we extend the mutation–selection–drift balance framework of directional selection on disease from Berg et al. 2025 by assuming that most risk variants are also subject to pleiotropic stabilizing selection on other quantitative traits, and analyze how this joint selection regime shapes genetic architecture, detectability of selection, and disease prevalence.

We show that when mutational input is symmetric—meaning risk-increasing alleles are equally likely to be derived or ancestral—directional selection against disease risk is completely undetectable from the site frequency spectrum unless ancestral states across sites are known. This symmetry regime appears plausibly consistent with GWAS observations for small-effect variants, implying many existing analyses were theoretically incapable of detecting selection. We further show that even when ancestral states are known, pleiotropic stabilizing selection actively erases the signature of directional selection. By symmetrically reducing the fixation probabilities of both risk-increasing and decreasing alleles, stabilizing selection effectively "freezes" the system when sufficiently strong, grinding the adaptive disease response to a halt and preventing the establishment of the mutational asymmetries required to generate directional signals.

Finally, we identify an unexpected effect of pleiotropy on disease prevalence. For selection against disease to be meaningfully suppressed by stabilizing selection, the mean population disease liability must shift far from the threshold, causing a decrease in prevalence. This creates a paradox though, as while strong pleiotropic stabilizing selection on traits can explain the absence of directional selection signals in disease GWAS data, it appears inconsistent with the persistence of many common diseases. This suggests that resolving the contradiction may require invoking a model where selection acts directly against individuals with low disease liability, effectively imposing stabilizing selection on liability itself. Together, these results clarify why directional selection against complex disease may have evaded detection thus far, and highlight fundamental limits on evolutionary inference from GWAS.

Rapidly estimating frequencies of known alleles in a pangenome graph from k-mer counts in pooled sequencing data

Miles Roberts

UC Berkeley

Pangenomic references, as opposed to single-genome references, are the new standard for population genetic analyses. However, genotyping new sequencing samples against a reference pangenome can be computationally costly as either the reference or sample scale in size. Genotyping can thankfully be sped up massively by counting short exact matches between sequencing reads and variants (i.e., k-mers) using tools like BayesTyper and Pangenie. However, these and similar tools are restricted to estimating only diploid genotypes while many population genetics datasets rely on pooling DNA from many genomes into one sequencing library (i.e., pool-seq) to cut labor costs. We present a method for counting allele-specific k-mers in pool-seq data that can estimate frequencies of known pangenomic alleles in a fraction of the time as vg giraffe. We benchmark the approach on both simulated and empirical datasets and find that vg giraffe and k-mer counting give similar estimates of allele frequencies. Further development of similar k-mer-based approaches will ideally provide preliminary results before committing to pangenome alignments or allow scaling to larger datasets.

Dynamics of recessive mutator alleles: model and data

Matin Saeidi

Columbia University

Germline mutation rates differ across individuals and evolve over time. Among humans, the variation in rates is largely explained by age and sex, but has also been shown to be heritable. There should therefore be "mutator alleles" segregating in humans, which increase the mutation rate. To date, a handful of such mutators have been identified in mammals. Intriguingly, most appear to be recessive in their effects on mutation rates. Yet existing "drift-barrier" models for the evolution of mutation rates assume a single copy of a mutator allele increases the mutation rate (i.e., semi-dominance). We extend the theory to allow for arbitrary dominance, ranging from fully recessive to semi-dominant. As is standard, our model incorporates genetic drift and purifying selection arising from the excess deleterious mutations generated by mutator alleles. We solve the model analytically for a constant population size and validate our results using forward simulations. If we parametrize the fitness effects of a mutator allele in terms of the cumulative fitness effects of the additional mutations that it introduces, we find that mutator alleles behave as would deleterious alleles with equivalent dominance and selection coefficients. Using realistic demographic models for European populations and empirical estimates of effect sizes for human mutator alleles, we infer the dominance coefficients that are most likely given their observed population frequencies. Finally, we characterize the power to identify mutator alleles in pedigree data, in terms of the properties of mutator alleles (i.e., their dominance and effect sizes).

A deep learning approach to detecting negative frequency-dependent selection

Cindy Gilda Santander

Florida Atlantic University

Balancing selection is a mode of natural selection that maintains genetic diversity through various mechanisms, including negative frequency-dependent selection (NFDS). However, distinguishing the genomic signature of NFDS from those of other balancing selection modes, such as overdominance, remains a significant challenge. Here we outline strategies to improve the modeling of genomic patterns expected under NFDS, with the goal of better differentiating them from signals of neutrality and alternative selection processes. We demonstrate how resource-efficient deep transfer learning, combined with novel data preprocessing and the modeling of genomic autocovariation, can effectively detect and characterize NFDS using either phased or unphased genotypes, and with or without temporal data from ancient DNA. Finally, we offer practical recommendations for both empiricists and method developers on advancing the detection of NFDS in genomic data.

Pseudo-likelihood kmer-based calculation of regional genomic distances and applications to phylogenetics

Ali Osman Berk Sapci

UC San Diego

Calculating similarity between DNA sequences is a fundamental problem in computational genomics, arising in many forms such as read-to-genome mapping, genome-to-genome comparison, and the identification of abnormally similar or divergent regions. Such distances can be used in downstream tasks such as phylogenetic placement, homology identification, and the detection of horizontal gene transfer. Traditional approaches rely on sequence alignment, which, while accurate, are computationally expensive and do not scale well to large datasets or high-throughput applications. Scalable alternatives based on k-mer sketching exist; however, they fail to model uncertainty and the evolutionary relationships between references.

We present a principled, likelihood-based framework for estimating local Hamming distances between sequences using k-mers. Moving beyond simple presence/absence signatures, our approach integrates locality-sensitive hashing for inexact k-mer matching and a maximum pseudo-likelihood (MPL) formulation for distance estimation, supported by likelihood-based statistical tests for distinguishing similar distances. Together, these components enable accurate estimation of both local and global distances at scale, extending to distances where alignment is no longer feasible.

By incorporating an existing backbone phylogeny into this framework, the resulting distance-based patterns enable several downstream analyses. First, we demonstrate how these MPL distances can be used for phylogenetic placement of sequences as short as individual reads, improving upon existing tools for metagenomic sample differentiation and comparison, particularly for samples from novel environments. Then, we explore how analyzing pseudo-likelihoods of distances locally along sequences can identify genomic regions exhibiting distinctive evolutionary signals such as horizontal gene transfer and ultra-conserved elements.

Pleiotropic stabilizing selection shapes genetic architecture of complex traits

Joshua Schraiber

Univeristy of Southern California

Decades of GWAS have revealed that many complex traits in humans are highly polygenic, shaped by many loci of very small effect. The genetic architecture of these traits, that is, the genome-wide distribution of causal variants, their effect sizes, and their allele frequencies, is shaped by the combined action of evolution forces, including mutation, natural selection, and genetic drift. An understanding of how these forces shape genetic architectures crucial to the application of standard tools in statistical genetics, such as the genetic relatedness matrix and LD-score regression, which are widely used to estimate heritability, identify loci in mapping studies, and other analyses. However, the canonical approaches to using these tools relies on a model of evolution that implicitly assumes no drift and very strong selection. Moreover, there is increasing evidence that many loci are associated with more than one trait, indicating a substantial role of pleiotropy in shaping the genetic architecture of traits. Thus, classical evolutionary and statistical genetic theory and models are inadequate for understanding the forces shaping the genetic architecture of complex traits. Here, we present theoretical and empirical results demonstrating the importance of pleiotropic stabilizing selection in determining the relationship between allele frequency and effect size. First, we find that, compared to the single trait case, the genetic variance is distributed among rarer alleles in the pleiotropic case. Nonetheless, among causal variant, the relationship between allele frequency and effect size does not depend on pleiotropy. However, when causal variants are sparse and not precisely identified, as is the case when building a genetic relatedness matrix or performing LD-score regression, the impact of pleiotropy is visible. In that case, when mutations are highly pleiotropic, concentration of measure results in a simple form for the relationship between effect sizes and allele frequencies, which can be exploited to easily model effect sizes that

span orders of magnitude. By applying this model in a framework analogous to LD-score regression, we find substantial evidence for pleiotropic stabilizing selection shaping the genetic architecture of human traits.

A coalescent latent space model for flexible depiction of structure through time

Ahmed Selim

University of Chicago

Many methods describe how genetic ancestry is structured either in the present-day or through time through a population history model. These methods often model the structure in terms of discrete panmictic units (populations) with some migration/admixture between them. Alternative approaches conceptualize structure in a more continuous manner (e.g. due to geographically structured mating / “isolation by distance”). Currently, there is the need for new tools that can reveal the structure of genetic ancestry in a time-varying way, without strong assumptions about whether the structure has discrete units or continuous gradients. and without necessarily assigning samples to fixed population labels. Here, we present progress on this problem, by developing a method that provides a time-varying latent space representation of coalescent structure present in an inferred Ancestral Recombination Graph (ARG).

The method is based on a latent space model, where the distance between the embeddings of two samples at some point in time models their instantaneous pairwise coalescent rate through a link function. We derive a loss that is a function of the latent space embeddings of the samples by invoking properties of the coalescent process. We scan the inferred ARGs to compute summary statistics that are inputs to this loss function, and find the configuration of positions through time that minimize the loss given the observed statistics.

We apply our method to a range of simulations of increasing complexity, and verify that reproduces time-resolved representations of ancestral relationships that reflect the simulated demographic history, revealing features such as split times, population size change, gene flow and admixture. Additionally, we apply the method to ARGs inferred from Human and Chimpanzee datasets, and are able to verify existing results and additionally unveil new insight about their

population history.

Limits on the number of traits maintained by stabilizing selection: revisiting Barton's paradoxes

Guy Sella

Columbia University

It's long been thought that numerous complex traits are highly polygenic and maintained near an optimal value by stabilizing selection, and this view is supported by past and recent evidence. Nonetheless, over three decades, Nicholas Barton published a series of theoretical arguments that challenge this view. He argued that the increase in genetic load and in variance in fitness with the number of traits under stabilizing selection impose unrealistically low bounds on the number of traits. Here, we revisit these paradoxes and argue that some of the bounds are relaxed by the effects of pleiotropic selection on quantitative genetic variation and others might not be unrealistically restrictive. In so doing, we clarify how stabilizing selection on multiple traits shapes the mean and variance of complex traits and of fitness in natural populations.

A phylogeny-based framework to uncover the evolution of mutational processes in primates

Vladimir Seplyarskiy

UTsouthwestern

Germline mutation is the ultimate source of genetic variation, yet comparative characterization of mutational processes across species has been constrained by the scarcity and cost of de novo mutation data. Although recent work suggests that both lifetime somatic mutation burdens and per-generation germline mutation rates vary within surprisingly narrow ranges across mammals, the evolutionary dynamics of the underlying mutational mechanisms remain mostly obscure.

Here, we introduce a phylogeny-aware, multi-step framework to infer germline mutational processes directly from multispecies genome alignments. We estimate branch- and region-specific trinucleotide substitution spectra and apply Reciprocal Principal Components Analysis (RPCA) to jointly analyze species within clades, performing independent decompositions for five primate clades and a rodent outgroup. This approach identifies nine distinct mutational processes, including five that are conserved across all clades.

We validate inferred processes using both human and non-human polymorphism data and link them to known genomic and molecular features. For example, processes associated with biased gene conversion covary with species-specific recombination rates, while transcription-coupled processes reflect strand asymmetries consistent with bulky lesion repair. Strikingly, we detect rapid evolutionary shifts in mutational spectra, including a marked reduction in sensitivity to bulky DNA lesions in Callitrichidae (marmosets and tamarins).

Together, this framework enables inference of mutational processes at a phylogenetic scale without reliance on pedigree data, providing a population-genetic toolkit to study the evolution of DNA damage and repair mechanisms across the tree of life.

Beyond flat clusters: Hierarchical IBD Models for fine-scale population structure

Ruhollah Shemirani

Icahn School of Medicine at Mount Sinai

Population structure is a pervasive source of confounding in genomic research. Common adjustment methods, such as principal components, capture coarse-scale structure but inadequately represent the fine-scale and recent demographic processes that shape the landscape of genetic similarity. While methods based on clustering genome segments shared Identical-By-Descent (IBD) offer greater sensitivity to recent ancestry, they typically impose flat, discrete partitions that fail to reflect the continuous and hierarchical nature of population structure. As a result, population structure is often incompletely modeled across resolution levels, limiting interpretability, robustness to cohort composition and to recent admixture, and effective control of confounding in large biobank-scale datasets, while oversimplifying genetic diversity. These limitations become more pronounced as datasets grow in size and increasingly reflect the true demographic complexity of target populations.

Here, we present a novel, scalable, unsupervised framework to extract, organize, and interpret fine-scale population structure using an explicitly hierarchical model, simultaneously representing structure across multiple resolution levels. We explore and evaluate this hierarchy asymmetrically using established and novel network-based metrics that identify robust clusters across resolution levels and enable classification by demographic history even in the presence of common ascertainment biases and recent admixture. Applying this framework to 245,395 participants from the All of Us dataset, we scaled to ~15 billion pairwise haplotypes and recovered nested population structure associated with self-reported identity, geographic residence, health outcomes, accuracy of risk prediction models, and segregation of clinically relevant rare variants, with measurable gains in predictive power compared to standard genetically inferred ancestry representations. As sampling depth increases, the framework resolves progressively finer population granularity, particularly among European and Hispanic groups. At the finest resolutions, the hierarchy reveals widespread founder-like substructure embedded within broader population groups, spanning 19% (N=46,236) of participants across 14 clusters. This highlights the inadequacies of existing flat representations of population structure in confounding adjustment, with measurable consequences for genetic association studies and risk prediction calibration.

Fine-resolution asymmetric migration estimation

Hao Shen

University of Chicago

The inference of gene flow has been a central task in spatial population genetics since its inception. Early research focused on the theoretical properties of migration networks with specific topologies, typically restricted to small networks. Advances in large-scale gene flow inference has been facilitated by using approximations to pairwise coalescent times using commute times in random walks (which can be computed using "resistance distances"; for example, as in the EEMS and feems software packages). These approaches enable rapid inference across networks with hundreds of demes. However, the commute time approximation breaks down in the presence of strongly asymmetric migration and in such cases solving the pairwise coalescence time exactly is preferable (Lundgren and Ralph 2019). Due to its high computational complexity $O(d^6)$, standard solutions for exact pairwise coalescent times are limited to small scales in practice (e.g. <20 demes). To overcome these challenges, we introduce FRAME (Fine-Resolution Asymmetric Migration Estimation). FRAME employs techniques from computational linear algebra and solves the structured coalescent equation exactly with computational complexity of $O(d^4)$. By integrating this approach with penalized-likelihood methods in FEEMS, FRAME enables the inference of fine-resolution (hundreds of demes) asymmetric migration patterns with significantly improved computational efficiency. We validate the method using both equilibrium and non-equilibrium simulations generated with msprime and SLiM. Applied to empirical datasets of poplar trees, North American gray wolves, and ancient humans, FRAME produces the first-ever fine-resolution asymmetric gene flow maps and recovers strong asymmetric signatures of gene flow, such as signatures of the Neolithic and Steppe expansions in humans.

High dimensional confounder adjustment for multivariable Mendelian randomization using genetic factor analysis

Ruoyao Shi

University of Michigan

Mendelian randomization (MR) is a causal estimation method using genetic variants as instruments. A major challenge in MR is heritable confounding, which occurs when instruments affect traits that confound both the exposure and outcome. Multivariable MR (MVMR) can adjust for heritable confounding when GWAS summary data are available for known confounders. However, most existing MVMR methods can only adjust for a moderate number of confounders, and each requires strong, independent genetic instruments. In reality, true heritable confounders are often unmeasured, and we instead measure proxy traits sharing genetic architecture with underlying confounding sources. For example, there may be heritable confounding through an adiposity-related pathway contributing to traits such as BMI and waist circumference. Including all such traits in MVMR is often infeasible due to limited independent instruments. However, selecting only one adiposity-related trait discards information and may weaken confounding control. To address this, we propose a novel approach for high-dimensional confounder adjustment in MR, Factor-based Multivariable Mendelian Randomization (FarMR). We first apply Genetic Factor Analysis to identify latent genetic structure among correlated confounders. We then use a hierarchical shrinkage model to estimate the causal effect of the main exposure on the outcome while adjusting for both shared latent factors and trait-specific factors capturing residual genetic effects not mediated by the shared structure. Simulations show that FarMR substantially reduces bias and MSE compared to univariable MR, MVMR adjusting for all candidate traits, and MVMR adjusting for an independent set of traits. We apply FarMR to estimate the causal effects of C-reactive protein level (CRP) on diseases. We reduce 43-90 candidate confounders to 23-35 latent factors and find no evidence of a causal effect of CRP on coronary artery disease, type 2 diabetes, or knee

osteoarthritis. An additional advantage of FarMR is that it eliminates manual selection among correlated candidate confounders. It reduces researcher degrees of freedom, improving replicability and reliability of MR findings. Finally, we present an integrated pipeline allowing automatic detection of latent factors for confounding traits and feeding them directly into causal estimation. It regularizes and systematizes MR estimation with phenome-wide confounder adjustment, enabling accurate, reproducible causal inference.

REECAP: Contrastive learning of retinal aging reveals genetic loci linking morphology to eye disease

Liubov Shilova

Helmholtz Munich

Deep learning foundation models excel at disease prediction from medical images, yet their potential to bridge tissue morphology with the genetic architecture of disease remains underexplored. We present REECAP (Representation learning for Eye Embedding Contrastive Age Phenotypes), a framework that fine-tunes the RETFound retinal foundation model using a contrastive objective guided by chronological age. Applied to 87,478 fundus images from 52,742 UK Biobank participants, REECAP aligns image representations along the aging axis, yielding multivariate ageing phenotypes for genome-wide association studies (GWAS). GWAS of REECAP embeddings identifies 178 loci, including 27 that colocalize with risk loci of age-related eye diseases, 14 of which remained undetected by conventional disease-label GWAS. By enabling conditional image synthesis, REECAP further links genetic variation to interpretable anatomical changes. Benchmarking against alternative embedding models, we show that REECAP enhances both locus discovery and disease relevance of genetic associations, suggesting that aging-informed tissue embeddings represent a powerful intermediate phenotype to discover and interpret disease loci.

Inferring epistasis from temporal genetic sequence data and the limitations of inference

Kai Shimagaki

University of Pittsburgh

Epistatic interactions, or non-additive fitness effects between mutations, are a fundamental evolutionary factor and are prevalent across a wide range of organisms. Quantitatively understanding these interactions provides insights into complex genotype–phenotype relationships and the predictability of genetic population evolution. Temporal genetic sequence data provide a unique window for inferring epistatic interactions, and such data have become increasingly accessible. To accurately infer epistasis, it is essential to account for genetic linkage, recombination, and genetic drift, and population genetics theory provides a principled framework to address this task. Here, we propose a framework based on the Wright–Fisher diffusion model that infers epistatic interactions as well as selection coefficients from temporal genetic data. We examine the accuracy of the inference under multiple evolutionary regimes: under strong selective pressure with low recombination frequency, the method achieves higher accuracy, whereas frequent recombination and genetic drift limit accurate epistasis inference. This framework is computationally efficient, enabling applications to long genetic sequences (>5 kb), and we apply it to infer epistatic interactions in the HIV-1 surface protein using multiple inter-host HIV-1 evolutionary data sets.

Robust detection of selection from ancient DNA time series using mixed models

Lucas Sort

Mathematical Genomics Research Unit, iTHEMS, RIKEN

Over the past decade, the emergence of ancient DNA has opened new opportunities for studying evolutionary processes. However, inferring signals of selection from such data remains a methodological challenge since controlling for genetic drift, population stratification, admixture, and dynamically changing demographic histories, among other confounding evolutionary processes, is difficult. The two main frameworks for modeling ancient DNA time series, Hidden Markov Models and Generalized Linear Mixed Models, have individual strengths in modeling genetic drift and population structure, respectively. Here, we develop a new approach that can produce calibrated p-values while accounting for both drift and population structure, motivated by theory that clarifies how these frameworks relate to the classical Wright–Fisher model, and we present relevant comparisons across methods.

Joint phylogenetic and transmission inference with JUNIPER

Ivan Specht

Stanford University

Transmission reconstruction—the inference of who infects whom in disease outbreaks—offers critical insights into how pathogens spread and provides opportunities for targeted control measures. We developed JUNIPER (Joint Underlying Network Inference for Phylogenetic and Epidemiological Reconstructions), a highly-scalable pathogen outbreak reconstruction tool that samples the space of phylogenetic trees decorated with transmission events. In contrast to the coalescent model, JUNIPER infers the phylogenetic tree and transmission network structure under an SIR or SEIR epidemic model with time-dependent reproductive number, making it particularly well-suited for studying outbreaks. We achieved state-of-the-art statistical efficiency by (1) implementing a Hamiltonian Monte Carlo sampler for effective population size trajectories with analytic, linear-time gradient computations, (2) parallelizing Metropolis-Hastings tree updates over subtrees of the transmission network, and (3) representing hosts and mutations explicitly on the phylogeny. We benchmarked JUNIPER on synthetic and real outbreaks in which transmission links were known or epidemiologically confirmed. We demonstrated JUNIPER's real-world utility on two large-scale datasets: over 1,500 bovine H5N1 cases and over 13,000 human COVID-19 cases. Based on these analyses, we quantified the elevated H5N1 transmission rates in California and identified high-confidence transmission events, as well as demonstrated the efficacy of vaccination for reducing SARS-CoV-2 transmission. By overcoming computational and methodological limitations in existing outbreak reconstruction tools, JUNIPER provides a robust framework for studying pathogen spread at scale.

Collateral mutagenesis funnels multiple sources of DNA damage into a ubiquitous mutational signature

Natanael Spisak

Institut Imagine in Paris

Mutations reflect the net effects of myriad types of damage, replication errors, and repair mechanisms, and thus are expected to differ across cell types with distinct exposures to mutagens, division rates, and cellular programs. Yet when mutations in humans are decomposed into a set of signatures, one single base substitution signature, SBS5, is present across cell types and tissues, and predominates in post-mitotic neurons as well as male and female germlines. The etiology of SBS5 is unknown. By modeling the processes by which mutations arise, we infer that SBS5 is the footprint of errors in DNA synthesis triggered by distinct types of DNA damage. Supporting this hypothesis, we find that SBS5 rates increase with signatures of endogenous and exogenous DNA damage in cancerous and non-cancerous cells and co-vary with repair rates along the genome as expected from model predictions. These analyses indicate that SBS5 captures the output of a “funnel”, through which multiple sources of damage result in a similar mutation spectrum. As we further show, SBS5 mutations arise not only from translesion synthesis but also from DNA repair.

Systematic estimation of the number of mutational origins in large population cohorts

Remus Stana

UT Southwestern Medical Center

Expansion of sequencing data leads to deviation from the infinite site model as can be seen through the association between site frequency spectrum and mutation rates highlighting pervasive recurrence. For example, in gnomADv4 the most common frequency class of non-CpG mutations is singletons as expected without substantial contribution of recurrent mutations, while for CpG mutations it is eleven. Despite prevalence of recurrent mutations there is currently no method to estimate the number of origins for single nucleotide variants. Conceptually it is possible to estimate the number of mutational origins by analyzing local genetic similarity of individuals carrying a segregating mutation. To estimate recurrence in biobank-scale data we developed an approach based on ancestral recombination graphs while accounting for limitations of haplotype inference. We validated the accuracy of our methodology with simulations. Finally, application of our approach to All of Us yields discovery of unknown hypermutable sites.

Bayesian inference of demographic history and structure from the distribution of heterozygous sites distances

Tommaso Stentella

Max Planck Institute for Molecular Genetics

In the last two decades several methods to infer the demographic history of a population from whole genome sequence data have been developed. Despite these efforts, two outstanding challenges remain unsolved: how to efficiently explore the space of possible demographic models and how to estimate confidence intervals of all model parameters.

Here we present new theoretical insights into the demographic inference problem. We consider the distribution of distances between consecutive heterozygous sites (IBS tracts lengths) and show that, under a few simplifying assumptions, it is in one-to-one correspondence with the history of effective population size. The simplicity of our analytical result allows us to perform Bayesian inference and efficiently fit and compare several parametric models of varying complexity. Our procedure yields the model which partitions the demographic history into an optimal number of epochs, each having arbitrary duration. Both duration and population sizes are estimated together with confidence intervals.

We show that our methodological developments allow for a highly robust, scalable and sensitive implementation for inference on real data. Specifically, it is well suited to infer thousands of unique demographic histories from single genomes in the 1000 Genomes Project, revealing fine variation at the individual level within and across populations. We also demonstrate that our method can be applied to archaic human genomes. Our analysis enhances our understanding of the extent to which demographic history can be reconstructed from whole genome sequence data.

Authors: Tommaso Stentella, Paul Etheimer, Florian Massip, Michael Sheinman, Peter F. Arndt

The impact of associative overdominance on genetic diversity

Steven Sun

UC San Francisco

The impact of purifying selection on neutral diversity at linked sites has been extensively

explored within the framework of codominance. When codominant deleterious variants are subject to weak to moderate selection, there's a reduction in variation at linked neutral sites, known as background selection (BGS). However, if deleterious variants act recessively, an unexpected increase in variation can occur, akin to classic overdominance effects. This phenomenon, termed associative overdominance (AOD), is less thoroughly understood compared to BGS. Using forward simulations incorporating a gamma distribution of fitness effects, we demonstrate how AOD influences genetic diversity via the pairwise nucleotide diversity (π) and the site frequency spectrum (SFS) statistics. Our analysis reveals that the SFS resulting from classic overdominance models differs from those influenced by AOD. Additionally, we illustrate how bottleneck events and the presence of concurrent codominant alleles further alters the effects of AOD. Notably, bottleneck events or the inclusion of codominant alleles alongside fully recessive ones transforms the SFS shaped by AOD to resemble those produced under a milder selection regime.

Detecting deep-time selection sweeps using coalescent-based inference

Tianyi Wang

Harvard University

What makes humans human is a central question in evolutionary biology. Understanding how natural selection has shaped human genomes since the split from chimpanzees over deep evolutionary time remains a critical challenge. While numerous methods have been developed to detect recent positive selection, signals of selection acting hundreds of thousands of years ago are often eroded by recombination and confounded by complex demographic histories, fragmenting haplotypes and obscuring classical sweep signatures.

To address this limitation, we shift the focus from genome-wide extreme signals to regional deviations in local coalescent patterns. In this study, we develop a selection scan based on PSMC-inferred coalescent profiles, aiming to identify genomic regions that coalesce significantly faster than the genome-wide background across different coalescent time depths. Such compressed genealogies are expected under ancient selective sweeps whose genealogical signatures persist over deep evolutionary time but are no longer detectable using haplotype-based approaches.

Applying this framework to multiple individuals from diverse human populations, we identify genomic regions exhibiting accelerated coalescence consistent with selective events occurring deep in the Pleistocene. These candidate regions are enriched for genes with diverse biological functions, highlighting the role of long-term selective pressures in shaping human adaptation.

Phylogenetic deconvolution of wastewater sequencing data to detect influenza reassortment

Audrey Li-Wen Wang

University of California, Berkeley

Influenza reassortment occurs when two or more influenza viruses co-infect the same host cell and exchange entire RNA segments, generating novel viral genotypes with mixed ancestry. This process can rapidly produce viruses with altered antigenic properties or host adaptation, and all past influenza pandemics have resulted from reassortment events of influenza A viruses (IAVs). CDC's National Wastewater Surveillance System monitors influenza concentrations in wastewater to infer disease risk and population health, complementing existing clinical surveillance systems that are often constrained by resources, access, and time. A major challenge in detecting reassortants through wastewater-based genomic surveillance is that samples often contain mixtures of IAV clades derived from multiple hosts. In addition, the segmented nature of the IAV genome complicates downstream bioinformatic analysis, particularly the assignment of individual segments to their corresponding clades in the presence of reassortants. To address this, we developed a probabilistic deconvolution method to demix IAV clades in wastewater and identify potential reassortment events using whole-genome sequencing data. The method leverages tronko, a phylogenetic placement method for assignment of reads, to calculate read-to-clade likelihoods. Following this, we use an expectation–maximization (EM)–based mixture model to estimate relative abundances. A likelihood ratio test is then used to assess whether clade proportions differ significantly across segments, providing evidence of a circulating reassortant. To verify the method's deconvolution accuracy and its sensitivity of detecting reassortant, simulated Illumina reads were generated to assess performance. In addition, a gBlock spike-in experiment was performed by mixing four different IAV strains, including H1N1pmd09, H3N2, H5N1 2.3.4.4b, and artificially-created reassortant strain H5N2 at different mixing ratios. Extracted wastewater RNA was added to the

IAV mixture to create an intentional noise. Positive controls with only gBlock mixture and negative control with only wastewater were included. All samples were sequenced and validated by the developed method. We show that the method successfully deconvoluted complex mixtures and identified reassortments even in high-noise environments. This method provides a rigorous statistical basis for tracking evolutionary dynamics and detecting emerging pandemic threats via wastewater surveillance.

Concentrating association power along specific biological pathways by controlling for heritable covariates in proxy GWAS

Jeremy Wang

UCLA

Modern studies have documented extensive heterogeneity in the genetic architecture of complex traits. Although heterogeneity biases and reduces the power of genome-wide association analyses (GWAS), it has recently been leveraged to explain cross-ancestry differences in disease risk as well as inter-individual differences in comorbidities and treatment response. This utility has dramatically increased interest in resolving genetic heterogeneity by sorting causal variants into mechanistic pathways. However, the success of these efforts is inherently shaped by the amount of heterogeneity ascertained in GWAS discovery cohorts and limited by the biases incurred in GWAS of heterogeneous traits. Here, we introduce a new strategy, N-SEVER, to directly approximate GWAS of pathways, circumventing these issues. N-SEVER converts a target trait, a set of endophenotypes, and a set of heritable covariates into a new phenotype that mimics the genetic pathways shared by the target and endophenotypes while avoiding genetic pathways shared with the covariates. Because N-SEVER functions using only summary statistics for the target trait and covariates, we are able to study rarer traits such as type 2 diabetes (T2D) and major depressive disorder (MDD) in the UK Biobank while controlling for hard-to-measure covariates such as MRI-derived liver fat. In T2D, we define a blood biochemistry phenotype that matches a known insulin deficiency pathway on cell type enrichments and

genetic correlations with auxiliary traits. GWAS of this trait increases the number of known insulin deficiency-associated T2D loci by 10% despite sample sizes an order of magnitude smaller than in the latest T2D meta-analyses. A diffusion MRI (dMRI) N-SEVER phenotype targeting T2D recovers a known

locus associated with behavioral risk for T2D while discovering novel loci with pleiotropic effects on brain structure and T2D risk. Finally, a dMRI N-SEVER phenotype targeting MDD shows substantially higher genetic correlation with MDD (rg = 0.43) than all previously known dMRI phenotypes and shows $\sim$ 100× greater power to detect MDD loci with effects on brain structure.

Uniform bacterial genetic diversity along the gut

Michael Wasney

UCLA

Environmental gradients throughout the digestive tract shape spatial variation in the composition and abundance of bacterial species along the gut. However, much less is known about how genetic diversity within bacterial species is distributed along the gut. Understanding this distribution is important because bacterial genetic variants confer traits that influence both microbiome function and host physiology, including local inflammation and nutrient metabolism. Thus, to understand how the microbiome functions at a mechanistic level, it is essential to understand how genetic diversity is organized along the gut. In this study, we profiled genetic diversity of approximately 30 common gut commensal bacteria in five regions along the gut lumen in germ-free mice colonized with the same healthy human stool sample. Although species composition varied considerably along the gut, genetic diversity within species was substantially more uniform. Driving this uniformity were similar strain frequencies along the gut, implying that multiple, genetically divergent strains of the same species can coexist within a host without spatially segregating. Additionally, the approximately 60 unique evolutionary adaptations arising within mice tended to sweep throughout the gut, showing little gut region specificity. We then analyzed metagenomic samples collected along the guts of conventional mice and healthy humans and found similar dynamics with their natural microbiomes, suggesting that uniform bacterial genetic diversity may be common to multiple host species. Together, our findings demonstrate that uniform spatial distribution of genetic diversity along the gastrointestinal tract is a robust feature of mammalian gut ecosystems.

GPU-accelerated coordinate ascent variational inference for scalable prediction and association analyses

Marius Weidmann

Oxford University, Department of Statistics

GWAS analyses of modern biobank datasets, such as UK Biobank and All of Us, require frameworks that can scale to millions of variants and individuals without sacrificing statistical power. Existing linear mixed model approaches trade modeling complexity for computational cost. To improve these tradeoffs, we introduce a highly scalable GWAS framework that leverages GPU acceleration and coordinate ascent variational inference. Our approach relies on several additional computational optimizations, including genotype block streaming and asynchronous ping-pong buffering, allowing us to fit dense predictors at biobank scale and millions to tens of millions of variants on a single A100 GPU.

We compare this approach to Quickdraws, a recent method based on stochastic variational inference, and obtain $\sim 18\times$ faster convergence and a $+1.7\%$ relative gain in prediction R^2 in simulations using UK Biobank data ($N_{train} = 293\text{k}, N_{test} = 32\text{k}$) across varying levels of polygenicity. We further leverage transfer learning to implement a leave-one-chromosome-out approach that enables modeling of 50 traits ($N \approx 293\text{k}, M = 9.3\text{M}$) in ~ 28 hours, suggesting that this framework can be used to perform GWAS analyses on modern biobank-scale datasets without sacrificing statistical power relative to state-of-the-art approaches.

The Poisson recombination model artificially increases the spatial dependence of coalescence and ignores sex-biased transmission

Amy Williams

Brigham Young University

Analyses of human genetic data, including ancestral recombination graph (ARG) inferences, most commonly leverage a Poisson recombination model. While known to differ markedly from more realistic crossover interference models, mathematical convenience and assumed realism beyond a few generations of meiosis have entrenched this model for nearly all analyses and most simulations.

The physical constraint limiting recombination to occur only between the two haplotypes an individual carries induces correlation in the haplotypic origin of DNA across sites. That is, individuals can inherit two or more genetic segments transmitted by the same ancestor, even from many generations ago. This dependence across genomic sites is non-Markovian, which stymies methods for ARG inference and other analyses.

We simulated 200 genomes, recording the transmissions from all 2{ˆ}20 ˜= 10{ˆ}6 ancestral haplotypes under both a Poisson model with a sex averaged genetic map and one incorporating crossover interference and sex-specific maps. Under the Poisson model, an average of 36.7 segments are 'repeats'—i.e., descend from the same haplotype as another segment. Strikingly, use of the more realistic model produces only 6.4 repeats, an 83\{}% reduction (P < 10{ˆ}-15). This substantial difference implies that Markovian ARG models are better approximations than previously known, and it carries implications for a host of other genetic analyses.

Another feature we will characterize is the impact of sex on the lineages that transmit DNA. Standard models assume that DNA is equally likely to descend from any lineage, but ancestor sex is likely to impact this distribution since females transmit ~1.6× as many crossovers as males. Because a crossover joins both of a person's paternal and maternal lineages, both of a female's parents' lineages are more likely to be transmitted to the next generation than both of a male's lineages.

ARGs of multiple chromosomes

Yan Wong

University of Oxford

Modern population genetic simulators have reached the point where they can simulate all the chromosomes in a genome, for a large number of individuals. For example, Anderson-Trocme et al. (2023) simulate all chromosomes from 1.4 million humans on the basis of a reconstructed pedigree, and recent versions of SLiM support simulation and recording a single ARG of multiple chromosomes as separate tree sequence files that share node identities. However, there is no current agreement on how to encode ARGs constructed from inferred diploid data such that non-independence between chromosomes is recorded. I present a set of proposals for storing and analysing whole-genome data in tskit as a series of linked tree sequences, which allows integrated whole-genome approaches to ancestral analysis. In addition I present initial results on approaches to phase across chromosomes, which is required e.g. for carrying out haploid genome-wide analysis, and can serve as a basis for inferring pedigrees from highly sampled datasets.

Comparison of methods for non-negative covariance matrix decomposition

Annie Xie

University of Chicago

The covariance decomposition problem involves decomposing a covariance matrix $\{}Sigma$, into a sum of parts, $\{}Sigma \{}approx \{}sum_{k=1}^{}{K} l_k l_k'$. We are particularly interested in decompositions where each component $l_k l_k'$ has a physical or scientific interpretation. While non-negative matrix factorization has been widely used to obtain interpretable "parts-based" representations of data matrices, covariance matrix decomposition is more commonly performed with Factor Analysis (FA) or Principal Components Analysis (PCA). We propose non-negative covariance matrix decomposition as it provides a generalization of clustering and often yields more interpretable components than FA and PCA. In our study, we will review several methods for non-negative covariance decomposition based on symmetric non-negative matrix factorization, "additive clustering", and Empirical Bayes matrix factorization. We will also highlight connections between non-negative decompositions of the covariance matrix and semi-nonnegative decompositions of the data matrix. Furthermore, using simulations, we will systematically assess and compare the methods; we will investigate the types of structure each method can find and identify settings in which it struggles.

keju: powerful and accurate enhancer effect estimation in Massively Parallel Reporter Assays

Albert Xue

UCLA

Massively Parallel Reporter Assays (MPRAs) interrogate the regulatory function of thousands of candidate enhancers in parallel through linked DNA and RNA readouts with an engineered construct and attached minimal reporter. Given the complexity of MPRA experimental designs, several different sources of uncertainty complicate inference. We show that previous methods do not account for substantial differences in uncertainty levels between the DNA and RNA counts and between batches. Accordingly, we present keju, a hierarchical statistical model that estimates enhancer transcription rate, differential activity between conditions, and minimal promoter effects on transcription rate for MPRA data. To maximize statistical power, keju conditions on the DNA counts to model batch-specific and modality-specific uncertainty in the RNA counts. keju shows vastly improved sensitivity (59%) in simulations compared to previous methods (31% for MPRAnalyze and 9% for BCalm). keju also has lower, more robust false positive rates, calling only 6.8% of unlabeled negative controls significant in real data (compared to 34% for MPRAnalyze and 12% for BCalm).

General moment closure for the neutral two-locus Wright-Fisher dynamics

Sean Yetter

University of Chicago

The Wright-Fisher diffusion and its dual process, the coalescent, are at the core of many results and methods in population genetics. Approaches have been developed to study the dynamics of the moments of the diffusion under genetic drift, mutation, and recombination using ordinary differential equations (ODEs). The dynamics of these moments can be used to study population genetic processes and are key building blocks of efficient methods to infer population genetic parameters, like demographic histories or fine-scale recombination rates. They can also be used to characterize linkage disequilibrium, which is essential for accurate interpretation of genome-wide association studies. However, the system of ODEs does not close under recombination; that is, computing moments of a certain order requires knowledge of moments of higher orders, thus, the system cannot be solved directly. By applying a coordinate transformation to the diffusion generator, we show that the canonical moments in these alternative coordinates yield a closed system. Compared to previous approaches in the literature, we believe that this approach can be readily extended to more general scenarios. Through simulations, we verify that the derived closed system accurately captures the dynamics of the moments, and can be used to efficiently compute expected diversity and linkage statistics in population genetic samples.

Cross-species gene expression encodes signatures of gene essentiality

Matteo Zambon

CRG, Centre for Genomic Regulation

Evolutionary constraint at the sequence level has long been central to understanding the genetic basis of disease. Measures of loss-of-function tolerance in human populations have proven powerful indicators of dosage sensitivity and essentiality, while comparative sequence conservation across species underpins most models for variant pathogenicity.

While correctly identifying genes critical for organismal fitness, sequence constraint provides limited insight into disease mechanisms or the tissues in which pathology manifests. In this regard, transcription offers a complementary perspective by embedding genes within their functional and regulatory contexts under which they are exposed to evolutionary forces. Variation in expression levels across tissues and species reflects how selective pressures shape phenotype, yet its relationship to gene essentiality and disease relevance remains poorly characterized.

Here, we develop a modeling framework leveraging cross-species transcriptional profiles across major tissues from multiple vertebrate species to derive gene-level representations that interpret sequence constraint through expression-based phenotypic and regulatory context. Adopting a semi-supervised approach, this architecture enables the prediction of gene essentiality, the prioritization of disease-relevant tissues, and the probing of the relationship between sequence-based and regulatory constraint. We find that cross-species expression patterns can be used to predict sequence constraint, suggesting a systematic coupling between the two. These findings support the hypothesis that sequence and expression constraint capture distinct yet complementary aspects of evolutionary pressure, and their integration provides a richer framework for interpreting gene essentiality and the molecular basis of hereditary disease.

A flexible empirical Bayes framework for inference of large-scale functional effects with application to dynamic eQTL analysis

Ziang Zhang

University of Chicago

In this work, we introduce functional adaptive shrinkage (FASH), a novel Empirical Bayes (EB) method for the joint analysis of genomic observation units where each unit measures an effect function at several levels of a continuous condition variable. The method is motivated by dynamic expression quantitative trait loci (eQTL) studies, which seek to characterize how genetic effects on gene expression vary continuously over conditions such as time. FASH integrates a broad family of Gaussian processes, defined through linear differential operators, into an empirical Bayes framework, enabling improved estimation of effect functions and providing principled measures for large-scale hypothesis testing, including false discovery and false sign rates. To further ensure conservative inference, we develop a Bayes factor–based adjustment to the FASH prior that can be incorporated into any EB shrinkage procedures with negligible computational cost. We illustrate FASH in a re-analysis of a dynamic eQTL dataset spanning 16 days of cardiomyocyte differentiation from induced pluripotent stem cells, where it identifies novel dynamic eQTLs, reveals diverse temporal effect patterns, and offers improved power and flexibility for characterizing effect functions compared with the original parametric interaction analysis. Beyond the context of dynamic eQTLs, FASH provides a general framework for the joint analysis of large-scale functional effects, with potential applications well beyond genomics and a readily available implementation in an open-source R package.

Identification of early exhaustion-specific regulatory programs in T Cells

Xinru Zhang

Gladstone Institutes, UCSF

T cell exhaustion is a dysfunctional state characterized by reduced proliferative capacity and impaired effector functions, including reduced ability to eliminate virus-infected or cancerous cells. Exhaustion represents a major barrier to effective immunotherapy. While terminally exhausted T cells are largely refractory to functional restoration, accumulating evidence suggests that early exhausted T cells retain a degree of plasticity and may be reverted to functional effector states. Characterizing the DNA sequences and upstream transcription factors that regulate early exhaustion would open the door to therapeutic epigenetic editing of T cells.

Towards this goal, we used single-cell multiomics data to train sequence-to-activity models for different T cell subtypes and states. After confirming model accuracy on held-out data, we performed a systematic screen to comprehensively identify regulatory regions uniquely accessible in early exhausted T cells. Then, we conducted in silico perturbation analyses to (1) identify transcriptional regulators binding these sequences, (2) assess how sensitive chromatin accessibility is to the binding motifs of these transcription factors, and (3) predict of perturbation of the regulatory elements could modulate the expression of nearby genes. Our findings provide new insights into the regulatory landscape underlying early T cell exhaustion and may inform strategies to overcome T cell dysfunction in cancer immunotherapy. More broadly, this project demonstrates the power of cell type specific deep learning models to inform genetic and epigenetic editing experiments.

Exploring selective scanning by LD statistic Dz: simulations and empirical studies

Alouette Zhang

McGill University

The genetic diversity seen in present-day organisms reflects a long history of evolution. Selective sweeps are key evolutionary events in which beneficial variants increase rapidly in frequency and reach fixation in the population. Many existing methods leveraged the excess linkage disequilibrium (LD) pattern left by sweeps for detection. In this project, we re-visited the LD statistic Dz to characterize its behaviour under classical selective sweeps. Originally introduced by Hill and Robertson as a stepping stone to compute D{ˆ}2, Dz was defined to compute the degree of LD between low frequency variants and has recently emerged as a useful empirical measure of diversity. We show that the expectation of Dz is informative for recent and ancient sweeps, producing a distinct “volcano” shape that persists over time near the selected locus. Using simulations with human-like parameters, expected Dz retains substantial power to detect sweeps occurring up to 60 thousand years ago, overlapping with the out-of-Africa event. Applying expected Dz in a scanning window to the 1000 Genomes Project data recapitulates well-studied sweeps and identifies candidates of early sweeps shared across AFR, EUR, and EAS populations.

Together, these results demonstrate Dz’s potential for incorporation into the current line of sweep detection techniques, taking advantage of genetic patterns not considered by most current methods.

Beyond purifying selection: A quantitative genetic model of background selection due to stabilizing selection

Sharon Zhou

University of Chicago

Linked selection substantially shapes patterns of genetic variation across the tree of life. Motivated by recent evidence in human genetics that stabilizing selection is widespread in complex traits, we ask how stabilizing selection contributes to linked selection. As a first step towards understanding the full effect of stabilizing selection on linked variation, we study underdominance as a proxy for the allele-frequency dynamics induced by stabilizing selection. We extend a quantitative genetic framework for linked selection, first developed by Santiago & Caballero, to underdominant selection and derive analytical expressions for the expected reduction in neutral diversity. Surprisingly, forward simulations of linked underdominant selection show a significantly stronger reduction in neutral diversity than predicted by the theory. We hypothesize that this may be the result of (1) frequency-dependent conversion between additive and non-additive (dominance) variance under nonadditive selection, which may violate key assumptions of the framework, and (2) positive LD among minor alleles that inflates additive variance.

Genetic architectures of human complex traits reveal stronger selective constraint on variants for brain-related traits

Huisheng Zhu

Stanford University

Genome-wide association studies (GWAS) have identified hundreds of significant loci for some psychiatric disorders, yet the strength of these associations remains modest compared to other human complex traits with similar numbers of hits. Whether this pattern reflects statistical artifacts or real biological differences — and, if the latter, what underlies it — remains unclear. In addition to psychiatric disorders, we find that traits with functional enrichment in the central nervous system (CNS), whether binary or quantitative, share similar genetic architectures, characterized by GWAS hits of limited statistical significance and generally higher allele frequencies. To robustly compare traits that differ in GWAS statistical power, we demonstrate how binarizing a quantitative trait reduces power. This loss of power can be replicated by downsampling the same quantitative trait to a matched "effective sample size". After matching "effective sample sizes", we show that CNS-enriched traits have large mutational target sizes, with contributing variants and genes experiencing stronger selection than those for other traits. Our findings reveal heterogeneity among diseases and provide insights into traits that more effectively capture fitness-relevant processes.