

54th European Mathematical  
Genetics Meeting  
Davos, 2026



---

# Imprint

**EMGM Davos 2026**

**Book of Abstracts**

**54th European Mathematical Genetics Meeting**

**Davos, Switzerland**

**14–15 April 2026**

**Edited by**

Cristian Riccio, Inés del Carmen Mora García, Andreas Ziegler

**Published by**

Cardio-CARE AG

**Place and date of publication**

Davos, Switzerland, 9 April 2026

**How to cite this volume**

Cristian RICCIO, Inés del Carmen MORA GARCÍA, and Andreas ZIEGLER, editors. *Book of Abstracts: 54th European Mathematical Genetics Meeting*. Davos, Switzerland: Cardio-CARE AG; 2026.

**Disclaimer**

This volume includes the abstracts submitted by the authors, accepted following review by the Scientific Committee, and corresponding to contributions whose authors had completed their registration for the conference. Editorial changes were limited to formatting, stylistic harmonisation, standardisation of affiliations, and correction of major spelling and grammatical errors. The authors remain responsible for the content of their abstracts.

**Copyright**

© 2026 Cardio-CARE AG for the compilation and layout.

Copyright of the individual abstracts remains with the respective authors.

---

## Committees

### Scientific Committee

**Robin Hofmeister**

Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

**Vivian Link**

Cardio-CARE, Medizincampus Davos, Davos, Switzerland

**Cristian Riccio**

Cardio-CARE, Medizincampus Davos, Davos, Switzerland

**Aleksejs Sazonovs**

Center for Molecular Prediction of Inflammatory Bowel Disease, Aalborg University, Aalborg, Denmark

**Pascal Schlosser**

Institute of Genetic Epidemiology, Department of Data Driven Medicine, Faculty of Medicine and Medical Center, Albert Ludwig University of Freiburg, Freiburg im Breisgau, Germany

**Sebastian Schönherr**

Institute of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria

**Adriaan van der Graaf**

Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

**Andreas Ziegler**

Cardio-CARE, Medizincampus Davos, Davos, Switzerland

### Organising Committee

Pascale Dillier, Vivian Link, Cristian Riccio, Anna Schuster, Andreas Ziegler

All members of the Organising Committee are affiliated with Cardio-CARE, Medizincampus Davos, Davos, Switzerland.

---

---

# Oral Presentations

## **Using Genomic structural equation models, genome wide association study data and simple clinical measures to provide an improved metric of obesity**

Daeun Kim<sup>1</sup>, Chi Zhao<sup>2</sup>, Emmaleigh Wilson<sup>3</sup>, Baiyu Qi<sup>1</sup>, Anne Justice<sup>4</sup>, Tuomas O. Kilpeläinen<sup>5</sup>, Cecilia Lindgren<sup>6</sup>, Stravuola Kanoni<sup>7</sup>, Ruth Loos<sup>8</sup>, Sonja I. Berndt<sup>9</sup>, Laura M. Raffield<sup>3</sup>, Kari North<sup>10</sup>, Kristin Young<sup>10</sup>, Karen L. Mohlke<sup>3</sup>, Gina M. Pelosso<sup>11</sup>, Cassandra N. Spracklen<sup>2</sup>, Mariaelisa Graff<sup>1</sup>, Tim M. Frayling<sup>12,\*</sup>

<sup>1</sup>Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, USA, <sup>2</sup>Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, Amherst, USA, <sup>3</sup>Department of Genetics, University of North Carolina, Chapel Hill, USA, <sup>4</sup> Department of Population Health Sciences, Geisinger College of Health Sciences, Geisinger, Danville, USA, <sup>5</sup>Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark, <sup>6</sup>Medical and Population Genetics Program & Type 2 Diabetes Systems Genomics Initiative, Broad Institute of MIT and Harvard, Cambridge, USA, <sup>7</sup>Clinical Pharmacology and Precision Medicine, William Harvey Research Institute, Faculty of Medicine, Queen Mary University of London, London, United Kingdom, <sup>8</sup>Department of Health and Medical Sciences, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark, <sup>9</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, USA, <sup>10</sup>Department of Epidemiology, School of Public Health, Brownsville, USA, <sup>11</sup>Department of Biostatistics, School of Public Health, Boston University, Boston, USA, <sup>12</sup>Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland, \*presenting author

People of the same weight, sex, age and ancestry can have very different risks of obesity related disease, especially metabolic diseases such as type 2 diabetes and heart disease. These differences have led to proposals to improve the definition of obesity so that it includes a better measure of disease risk. Previous epidemiological studies have incorporated cut offs for waist circumference and circulating factors such as glucose and triglyceride levels to define “metabolic syndrome”. However, relying on thresholds for multiple risk factors wastes information when those risk factors are continuous variables. Genomic structural equation models have become a popular and robust method to incorporate information from multiple individual phenotypes using summary data from large genome wide association studies (GWAS). For example, a recent study used summary

GWAS data from seven metabolic traits and 4.9 million people to define a common latent trait for metabolic syndrome. However, no such approaches have been designed to explicitly capture the adverse effects of excess weight.

In this presentation I will discuss our Genomic SEM approach to provide a better measure of the adverse effects of excess weight compared to BMI. I will discuss how we used large scale summary GWAS data from four key traits, BMI, waist hip ratio, circulating triglycerides and circulating High density lipoprotein levels, and Genomic SEM to generate a latent trait for metabolically “unfavorable adiposity”. I will describe how we used sex and ancestry stratified data and will show how we tested the genetics of “unfavorable adiposity” in independent UK Biobank data. Finally, I will describe how we used downstream GWAS approaches to characterize the genetics of this trait and compare it to BMI.

## **Improving GWAS power by exploiting phenotype severity**

Jasper Hof<sup>1</sup>, Chao Ning<sup>1</sup>, Liam Quinn<sup>2</sup>, and Doug Speed<sup>1</sup>

<sup>1</sup>Aarhus University, Aarhus, Denmark, <sup>2</sup>Zealand University Hospital, Køge, Denmark

**Introduction:** common complex diseases are clinically heterogeneous, yet most genome-wide association studies (GWAS) assume genetic homogeneity among cases. Growing evidence suggests that phenotypic heterogeneity reflects underlying variation in genetic architecture. To address this, we developed StratGWAS, a scalable framework that leverages clinically relevant measures of heterogeneity to construct a new phenotype that better reflects genetic liability within diseases.

**Methods:** StratGWAS stratifies cases using auxiliary variables (e.g., age at onset and medication burden), estimates genetic covariance between strata, and derives an optimally weighted transformed phenotype that maximizes heritable signal. This enables GWAS to assign greater weight to individuals with higher genetic liability, creating a more powerful phenotype. We evaluated the performance of StratGWAS through simulations (N = 100k) and application to 21 common traits in the UK Biobank (N = 368k).

**Results:** In simulations, StratGWAS consistently outperformed existing methods, such as ADuLT and GenomicSEM, with greatest power gains observed when shared genetic liability between auxiliary and target variable was high. StratGWAS also identified scenarios where the auxiliary variable introduces false positives by computing an “inflation criterion”, which correlated highly ( $\rho = 0.985$ ) with type 1 error. Applied to 21 common diseases in the UK Biobank, StratGWAS upweighted subgroups with earlier onset and higher medication burden, reflecting severe cases with greater genetic burden. This resulted in a 17% increase in independent genome-wide significant loci compared to standard GWAS.

**Conclusions:** StratGWAS provides a powerful approach to incorporate clinical heterogeneity into genetic studies, improving locus discovery and enabling more precise dissection of disease architecture beyond case-control phenotypes.

## **Limitations and alternatives for genome-wide association studies (GWAS) for complex traits**

Stefanie Muff\*<sup>1</sup> and Gard Gravdal<sup>1</sup>

<sup>1</sup> Department of Mathematical Science, Norwegian University of Science and Technology, Trondheim, Norway

The identification of genetic markers associated with phenotypic traits remains a key objective in medical research, as well as in animal and plant breeding and the study of wild animal and plant systems. Genome-wide association studies (GWAS) are the standard tool to identify genomic regions associated with a phenotypic trait of interest. Despite the success of GWAS, the identified loci usually only explain a small percentage of the genetic variation, known as the problem of “missing heritability”. Despite the fact that p-values blend the estimated effect size with the uncertainty of the estimate, which renders them into an unreliable measure for variable importance, GWAS are exclusively based on p-values. In addition, it has become clear that the genetic architecture of complex traits is often even more complex than originally thought. p-value thresholding to find “significant” genomic markers (e.g., SNPs) has therefore intrinsic limitations, and these are aggravated when sample sizes are small. Here, we investigated alternative measures to identify associations between SNPs and phenotypic traits. In contrast to GWAS, we used models designed for genomic prediction that jointly include all SNP effects, and then calculated variable importance from approximations of so-called “Shapley values”, which were constructed to capture the contribution of each individual variable to the amount of variance explained by the total model. We did this for two types of approaches, namely for 1) a machine learning-based methods using boosted regression trees, and 2) Bayesian linear models, in particular BayesR. Advantages and limitations are discussed.

## **Evaluating generalizability of causal gene prioritization with biologically informed datasets**

Leonhard Kohleick\*<sup>1</sup>, Kurt Herzog<sup>1</sup>, Martijn Zoodma<sup>1</sup>, Benjamin Wild<sup>1</sup>, Claudia Langenberg<sup>1,2</sup>, and Maik Pietzner<sup>1,2</sup>

<sup>1</sup>Berlin Institute of Health, Berlin, Germany, <sup>2</sup>Precision Healthcare University Research Institute (PHURI), Queen Mary University of London, London, United Kingdom

Assigning the causal gene at the thousands of GWAS loci reported to date remains a major bottleneck for clinical translation. Because truly representative gold-standard variant-to-gene (V2G) datasets are scarce, recent prioritization frameworks have improved performance by training on hybrid labels that combine manual curation with data-derived sets. Many of these strategies implicitly assume that proximity to coding variants helps identify informative functional annotations for nearby, distinct credible sets. While useful, this assumption may not reflect the full diversity of biological mechanisms that drive V2G causality across diseases, biomarkers, intermediate phenotypes, and health-related behaviors. We address this limitation by creating three biologically grounded (rather than proximity based) benchmarks for gene prioritization: 105 fine-mapped trans-pQTL loci where the putative causal gene encodes an interaction partner of the trans-associated protein, 29 GWAS loci implicating FDA-approved drug target genes, and 28 fine-mapped mQTL credible sets supported by multiple lines of evidence.

We find that recently proposed prioritization methods perform poorly on these biologically informed benchmarks, although retraining on similar feature sets improves performance. Overall, models trained under one “gold-standard” labelling assumption generalize weakly to orthogonal high-confidence V2G assumptions, with an average reduction of 0.2 (range 0.16– 0.43) in area under the precision–recall curve. Consistent with this, analysis of large-scale pan-biobank results shows that annotator choice can lead to divergent biological conclusions. These results highlight the need for diverse evaluation criteria to develop more robust and generalizable gene prioritization methods.

## Deconvoluting linkage disequilibrium to refine GWAS associations

Martin Tournaire\*<sup>1,2</sup>, Asma Nouira<sup>2,3</sup>, Mario Favre-Moiron<sup>1,2</sup>, and Marie Verbanck<sup>1</sup>

<sup>1</sup>Oncologie Computationnelle, Institut Curie, Paris, France <sup>2</sup>UAR3612 CNRS, US25 Inserm, Team BioSTM, Université Paris Cité, Centre national de la recherche scientifique, CNRS, Institut National de la Santé et de la Recherche Médicale (Inserm), Faculté de pharmacie de Paris, Paris, France, <sup>3</sup>Radiobiology Laboratory for Accidental Exposure, Autorité de Sécurité Nucléaire et de Radioprotection (ASNR), PSE-SANTE, SERAMED, LRMed, Montrouge, France

Genome-wide association studies (GWAS) have identified over a million associations between genetic variants and complex traits. However, the elucidation of the “causal” variant and its biological interpretation remains a formidable challenge, despite being crucial for downstream analyses. This difficulty is notably driven by linkage disequilibrium (LD), which induces correlation between physically close variants. Fine-mapping methods, mostly based on Bayesian frameworks, were developed to prioritize likely causal variants. However, they only output probabilities and are highly sensitive to misspecified priors such as the genomic window.

In contrast, we present an LD-aware summary-statistic framework that explicitly deconvolutes LD to recover more interpretable variant-trait effect estimates. We first construct a stabilized pseudo-inverse of the LD matrix using block-wise truncated singular value decomposition (TSVD). By sliding across LD blocks, this step stably inverts LD structure even in high-correlation regions. We then compute LD-adjusted variant effect estimates by applying this pseudo-inverse directly to the GWAS summary statistics.

We first validated the approach in large-scale simulations varying heritability, polygenicity, and sample size. LD deconvolution consistently improved precision for identifying causal variants and reduced LD-induced false positives relative to GWAS, with a trade-off of reduced recall. We then applied the method to UK Biobank summary statistics for lipid and metabolic traits. For circulating triglycerides, LD deconvolution prioritizes variants in biologically established genes such as *APOE* and *APOC1*, while deprioritizing less plausible correlated variants. By deconvoluting LD, this framework provides LD-adjusted variant

effect estimates that strengthen the interpretability and reliability of downstream post-GWAS analyses, such as Mendelian randomization.

## **Using large language models for rare variant association testing in large-scale biobanks**

Joelle Mbatchou<sup>1</sup>, Christopher Gillies<sup>1</sup>, Andrey Ziyatdinov<sup>1</sup>, Jack Kosmicki<sup>1</sup>, Manuel Allen Revez Ferreira<sup>1</sup>, Goncalo Abecasis<sup>1</sup>, Lukas Habbeger<sup>1</sup>, Maya Ghousaini<sup>1</sup>, Jonathan Marchini<sup>1</sup>

<sup>1</sup>Regeneron Genetics Center, Tarrytown, United States of America

The application of whole exome sequencing in studying of rare genetic variation has been well-established as a powerful and cost-effective strategy for novel drug target discovery. The study of rare genetic variation, potentially important in the development of complex diseases, has been increasingly performed thanks to advances in sequencing technologies. Gene-based tests have been developed to address the challenges with single variant tests caused by the rarity of these variants and the need for large sample sizes. These tests aggregate information across many variants and can integrate external functional annotations to improve the power of rare variant analysis. In recent years, large language models (LLMs) have been used to predict the functional impact of genetic mutations, potentially enhancing the power of rare variant association tests, and complementing functional prediction approaches based on in-silico algorithms. We showcase the integration of functional scores leveraging LLMs for large-scale gene-based association testing in the UK Biobank, highlighting their potential to improve the detection of rare variant associations and advance our understanding of complex genetic diseases.

## Identifying shared genetic associations in fibrosis: a multi-organ rare variant analysis

Dominic Sayers<sup>1,2</sup>, Ebrima Joof<sup>1,2</sup>, Nick Shrine<sup>1,2</sup>, Georgie Massen<sup>3</sup>, Jennifer Quint<sup>3</sup>, Hilary Longhurst<sup>4</sup>, Olivia Leavy<sup>1,2</sup>, Gisli Jenkins<sup>5</sup>, Louise Wain<sup>1,2</sup>, Katherine Fawcett<sup>1,2</sup>, Richard Allen<sup>1,2</sup>

<sup>1</sup>Division of Public Health and Epidemiology, University of Leicester, Leicester, United Kingdom, <sup>2</sup>NIHR Leicester Biomedical Research Center, University of Leicester, Leicester, United Kingdom, <sup>3</sup>School of Public Health, Imperial College London, London, United Kingdom, <sup>4</sup>DC action, London, United Kingdom, <sup>5</sup>National Heart Lung Institute, Imperial College London, London, United Kingdom

Fibrotic diseases contribute to one-third of global mortality. Identifying genes that harbour rare variation associated with fibrosis across multiple organ systems could reveal shared pathological mechanisms and novel therapeutic targets. We assigned fibrotic diseases to 12 organ-systems for analysis (1): biliary, cardiovascular, diabetes, skin, intestinal/pancreatic, liver, pulmonary, reproductive, skeletal, systemic, renal, and lymphatic. We included 417,814 individuals of European ancestry with whole-exome sequencing data from UK Biobank, defining cases using hospital episode statistics and mortality records and selecting 10 controls-to-case matched by age and sex. Rare variant gene-based collapsing analyses were performed using Regenie (v3.5) separately in each organ-system. An omnibus p-value was calculated using ACAT to combine ACAT-V, SKAT-O, and burden tests (2). Significant genes were those reaching a Bonferroni corrected p-value of  $8.97 \times 10^{-5}$ . Significant genes were investigated in the other organ groups with significance defined as those reaching a Bonferroni corrected threshold for the number of genes followed up. Replication of this study is being performed in the All of Us biobank.

In total, 89 unique significant genes were identified across all phenotypes. *PKD1* was significantly associated with liver and renal fibrosis and reached Bonferroni corrected significance with diabetes. *PKD2* and *OR6C70* also reached Bonferroni corrected significance with liver and renal.

Taking an organ-level approach to improve power for rare variant collapsing analyses, I identified 3 variants (*PKD1*, *PKD2*, and *OR6C70*) that showed significance in multiple fibrotic

phenotypes, suggesting potential shared genetic pathways with fibrosis across different organ systems. These genes may represent promising targets for future therapeutic interventions in fibrotic diseases.

## Rare variant burden tests in the Fenland cohort

Jack Murzynowski<sup>1</sup>, Nicola Kerrison<sup>1</sup>, Ken Ong<sup>1</sup>, Nicholas Wareham<sup>1</sup>, and John Perry<sup>1</sup>

<sup>1</sup>MRC Epidemiology Unit, School of Clinical Medicine, Institute of Metabolic Science, University of Cambridge, Cambridge, United Kingdom

Large-scale whole-exome sequencing (WES) studies in population biobanks show that rare protein-coding variants can have large effects on human phenotypes. However, many associations remain difficult to replicate, and the size and design of most biobanks limit deep phenotyping and genotype-based follow-up. To address this, we present new WES in the Fenland cohort, a population of 11,458 individuals with unique phenotypes. Using these data, we developed a comprehensive gene-burden testing pipeline and applied it to identify rare-variant associations with metabolic phenotypes unique to Fenland. Here, we describe the quality control and framework underpinning the burden testing pipeline and report novel gene-trait associations including *BRSK2* with fasting glucose, *NID2* with post-oral glucose tolerance test (OGTT) glucose, *PC* with post-OGTT insulin, and *CASQ1* with resting energy expenditure. We also replicate known signals including *G6PC2* with fasting glucose, and *MAP3K15* with HbA1c. Together, these results position Fenland as a resource for both replication and discovery, with recall-by-genotype enabling future functional follow-up.

## **Inferring stabilizing selection acting on protein levels using rare variant associations**

Mihaela Diana Zanoaga\*<sup>1</sup> and Zoltán Kutalik<sup>1</sup>

<sup>1</sup>University of Lausanne, Lausanne, Switzerland

**Introduction:** Stabilizing selection shapes the level of molecular readouts, such as protein abundance, yet remains difficult to quantify. State-of-the-art approaches designed for complex human traits yield unintuitive results when applied to molecular traits.

**Data and Methods:** Leveraging the observation that, under stabilizing selection, large burden effects must be rare, we devised a new maximum-likelihood-based method that infers selection by linking the frequency and the effect size of rare loss-of-function variant burden of ~18'000 genes on ~3'000 proteins in ~54'000 participants in the UK Biobank. We compared our rare-variant-based estimates with those inferred from common variants using LDpred2. We also evaluated the selection coefficients against independent measures of genetic and evolutionary constraints, including protein heritability, loss-of-function intolerance (pLI), and cross-species transcript evolutionary rates.

**Results:** Our framework identified 397 proteins under selection ( $p < 0.05$ ), 92% of which exhibited signatures of stabilizing selection. In line with theoretical expectations, stronger inferred selection was associated with lower protein heritability ( $\rho = 0.20$ ,  $p = 5.3 \times 10^{-6}$ ). Our estimates correlated well also with loss-of-function intolerance scores ( $\rho = 0.19$ ,  $p = 1.2 \times 10^{-6}$ ), and cross-species transcript evolutionary rates ( $\rho = 0.14$ ,  $p = 0.05$ ). We finally observed a mild, but nominally significant correlation with common-variant-based estimates by LDpred2 ( $\rho = 0.059$ ,  $p = 0.019$ ). Interestingly, proteins under strongest inferred selection, including ARSA and TPP1, were enriched in lysosomal pathways, compatible with their essential, dosage-sensitive cellular functions.

**Conclusions:** Our likelihood-based rare-variant burden framework reveals stabilizing selection through rare variants, providing a complementary and underexplored lens on the evolutionary constraints shaping molecular traits.

# **Omics and Mendelian Randomization: A Journey into Biological Mechanisms**

Eleonora Porcu<sup>1</sup>

<sup>1</sup>Nestlé Institute of Health Sciences, Lausanne, Switzerland

The increasing number of studies incorporating data across multiple biological levels, such as genomics, transcriptomics, proteomics, and metabolomics, raises critical biomedical questions regarding the systematic integration of these diverse datasets to uncover new biological mechanisms that elucidate the processes of health and disease. Statistical causal frameworks, particularly Mendelian randomization (MR), provide a robust foundation for integrating these data and facilitating novel biological discoveries.

Mendelian randomization emerges as a valuable strategy for examining causality within complex biological and omics networks, offering insights that can inform drug development and prioritize intervention targets for disease prevention. Among the various omics fields, transcriptomics is the most extensively studied, with numerous investigations employing MR to identify causal genes associated with complex traits and to differentiate between mere correlations and true causal effects.

Recent advancements in multi-omics MR approaches have further integrated additional omics layers, allowing researchers to explore the biological pathways underlying gene-trait relationships. As these multi-omics methodologies become increasingly prevalent, a more systematic approach is essential to manage the growing complexity of data. Although combining MR results with observational data can enhance the robustness of causal inferences and provide a more comprehensive understanding of the relationships between exposures and outcomes, biological validation remains a critical step in confirming findings derived from these analyses.

While MR provides powerful tools for understanding causal relationships, especially when randomized controlled trials are not feasible, the interpretation of results must be approached with caution. Here I will highlight the potential of multi-omics integration and MR in advancing our understanding of health and disease, emphasizing the importance of rigorous validation and careful interpretation in the pursuit of biological insights.

## **Rare variant burden Mendelian randomization provides orthogonal evidence for protein causal effects**

Théo Cavinato<sup>\*1,2</sup>, Robin Hofmeister<sup>1,2,3,4</sup>, Mihaela-Diana Zanoaga<sup>1,2</sup>, Adriaan Van Der Graaf<sup>1,2,3</sup>, and Zoltán Kutalik<sup>1,2,3</sup>

<sup>1</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, <sup>2</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland, <sup>3</sup>Center universitaire de médecine générale et santé publique, Lausanne, Switzerland, <sup>4</sup>Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

Large-scale exome sequencing in population biobanks enables the systematic study of rare variant effects. Rare loss-of-function variants can be aggregated into gene-level burden scores, providing proxies for gene inactivation. Here, we used burden scores as a novel class of instrumental variables for Mendelian randomization (MR) to estimate causal effects between proteins, and compared them to cis pQTL-based MR. We performed burden testing across all gene-protein expression pairs in the UK Biobank. Burden scores were constructed using high-confidence loss-of-functions and an identified optimal minor allele frequency (MAF) threshold balancing variant rarity and statistical power. Burden scores significantly associated with their corresponding protein expression were used as instruments in burden-MR. A MAF threshold of  $1 \times 10^{-3}$  maximized significant negative burden effects (747 genes), but a more stringent threshold of  $1 \times 10^{-4}$  was selected to ensure independence from pQTLs. At this threshold, 732 genes showed significant effects on their corresponding protein, 98% of which were negative and were significantly stronger than their effect on unrelated proteins. The causal effects estimated by burden-MR and pQTL-MR were significantly correlated, with a Pearson correlation ranging from 0.70 to 0.88, increasing as the burden-MR significance threshold decreased. The signs of the effects were significantly more concordant than expected by chance (one-sided binomial-test; ORs > 580.5). We introduced Burden-MR, an orthogonal line of causal evidence for molecular exposures, yielding estimates concordant with pQTL-MR. This approach can be extended to complex traits as exposures or outcomes, providing independent evidence for causal inference in a wide range of settings.

# **Separating social influence from shared geographic environments in complex traits: A spatial Mendelian randomization approach**

Tabea Schoeler\*<sup>1</sup> and Zoltan Kutalik\*<sup>1</sup>

<sup>1</sup>University of Lausanne, Lausanne, Switzerland

Health and behavior are shaped by both geographic environments (e.g., healthcare access) and social interactions. Because both influences induce phenotypic similarity among nearby individuals, separating social transmission from geographic confounding is challenging. To estimate the influence of an individual's trait ( $Y_j$ ) on the same trait in another individual ( $Y_i$ ), we developed a spatial Mendelian Randomization (MR) framework incorporating genetic instruments ( $G$ ). Assuming that nearby individuals are more likely to influence each other, we quantified social interaction as a function of distance decay:  $Y_i \sim \lambda \cdot \exp(-\rho \cdot d_{ij}^2) \cdot G_j$ , where the effect ( $\lambda$ ) between two individuals at minimal geographic (Euclidean) distance ( $d_{ij}=0$ ) diminishes at a rate of  $\rho$  as distance increases. We validated the framework in simulations and applied it to 43 geographically clustered traits (e.g., BMI, lifestyle behaviors) in the UK Biobank. Correlational analyses identified social interaction effects for  $k=35$  traits, whereas MR analyses detected fewer ( $k = 6$ ) and substantially attenuated effects. For example, alcohol use showed strong evidence of social interaction in correlational analyses, but markedly attenuated estimates in MR analyses ( $\rho = 6.61 \times 10$  vs.  $1.81 \times 10$ ), suggesting a substantial contribution from shared geographic environments. Traits identified only in correlational analyses (e.g., smoking) are therefore likely driven by geographic confounding rather than social influence. Overall, spatially informed MR identifies social transmission effects independent of shared geography while demonstrating that geographic context plays a dominant role in shaping complex traits. Accounting for spatial confounding is therefore essential when studying social interaction effects in geographically connected populations.

## **Large-scale mapping of polygenic risk to disease phenotypes in the Estonian Biobank**

Merli Koitmäe<sup>\*1,2</sup>, Triin Laisk<sup>2</sup>, Jelisaveta Dazigurski<sup>2</sup>, Oliver Aasmets<sup>2</sup>, Krista Fischer<sup>2,1</sup>, Reedik Magi<sup>2</sup>, and Kristi Läll<sup>2</sup>

<sup>1</sup>Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia,

<sup>2</sup>Institute of Genomics, University of Tartu, Tartu, Estonia

The rapid expansion of freely available genomic resources has enabled large-scale analyses that were previously impossible. A prominent example is the PGS Catalog, which hosts over 5,000 polygenic risk scores (PGSs) for hundreds of phenotypes, providing unprecedented opportunities to investigate genetic predisposition across human diseases. In this study, we systematically assessed associations between all available PGSs and all ICD-10 diagnosis codes in the Estonian Biobank using high-throughput logistic regression. Analyses were performed under multiple scenarios, including adjusted (age, sex, BMI) and unadjusted models, as well as sex-specific and age group-specific analyses. Participants were analyzed separately in two recruitment waves of ~50,000 and 150,000 individuals, followed by meta-analysis to account for cohort differences. Bonferroni correction was applied based on number of independent PGSs and ICD-10 codes estimated through principal component analysis explaining 99.5% of variance. We identified substantial heterogeneity in PGS effects across subgroups, with numerous PGS-specific interaction effects. PGS × sex interactions were frequently observed for musculoskeletal and circulatory diseases, while cohort × PGS interactions were most prominent for digestive and endocrine/metabolic conditions. Age-stratified analyses revealed differential associations, with distinct patterns in younger individuals for health status factors and reproductive disorders, and in older individuals for circulatory and endocrine/metabolic diseases. Although covariate adjustment attenuated effect sizes, many subgroup-specific patterns remained. Overall, this comprehensive analysis maps PGS–disease associations at scale, highlighting clinically relevant subgroup-specific effects and potential modifiers of genetic risk. All results will be publicly released through a dedicated database to facilitate future research and precision medicine efforts.

## **Genetic features that predict susceptibility to immune-mediated diseases also predict ability to control viral load of chronic infections**

Wei-Yu Lin\*<sup>1</sup> and Chris Wallace<sup>1,2</sup>

<sup>1</sup>MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom, <sup>2</sup>Cambridge Institute of Therapeutic Immunology and Infectious Disease (CITIID), University of Cambridge, Cambridge, United Kingdom

Background: Our immune system protects us from pathogens, but immune dysregulation can lead to immune-mediated disease (IMD), such as multiple sclerosis (MS). Recent evidence that Epstein-Barr Virus (EBV) infection is causally associated with MS has linked these dual aspects of immunity. To date, shared risk factors in the MHC have been identified, but evidence outside the MHC has been inconclusive. We hypothesised that distinguishing different components of genetic risk could identify non-MHC shared genetics. Methods: We constructed 13 genetic features that capture patterns of genetic risk across different IMDs excluding the MHC using modified principal component analysis. These features can be thought of as partial polygenic scores. We projected summary statistics for independent IMD studies and viral load traits onto this basis to quantify and characterize shared genetic relationships. Results: Several features were associated with a variety of IMDs and viral load traits. EBV was genetically closest to autoimmune diseases such as rheumatoid arthritis. In contrast, most significant EBV features showed opposite effects on MS risk. Conversely, HHV7 showed concordance with MS across all shared features. Conclusion: A subset of components of genetic risk on IMD are associated with viral load, with autoimmune diseases showing the strongest overlap. Through reducing the multiple testing burden, we identified genetic features associated with both EBV and MS outside the MHC which act in opposing directions. Although EBV is causal for MS, more work is needed to determine direction for the shared associations we detect: disease may alter ability to control viral load.

## **A phenome-wide map of pleiotropy: connecting genetic signals to disease mechanisms and drug discovery**

Anna O'Carroll (née Hutchinson)<sup>1</sup>, Zhana Kuncheva<sup>2</sup>, Daniel Crouch<sup>1</sup>, Soroosh Afyouni<sup>2</sup>, Manju Dissanayake<sup>2</sup>, Peter Scanlon<sup>2</sup>, Adrian Cortes<sup>3</sup>, Chris Foley<sup>2</sup>

<sup>1</sup>GlaxoSmithKline, Stevenage, United Kingdom, <sup>2</sup>Optima Partners, bioXcelerateAI, Edinburgh, United Kingdom, <sup>3</sup> GlaxoSmithKline, Heidelberg, Germany

Genetic datasets are expanding, yet mechanistically understanding how genetic variation shapes human biology remains challenging. Existing methods like phenome-wide association studies (PheWAS) are confounded by linkage disequilibrium, while colocalisation is computationally intensive. We present a comprehensive genome- and phenome-wide map of pleiotropy, offering an in-depth view of shared causal architecture across the human phenome. We used a novel, scalable pipeline to analyse a wide range of data, including over 3000 traits from UK Biobank whole-genome sequencing (WGS) and VA Million Veteran Program genotyping data, as well as 2940 Olink proteins from UK Biobank WGS data. To overcome analytical bottlenecks, we developed Switch-Step, a fine-mapping algorithm more than 80 times faster than SuSiE, and Imp-Map, a summary statistic imputation method more than 50 times faster than Imp-G. These innovations, combined with cloud infrastructure, enabled genome-wide multi-trait colocalisation using HyPrColoc across more than 580,000 genetic signals. The resulting map identifies over 50,000 pleiotropic loci with high posterior probability ( $> 0.7$ ), showing a median of three colocalising traits (IQR 2 to 6). By integrating GWAS with expression, proteomic, and metabolomic QTLs, we prioritize causal genes and biomarkers, clarifying molecular links to complex diseases. This scalable framework has significant implications for drug development, including indication extension, repurposing, adverse association prediction, and identifying novel targets and biomarkers. Our approach exemplifies how method innovation, large-scale data, and scalable infrastructure can transform pleiotropy into a powerful discovery tool.

## **Leveraging cis- and trans-variants to improve protein expression level prediction for proteome-wide association studies**

Suzanne Leal\*<sup>1</sup>, Rui Dong<sup>1</sup>, Derek Lamb<sup>1</sup>, Gao Wang<sup>1</sup>, and Andrew Dewan<sup>2</sup>

<sup>1</sup>Columbia University, New York, United States, <sup>2</sup>Yale University, New Haven, United States

Since genetic effects are often mediated through proteins, the analysis of proteomic data can provide insights into disease etiology. However, most studies lack proteomic data. To address this problem, we developed TransCisPredict to perform proteome-wide association studies (PWAS) at a biobank scale. TransCisPredict reduces computational burden through linkage-disequilibrium block selection which facilitates incorporating cis- and trans variants to predict protein expression and performs protein-phenotype association analyses. To account for differences in protein regulatory architecture, four prediction methods are used for weight estimation, i.e., BayesR, Elastic Net, LASSO, and SuSiE. Five-fold cross-validation (CV) is used to select the optimal method for each protein. Weight-estimation was performed using White British UK Biobank study subjects (N=42,644) with proteomic and genotype-array data. Of the 2,920 available protein expression levels, 2,339 could be predicted with a CV-R<sup>2</sup>> 0.05 when cis- and trans-variants were used. When analysis was limited to cis-variation, expression levels could only be predicted for 466 proteins. A PWAS was performed for 2,339 predicted protein expression levels and type 2 diabetes (T2D) using White British UK Biobank study subjects without proteomic data (N=364,132) followed by two-sample Mendelian randomization using a method that controls for horizontal pleiotropy for validation. Forty proteins were associated with T2D and validated. For the 466 cis-only predicted protein expression levels, three proteins were associated with T2D and validated. Incorporating both cis- and trans-variation using TransCisPredict facilitates the prediction of many more proteins compared to using cis-only variants thereby increasing the power of PWAS.

## **Decoding the genetic basis of nurture using untransmitted parental alleles**

Leona Knusel<sup>1,2</sup>, Robin Hofmeister<sup>1,2,3,4</sup>, Liza Darrous<sup>5</sup>, and Zoltán Kutalik<sup>1,2,3</sup>

<sup>1</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, <sup>2</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland, <sup>3</sup>University Center for Primary Care and Public Health, Unisanté, Lausanne, Switzerland, <sup>4</sup>Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia, <sup>5</sup>Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland

Parental UnTransmitted Alleles (UTAs) can influence offspring traits through rearing mechanisms. Therefore, estimating these indirect genetic effects allows us to study how parental characteristics shape offspring traits. However, systematic evaluation of the effect of UTAs on complex traits has been limited due to the absence of data on UTAs. We developed a method for estimating UTAs using close relatives as surrogate parents to increase sample size. This enabled analysis of how UTAs affect offspring traits in a GWAS framework, followed by Mendelian Randomisation (MR) to investigate genetic rearing effects. In the UK Biobank, we inferred untransmitted alleles for ~130,000 individuals (vs ~5,000 in traditional parent-offspring studies). We tested associations between untransmitted parental alleles and 28 traits. The genetic correlation between UTA GWAS and population GWAS effects for the same trait was significantly smaller than 1 for 20 out of 28 traits, suggesting that rearing mechanisms do not only involve the same trait. MR results support this finding. For example, parental education ( $s = -0.10$ ,  $SE = 0.01$ ,  $p = 2.56e-13$ ), but not parental BMI ( $s = 0.02$ ,  $SE = 0.01$ ,  $p = 0.06$ ) significantly affects offspring BMI. Exceptions to this were geographic and SES related traits that seemed to be the main determinant of these same traits in the offspring. Overcoming the traditional requirement for parental genetic information by leveraging related samples in biobanks allowed us to evaluate UTAs at an unprecedented scale. Using those UTAs, we investigated how parental behaviour affects offspring traits independent of direct genetic effects.

# **Integrating variable number tandem repeats into association and fine-mapping reduces missing heritability: a framework for lipoprotein(a)**

Silvia Di Maio\*<sup>1</sup>, Johanna Franziska Schachtl-Riess<sup>1</sup>, Hansi Weissensteiner<sup>1</sup>, Lukas Forer<sup>1</sup>, Stephan Amstler<sup>1</sup>, Cathrin Pfuertscheller<sup>1</sup>, Anna Köttgen<sup>2</sup>, Florian Kronenberg<sup>1</sup>,  
Stefan Coassin<sup>1</sup>, and Sebastian Schönherr<sup>1</sup>

<sup>1</sup>Institute of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria, <sup>2</sup>Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau, Germany

Background: Variable number tandem repeats (VNTRs) are polymorphic human variants, with > 100 in protein-coding, medical genes. Genotyping arrays miss VNTR repeat number and intrarepeat mutations. Thus, GWAS overlook their effects and mischaracterize local LD, limiting finemapping. A prominent example is the KIV2 VNTR in *LPA* gene. *LPA* nearly monogenically regulates lipoprotein(a) (Lp(a)), major cardiovascular risk factor. KIV2 VNTR complicates *LPA* genetic characterization, with ancestry disparities, leaving Lp(a) heritability largely unexplained. Results: We called KIV2 intrarepeat variants and developed a coverage-based approach to estimate KIV2 copy number. In UK Biobank (UKB), KIV2 number explained ~30% of Lp(a) variance in White and Black participants. We implemented a locus-wide, repeat-aware association and finemapping framework that jointly evaluates SNPs within and outside KIV2 while adjusting for KIV2 number. In Whites (n=140,015), finemapping identified 61 likely causal variants (32 were in singleSNP credible sets). 11 SNPs were within KIV2. We recovered 11 known functional SNPs and revealed novel signals: two missense variants in nonrepetitive *LPA* regions and three KIV-2 SNPs (one missense and one splice mutation). In Blacks (n=2,037), 29 likely causal variants were finemapped (5 in KIV2), including unreported ancestry-specific missense (three) and splice (one) variants. Conclusion: Integrating VNTRs into association and finemapping uncovers ancestry-specific variants and improves causal resolution. This can improve polygenic scores and provides genetic instruments for Mendelian randomization to investigate Lp(a) causality across

ancestries, and assess safety and efficacy of Lp(a)-lowering therapies. Our framework offers a generalizable template for incorporating VNTRs into biobank studies.

## **Estimating the genetic basis of quantitative traits across populations**

Ekaterina Maksimova\*<sup>1</sup>, Ilse Krätschmer<sup>1</sup>, Gasper Tkacik<sup>1</sup>, and Matthew Robinson<sup>1</sup>

<sup>1</sup>Institute of Science and Technology Austria, Klosterneuburg, Austria

Genome-wide association studies have been disproportionately skewed towards analyses in the European population, which induces health disparities for underrepresented groups and hinders understanding of underlying genetic mechanisms. However, these issues are not resolved by simply increasing the diversity of participants, while applying conventional methodology that does not account for differences in allele frequencies, linkage disequilibrium (LD) patterns, and environmental exposures across populations. We propose an eigendecomposition– based multivariate Bayesian approach that jointly models effect sizes across the genome and across populations, explicitly incorporating population-specific allele frequencies, LD, and functional annotations. We evaluated both discovery and genomic prediction performance of this approach against state-of-the-art methods using simulated data and height data from the UK Biobank and All of Us. In simulations, we show that multivariate modelling of effect sizes enables accurate estimation of genetic correlations, improved fine-mapping of shared and population-specific effects, and increases cross-ancestry polygenic prediction accuracy by 23-535% in both EUR and non-EUR ancestries relative to the state-of-the-art multi-ancestry methods. In real data for height, our model estimates genetic correlations of 0.99 (EUR–AFR), 0.83 (EUR–AMR), and 0.63 (AFR–AMR) and improves test prediction accuracy by 5–150% in non-EUR ancestries and 5–9% in EUR group relative to other methods. Overall, we provide empirical evidence that discovery and genomic prediction are greatly enhanced by analyzing effect sizes across genome and populations jointly. Therefore, multivariate analysis incorporating multi-ancestry data should be commonplace, improving our ability to infer shared and population-specific genetic architecture patterns of human complex traits.

## **Improved biomarker selection and disease onset prediction in proteomics survival models via vector approximate message passing**

Jakub Bajzik\*<sup>1</sup>, Al Depope<sup>1</sup>, Yasaman Zolfimoselo<sup>2</sup>, Alexander Sharipov<sup>2</sup>, Marco Mondelli<sup>1</sup>, and Matthew Robinson<sup>1</sup>

<sup>1</sup>Institute of Science and Technology Austria (IST Austria), Klosterneuburg, Austria, <sup>2</sup>Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

For many common complex human diseases, incidence increases exponentially with age. In large-scale biobanks, linking time-to-diagnosis information to multiomics measures can reveal biological pathways involved in disease onset and progression. However, multiomics association testing is predominantly conducted using marginal Cox proportional hazards (CoxPH) models that test one variable at a time, ignoring correlation structure and inflating false discoveries. We introduce vampW, a novel parametric method based on Bayesian Vector Approximate Message Passing, which jointly models correlated features to select associated variables and predict disease onset age. Using the UK Biobank proteomics data (2,924 proteins from 53,018 participants), we first show in simulations that vampW outperforms standard and deep-learning forms of CoxPH variable selection and prediction methods. We then apply vampW to 24 observed health outcomes, identifying 219 protein associations, which are mostly not the top CoxPH marginal discoveries. The associated proteins are a mixture of well-established and novel potential biomarkers, which replicate across other cohorts and technologies. We highlight the importance of exponential age correction to identify disease-associated proteins beyond those that covary with age. VampW reduces the root mean squared error when predicting disease onset times by over 32% and 26%, when compared to CoxPH variants and the deep learning approach DeepSurv, respectively. In summary, vampW offers accurate and interpretable variable selection and out-of-sample prediction within a single computational framework, making it a powerful tool for dissecting the proteomic architecture of human disease onset.

## **A comparison of the Cox mixed and Andersen-Gill models for the genome-wide analysis of recurrent events: malaria data and a set of simulated data**

Jean-Josue Tokpo\*<sup>1,2</sup>, Philippe Broët<sup>1</sup>, Hervé Perdry<sup>1</sup>, and Jacqueline Milet<sup>2</sup>

<sup>1</sup>Centre de recherche en épidémiologie et santé des populations, Université de Versailles, Saint-Quentin-en-Yvelines, Université Paris-Saclay, Villejuif, France, <sup>2</sup>Mère et enfant en milieu tropical : pathogènes, système de santé et transition épidémiologique, Institut de Recherche pour le Développement, Université Paris Cité, Faculté de pharmacie, Laboratoire de parasitologie, Paris, France

The most common approach used for conducting genome-wide association studies (GWAS) on recurrent events is to use the Cox mixed model with normal frailty (Hougaard, 2000), possibly using a two steps approximation (Milet et al., 2019; Hof et al., 2023). Another possible approach would be to use the Andersen-Gill model with robust variance estimates (Lin and Wei, 1989) which, to our knowledge, has not been tested in the GWAS context. We compared the two approaches on malaria data from a cohort of newborns in Benin. The Andersen-Gill model allowed to identify at the genome wide significance level nine peaks of association, eight of which were missed by Cox's mixed model.

We also compared these methods on simulation data. Simulations were conducted considering several models in which the genotype has an additive or multiplicative effect on the risk of infection, under several genetic models. The analysis shows that when the proportional hazards assumption is satisfied, both methods have similar power in the majority of simulation scenarios. However, when the simulations depart from this assumption, the Andersen-Gill model proves to be more powerful in several scenarios, leading to a gain of up to 20% of power when the risk allele is dominant and has an additive effect on risk. Our results highlight the potential advantages of the Andersen-Gill model for GWAS of recurrent events, notably when the proportional hazard assumption is not satisfied, without a significant loss of power in the opposite case.

## **PMCN: an R-package for phenome-wide molecular causal inference network analysis at Biobank scale**

Zulema Rodriguez-Hernandez\*<sup>1</sup>, Dmitrii Zakharenko<sup>1</sup>, Oleg Borisov<sup>1</sup>, Anna Köttgen<sup>1</sup>,  
and Pascal Schlosser<sup>1</sup>

<sup>1</sup>Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Background: High-dimensional phenotypic data from biobanks and molecular prediction frameworks (FUSION, PrediXcan, BLISS) have transformed genetic screens, including transcriptome- and proteome-wide association studies (TWAS/PWAS). Together, they enable causal inference from molecular layers to phenotypes, but scaling and integration remain challenging.

Methods: We developed PMCN, an R package for Phenome-wide Molecular Causal inference Network analysis, integrating TWAS/PWAS with SuSiE-based colocalization to address genetic confounding. PMCN includes unsupervised biclustering methods (Bibit, nsNMF, FABIA) to simultaneously identify groups of molecular features and traits sharing association patterns. We applied PMCN across 45 tissues using GTExv8 and 150 GWAS meta-analysis summary statistics combining three biobanks (Europeans=88.4%): MVP (N=629,267), FinnGen (N=500,349), and UK Biobank (N=420,531). TWAS and colocalization analyses (Posterior Probability  $H_4 \geq 0.5$ ) were followed by cross-tissue BiBit biclustering ( $\geq 75$  models;  $\geq 4$  traits), along with GO-pathway (BH-adjusted  $P < 0.05$ ;  $\geq 3$  genes), and trait-category enrichment.

Results: We identified an average of 3,716 (IQR:2,707-4,534) significant gene-trait associations ( $P < 0.05/N_{\text{models}}$ ) per tissue, with 38.2% (IQR:37.3-39.1%) supported by colocalization. We found 9,303 biclusters (mean  $N_{\text{models}}=139.1$ ;  $N_{\text{traits}}=6.4$ ; median  $NGO\text{-enrichment}=72$ ) spanning multiple tissues (range:31-45; mean=42.7) and categorized them as category-specific (0.1%), dominated (61.5%), and heterogeneous (38.3%). A trait-specific bicluster ( $N_{\text{tissues}}=45$ ;  $N_{\text{genes}}=12$ ; 4 thyroid-related traits) included well-established thyroid genes (*CAPZB*, *FGF7*, etc.) and evidence linking 10 of 12 genes to thyroid. A heterogeneous bicluster ( $N_{\text{tissues}}=44$ ;  $N_{\text{genes}}=27$ ;  $N_{\text{traits}}=4$ ; colocalization=85%) revealed

potential pan-disease processes in metabolism, signaling, translation, and lysosome/vacuole organization.

Conclusion: PMCN accelerates causal networks discovery, uncovering known and novel pathophysiological pathways and nominates shared molecular targets across diseases for functional follow-up.

## 18 Perturb-seq Experiments as A Source of Causality Orthogonal to Genetic Methods

Sydney Fleming<sup>1</sup>, Théo Cavinato<sup>1,2</sup>, Samuel Moix<sup>1,2</sup>, Mihaela-Diana Zanoaga<sup>1,2</sup>, Zoltán

Kutalik<sup>1,2,3</sup>, and Adriaan Van Der Graaf\*<sup>1,2</sup>

<sup>1</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, <sup>2</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland, <sup>3</sup>Center universitaire de médecine générale et santé publique, Lausanne, Switzerland

Perturb-seq couples CRISPR knockdowns with single-cell RNA-seq to create genome-wide transcriptomic regulatory networks, serving as orthogonal ground truth for genetic causal inference techniques like Mendelian randomization (MR). We harmonized 18 different Perturb-seq experiments from 11 individual studies across 9 cell types. This dataset comprises 3.7 billion total perturbation-expression combinations across 19,129 distinct perturbed genes. We identified 39 million significantly perturbed gene combinations (DESeq2 adjusted  $P < 0.05$ ), representing causal effects onto 17,266 distinct expression phenotypes. We validated the dataset by comparing perturbation effects between experiments; while quantitative correlations were modest (median Pearson  $R=0.08$ ) due to the sparse nature of single cell RNA-seq, experimental and tissue-specific heterogeneity, the direction of correlation was exclusively positive (Pearson  $R$  range: 0.01-0.74).

The resulting networks successfully recapitulate known biology. For example, 186 genes show differential expression after perturbation of the statin target *HMGCR* across 7 experiments, and these genes are strongly enriched for cholesterol pathways (Reactome  $OR=124.7$ ,  $P=3.6 \times 10^{-20}$ ). Notably, *LMNA*, the causative gene for familial partial lipodystrophy type 2 (OMIM: 151660) and dilated cardiomyopathy 1A (OMIM: 115200), was significantly upregulated by *HMGCR* knockdown ( $\log_2$  fold change=0.32, adjusted  $P=0.045$ ), consistent with the cardiovascular benefits of statin treatment. Comparing Perturb-seq to cis-MR (UKB-PPP) revealed strong alignment (86% directional consistency of Pearson correlations), highlighting Perturb-seq's utility in characterizing genetic loci with orthogonal biases. This methodology identifies triangulated consensus regulatory networks valuable for: i) benchmarking AI models, ii) characterizing drug target effects (therapeutic vs. side effects),

and iii) identifying regulatory blind spots caused by context-specificity or stubborn genetic architecture.

## **Too many omics? There's an app (Algorithm) for that**

Marylyn Ritchie<sup>1</sup>

<sup>1</sup>Center for AI, Division of Computational Health Sciences and AI, College of Medicine, Medical University of South Carolina, Charleston, United States of America

The rapid growth of high-throughput technologies has produced an unprecedented wealth of multi-omics data, spanning genomics, transcriptomics, proteomics, metabolomics, epigenomics, and beyond. While each individual layer offers valuable insights, the true promise of precision medicine and complex trait discovery lies in their integration. Yet, bringing these heterogeneous, high-dimensional, and often noisy data sources together remains one of the central challenges in mathematical genetics and computational biology.

In this talk, I will explore a spectrum of machine learning approaches designed for multi-omics integration, highlighting different machine learning approaches with practical applications. I will begin with descriptions of the different strategies for integration such as meta-dimensional and multistaged analyses. I will then move to specific applications where these approaches have been valuable in identifying important biological insights. Finally, I will discuss recent advances in deep learning, variational autoencoders, and other latent factor models that offer flexible, nonlinear frameworks for integration while grappling with issues of interpretability and generalizability. Throughout, I will emphasize the trade-offs between simplicity and complexity, interpretability and predictive power, as well as supervised versus unsupervised strategies. Case studies from population health and disease genetics will illustrate how integrative analyses can uncover novel biology that is invisible to single-omics approaches.

By the end, I hope to leave the audience with a practical “algorithmic toolkit” for multiomics integration, along with a sense of where the field is heading. Whether the goal is risk prediction, biomarker discovery, or mechanistic insight, machine learning offers increasingly powerful ways to make sense of the overwhelming—and overlapping—layers of omics data.

## **Multi-Trait, gene-based and multi-omics integration broadens spontaneous coronary artery dissection biology and prioritizes arterial-wall pathways for follow-up**

Takiy-Eddine Berrandou<sup>1</sup>, Adrien Georges<sup>1</sup>, Ingrid Tarr<sup>2</sup>, Robert Graham<sup>2,3,4</sup>, Eleni Giannoulatou<sup>2,3</sup>, Doug Speed<sup>5</sup>, and Nabila Bouatia-Naji<sup>1</sup>

<sup>1</sup>Université Paris Cité, PARCC, Inserm, Paris, France, <sup>2</sup>Victor Chang Cardiac Research Institute, St. Vincent's Hospital, Sydney, Australia, <sup>3</sup>University of New South Wales, Sydney, Australia, <sup>4</sup>St. Vincent's Hospital, Sydney, Australia, <sup>5</sup>Center for Quantitative Genetics and Genomics, Aarhus, Denmark

Spontaneous coronary artery dissection (SCAD) is a non-atherosclerotic cause of acute myocardial infarction in women < 50 years and remains genetically underpowered. We used multi-trait and gene-based inference to boost discovery and map pathways. We harmonized GWAS summary statistics for SCAD (1,917 cases; 9,293 controls) and seven related traits (fibromuscular dysplasia, intracranial aneurysm, cervical artery dissection, migraine, coronary artery disease, abdominal aortic aneurysm, thoracic aortic aneurysm/dissection), imputing missing SNP-trait Z-scores with GAUSS to obtain 4.85 million shared autosomal variants. We applied SCAD-centered MTAG and leave-one-out MTAG, and evaluated lead variants at novel loci in an independent cohort (293 cases; 1,127 controls). Gene-level association used LDK-GBAT with a 10,000-sample UK Biobank European LD reference. We integrated coronary-artery open-chromatin enrichment, cis-eQTL mapping and colocalization. MTAG identified 40 independent SCAD loci, including 24 novel; 17/24 lead variants had concordant directions in validation. Signals were enriched in open chromatin of coronary-artery smooth muscle cells and fibroblasts. LDK-GBAT detected 46 Bonferroni-significant genes, including 12 outside any genome-wide significant single-variant locus. Multi-omics annotation prioritized 56 candidate genes, highlighting arterial-wall integrity alongside vasoactive and coagulation biology. Gene-set analyses confirmed strong enrichment of extracellular matrix organization and revealed novel enrichment for bone mineralization and TGF- $\beta$  signaling terms. Integrating multi-trait GWAS, gene-based testing, epigenetic and transcriptomic data substantially expand the SCAD genetic landscape. Our findings cover key arterial-wall pathways beyond extracellular matrix organization, and

point at relevant biological mechanisms in non-atherosclerotic dissection. These findings nominate tractable targets for experimental follow-up and support future efforts toward SCAD risk stratification in women.

## **Integration of metabolome-wide CNV-GWAS and PacBio long-read sequencing uncovers a complex, high-impact protective**

### **LDLR multiplication**

Maarja Jõeloo<sup>1,2</sup>, Adriaan Van Der Graaf<sup>3,4</sup>, Nele Taba<sup>2</sup>, Chiara Auwerx<sup>3,4</sup>, Kristi Krebs<sup>2</sup>, Reedik Mägi<sup>2</sup>, Zoltán Kutalik<sup>3,4,5</sup>, Simone Rubinacci<sup>1</sup>, Kaur Alasoo<sup>6</sup>, and Lili Milani<sup>2</sup>

<sup>1</sup>Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland, <sup>2</sup>Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia, <sup>3</sup>Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, <sup>4</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland, <sup>5</sup>University Center for Primary Care and Public Health, Unisanté, Lausanne, Switzerland, <sup>6</sup>Institute of Computer Science, University of Tartu, Tartu, Estonia

Circulating metabolites regulate essential physiological processes and provide critical insights into disease mechanisms. While SNV-based genome-wide association studies (GWAS) have mapped many genetic effects on metabolite levels, the impact of larger rearrangements, such as copy-number variations (CNVs), remains largely unexplored. We performed the largest CNV-GWAS meta-analysis to date focusing on 249 circulating metabolic markers and rare microarray-based CNV events from 376,417 participants across the Estonian and UK Biobanks. Our analysis revealed 493 genome-wide significant ( $p < 1.7 \times 10^{-6}$ ) associations involving 41 distinct CNV loci. As a primary example of the impact of these associations, we identified a rare *LDLR* whole gene multiplication enriched in the Estonian population (AF EST=0.019%, AF UK=0.001%) and highly associated with LDL-related traits such as ApoB (effect size: -1.83 s.d. units,  $p=4.4 \times 10^{-87}$ ) and non-HDL cholesterol (-1.74 s.d. units,  $p=3.5 \times 10^{-77}$ ). We reconstructed the full variant structure by leveraging the PacBio long-read data from the Estonian Biobank and revealed a complex ~104 kbp inverted triplication, resulting in the total of three functional *LDLR* copies per allele. Carriers of this multiplication exhibit a 44% lifelong reduction of non-HDL cholesterol levels and a significantly lower prevalence of hyperlipidaemia (3.85% compared to 31.9% in copy-neutral controls; Fisher's exact  $p=0.001$ ). To our knowledge, this triplication represents the first known high-impact protective *LDLR* coding variant identified across European populations. In summary, this work provides a valuable resource for advancing our understanding of the genetic basis of

metabolic biomarkers and showcases the power of integrating large-scale microarray-based discovery analysis with long-read sequencing to resolve complex disease associated variants.

## **Identifying sub-threshold trans-pQTLs via data-driven protein clustering using rank-rank hypergeometric overlap**

Mario Favre-Moiron<sup>1,2</sup>, Marie Verbanck<sup>1</sup>, and Céline Lefebvre<sup>2</sup>

<sup>1</sup>Oncologie Computationnelle, Institut Curie, Institut National de la Santé et de la Recherche Médicale, Paris, France, <sup>2</sup>Institut de Recherches SERVIER, Institut de Recherche Servier, Gif-sur-Yvette, France

Protein Quantitative Trait Loci (pQTL) can provide a proximal and functionally interpretable link between genetic variation and disease biology. However, only a limited number of trans-pQTLs are detected at standard genome-wide significance thresholds, likely due to statistical power limitations and the complexity of distal regulatory mechanisms. Additionally, current approaches to recover these missing signals heavily rely on existing knowledge such as interaction databases, limiting the discovery of novel pathways. Here, we present a strategy to uncover trans-acting pQTLs using “co-pleiotropy”, defined as the significant overlap of proteins associated with pairs of variants. We applied this method to the UK Biobank Pharma Proteomics Project (UKB-PPP) dataset. Specifically, we adapted Rank-Rank Hypergeometric Overlap (RRHO) to compare the ranked protein association profiles between variant pairs. The proteins driving these significant overlaps directly constitute the functional clusters, enabling the identification of biological modules without a priori assumptions. Crucially, by aggregating subtle signals across these data driven protein clusters, our approach detects variants that individually fall below standard genome-wide significance thresholds. Preliminary results indicate that these “sub-threshold” signals are biologically relevant, recapitulating known Protein-Protein Interaction (PPI) networks and distinguishing coherent regulatory modules. We are confident that such a framework may offer a powerful tool to decode the post-transcriptional landscape of GWAS loci beyond the reach of conventional single-protein association studies.

# Posters

## **P01. Deep learning architectures for EEG-based classification of Dravet Syndrome: a comparative study of LSTM, LSTM-GRU, and hybrid CNN-LSTM Models**

Shahid Bashir\*<sup>1</sup>, Sikandar Hussain<sup>2</sup>, Soyiba Jawed<sup>2</sup>, Ali Mir<sup>1</sup>, Raidah Al-Baradie<sup>1</sup>,  
Mona Ibrahim Ali<sup>1</sup>, and Motasem Sager<sup>3</sup>

<sup>1</sup>Research Center, King Fahad Specialist Hospital, Dammam, Saudi Arabia, <sup>2</sup>National University of Sciences and Technology, Islamabad, Pakistan, <sup>3</sup>Birmingham Heartlands Hospital, Birmingham, United Kingdom

**Introduction:** Dravet syndrome (DS) is a rare and severe developmental and epileptic encephalopathy, most commonly associated with pathogenic *SCN1A* mutations. It is characterized by early-onset, drug-resistant seizures and neurodevelopmental impairment. Early diagnosis remains challenging, and current diagnostic tools offer limited real-time clinical support. Artificial intelligence (AI) applied to electroencephalography (EEG) provides a promising non-invasive solution.

**Objectives:** This study evaluated the performance of deep learning models, particularly a hybrid convolutional neural network–long short-term memory (CNN-LSTM) architecture, for automated EEG-based classification of DS.

**Methods:** The cohort included nine patients with DS confirmed by pathogenic *SCN1A* variants or clinical diagnostic criteria and twenty age- and sex-matched healthy controls. Demographic, clinical, developmental, cognitive, and genetic data were extracted from medical records. Preprocessed EEG signals were analyzed using three architectures-LSTM, hybrid LSTM-GRU, and hybrid CNN-LSTM-designed to capture temporal and spatial EEG features.

**Results:** From 2,823 EEG samples (1,625 DS; 1,198 controls), the CNN-LSTM model achieved the best and most stable performance, with accuracies of 94.74% (training), 92.92% (validation), and 91.86% (testing). The model showed strong discriminative ability (precision, recall, and F1-score = 0.92; AUC-ROC = 0.973). t-SNE visualization demonstrated distinct EEG clustering between DS and controls, and the confusion matrix confirmed robust

classification. The mean seizure onset age was 7.04 months, consistent with previous reports.

Conclusion: CNN-LSTM-based EEG analysis enables accurate automated classification of DS. These findings highlight the potential of AI-driven EEG tools to support early diagnosis and precision epilepsy care, warranting validation in larger, real-world clinical datasets.

## **P02. Pleiotropic pattern inference from univariate GWAS**

Gloria Benoit\*<sup>1</sup>, Léo Henches<sup>1</sup>, Hanna Julienne<sup>1</sup>, Junsun Yu<sup>2</sup>, Courtney Tern<sup>2</sup>,  
Michael

Cho<sup>2</sup>, and Hugues Aschard<sup>1,3</sup>

<sup>1</sup>Institut Pasteur de Paris, Paris, France, <sup>2</sup>Brigham and Women's Hospital, Boston, United States, <sup>3</sup>Harvard T.H. Chan School of Public Health, Boston, United States

The increase of publicly available GWAS summary statistics offer opportunities to understand the shared and specific genetic components of multifactorial traits and diseases. Thus, numerous methods have been proposed to decipher pleiotropic patterns from GWAS. They cover various scales, including global and region-based genetic correlation methods, and matrix factorization approaches that aim at inferring latent genetic factors, that is, clusters of variants displaying the same multitrait effects. Those methods are built on different assumptions, addressing computational challenges through approximations, and being subject to heterogeneous constraints in the number of variants and phenotype they can analyze jointly. Here, we built a pipeline to detect pleiotropic patterns across using six approaches: SUPERGENOVA and HDL-L, which estimate local genetic correlations; and GFA, Guide, GLEANR and FactorGo which derive latent factors. We used this pipeline to assess their similarity and specificity in the inference of pleiotropic patterns with a set of 13 respiratory-related GWAS. We compared the methods by estimating the correlation in the variants clustering, defined as either factor's weights or pairwise local genetic correlation. Overall, FactorGo was the most consensual method, displaying high correlation with both estimated local genetic correlation and other matrix factorization methods. On the other hand, GFA displayed the lowest correlations, suggesting those methods still capture orthogonal component. We aim to expand the number of phenotypes compared, investigate how to better harmonize the set of SNPs used, and evaluate metrics to combine the results across methods, to form robust pleiotropic pattern inference.

## **P03. Comparison of secondary analysis pipelines for whole genome sequencing**

Raphael Betschart<sup>1,2</sup>, Stefan Blankenberg<sup>1,3,4,5</sup>, Raphael Twerenbold<sup>3,4</sup>, Tanja Zeller<sup>2,4,5</sup>, and Andreas Ziegler<sup>1,3,5,6</sup>

<sup>1</sup>Cardio-CARE, Medizincampus, Davos, Switzerland, <sup>2</sup>Institute of Cardiogenetics, University of Lübeck, Lübeck, Germany, <sup>3</sup> University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>4</sup> German Centre for Cardiovascular Research (DZHK), Hamburg/Kiel/Lübeck site, Germany, <sup>5</sup> Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>6</sup>Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Introduction: The accuracy of secondary analysis pipelines to preprocess data of whole genome sequencing (WGS) studies is constantly improving. This study compares four mapping and alignment algorithms (Nvidia Parabricks FQ2BAM 4.5.1, DRAGEN 3.8.4, DRAGEN 4.3.4, and Nvidia Parabricks Giraffe 4.5.1), and five variants calling algorithms (Nvidia Parabricks DeepVariant 4.5.1, Nvidia Parabricks GATK Haplotypecaller 4.5.1, DRAGEN 3.8.4, DRAGEN 4.3.4, and DeepVariant 1.9.0). All these algorithms run on specialized hardware, such as fieldprogrammable gate array (FPGA) cards in case of DRAGEN, and graphics processing units (GPU) for Nvidia Parabricks.

Comparisons were based on three Genome-in-a-Bottle samples (HG002, HG003, HG004). Sample HG002 was sequenced 69 times, HG003 four times, and HG004 eight times. The publicly available truth set of the GIABs was used for comparison, and performance was assessed by F1 score, precision, and recall.

Results: Mapping and alignment performed with DRAGEN 4.3.4 outperformed Nvidia Parabricks FQ2BAM 4.5.1, mainly due to the fact that DRAGEN 4.3.4 uses a pangenome as its reference genome, instead of a linear one. Nvidia Parabricks Giraffe 4.5.1 also outperformed Nvidia Parabricks FQ2BAM, but did not outperform both DRAGEN 3.8.4 and DRAGEN 4.3.4. Among the variant calling algorithms, DRAGEN 4.3.4 outperformed the other four variant callers. When mapping to a pangenome, it is important to choose a pangenome-aware

variant caller, as indicated by the pangenome-aware variant caller DeepVariant 1.9.0 outperforming Nvidia Parabricks DeepVariant 4.5.1, which is not pan-genome aware.

## Poster highlight • P04. GWAS for recurrent pregnancy loss in Uzbekistan identifies novel association near *PAPPA* gene.

Chia-Yi Chu<sup>1</sup>, Yevheniya Sharhorodska<sup>2,3</sup>, Anastasiya Punko<sup>4</sup>, Zebinisa Mirakbarova<sup>5,6</sup>,  
 Khurshid Meylikov<sup>4,5</sup>, Abdushukur Rakhmatullayev<sup>4,5</sup>, Yuliya Kapralova<sup>4,5</sup>, Konstantin Rudometkin<sup>7</sup>, Abrorjon Abdurakhimov<sup>5</sup>, Dilbar Dalimova<sup>4</sup>, Shahlo Turdikulova<sup>4</sup>,  
 Alisher Abdullaev<sup>4</sup>, and Inga Prokopenko<sup>1,7</sup>

<sup>1</sup>People-Centered Artificial Intelligence Institute, University of Surrey, Guildford, United Kingdom, <sup>2</sup>Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy, <sup>3</sup>Department of Clinical Genetics, Institute of Hereditary Pathology, National Academy of Medical Sciences, Lviv, Ukraine, <sup>4</sup>Center for Advanced Technologies, Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan, <sup>5</sup>Institute of Biophysics and Biochemistry, National University of Uzbekistan, Tashkent, Uzbekistan, <sup>6</sup>Central Asian University, Tashkent, Uzbekistan, <sup>7</sup>Department of Clinical Medicine, University of Surrey, Guildford, United Kingdom

Background: Pregnancy loss is a complex reproductive condition affecting 15% of women worldwide. Genome-wide association studies (GWAS) investigated its genetic architecture, but identified only a small number of robust risk loci. Furthermore, most GWAS focused on individuals of European ancestry, limiting insights into genetic susceptibility in underrepresented populations. The Advanced Pregnancy Loss Study in Uzbekistan (ALSU) aims to explore the genetic basis of pregnancy loss in population of Uzbekistan.

Materials and methods: We collected epidemiological and sociodemographic information from 1,815 women participating in ALSU and acquired genome-wide imputed data for 998 participants. We selected 289 recurrent pregnancy loss cases with two or more miscarriages or non-developing pregnancies before 24 weeks, and 307 controls with live birth and no history of pregnancy loss. We performed a GWAS for 7,724,733 single nucleotide variants (SNVs) with minor allele count (MAC)≥20 from genome-wide imputed data, adjusted by the first three genetic principal components, using SAIGE under generalized mixed model accounting for relatedness.

Results: We identified a novel genome-wide significant association near Pregnancy Associated Plasma Protein A (*PAPPA*) gene (rs3037402-C, effect allele frequency (EAF) = 0.18,

OR (95%CI) = 2.50(1.81- 3.47),  $P = 3.8 \times 10^{-8}$ ). We also observed a nominal association at an established locus near *TSHZ3*, with effect direction consistent with previous reports (rs2032905-G, EAF=0.53, OR (95%CI) = 1.30(1.01-1.66),  $P = 0.039$ ).

Conclusion: This first GWAS of recurrent pregnancy loss in Central Asian population identified a novel associated locus and highlighted the importance of ancestry-diverse studies for understanding the genetic basis of pregnancy loss.

Funding: World Bank (REP-03032022/192); AID, Uzbekistan (FL-8323102086); budgetary funding to CAT, Uzbekistan.

## **P05. Genetic architecture of imaging-derived cardiac phenotypes in cardiomyopathies: a systematic review**

Amra Dhabalia Ashok<sup>\*1</sup>, Maria Luisa Benesch Vidal<sup>2</sup>, Maximilian Lackner<sup>2</sup>, Christina Magnussen<sup>2,3,4</sup>, Andreas Ziegler<sup>1,2,5,6</sup>, and Peter Moritz Becher<sup>2</sup>

<sup>1</sup>Cardio-CARE, Medizincampus Davos, Davos, Switzerland, <sup>2</sup> University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>3</sup> German Centre for Cardiovascular Research (DZHK), Hamburg/Kiel/Lübeck site, Hamburg, Germany, <sup>4</sup>Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>5</sup>German Center for Cardiovascular Research, Hamburg/Kiel/Lübeck, Germany, <sup>6</sup> Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Cardiomyopathies are myocardial disorders characterized by structural and functional heart abnormalities, often causing heart failure, arrhythmias, or sudden cardiac death. These conditions arise from environmental, systemic, or genetic factors, with hereditary components particularly prominent in dilated and hypertrophic cardiomyopathies. Advances in cardiac imaging including echocardiography, myocardial strain analysis, and cardiac MRI now enable extraction of precise image-derived cardiac phenotypes (IDCPs) that quantify ventricular volumes, wall thickness, ejection fraction, and myocardial strain. Concurrently, cardiovascular genetics has identified numerous cardiomyopathy-associated variants through genome-wide association studies and high-throughput sequencing. Integrating genetic data with IDCPs promises to unravel disease mechanisms, identify subclinical phenotypes, and improve patient stratification. This systematic review synthesizes current evidence on genetic associations between cardiomyopathies and IDCPs, evaluating how genetic variation influences imaging biomarkers to advance precision cardiology. Eighty-four studies were included, examining genetic variants associated with IDCPs across cardiomyopathy subtypes. Key findings revealed strong associations between sarcomeric gene mutations and increased left ventricular wall thickness plus abnormal myocardial strain in hypertrophic cardiomyopathy. In dilated cardiomyopathy, TTN and LMNA variants consistently correlated with ventricular dilation and reduced ejection fraction. Advanced imaging, particularly cardiac MRI and strain

echocardiography, detected distinct phenotypic signatures in mutation carriers before clinical symptom onset. Longitudinal studies indicated certain variants predict progressive cardiac deterioration.

Integrating genetic information with quantitative imaging phenotypes significantly enhances early diagnosis, risk stratification, and precision medicine implementation in cardiomyopathy management. Open Science Forum registration: <https://doi.org/10.17605/OSF.IO/VPKS2> (April 20, 2025).

**Poster highlight • P06. Efficient phasing strategies and optimal reference panel selection improve genotype imputation for Whole-Exon Sequencing (WES) data**

Carola Di Meo<sup>\*1</sup>, Francesca Rosamilia<sup>1</sup>, Marta Rusmini<sup>1</sup>, Giovanni Fiorito<sup>1</sup>, Serena Sanna<sup>2</sup>, and Paolo Uva<sup>1</sup>

<sup>1</sup>Clinical Bioinformatic Unit, IRCCS Istituto Giannina Gaslini, Genoa, Italy,

<sup>2</sup>Department of Genetics, University of Groningen, Groningen, Netherlands

Whole-Exome sequencing (WES) is a cost-effective strategy for investigating rare coding variants, but stringent quality control and limited genomic coverage may reduce statistical power, particularly in regions flanking. Genotype imputation of WES data using Whole-Genome sequencing (WGS) reference panels is commonly applied to mitigate the loss of relevant association signals. However, previous studies consistently report poor performance of rare variant imputation, likely reflecting challenges in haplotype reconstruction and the need for ancestry-matched reference panels. We systematically evaluate WES imputation performances across multiple combinations of reference panels, phasing, and imputation tools to assess the benefit of an internal ancestry matched WGS reference panel. A total of 1,599 WES samples from the IRCCS Gaslini Institute, were imputed using two reference panel: i) the 1000 Genomes Project and ii) an internal WGS dataset including 1,561 samples. Two imputation pipelines were applied: Minimac4 + Shapeit4 and GLIMPSE2, which leverages off-target reads to improve haplotype reconstruction and reduce random phasing. Imputation performance of each reference panel–pipeline combination was assessed using dosage R<sup>2</sup> between masked true and imputed genotypes. Our results indicate that imputation using the internal reference panel outperforms the 1000 Genomes reference, specifically in individuals of European ancestry. Average dosage R<sup>2</sup> reached 0.45 in the 0.5–1% MAF interval and was slightly higher in European samples. Overall, our findings demonstrate that optimizing phasing strategies and leveraging off-target reads substantially improve WES imputation, supporting robust detection of rare variant associations and enhancing the utility of exome sequencing for large-scale genetic studies.

## **P07. Back-in-time reconstruction of population structure using reconstructed haplotypes**

Jin Du<sup>1</sup> and Stefan Böhringer\*<sup>1</sup>

<sup>1</sup>Leiden University Medical Center, Universiteit Leiden, Leiden, Netherlands

Analysis of population structure (PS) is important for many applications. Haplotypes (HTs) can be used to increase resolution of PS analysis. We develop a consistent, closed-form method-of-moments estimator for HT frequencies obviating the need for an expectation maximization (EM) algorithm. To make the estimator efficient, a fixed number of Newton-Raphson steps is used. We prove consistency and show efficiency by comparing the estimator with EM results empirically. To control the complexity of reconstructions, grouping strategies based on HT frequencies and HT similarity are employed. Genome wide data can be analyzed.

To analyze population history, the relationship of age and frequency of variants is exploited. First, the genome is clustered into independent groups using hierarchical clustering. Second, count matrices (valued between 0 and 2) are constructed, containing expected counts of HTs. Third, weighted principal component analyses (PCAs) are performed where weights emphasize different regions of the haplotype frequency spectrum. Each weighing scheme corresponds to a certain average age of HTs. Finally, trajectories for individuals are derived from the different PCAs.

We perform simulations to characterize statistical and run-time characteristics of our HT reconstruction algorithm. To perform a back-in-time reconstructions, hapmap 2 data is used. We show that the use of HTs offers additional insight compared to a genotype-based PS analysis. The use of genetic trajectories can help to better interpret the data, improve genetic association analyses, and also allows novel analyses investigating effects of trajectories.

## **P08. Mathematical modelling of ancestry-aware Bayesian rare-variant association using expression-informed functional priors in admixed populations and ultra-rare genetic architecture inference framework**

Rifaldy Fajar\*<sup>1</sup>, Prihantini Prihantini<sup>2</sup>, and Rini Winarti<sup>3</sup>

<sup>1</sup>Computational Biology and Medicine Laboratory, Yogyakarta State University, Yogyakarta, Indonesia, <sup>2</sup>Department of Mathematics, Bandung Institute of Technology, Bandung, Indonesia, <sup>3</sup>Department of Biology, Yogyakarta State University, Yogyakarta, Indonesia

Background/Aim: Rare-variant association analysis remains mathematically fragile under population admixture, ultra-rare allele spectra, and noisy functional annotations, frequently leading to unstable inference and reduced discovery power. We aimed to develop a principled Bayesian mathematical framework that integrates ancestry-aware genetic structure with tissue-informed functional priors to enable robust gene-level rare-variant inference with explicit uncertainty quantification.

Methods: We developed ARV-EBayes, an ancestry-aware hierarchical Bayesian model, using whole-genome sequencing data from the 1000 Genomes Project Phase 3 and the Human Genome Diversity Project (4,094 individuals). Gene-level associations were inferred by aggregating variant effects under spike-and-slab priors, with variant weights modelled by ancestry-stratified minor allele frequency bins and continuous tissue expression scores from the Illumina Human BodyMap 2.0 atlas (GEO GSE30611). Heavy-tailed components addressed annotation misspecification. Posterior probabilities of association and Bayes factors were estimated via Markov chain Monte Carlo, with error control assessed using posterior predictive null simulations and ancestry-preserving permutations.

Results: Across 200 null simulations ( $\sim 1 \times 10^6$  gene-level tests), empirical type-I error at  $1 \times 10^{-5}$  was  $1.1 \times 10^{-5}$  (95% CI  $0.9\text{--}1.3 \times 10^{-5}$ ). With 8–12 causal variants per gene (median MAF 0.10–0.15%), ARV-EBayes improved power by 18.6 percentage points over SKAT-O (95% CI 12.4–24.1;  $p < 1 \times 10^{-5}$ ) and reduced the false discovery proportion by 22% at matched recall.

Posterior 95% credible intervals achieved 93–97% coverage, with < 3% power loss under 30% mis-specified expression-informed priors.

Conclusions: This ancestry-aware Bayesian model enables stable and interpretable rare variant inference under realistic population structure and ultra-rare genetic architectures, offering a robust framework for gene-level discovery across diverse populations.

## **P10. Same model, better performance: the impact of shuffling on DNA Language Models benchmarking**

Davide Greco<sup>1</sup> and Konrad Rawlik<sup>1</sup>

<sup>1</sup>Baillie Gifford Pandemic Science Hub, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh, United Kingdom

Large Language Models are increasingly popular in genomics due to their potential to decode complex biological sequences. Hence, researchers require a standardized benchmark to evaluate DNA Language Models (DNA LMs) capabilities. However, evaluating DNA LMs is a complex task that intersects genomics' domain-specific challenges and machine learning methodologies, where seemingly minor implementation details can significantly compromise benchmark validity. We demonstrate this through BEND (Benchmarking DNA Language Models), where hardware-dependent hyperparameters – number of data loading workers and buffer sizes – create spurious performance variations of up to 4% for identical models. The problem stems from inadequate data shuffling interacting with domain specific data characteristics. Experiments with three DNA language models (HyenaDNA, DNABERT-2, ResNet-LM) show these artifacts affect both absolute performance and relative model rankings. We propose a simple solution: pre-shuffling data before storage eliminates hardware dependencies while maintaining efficiency. This work highlights how standard ML practices can interact unexpectedly with domain-specific data characteristics, with broader implications for benchmark design in specialized domains.

## **P11. Dynamic machine-learning modelling of multi-omics features for predicting weight loss maintenance**

Tingyu Guo<sup>\*1</sup>, Zhanna Balkhiyarova<sup>1,2</sup>, Mari Näätänen<sup>3</sup>, Carlos Gómez-Gallego<sup>3</sup>,  
Anna

Karlund<sup>3,4</sup>, Leila Karhunen<sup>3</sup>, Marjukka Kolehmainen<sup>3</sup>, Marika Kaakinen<sup>5,6</sup>, Ayse  
Demirkan<sup>1,2</sup>, and Inga Prokopenko<sup>1,2</sup>

<sup>1</sup>People-Centered Artificial Intelligence Institute, University of Surrey, Guildford, United Kingdom, <sup>2</sup>Department of Clinical Experimental Medicine, University of Surrey, Guildford, United Kingdom, <sup>3</sup>Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland, <sup>4</sup>Department of Life Technologies, Food Sciences Unit, University of Turku, Turku, Finland, <sup>5</sup>Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom, <sup>6</sup>Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

Obesity poses a major public health challenge, with long-term weight loss maintenance (WLM) remaining difficult. To systematically assess omics predictors contributing to WLM, we analyzed multimodal longitudinal data from the Finnish ELIPA food intervention study using dynamic, machine-learning (ML) based models. The 33-week ELIPA study included measurements at baseline, post-weight loss (WL), and during and after weight maintenance (WM). 82 participants with baseline BMI > 40 underwent a 7-week very-low-calorie diet, a 2-week transitional period and a subsequent 24-week WM period. Multi-omics data, including anthropometrics, biochemical, metabolomics, microbiota, and transcriptomics were collected. To capture within-individual dynamics, we calculated velocity of change in features over WL and WM. We applied random forest ML models to predict percentage BMI change during WM using velocity-based multi-omics features, and variable importance (VI) was used to rank predictors. Finally, we used 5-fold cross-validation to evaluate the model performance. All predictors were evaluated for their contribution to WLM. The primary model suggested that the rate of fat mass change during the WM period was the strongest predictor (VI=93.2%). Metabolomic velocity features during WL, including tryptophan (VI=29.43%) and phosphatidylcholine-related compounds (VI=21.42%) were also identified. A subset model indicated transcriptomic velocity features during WL as leading predictors, including EPM2A (VI=90.41%) and C6orf106 (VI=80.93%). Microbial velocity features, including *Flavonifractor plautii* and *Aeromonas* during WM, were also highlighted.

Our results demonstrate that velocity-based modelling of longitudinal multi-omics data helps identifying dynamic predictors of WLM, highlighting the opportunities for personalized WLM strategies through targeted multi-omics adaptation using ML approaches.

## **Poster highlight • P12. Sparse-prior variational autoencoders for improved biological Large Language Model interpretability**

Rachael Harkness<sup>1</sup>, Konrad Rawlik<sup>1</sup>

<sup>1</sup> Baillie Gifford Pandemic Science Hub, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh, United Kingdom

Sparse autoencoders (SAEs) enable dictionary learning over polysemantic features in large language models (LLMs), extracting interpretable, monosemantic representations. These techniques are valuable for both natural language and biological sequence models, where interpretable features may correspond to regulatory elements, structural patterns, or functional domains governing cellular processes. However, SAE training faces a critical challenge: “dead” neurons that remain inactive throughout training. Current solutions rely on complex neuron resampling strategies that are memory-intensive, reduce training efficiency, and exhibit high sensitivity to hyperparameters. We propose sparse-prior variational autoencoders (VAEs) as an alternative framework. By replacing deterministic sparse representations with stochastic sampling from sparse posterior distributions, our approach addresses the dead neuron problem without requiring resampling. The variational framework offers theoretical advantages: (1) stochastic sampling ensures neurons can activate even when their posterior means are low, preventing permanent inactivation, (2) sampling provides inherent regularization against overfitting, and (3) the stochastic nature ensures rarely active neurons continue receiving gradient updates, encouraging learning of more diverse interpretable features. We evaluate our approach in DNA and protein sequence domains against SAEs with resampling. We train logistic regression classifiers on learned representations to predict biologically significant functional elements in both domains, while also exploring steering methods for protein design. Linear separability with respect to functional classes serves as a quantitative indicator of successful feature decomposition. This work demonstrates that variational inference provides a principled alternative to resampling while improving diversity and quality of learned interpretable features in biological language models.

## **P13. Metagenome pipeline for taxonomic-free genetic screening**

Léo Henches\*<sup>1</sup>, Raphaël Malak<sup>1</sup>, Arthur Frouin<sup>1</sup>, Antoine Auvergne<sup>1</sup>, Christophe Boetto<sup>1</sup>, Harry Sokol<sup>2,3,4</sup>, Rayan Chikhi<sup>5</sup>, and Hugues Aschard<sup>1,6</sup>

<sup>1</sup>Unité de génétique statistique, Département de Biologie Computationnelle, Institut Pasteur, Institut Pasteur de Paris, Paris, France, <sup>2</sup>Institut National de la Santé et de la Recherche Médicale, Université Paris Cité, Université Sorbonne Paris Nord, Inserm, Paris, France, <sup>3</sup>Assistance publique, Hôpitaux de Paris (AP-HP), AP-HP, Paris, France, <sup>4</sup>Département Microbiologie et Chaîne Alimentaire, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Jouy-en-Josas, France, <sup>5</sup>Institut Pasteur, Institut Pasteur de Paris, Paris, France, <sup>6</sup>Harvard T.H. Chan School of Public Health, Boston, United States of America

Genome-wide association studies (GWAS) have been pivotal for uncovering the genetics of human phenotypes, and there is now growing interest in applying GWAS-like methods to explore genetic associations in metagenomic analyses. Here, we developed a taxonomy free GWAS approach that uses k-mers, i.e. DNA words of length k that can capture single nucleotide polymorphisms, insertion or deletion events, and gene presence-absence, to analyse metagenome data. We release MOGS (Metagenome Occurrence-based Genetic Screening), a software to run association analysis in metagenomic data, and incorporate it into a pipeline that starts from raw sequencing data and produces GWAS results. We make available a comprehensive tutorial to apply this GWAS approach on any metagenomic data. To show the power of this approach, we applied this method on human microbiome data to 26 traits spanning demographics, physiological measurements, health, and lifestyle in 938 healthy participants from the Milieu Intérieur cohort. We generated a k-mer abundance matrix encompassing 97million distinct k-mers. GWAS of the 26 traits identified significant associations for seven of them: age, sex, depression, appetite, cooked meat consumption, soda intake, and smoking. By modelling the correlation structure among k-mers, we identified a modest number of independent signals and conducted a comprehensive in silico functional annotation of these signals, revealing potential mechanisms of host-microbiota interaction. Overall, our analyses demonstrate that k-mers can capture biologically relevant functions shared across multiple taxa and provide a refined modelling framework that complements the standard taxonomic-based screening approach.

## **Poster highlight • P14. A practical end-to-end pipeline for KnockoffGWAS with application to primary biliary cholangitis**

### **GWAS data**

Richard Howey\*<sup>1,2</sup> and Heather Cordell<sup>2</sup>

<sup>1</sup>Research Software Engineering, Newcastle University, Newcastle upon Tyne, United Kingdom, <sup>2</sup>Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

KnockoffGWAS, introduced by Sesia et al., provides a powerful framework for genome wide association studies that rigorously controls the false discovery rate (FDR) while accounting for the complex linkage disequilibrium structure inherent to genetic data. The approach constructs synthetic “knockoff” copies of genomic variants that mirror the correlation patterns of the original genotypes, enabling direct comparison of feature importance between true and synthetic variants. This contrast-based strategy reduces the rate of false positives that commonly arise in traditional GWAS and offers more reliable identification of causal signals. However, despite its methodological strengths, the practical application of KnockoffGWAS has been hindered by challenging data-format requirements, multi-step computational workflows, and the lack of an end-to-end, user-friendly implementation. In this work, we present a fully operational pipeline that applies KnockoffGWAS to a large Primary Biliary Cholangitis (PBC) dataset, starting solely from standard genetic data and genetic map files. The pipeline includes all necessary Bash and R scripts to convert genotype formats, generate knockoff variables, run the inference procedure, and produce final association results. It automates the incorporation of multiple external tools, resolves compatibility issues, and standardized each stage into a clear, reproducible workflow.

To support broader adoption, we also provide an accompanying website offering a complete step-by-step guide-from data preparation and quality control to knockoff construction, inference, and interpretation. Demonstrating the pipeline on the PBC data illustrates that KnockoffGWAS can be feasibly applied at scale to real-world complex genetic disease cohorts. This work transforms a theoretically innovative yet practically inaccessible

method into a usable and reproducible tool, enabling wider use of the KnockoffGWAS approach.

## **P15. Genetic correlations between asthma subtypes and neuropsychiatric disorders**

Haibo Huang<sup>1</sup>, Raphaël Vernet<sup>1</sup>, Lucie Troubat<sup>1</sup>, Florence Demenais<sup>1</sup>, Christophe Linhard<sup>1</sup>, Stephan Fischer<sup>2</sup>, Yuka Suzuki<sup>2</sup>, Hanna Julienne\*<sup>2</sup>, and Emmanuelle Bouzigon<sup>1</sup>

<sup>1</sup>Inserm, HealthFex, group of Genomic Epidemiology of Multifactorial Diseases, Université Paris Cité, Paris, France, <sup>2</sup>Department of Computational Biology, Institut Pasteur de Paris, Paris, France

Asthma patients suffer more frequently than general population from anxiety and depression. However, the link between asthma and neuropsychiatric disorders is poorly understood. To clarify the interplay between asthma and neuropsychiatric disorders, we computed global and local genetic correlations ( $r_g$ ) among four asthma subtypes and twelve neuropsychiatric disorders using linkage disequilibrium score regression (LDSC) and local analysis of covariant association (LAVA) on full GWAS summary statistics of European population. We detected significant global  $r_g$  (after Bonferroni correction) between ten trait pairs, restricted to three asthma subtypes (asthma, adult onset, moderate to severe) and four neuropsychiatric disorders: major depression (MDD,  $r_g=0.26-0.37$ ), post-traumatic stress ( $r_g=0.28-0.40$ ), bipolar (BIP,  $r_g=0.10$ ), and attention deficit hyperactivity disorder ( $r_g=0.26-0.36$ ). No significant  $r_g$  was found for childhood asthma. At the local level, we identified 96 significant shared regions out of 2108 genomic regions, with a maximum of 19 shared regions between asthma and MDD. Among these trait pairs with significant local  $r_g$ , 49 trait pairs had non-significant global  $r_g$ . Notably, asthma and schizophrenia pair exhibited 11 significant local correlations (global  $r_g=0.06$ ), and childhood asthma and BIP pair exhibited 4 significant local correlations (global  $r_g=0.07$ ). Across these trait pairs, the proportion of positive and negative local correlations was balanced, which may explain the absence of significant global  $r_g$ . To further understand their specific genetic links, we will 1) perform multi-trait analysis and fine mapping across asthma subtypes and neuropsychiatric disorders; 2) conduct functional analyses to identify pathways and cell types tied to their significant variants.

## **P16. CARAVAN: A Nextflow pipeline that implements additional models for common and rare variants analysis**

Georgios Koliopoulos<sup>1</sup>, Nicholas Toda<sup>1</sup>, Felicia Sandberg<sup>1</sup>, Hugo Solleder<sup>1</sup>, Amra Dhabalia Ashok<sup>1</sup>, Raphael Betschart<sup>1</sup>, and Andreas Ziegler<sup>1,2,3,4</sup>

<sup>1</sup>Cardio-CARE, Medizincampus Davos, Davos, Switzerland, <sup>2</sup>University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>3</sup>Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>4</sup> Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Common and rare variant analysis have been widely used to study the genetic basis of many different phenotypes. Statistical methods have been developed to carry out these analyses for numerous theoretical frameworks and trait types, and these methods have been previously integrated into reproducible analysis pipelines. However, these pipelines have not yet included some approaches, such as Tobit models, survival analysis, or ordinal traits. Here we present CARAVAN, a single user-friendly pipeline that extends the models available to bridge the gap in missing analytical methods. CARAVAN is a Nextflow pipeline that runs genome-wide survival analysis and association analysis on numerous trait types for common variants, rare variants with Leave One Variant Out analysis, and short tandem repeats with REGENIE. Association analysis for ordinal traits has been integrated into the pipeline using mixed model association test analysis with POLMM. Survival analysis is implemented using the R package “survival”. The pipeline provides an R package TARO to create testing units for rare variant analysis and define variant weights within a testing unit. A key element of the pipeline is its comprehensive results report module. The additional models provided by CARAVAN extend the capabilities of research groups to deepen our understanding of complex traits.

**P17. *NOTCH3* R1231C phenotypic variability across three independent cohorts: roles of genetic background, environment, and ascertainment bias in CADASIL cerebral small vessel disease**

Anne-Louise Leutenegger\*<sup>1</sup>, Matthieu Pluntz<sup>1</sup>, Teresa Nutile<sup>2</sup>, Louis Lambert<sup>3</sup>, Daniela

Ruggiero<sup>2</sup>, Jessica Lebenberg<sup>3</sup>, Hervé Perdry<sup>1</sup>, Elisabeth Tournier-Lasserre<sup>3,4</sup>, Hugues

Chabriat<sup>3,4,5</sup>, and Marina Ciullo<sup>2</sup>

<sup>1</sup>CESP UMR1018, Institut National de la Santé et de la Recherche Médicale (Inserm), Hôpital Paul Brousse, Villejuif, France, <sup>2</sup>CNR IGB-ABT, Institute of Genetics and Biophysics A. Buzzati-Traverso, CNR, Naples, Italy, <sup>3</sup>ICM UMR1127 GenoVasc – Institut National de la Santé et de la Recherche Médicale (Inserm), GenoVasc, Hôpital de la Pitié-Salpêtrière, Paris, France, <sup>4</sup>APHP, Translational Neurovascular Center and CERVCO, Assistance publique, Hôpitaux de Paris, Translational Neurovascular Center and CERVCO, Hôpital Lariboisière, Paris, France, <sup>5</sup>Université Paris Cité, FHU Neuro-Vasc, Université Paris Cité, Paris, France

Background: CADASIL *NOTCH3* mutations show variable phenotypic expression. We compared three independent carrier cohorts to disentangle genetic, environmental, and ascertainment factors contributing to phenotypic variability.

Methods: We focused on the R1231C mutation, the most frequent in European populations. We analyzed 53 carriers from an isolated Italian population (CILCAD), compared with hospital-based CADASIL patients from Lariboisière Hospital (LRB, N=17) and population-based carriers from UK Biobank (UKB, N=264). Clinical, MRI, and biochemical parameters were systematically compared. Polygenic scores (PGS) for LDL were estimated across samples.

Results: LDL levels were significantly lower in CILCAD compared to both LRB and UKB subjects; this difference was abolished by statin treatment. CILCAD showed lower LDLPGS than LRB ( $p=2.6E-06$ ), which does not share the same haplotype in the 8.9 Mb region around the *NOTCH3* gene. Disease severity in CILCAD was comparable to UKB; both population-based samples were less severe than the hospital-based LRB sample. A negative association

between LDL and stroke (adjusting for age, sex, and statin treatment) was detected in CILCAD ( $p=0.02$ ) and replicated in UKB ( $p=0.03$ ).

Conclusions: Phenotypic variability persists even within genetically and environmentally homogeneous populations. Genetic background substantially influences CADASIL phenotype through linked genetic variants. Very low LDL may paradoxically increase stroke risk in CADASIL across populations. Ascertainment bias substantially affects disease severity estimates. Population-based samples reveal milder CADASIL phenotypes than hospital-based cohorts.

## **P18. Roadmap to a successful rare variant association study – a topic review**

Vivian Link<sup>\*</sup>,<sup>1</sup>, Amra Dhabalia Ashok<sup>\*</sup>,<sup>1</sup>, Andreas Ziegler<sup>1,2,3,4</sup>

<sup>1</sup>Cardio-CARE, Medizincampus Davos, Davos, Switzerland

<sup>2</sup> University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>3</sup> Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

<sup>4</sup>Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

<sup>\*</sup>equal contribution

Rare variants are crucial for understanding genetic contributions to complex traits. To gain statistical power, multiple rare variants are often collapsed into one testing unit. Conducting such collapsing Rare Variant Analyses (RVA) involves numerous decisions, including defining the testing unit and qualifying variants. Another important difficulty with collapsing methods is the loss of information about the effect size of individual variants. To navigate these questions, we conducted a topic review. We dissected papers performing RVA from the last 10 years, recording strategies for association testing, assessing robustness, and interpreting the association results biologically. Many studies employed similar association tests and annotation tools. The primary testing unit was the coding region of genes, with exons or domains and even non-coding regions used occasionally. Robustness was often assessed through replication in separate datasets with different ancestries. Synonymous mutations served as negative controls, presumed not to affect disease, while known phenotype-affecting genes acted as positive controls. Strategies for biological interpretation were diverse, but combining collapsing methods with the analysis of single variants yielded comprehensive insight. Specifically, qualifying variants within an associated gene were individually studied to define allelic series, assess penetrance, and investigate the mode of inheritance. In summary, strategies for interpreting the biological significance of RVA results are diverse, contrasting with the more homogeneous upstream analytical methods. Researchers have access to various techniques, depending on sample

sizes and phenotype numbers. Particularly compelling approaches combine collapsing methods with follow-ups on single rare variants.

## **P19. Model choice induced gene-environment interactions**

Märt Möls\*<sup>1</sup>

<sup>1</sup>Institute of Genomics, Tartu, Estonia

The choice of genetic model significantly influences gene-environment (GxE) interactions. When additive models are applied to multiplicative SNV effects, where alleles proportionally increase or decrease the phenotype, the resulting GxE effects emerge because the absolute change in phenotype depends on other variables. Conversely, using multiplicative models for additive SNV effects necessitates including GxE terms for accurate modelling. In Genome-Wide Association Studies (GWAS), detecting the correct model for each SNP is often underpowered, and model choice is considered less critical since it minimally affects the detection of genomic risk loci. However, when constructing Polygenic Risk Scores (PRS) for phenotype prediction, model misspecification and omission of GxE terms can substantially reduce predictive power, particularly in out-of-sample populations if the distribution of environmental variables is different compared to the main study population. This study examines the theoretical implications of model misspecification, identifying scenarios where it matters and where it can be overlooked. We propose a novel, straightforward model to address polygenic GxE interactions arising from partial misspecification and compare its effectiveness against existing strategies used for polygenic modelling of gene environment interactions.

## **P20. Genome-wide association study meta-analysis in 63,936 women identifies 14 novel loci regulating menstrual cycle length and reproductive health**

Emily Morbey<sup>\*1</sup>, Felix R. Day<sup>1</sup>, Marc Vaudel<sup>2</sup>, Jack Murzynowski<sup>1</sup>, Ken Ong<sup>1</sup>, Stefan Johansson<sup>2</sup>, Siri Haberg<sup>3,4</sup>, and John Perry<sup>1</sup>

<sup>1</sup>MRC Epidemiology Unit, University of Cambridge, School of Clinical Medicine, Box 285, Institute of Metabolic Science, Cambridge, United Kingdom, <sup>2</sup>Department of Clinical Science, University of Bergen, Bergen, Norway, <sup>3</sup>Norwegian Institute of Public Health, Center for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway, <sup>4</sup>Department of Global Public Health and Primary, University of Bergen, Bergen, Norway

The female menstrual cycle is a complex and coordinated sequence of events, which controls the selection, maturation and ovulation of an oocyte, on average lasting 28 days. Observational studies have shown links between menstrual cycle length and regularity with reproductive and overall health, but few studies use genetics to determine the biological associations between these traits. While genome-wide association studies have implicated 5 signals associated with menstrual cycle length, none have adjusted for hormonal contraceptive use, or considered only menstrual cycle length within a regular 21–35-day window. Here, we leverage data from two European cohorts to perform the largest GWAS meta-analysis of menstrual cycle length to date in 63,939 women of European ancestry. We identify 14 novel signals for menstrual cycle length, in addition to 5 that had been reported in a previous GWAS meta-analysis of menstrual cycle length. We also identify causal associations between continuous menstrual cycle length in the normal range and various reproductive traits. We show that normal cycle variation can be an indicator of fertility outcomes as well as hormonal health and risk of menstrual disorders. This information can empower women to make informed decisions about their fertility and overall reproductive wellbeing.

## Poster highlight • P21. Adaptive dynamics of HIV-1 populations over a six-year-long experimental evolution: a genomic perspective

Ali Movasati<sup>1,2,3</sup>, Christine Leemann<sup>1</sup>, Kathrin Neumann<sup>1</sup>, Rongfeng Chen<sup>1</sup>, Lygeri Sakellaridi<sup>4</sup>, Karin Metzner<sup>1,2</sup>, and Roland Regoes<sup>5</sup>

<sup>1</sup>Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland, <sup>2</sup>Institute of Medical Virology, University of Zurich, Zurich, Switzerland, <sup>3</sup>Life Science Zurich Graduate School, University of Zurich, Zurich, Switzerland, <sup>4</sup>Institute for Virology and Immunobiology, University of Wurzburg, Würzburg, Germany, <sup>5</sup>Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

Numerous experimental evolution studies have suggested that adaptation rate of microbial populations evolving in stable environments declines over time. To investigate the characteristics of adaptation deceleration in a fast-evolving virus, we propagated HIV-1 in two human T-cell lines (MT-2 and MT-4) for ~6 years and tracked its genome evolution through NGS. The sequencing data can be explored via LTEEviz, an interactive web application. Time-resolved sequencing data indicated that despite a constant fixation rate of 0.085 (MT-2) and 0.042 (MT-4) mutations per generation, the fixation kinetics of adaptive mutations changed considerably over time. The difference in sequence evolution rate between the two host cell lines can be partly attributed to the larger effective population sizes of MT-4 lines. Moreover, the rate of fixation of adaptive mutations decreased by 44% per 300 generations in both host cell lines, while their conferred fitness gain diminished by 27% (MT-2) and 18% (MT-4) per every added adaptive mutation in their genetic background. Furthermore, we identified unique yet consistent patterns of sequence evolution among different regions of the HIV-1 genome. In particular, *nef* and *vpr* accessory genes demonstrated patterns of random evolution, expected in the absence of selection. In line with this expectation, the evolving populations of HIV-1 acquired and fixed loss-of-function mutations in *nef* and *vpr* throughout the experiment. Additionally, *nef* gene repeatedly underwent large deletions, leading to the removal of approximately 400 bp (~4.3% of the HIV-1 genome). These large *nef* deletions increased in frequency faster than expected under neutrality and in a length-dependent manner.

## **P23. Systematic brain-region analysis of aging-associated lncRNAs links synaptic decline to glial-metabolic activation**

Amna Obaid<sup>1</sup>, Mohammad Uzair<sup>\*2</sup>, and Shahid Bashir<sup>\*3</sup>

<sup>1</sup>International Islamic University, Islamabad, Pakistan, <sup>2</sup>King Fahd University for Petroleum Minerals, Dhahran, Saudi Arabia, <sup>3</sup>Research Center, King Fahad Specialist Hospital, Dammam, Saudi Arabia

Long non-coding RNAs (lncRNAs) have been implicated in aging biology, yet brain region patterns of aging-associated lncRNA expression remain incompletely characterized. We performed a brain-region replication and extension of a GTEx-based aging framework using GTEx v11 bulk RNA-seq from 13 adult human brain regions (n=181-300 samples/ region). Within each region, we modelled gene expression as a function of age and sex and defined aging-associated lncRNAs using a false discovery rate (FDR) < 0.05 and —fold-change from 25 to 75 years— > 1.5. Brain-region specificity was quantified by the *tau* (Tau) index across regions. We performed GTEx-internal co-expression analysis for region- and direction-specific age-lncRNAs, retained the top 5% co-expressed coding genes expressed in the same region (median TPM > 1), and conducted GO/KEGG enrichment. We identified 1,192 region-level aging-associated lncRNA signals, out of which 625 were unique across 12 regions (none in spinal cord cervical C-1). The anterior cingulate cortex and frontal cortex had the highest number of age-lncRNAs, i.e., 209 and 207, respectively. lncRNAs were more region-restricted than other genes (median *tau* 0.741 vs 0.617; Wilcoxon  $p < 2.2 \times 10^{-16}$ ), whereas age-lncRNAs were significantly less region-specific than background lncRNAs (median *tau* 0.646 vs 0.744;  $p < 2.2 \times 10^{-16}$ ). Co-expression enrichment showed strong directionality. Downregulated age-lncRNA networks were consistently enriched for synaptic and neurotransmission processes. In contrast, upregulated networks were enriched for gliogenesis and lipid metabolic pathways, including fatty acid degradation. Therefore, these results highlight convergent region-level aging programs and motivate donor-level single-nucleus validation for cell-type attribution.

## **Poster highlight · P24. Using Bayes factor design analysis to quantify a minimum level of evidence in family-based studies**

Cathal Ormond<sup>1</sup> and Elizabeth Heron<sup>1</sup>

<sup>1</sup> Discipline of Psychiatry, School of Medicine, Trinity College Dublin, Dublin, Ireland

Family-based study designs can be useful to identify rare variants that increase individual liability for a genomic trait. However, robust statistical approaches to discover and prioritise such variants have been lacking. To address this, we developed a Bayesian inference model to evaluate the causality of rare variants in pedigrees (BICEP) for both Mendelian and complex genetic architectures. The co-segregation module of BICEP can be applied to pedigrees of any size. However, the resulting Bayes factors can be challenging to interpret for non-specialist users, and selecting an appropriate sufficient level of evidence is non-trivial. To mitigate this, we applied the principles of Bayes factor design analysis (BFDA, Schönbrodt and Wagenmakers 2018), to the BICEP model. Using BFDA, we can determine if a given pedigree structure and phenotypes can provide sufficient evidence to distinguish between rare causal and neutral variants. This also allows us to determine the evidence thresholds specific to individual pedigrees and to translate these thresholds to more intuitive metrics such as sensitivity, specificity etc. We use simulated data to explore how pedigree size and the genetic model as well as the Bayesian model parameters impact the BFDA results. This new theoretical framework requires only the pedigree structure and the binary phenotypes. This approach aids planning of pedigree analyses prior to generating, often costly, genetic data and can provide greater interpretability to users for rare-variant pedigree analysis.

## P25. Search for genetic interactions in Alzheimer's disease

Sagnik Palmal<sup>\*1</sup>, Eadb Consortia<sup>2</sup>, Eadi Consortia<sup>3</sup>, Bonn Consortia<sup>4</sup>, Gerad Consortia<sup>5</sup>, Demgene Consortia<sup>6</sup>, Gr@ace/degescos Consortia<sup>7</sup>, and Céline Bellenguez<sup>1</sup>

<sup>1</sup>Facteurs de Risque et Déterminants Moléculaires des Maladies liées au Vieillessement, U 1167, Institut National de la Santé et de la Recherche Médicale (Inserm), Université de Lille, Centre Hospitalier Régional Universitaire [CHU Lille], Institut Pasteur de Lille, Lille, France, <sup>2</sup>European Alzheimer Dementia Biobank, Institut Pasteur Lille, BP 245, Lille, France, <sup>3</sup>European Alzheimer Disease Initiative, Department of Neurology, Institute of Memory and Alzheimer's Disease (iM2A), Pitié-Salpêtrière Hospital, Ap-Hp, Paris, France, <sup>4</sup>Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany, <sup>5</sup>Genetic and Environmental risk in Alzheimer's Disease consortium, Division of psychological Medicine and Clinical Neurosciences, MRC Center for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, United Kingdom, <sup>6</sup>Demgene, Oslo, Norway, <sup>7</sup>GR@ACE/DEGESCO, Barcelona, Spain

Alzheimer's disease (AD) is the leading cause of dementia and results from both genetic and non-genetic factors: in addition to 14 modifiable risk factors of dementia, more than 90 genetic loci are associated with AD risk. However, gene-gene and gene-environment interactions have largely been overlooked in the context of AD. Such interactions are inherently challenging to detect given their modest effect sizes and multiple testing burden considering the large number of variants and environmental factors to be tested. To reduce the search complexity, we scanned the genome of AD cases to identify variants impacting age-at-onset variability, as such variants are expected to be enriched for putative genetic interactions in AD. We analyzed the autosomes of 23,591 individuals of European ancestry from the EADB, EADI, GERAD, Bonn and Demgene studies across 13 countries using both Levene's test (OSCA software) and a quantile integral linear model (QUAIL software). Analyses were adjusted for principal components and genotyping centers. However, we observed genomic inflation due to population structure while using Levene's test but not while using QUAIL. Therefore we performed meta-analysis of QUAIL results across the five studies and detected a genome-wide significant association in APOE ( $P < 5 \times 10^{-8}$ ) and two suggestive signals ( $P < 1 \times 10^{-6}$ ) on chromosomes 12 and 18. We now aim to extend this analysis to other studies available to us to obtain a comprehensive list of variants for GxG and GxE interaction testing in AD.

## Poster highlight • P26. Pharmacogenetic Signals in 140,000 UK Biobank Participants: Known Associations and Novel Gene-Drug Insights

Maria Pieczarka<sup>1</sup>, Paula Konowalska<sup>1</sup>, Jacek Hajto<sup>1</sup>, Sylwia Grubarek<sup>1</sup>, Pawel Pienkowski<sup>1</sup>, Bartosz Baszkiewicz<sup>1</sup>, Michal Bożyk<sup>1</sup>, Rafal Kafel<sup>1</sup>, Dzesika Hoinkis<sup>1</sup>, Marcin Piechota<sup>1</sup>, Malgorzata Borczyk<sup>\*,1</sup>, Michal Korostynski<sup>\*,1</sup>

<sup>1</sup>Laboratory of Pharmacogenomics, Maj Institute of Pharmacology, Polish Academy of Sciences, Kraków, Poland.

\*Corresponding authors

Pharmacogenetics is a key component of precision medicine, with several gene-drug pairs already embedded into clinical guidelines. Our study leverages carefully extracted large-scale prescribing records from the UK Biobank analysed for association with a curated set of CYP and HLA variants. While CYP genes are well-known pharmacogenetic loci that influence drug pharmacokinetics and pharmacodynamics, HLA alleles are less commonly studied in relation to gene-drug interactions, but may play a non-negligible role in immune-mediated drug responses, including hypersensitivity reactions and drug intolerance. Prescribed doses and treatment durations were algorithmically extracted from primary care prescription data, accounting for intervals and significant dose changes over time. Genetic predictors included 362 standardized HLA variants and pharmacogene activity scores for six cytochrome P450 (CYP) genes. We robustly replicated known associations for warfarin dose (*CYP2C9* and *CYP4F2*, BH-adjusted  $p = 2e-80$  and  $8e-5$ ) and amitriptyline (*CYP2D6*,  $p = 8.8e-3$ ). Novel findings include associations of diazepam dose and gabapentin treatment duration with CYP pharmacogenes, and dihydrocodeine dose with HLA variants, potentially reflecting differences in tolerability, clinical response, or prescribing practices. This work highlights the UK Biobank's unique value - combining deep phenotypic, genetic, and longitudinal primary care data - as a powerful resource for large-scale pharmacogenetic discovery.

The project is funded by the Medical Research Agency with funds from the KPO (KPOD.07.07-IW.07-0099/24). The KPO is financed from the Recovery and Resilience Facility (RRF).

## **P27. Mathematical modeling of multi trait fine mapping using single cell expression quantitative trait loci graphical priors and linkage disequilibrium uncertainty**

Prihantini Prihantini\*<sup>1</sup> and Rifaldy Fajar<sup>2</sup>

<sup>1</sup>Department of Mathematics, Bandung Institute of Technology, Bandung, Indonesia, <sup>2</sup>Computational Biology and Medicine Laboratory, Yogyakarta State University, Yogyakarta, Indonesia

**Background/Aim:** Modern fine mapping often produces overconfident credible sets due to linkage disequilibrium uncertainty, multi-trait heterogeneity, and biased tissue enrichment. We aimed to develop a mathematical modelling framework that jointly performs multi-trait fine mapping and cell-type enrichment using single-cell regulatory evidence while preserving validity.

**Methods:** We analyzed summary statistics from 18 immune and metabolic traits obtained via OpenGWAS, covering 2,314 independent loci, with multi-ancestry linkage disequilibrium references from the 1000 Genomes Project. A hierarchical probabilistic graphical model was specified in which genome-wide association summary statistics followed a multivariate normal likelihood, incorporating shrinkage priors on linkage disequilibrium matrices. Latent causal indicators were decomposed into shared and trait-specific components. Cell-type-specific prior probabilities were derived from single-cell expression quantitative trait loci measured in GEO GSE196830 and incorporated through a measurement-error layer. Inference was performed using variational Bayesian optimization with expectation propagation.

**Results:** In spike-in simulations, nominal 95% credible sets achieved 95.6% empirical coverage (95% confidence interval 94.2–96.8), compared with 86.9% (84.1–89.3) for benchmark methods ( $p < 1 \times 10^{-5}$ ), while reducing median credible set size by 34%. Posterior inclusion probability calibration improved (Brier score reduction 0.018,  $p = 0.0003$ ). When applied to observed trait data, the model identified 27 shared causal signals across related traits and yielded cell-type enrichment results that remained controlled under linkage

disequilibrium– matched null models (false discovery rate 5.0%, genomic inflation factor 1.02).

Conclusions: This framework offers a principled approach to multi-trait fine mapping with robust uncertainty control and interpretable cell-type enrichment, enabling more reliable causal variant prioritization in complex trait genetics.

## **P28. T-Rex: A cross-platform tool for rare variant detection in whole-exome TRIO studies**

Sara-Luisa Reh<sup>1,2</sup>, Carolin Walter<sup>3</sup>, Judith Lohse<sup>4</sup>, Tabita Ghete<sup>5,6,7</sup>, Markus Metzler<sup>5,6,7</sup>, Anne Quante<sup>2,8</sup>, Julia Hauer<sup>1,2,9</sup>, Franziska Auer<sup>1,2,6</sup>

<sup>1</sup>School of Medicine and Health, Department of Pediatrics, Technical University of Munich, Munich, Germany, <sup>2</sup>Bavarian Cancer Research Center (BZKF), Munich, Germany, <sup>3</sup>Institute of Medical Informatics, University of Münster, Münster, Germany, <sup>4</sup>Pediatric Hematology and Oncology, Department of Pediatrics, University Hospital and Faculty of Medicine Carl Gustav Carus, Dresden University of Technology (TUD), Dresden, Germany, <sup>5</sup>Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Erlangen, Germany, <sup>6</sup>Pediatric Oncology Network Bavaria, KIONET, Germany, <sup>7</sup>Bavarian Cancer Research Center (BZKF), Erlangen, Germany, <sup>8</sup>Institute of Human Genetics, TUM School of Medicine and Health, Klinikum rechts der Isar, TUM University Hospital, Technical University of Munich (TUM), Munich, Germany, <sup>9</sup>German Center for Child and Adolescent Health (DZKJ), partner site Munich, Germany

Whole-exome sequencing (WES) enables the identification of rare germline variants contributing to pediatric diseases. TRIO study designs, comparing affected children with their parents, are particularly effective for rare disease genetics. However, WES data analysis requires bioinformatics expertise, varies across institutions, and is often incompatible with clinical workflows. We developed T-Rex, a cross-platform desktop application for standardized and local analysis of WES germline TRIO data without requiring programming knowledge. T-Rex integrates state-of-the-art tools for alignment, dual-variant calling (GATK HaplotypeCaller + VarScan2), and SNPEff/SNPSift annotation, and implements statistical testing, including Transmission Disequilibrium Test (TDT) with multiple-testing correction. Performance evaluation showed linear time complexity and constant space complexity. User acceptance testing (n=13) confirmed that clinicians, researchers, and medical doctoral candidates with diverse technical backgrounds can learn to operate T-Rex quickly. Application to a published cohort of 121 pediatric cancer TRIOs, filtering for rare protein-coding variants (MAF  $\leq$  0.1% in gnomAD v4.0), reproduced all assessable previously reported pathogenic variants. T-Rex enables clinicians to analyze WES TRIO data in compliance with data protection regulations without requiring additional software licenses. The platform accurately reproduces previously reported pathogenic variants, demonstrating its reliability. As the first platform for WES analysis that requires no programming knowledge,

T-Rex has the potential to foster collaborative research between clinics, reducing reliance on external providers.

## **P29. Design and development of a results relational database for cardiovascular phenotypes**

Cristian Riccio\*<sup>1</sup>, Amra Dhabalia Ashok<sup>1</sup>, Linlin Guo<sup>2,3,4</sup>, Georgios Koliopoulos<sup>1</sup>, Tanja Zeller<sup>3,5</sup>, Raphael Twerenbold<sup>2,3,4</sup>, and Andreas Ziegler<sup>1,2,4,6</sup>

<sup>1</sup>Cardio-CARE, Medizincampus Davos, Davos, Switzerland, <sup>2</sup> University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>3</sup>German Center for Cardiovascular Research (DZHK), Hamburg/Kiel/Lübeck site, Germany, <sup>4</sup>Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, <sup>5</sup>Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany, <sup>6</sup>Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

**Introduction:** Advances in omics technologies have enabled multi-omics research of complex cardiovascular phenotypes. However, integrating and querying results across large public and in-house datasets remains challenging for interdisciplinary teams. The aim of this study was to design, implement, and evaluate a relational database system that facilitates the integrated exploration of large-scale cardiovascular genetic association results.

**Methods:** To address these challenges, we designed and implemented a relational database for the integration of large-scale genetic and functional genomics resources, including the GWAS Catalog, GTEx, and in-house association results. Product specifications were developed through close collaboration with biologists, clinicians, data scientists, and bioinformaticians to ensure usability, scalability, and compliance with data protection requirements. A commercial database solution was subsequently selected.

**Results:** The relational database was combined with a business intelligence and analytics platform to enable interactive exploration of genetic association results. The system supports structured storage of association results and metadata, together with querying via both SQL and a graphical user interface. Users can query results, visualise findings interactively, and download query outputs. Access permissions can be tailored at the user level, ensuring that each user can access the results relevant to their research.

Conclusion: The resulting relational database provides a scalable, collaborative platform for the integration and querying of multimodal cardiovascular phenotypes.

## **P30. Are controls needed? - HWE deviation test**

Albert Rosenberger\*<sup>1</sup>

<sup>1</sup>Department of Genetic Epidemiology, University Medical Center  
Göttingen, Göttingen, Germany

Comparing the genotype distribution of a genomic marker between cases and controls can be considered a standard procedure for estimating genotypic or allelic relative risks. The structural similarity of cases and controls with regard to other influencing characteristics is essential for an unbiased estimate. This is either assumed, achieved through matching, or must be approximated using analytical methods (e.g. via propensity score or target trial emulation). The validity of genotyping is checked, among other things, using HWE in controls. The odds ratio must be estimated as a surrogate for the RR. In contrast, for a study with a case-only design only the genotype distribution of cases is determined and analyzed. These have so far been used exclusively to detect multiplicative gene-gene or gene-environment interactions based on the work of Piegorsch 1994. The correctness of the independence assumption is crucial here.

It is possible to estimate RR from the deviation of the genotype distribution in cases from the HWE. Inbreeding or admixture can be taken into account. However, the minor allele frequency (MAF) should be approximately known. Thanks to publicly available databases, this is often known with sufficient accuracy for different genetic ethnicities. The model presented can be extended for GxG and GxE. The time-consuming and cost-intensive collection and genotyping of biological samples from control subjects, as well as the assumption of structural equality and the use of OR, are no longer necessary. Today, there are sufficient biobanks available for the replication and validation of genomic associations discovered in case-only designs.

## **P32. Independent genetic signals underlying recurrent pregnancy loss and smoking at a shared locus**

Yevheniya Sharhorodska\*<sup>1,2</sup>, Vincent Paskat<sup>3</sup>, Oleg-Roman Gnateyko<sup>2</sup>, Danuta Zastavna<sup>2</sup>, Halyna Makukh<sup>4,5</sup>, Scapoli Chiara<sup>1</sup>, and Inga Prokopenko<sup>6</sup>

<sup>1</sup>University of Ferrara, Ferrara, Italy, <sup>2</sup>Institute of Hereditary Pathology, National Academy of Medical Sciences, Lviv, Ukraine, <sup>3</sup>Imperial College London; London, United Kingdom, <sup>4</sup>Scientific Medical Genetic Centre “LeoGENE”, Lviv, Ukraine, <sup>5</sup>Ivan Franko National University of Lviv, Lviv, Ukraine, <sup>6</sup>University of Surrey, Guildford, United Kingdom

Background: Recurrent pregnancy loss (RPL) affects approximately 5% of women and remains unexplained in up to half of cases. Smoking is a causal risk factor for spontaneous miscarriage, supported Mendelian randomization studies. However, the pathophysiological mechanisms linking smoking behaviour to RPL remain poorly understood. We aimed to investigate the genetic architecture of RPL using GWAS meta-analysis and used colocalisation to estimate the shared mechanistic link with smoking.

Methods: We performed ancestry-specific GWAS of recurrent miscarriage across five ancestries using BOLT-LMM in the UK Biobank data and analysed within SNPTEST TOPMedimputed data from the Ukrainian LUCAR (Lviv Ukrainian Cohort for Advancing Reproductive Health) cohort, comprising women with clinically confirmed idiopathic RPL and fertile controls. GWAS results from six datasets (15,687 cases/136,716 controls) were combined using trans-ancestry meta-regression implemented in MR-MEGA. Colocalisation analyses were conducted to determine whether RPL-associated loci share causal variants with smoking related GWAS signals.

Results: Using trans-ancestry GWAS meta-analysis of RPL, we identified a significant association at rs4852592-G (OR=1.0018, 95% CI 0.976–1.029; P=3.52×10), mapping to locus near *CTNNA2* gene, next to an established variant rs1492546 for smoking status. Colocalisation analysis of the RPL lead variant rs4852592 indicated that the RPL and smoking association signals at this locus are likely driven by distinct causal variants, with strong support for independent associations (posterior probability for different causal variants, H3=0.96).

Conclusion: Our findings identify RPL-associated locus near *CTNNA2* that overlaps a smoking-associated genomic region but is driven by independent causal variants, highlighting complex genetic architecture linking smoking and RPL.

## **P33. Robust mendelian randomization estimation using quantile regression under correlated and uncorrelated pleiotropy**

Julien St-Pierre<sup>1,2</sup>, Mireille Schnitzer<sup>1,3</sup>, Marc-André Legault<sup>1,2</sup>

<sup>1</sup>Faculté de Pharmacie, Université de Montréal, Montréal, Canada, <sup>2</sup>Centre de recherche du CHU Sainte-Justine, Université de Montréal, Montréal, Canada, <sup>3</sup>École de santé publique, Département de médecine sociale et préventive, Université de Montréal, Montréal, Canada

The validity of Mendelian randomization (MR) studies strongly depends on three assumptions required for a valid instrumental variable (IV), which are: (A1) The IV is associated with the exposure X; (A2) The IV is not associated with the outcome Y conditional on X and the unmeasured confounder U; and (A3) The IV is not associated with U. However, GWAS studies have shown evidence that hundreds of individual genetic variants have pleiotropic effects. Uncorrelated pleiotropy occurs when genetic variants have a direct effect on the outcome that is not mediated by the exposure, violating assumption A2, while correlated pleiotropy occurs when genetic variants affect the exposure and outcome via shared heritable confounders, violating assumption A3. In this work, we propose a MR method based on quantile regression (QR) that is robust to both violations of A2 and A3 and depends only on the plurality-valid assumption. Our method is a generalization of the median estimator to cases where possibly more than 50% of IVs are invalid. We propose a procedure for selecting the optimal quantile of the ratio estimates through a likelihood-based formulation of QR using the Asymmetric Laplace distribution (ALD). Due to its heavier tails, the ALD is more robust than the Gaussian distribution to outlying ratio estimates arising from invalid instruments with pleiotropic effects. If the sample distribution of MR ratios contains a dominant mode corresponding to valid instruments, then the ALD provides a working likelihood in which maximum-likelihood estimation targets the tau-th quantile as the mode of this distribution.

## **P35. FastHer: A fast GREML-equivalent-accuracy estimator of local SNP heritability with per-iteration linear complexity**

Gulnara Svishcheva<sup>1,2</sup>, Yakov Tsepilov<sup>1,3</sup>, and Tatiana Axenovich<sup>1</sup>

<sup>1</sup>Institute of Cytology and Genetics, Novosibirsk, Russia, <sup>2</sup>Vavilov Institute of General Genetics, Moscow, Russia, <sup>3</sup>Wellcome Sanger Institute, Hinxton, United Kingdom

Accurate estimation of local SNP heritability ( $h^2$ ) is fundamental for dissecting the genetic architecture of complex traits. The gold standard for this estimation is GREML (Yang et al., 2011), which, however, requires individual-level data. To leverage only summary statistics, HEELS was recently developed (Li et al., 2023). While HEELS preserves the key statistical properties of GREML, its cubic computational complexity per iteration renders the analysis of extended genomic regions computationally prohibitive.

We present FastHer, a novel summary-statistics-based maximum-likelihood estimator that provides statistical accuracy equivalent to HEELS while achieving linear per-iteration complexity. The key innovation is an exact analytical reformulation of the likelihood. Specifically, a one-time eigen-decomposition of the LD matrix ( $U = V\Lambda V'$ ) transforms the optimization problem from the matrix space ( $U, z$ ) to the vector space of eigenvalues ( $\Lambda$ ) and squared projected z-statistics ( $(V'z)^2$ ). This reformulation eliminates the need for matrix inversions during iterative search, reducing each iteration to simple vector operations and thus guaranteeing linear complexity.

Using simulations on real genotypes, we showed that FastHer produces results consistent with GREML and HEELS. Furthermore, using summary statistics, FastHer achieved complete agreement with HEELS for both the  $h^2$  estimate and its standard error. Crucially, FastHer delivers a dramatic practical speedup. For a genomic region with ~10,000 SNPs, it achieved an approximately 100-fold runtime reduction.

In summary, FastHer solves the critical computational bottleneck, enabling rapid and accurate genome-wide estimation of local SNP heritability directly from summary statistics. The tool is freely available as an R package: <https://github.com/gulsvi/FastHer>.

## **P36. Protein language model embeddings for gene-level association testing**

Morgan Thomas\*<sup>1</sup>

<sup>1</sup>University of Edinburgh, Edinburgh, United Kingdom

Genetic association testing is currently fragmented by allele frequency and effect size. Genome-wide association studies (GWAS) test millions of common variants with stringent multiple-testing correction, requiring large sample sizes. Rare variant aggregation tests collapse variants within a gene but rely on restrictive variant filtering with assumptions about deleteriousness and cannot naturally incorporate common variants. In addition, both approaches assume variants act independently, missing compounding of multiple variants within the same gene. We propose lifting association testing to protein functional space using protein language model embeddings of personal protein sequences from phased haplotypes from whole genome sequencing data. These embeddings encode predicted structural and functional properties, which we test against phenotype using kernel aggregation methods. This unified gene-level framework jointly models common and rare variants without variant filtering or functional prediction pipelines, reducing testing from millions of variants to ~20,000 genes while capturing epistatic effects. Unlike existing PLM applications that filter variants by deleteriousness for standard tests, we directly test gene-level associations analogous to GWAS. As a proof of concept, we analysed m6 candidate genes previously identified as genome-wide significant through standard GWAS in the GenOMICC COVID-19 cohort (N=6,641 severe, N=9,857 mild). Using Profluent-E1 embeddings, we successfully detected nominally significant associations for all candidate genes, including 2 driven by rare variants, illustrating our method's ability to capture both common and rare variant effects within a unified framework. In ongoing work, we are expanding this analysis to the whole proteome to systematically compare power with variant-level GWAS and identify novel protein-disease associations.

## **P37. Recent PRS developments in ancestry and context-specificity**

Elia Tiso\*<sup>1</sup>, Linda Repetto<sup>1</sup>, Kristi Läll<sup>1</sup>, Märt Möls<sup>1</sup>, Massimo Mezzavilla<sup>2</sup>, Luca Pagani<sup>1,2</sup>, Reedik Magi<sup>1</sup>, and Mait Metspalu<sup>1</sup>

<sup>1</sup>Institute of Genomics, Tartu, Estonia, <sup>2</sup>Dipartimento di Biologia, Padua, Italy

Polygenic Risk Scores (PRS) are increasingly used in precision medicine, yet their predictive performance varies across ancestry, sex, and other contexts, raising concerns about fairness and clinical utility. Recent methodological advances aim to improve portability and fairness through ancestry-aware modeling, privacy-preserving frameworks, and more interpretable designs, on top of increasing the overall performance. Despite progress, challenges remain in balancing accuracy, scalability, and transparency, which are essential for clinical implementation. Given the rapid pace of methodological development, existing reviews may no longer fully capture the state of the field. We present a systematic review that addresses this gap by highlighting strategies to improve PRS portability and fairness, methodological innovations and future directions for clinical translation. We focused on methodological developments from mid-2023 to late 2025, prioritising approaches that address context-specificity, especially ancestry. Using a structured PubMed query, we retrieved methods that introduced innovative frameworks for ancestry-aware or context-calibrated prediction which we classified into focus-related categories. For each method, we extracted core features and commented on the improvements over the PRS landscape. We then concluded with some notes on privacy-preserving strategies and ensemble learning pipelines, relevant challenges and opportunities to reach a broader application of PRS in disease prediction. We recognise notable improvements in ancestry-awareness and handling, and, overall, general improvements in input variability and framework's techniques are clear, narrowing the gap between PRS and clinical application. Despite these advances, challenges remain in balancing accuracy, interpretability, and scalability.

## **P38. Program-level heritability analysis prioritizes amygdala transcriptional modules in metabolic syndrome**

Mohammad Uzair\*<sup>1</sup>, Sadam Al-Azani<sup>1</sup>, and Moataz Dowaidar<sup>1</sup>

<sup>1</sup>King Fahd University for Petroleum Minerals, Dhahran, Saudi Arabia

Metabolic syndrome (MetS) is a polygenic cardiometabolic trait with emerging evidence for central nervous system contributions, yet the transcriptional programs concentrating common-variant risk remain unclear. We integrated human brain single-nucleus RNA-seq data (558,661 nuclei after quality control) across 14 meta-regions (105 dissections) to derive expression-specificity genomic annotations. Cell-type S-LDSC supported a neuronal component of MetS heritability ( $p < 0.001$ ), and regional S-LDSC highlighted the amygdala as the strongest signal ( $p = 0.0056$ ). To move beyond discrete cell types and regions, we performed region-stratified non-negative matrix factorization (NMF) in prioritised regions and converted gene loadings into continuous SNP annotations ( $\pm 100$  kb). Program-level S-LDSC identified multiple nominally enriched amygdala programs, led by Amygdala GP7 ( $p = 0.00135$ ) and GP3 ( $p = 0.022$ ). The additional programs (GP5, GP6 and GP8) showed supportive nominal signals, although none survived correction across programs (BH-FDR;  $q \approx 0.17$ ). In contrast, cerebellar programs showed no comparable SNP-level enrichment ( $p \geq 0.092$ ), despite MAGMA prioritising cerebellum at the gene-property level ( $p > 0.001$ ), which indicates method-dependent sensitivity across analytic scales. KEGG GSEA of MetS nominal amygdala programs highlighted neuromodulatory and synaptic signalling, e.g., retrograde endocannabinoid signalling in GP3 ( $p < 0.001$ ) and neuroactive ligand-receptor interaction in GP6 ( $p < 0.001$ ), alongside neuroendocrine/secretory modules (e.g., insulin secretion and GnRH secretion in GP5/GP7). These results prioritise limbic gene programs as candidate axes linking polygenic risk to neural and endocrine regulation in MetS.

## **P39. Simulation study of continuous outcome with genetics-by-age interaction**

Jason Young\*<sup>1</sup>, Luke Pilling<sup>1</sup>, Jane Masoli<sup>1</sup>, and Jack Bowden<sup>1</sup>

<sup>1</sup>University of Exeter, Exeter, United Kingdom

**Introduction:** Most genetic studies are cross-sectional and assume a linear effect of genotype across lifespan. Evidence is accumulating for gene-by-age interactions, with recent work by Winkler and Schoeler assessing ways to investigate this with cross-sectional or longitudinal data. Yet, there isn't yet a well-established framework for determining optimal types of data and analysis methods for a given dataset. This study aims to interrogate the statistical power and robustness of a range of different analysis options for detecting and estimating gene-by-age interactions.

**Methods:** Simulation analysis investigated the interaction model with gene-by-age interaction assumption, marginal model without the assumption, and change model estimating difference in outcome per year. Synthetic data consisting of outcome and age generated by underlying linear summations were linearly regressed by the three models in different setups with varying degrees of the measurement gaps, proportion of longitudinal data to the total sample size, and confounding.

**Results:** Longer time between two measurements increased the power of change model to identify gene-by-age effects. Change model was robust to time-invariant confounding. Marginal model was susceptible to type 1 error. Interaction model was robust against confounding in the presence of main effects.

**Conclusion:** Next steps of my PhD include investigating the nature and extent of the interaction model's protection effect. I will establish a framework to inform model choice for a given dataset to increase power and avoid type 1 error in gene-by-age analysis. I am developing the framework to consider other cohorts in the future with different levels of longitudinal data.

## **Poster highlight · P40. Genetic topic modelling for interpretable clustering: a demonstration on human complex diseases**

Leyi Zhang\*<sup>1</sup>, Christof Seiler<sup>1,2,3</sup>, Doug Speed<sup>4</sup>, Raphael Micheroli<sup>1</sup>, and Caroline Ospelt<sup>1</sup>

<sup>1</sup>Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland, <sup>2</sup>Department of Advanced Computing Sciences, Maastricht University, Maastricht, Netherlands, <sup>3</sup>Mathematics Center Maastricht, Maastricht University, Maastricht, Netherlands, <sup>4</sup>Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark

**Background and methods:** Genome-wide association studies (GWAS) often focus on individual traits, overlooking shared genetic architectures. We developed SNPic (SNP topic modeling for interpretable clustering), a generative probabilistic framework that applies Latent Dirichlet Allocation (LDA) to GWAS summary statistics. SNPic models diseases as mixtures of latent “genetic topics,” each representing aggregated gene sets from significant SNPs. We applied SNPic to FinnGen consortium data across four categories: Cancer, Autoimmune, Metabolic/Cardiovascular, and Psychiatric disorders.

**Results:** SNPic effectively grouped diseases by biological similarity. Topic distributions clustered autoimmune conditions (e.g., lupus, rheumatoid arthritis, psoriasis) and metabolic disorders (e.g., gout, type 2 diabetes). Top genes per topic aligned with known mechanisms: HLA-region genes dominated autoimmune topics, while others linked to lipid metabolism and neural development. SNPic+PCA visualization confirmed coherent clustering and revealed novel latent relationships, such as the genetic proximity between Marginal Zone Lymphoma and autoimmune diseases—a finding consistent with recent epidemiological evidence.

**Conclusion:** SNPic provides an interpretable, unsupervised approach to map the shared genetic landscape of complex traits. By moving beyond single-trait paradigms, this framework uncovers cross-disease biological mechanisms and offers a transparent tool for disease stratification and drug repurposing. SNPic can also reveal connections in plants or animal traits. Our results demonstrate SNPic as a generalizable model for integrative genomics, capable of identifying shared pathways that traditional GWAS might overlook.

## **P41. Investigating the impact of modelling SNP interactions on the performance of polygenic risk scores**

Jingqi Zhu\*<sup>1</sup>

<sup>1</sup>University of Manchester, Manchester, United Kingdom

Polygenic risk scores (PRSs) summarise an individual's genetic predisposition to disease and hold great potential for personalised disease risk prediction. Standard PRSs aggregate the effects of multiple single nucleotide polymorphisms (SNPs) across the genome using an additive linear model that assumes no interaction between SNPs. The predictive performance of PRSs is believed to improve if SNP interactions can be captured. Although machine learning models have been recently implemented to integrate SNP interactions into PRSs, the specific conditions under which they outperform standard PRSs have not yet been systematically evaluated. This work aims to use simulations to quantify the loss of predictive performance due to not modelling SNP interactions in PRSs, and the level of SNP interactions under which random forest and XGBoost outperform standard PRSs. Our findings suggest that the reduction in PRS predictive performance by not allowing for SNP interactions is only substantial when interaction effects are large. Random forest and XGBoost PRS only outperform standard PRS under strong interaction effects. The benefit of modelling SNP interactions therefore depends critically on the underlying genetic architecture of the disease.