Abstracts

INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

The 2025 Annual Meeting of the International Genetic Epidemiology Society

4

Genomic Determinants of Cervical Cancer Risk

Dhanya Ramachandran[1], Theresa Beckhaus[1], Dandan Liao[1], Rieke Eisenblätter[1], Peter Schürmann[1], Joe Dennis[2], Douglas Easton[2], Peter Hillemanns[3], Thilo Dörk[1], Cervigen Consortium
[1]Gynecology Research Unit, Hannover Medical School, Germany; [2]University of Cambridge, Centre for Cancer Genetic Epidemiology, Cambridge, United Kingdom; [3]Clinics of Gynecology and Obstetrics, Hannover Medical School, Germany

Cervical cancer is the second leading cause of cancer-related deaths in women worldwide in 2023. High-risk HPV infection causes lesions, with persistence increasing the risk of high-grade disease. The Cervigen Consortium, comprising of nine German centers, aims to identify risk factors through genetic case-control association studies in over 1,000 cases with invasive cervical cancer, over 1,500 cases with cervical dysplasias and over cancer-free 1,000 controls. Genome-wide association studies (GWASs) for cervical cancer have identified multiple genomic variants at the human leukocyte antigen (HLA) locus (6p21.32-33), important for immune response to HPV. In addition, genomic susceptibility loci at chromosomes 2 (*PAX8*), 5 (*TERT/CLPTM1L*) and 17 (*GSDMB*), have been reported. We have validated previously reported candidate gene loci and confirmed significant associations with multiple independent variants at the *HLA* locus (chr6), one variant at *PAX8* (chr2) and one variant at *GSDMB* (chr17) in our population. We further performed a genome-wide association study, identified, and replicated a novel signal at *PRKD1* (chr14). We investigated the functional relevance of variants by expression quantitative trait loci (eQTL) analysis in patient material, luciferase assays and gene transcript modulation in cervical cell lines. These results bring us closer to understanding key contributors to cervical cancer etiology.

Keywords: cervical carcinoma, GWAS, SNP, eQTL, HPV

5

Improved Methods for Analyzing Small Samples in High Dimensions With Application to Genome-Wide Association Studies of Rare Diseases

Anat Reiner-Benaim[1], Diana Valverde Pérez[2], Carlos López Solarat[2] and Sebastian Döhler[3]
[1]Ben-Gurion University of the Negev, Israel; [2]University of Vigo, Spain; [3]Darmstadt University of Applied Sciences, Germany

Background. Rare diseases (RDs) are a group of severe, chronic, degenerative, and often life-threatening conditions, which are typically complex genetic disorders. Common genetic variation, which modulates the effect of disease-causing genes, determines the RD expressiveness and severity. Studying this variation can help in assessing the risk of developing severe symptoms. Genome-wide association studies (GWAS) are used to statistically test numerous genetic variants for an association with severe RD symptoms. Multiple testing procedures are used to adjust for the inflation in type I error. However, only small samples of patients are typically available in RD studies. As a result, exact non-parametric statistical tests are used, which have low power due to their discreteness.

Objectives. This study aimed to develop statistical methodology for detecting weak and rare effects in high-dimensional and small-sample data, and to employ these methods for discovering variants associated with severe RD symptoms, using the Bardet-Biedl Syndrome (BBS) as an example.

Methods. We use methods for the discrete setting on simulated and real datasets and compare their performance. We apply them on BBS GWAS data to identify candidate SNPs for common variation effect.

Results. Incorporating discreteness into state-of-the-art multiple testing procedures such as the BH procedure increases their power and proves critical in the scenario of small counts and small and sparse effects. The discrete BH procedure enabled the discovery of 647 SNPs, compared to 129 SNPs by the classical BH method. Further research will include incorporating dependence into the discrete BH procedure and considering a permutation-based approach.

6

Familial Hypercholesterolemia: Is It Prime Time for Population-Wide Screening in Germany?

Cristian Riccio*[1], Natalie Arnold*[2,3,4], Georgios Koliopanos[1], Vivian Link[1], Linlin Guo[2,3,4], Raphael Betschart[1], Tanja Zeller[5,3], Stefan Blankenberg[2,3,4], Andreas Ziegler**[1,2,3,6], Raphael Twerenbold**[2,3,4]
[1]Cardio-CARE, Medicine Campus Davos, Davos Wolfgang, Switzerland; [2]Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [3]German Center for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Germany; [4]Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [5]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany; [6]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

** Contributed equally*
*** Contributed equally*

Background: The Healthy Heart Act suggests a general two-step screening for the early detection of familial hypercholesterolemia (FH) among children in Germany, where children with elevated LDL-C levels are screened in the first step. In the case of elevated LDL-C levels, a genetic screening is offered to identify monogenic forms of FH. However, data from population-based genetic screening of FH in Germany are still lacking.

Methods: We screened the participants of the population-based Hamburg City Health Study for genetic FH. The study included individuals aged 45 to 74 who underwent short-read whole-genome sequencing at planned 35x coverage. Based on mutation status, we compared LDL-C levels, adjusted for lipid-lowering medication.

Results: Among 7,373 participants, 23 individuals were identified as heterozygous FH, corresponding to a prevalence of 3.1‰. Just 50% of participants with genetic FH met the LDL-C threshold for severe hypercholesterolemia of 190 mg/dl. In turn, severe hypercholesterolemia was observed in 6.5% of participants, but only 2.3% of these cases were attributable to a pathogenic FH mutation. Consequently, 43 individuals would need to be screened to identify one case of genetic FH.

Conclusion: A population-based LDL-C-based screening strategy, as proposed in the Healthy Heart Act, appears limited. Our screening results indicate that it would miss half of FH cases and require 43 genetic tests to identify each case of genetic FH. There is an urgent need for screening results in children, such as those from the VRONI and the Fr1dolin studies, before country-wide screening for FH should be considered.

Keywords: familial hypercholesterolemia, LDLR, LDL, cholesterol

9

Spatial Genomics Profiling of Metabolic Dysfunction-Associated Steatohepatitis Biopsy Tissues for Liver Cancer Risk Prediction

Samuel O. Antwi*[1], Ampem Darko Jnr. Siaw[1], William Sherman[2], E. Aubry Thompson[3], Raouf E. Nakhleh[4], Tushar Patel[3,5]
[1]Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, Florida, United States of America; [2]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America; [3]Department of Cancer Biology, Mayo Clinic, Jacksonville, Florida, United States of America; [4]Department of Laboratory Medicine and Pathology, Mayo Clinic, Jacksonville, Florida, United States of America; [5]Department of Transplantation, Mayo Clinic, Jacksonville, Florida, United States of America
*Presenting author

Metabolic dysfunction-associated steatohepatitis (MASH) is a fast-rising cause of hepatocellular carcinoma (HCC), but the molecular processes underlying the MASH to HCC transition are unclear. The NanoString digital spatial profiler (DSP) offers flexibility of user-defined regions-of-interest (ROIs) for insights into tissue transcriptome within a spatial context that might underlie disease development. We used the DSP to investigate whether spatially resolved RNA-based gene expression signatures in the epithelium or stroma regions of MASH biopsies can predict HCC risk.

A nested case-control analysis was performed using MASH diagnostic biopsies of 20 patients. This comprised five non-cirrhotic MASH patients who later developed HCC (MASH-to-HCC), five patients with MASH-cirrhosis who later developed HCC, five MASH patients who later developed cirrhosis, and five MASH patients who never progressed to cirrhosis or HCC (MASH-only). The biopsies were stained with CK8/18 and α-SMA+ to visualize the epithelium and stroma, with 257 ROIs selected (175 epithelium and 82 stroma). Differentially expressed genes were identified. Logistic regression with 10-fold cross-validations was used, calculating AUC, sensitivity, and specificity.

In the stroma, we identified a 22-gene panel that is unique to the non-cirrhotic MASH-to-HCC transition. Modeling the 22-gene panel by comparing the non-cirrhotic MASH-to-HCC (cases) to all others (controls) yielded AUC=0.91, sensitivity=0.80, specificity=0.90. Similar analysis with a 19-gene panel from the epithelium yielded AUC=0.68, sensitivity=0.66, specificity=0.74.

We found a 22-gene panel in the stromal regions of non-cirrhotic MASH diagnostic biopsies that strongly predicts HCC risk. If validated in other studies, the panel could be targeted for clinical testing to enhance cancer risk prediction in MASH.

Keywords: MASH, spatial genomics, digital spatial profiling, liver cancer, HCC

10

Negative Control Outcomes and Selection Bias in Mendelian Randomization

Apostolos Gkatzionis*[1], Kate Tilling[1]
[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom

Mendelian randomization uses genetic variants as instrumental variables to investigate the causal effect of an exposure on an outcome of interest. The instrumental variable design allows Mendelian randomization studies to guard against confounding bias and reverse causation, but they remain susceptible to selection bias. Negative controls are commonly used as a sensitivity analysis to detect biases in observational studies. Here, we discuss under what conditions a variable can be used as a negative control outcome to detect selection bias in Mendelian randomization. As with other sources of bias, we show that the main requirement is that the negative control outcome shares confounders with the original exposure and outcome. The effect of the negative control on selection is of secondary concern; for example, a variable that does not affect selection can be a valid negative control for an outcome that does. We also argue that age and sex, sometimes used as negative control outcomes in Mendelian randomization analyses, are not valid negative controls in general but can still be useful under certain assumptions. In a real-data analysis, we investigate the pairwise causal relationships between 19 traits, utilizing data from the UK Biobank. Treating biological sex as a negative control outcome, we identify selection bias in analyses involving commonly used traits such as alcohol

consumption, body mass index and educational attainment.

Keywords: Mendelian randomization, selection bias, negative controls

12

Negative Controls to Evaluate the Sensitivity of Mendelian Randomization Estimates to Sample Selection Bias

Winfred N Gatua[1,2], Apostolos Gkatzionis[1,2], Matt Tudball[1,2,3,] Kate Tilling[1,2]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2]Population Health Sciences; Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Prime Biosciences, Canada

Background: Selection bias arises when the study sample is not representative of the population it is sampled from. Studies have shown that selection bias can impact genome wide association analysis. The selectioninterval package has been developed to help quantify the impact of selection bias on causal inference estimates by providing informative bounds.

Aim: We aimed to illustrate the presence of sex-differential participation bias in the UK Biobank (UKB) and to evaluate the use of the selectioninterval method to address this bias. Specifically, we applied negative control in selectioninterval to obtain more informative bounds on the causal effect of a genetic risk score for schizophrenia on body mass index (BMI).

Methods: We performed the GWAS of sex to ascertain sex-differential participation bias. Further we calculated the genetic risk score for schizophrenia in UKB and using known population parameters we used the selectioninterval package to estimate informative bounds for the effect of the genetic risk score for schizophrenia on BMI.

Results: We identified five autosomal variants associated with sex as a result of selection bias. A naive implementation of the selectioninterval package without constraints produced wide bounds for the effect of the schizophrenia risk score on BMI. Adding response rate and negative control constraints produced substantially narrower bounds. However, adding too many constraints adversely impacted the method's performance due to the complexity of the optimisation task it performs.

Conclusion: We have demonstrated the use of the selectioninterval package, which can be a useful tool to conduct sensitivity analyses for selection bias.

Keywords: Selection bias, Mendelian randomisation, Negative control, Informative bounds, Selectioninterval package

13

Phenome-Wide Association Study of Coronary Heart Disease Susceptibility Loci

Geetha Chittoor*[1], Yao Tu[2], Navya Shilpa Josyula[1], Zhe Wang[3], Alexandra Pereira4, Jennifer E. Below[5], Kari E. North[6], Misa Graff[6], Lindsay Fernandez-Rhodes[2], Anne E. Justice[1]

[1]Department of Population Health Sciences, Geisinger, Danville, Pennsylvania, United States of America; [2]Department of Biobehavioral Health, Pennsylvania State University, University Park, Pennsylvania, United States of America; [3]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine, Mount Sinai; New York, New York, United States of America; [4]Division of Aging, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [5]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [6]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Coronary heart disease (CHD) is a leading cause of death globally. Although several CHD-related genetic risk loci have been identified, largely in non-Hispanic White populations, their generalizability across multi-populations is understudied. Hence, to better understand the mechanisms of CHD risk, the spectrum of comorbidities, and potential early indicators of CHD, we conducted a phenome-wide association study (PheWAS) of 17 CHD-risk loci identified in a large genome-wide meta-analysis of ~130,000 Hispanic/Latino (HL) adults from >10 studies. Data included 108,369 (Females: 64,568; Males: 43,801; White/European Americans: 92.3%; HL: 3.5%; Black/African Americans: 2.7%; Asian Americans/Others: 1.5%) unrelated MyCode Community Health Initiative Study participants. PheWAS R package was used and covariates included age, sex (for non-sex specific PheCodes), genotyping array, EHR-reported race/ethnicity, and 20 PCs. We identified 57 significant PheCodes ($P<0.05/1,722=2.9E-05$), mainly belonging to circulatory and endocrine/metabolic groups including 411.4 (Coronary Atherosclerosis), 411 (Ischemic Heart Disease), 411.3 (Angina Pectoris), 411.2 (Myocardial Infarction), 272*(Lipid-related disorders), and 250.2 (Type 2 Diabetes) associated with variants near loci – *LPA, CDKN2B, PHACTR1*, and *LMOD1*. Interestingly, some variants have shown significant associations with PheCodes of pregnancy, dermatology, and genitourinary groups rather than circulatory or endocrine/metabolic, e.g., *RUNX3* with 654.2 (Rhesus Isoimmunization) and 696* (Psoriatic-related), and *NFKBIL1* with 599.3 (Dysuria). Furthermore, we found suggestive pleiotropic associations of *CDKN2B, LPA, LMOD1*, and *KANK3* variants with PheCodes not only from circulatory and endocrine/metabolic, but also from sense organs, respiratory, genitourinary, and digestive system groups. Our study demonstrated the utility of EHR/biobanks in generalizing HL/CHD-risk loci across populations and identifying novel shared genetic etiologies.

Keywords: CHD, PheWAS, Hispanic/Latino, pleiotropy, EHR-linked biobanks

14

A Mixture Model Approach for Correcting Systematic Measurement Errors in Sitting Height Data from UK Biobank

Jieyu Ge*[1], Jian Zeng[1], Loic Yengo[1], Peter Visscher[1,2]

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; [2]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

Background: Sitting height measurements in UK Biobank exhibit systematic discrepancies between imaging

visits, likely due to human error in box height data entry. Among 4,903 individuals with repeat DXA scans, the raw correlation between sitting height measurements was 0.863 (sex-adjusted: 0.779).

Methods: We developed a Stan-based mixture model to identify measurement errors, modeling observations as originating from nine potential error components that incorporate both shared and independent errors from manual data entry. We also employed Gaussian Mixture Model (GMM) clustering as an alternative approach, correcting identified individuals based on mixture probabilities and means. Both correction methods were compared against simple outlier removal by examining changes in SNP-based heritability and polygenic score prediction accuracy.

Results: Our mixture model identified 50 females and 128 males with probable measurement errors, similar to GMM clustering results (52 females, 124 males). After correction, the SNP-based heritability of sitting height was 0.522 ± 0.08, compared to 0.526 ± 0.08 for uncorrected data. Correlation with height polygenic scores was 0.4266 after correction versus 0.4296 before correction. Simple outlier removal showed slight improvements in prediction accuracy (0.4422).

Conclusion: Neither correction methods nor outlier removal substantially improved SNP-heritability or prediction accuracy. These approaches may prove more beneficial in larger datasets with repeated measurements than in our subset of fewer than 5,000 individuals.

Keywords: UK Biobank, measurement error, mixture model, Gaussian Mixture Model, sitting height, error correction

15

Estimating Gene Conversion Rates From Population Data Using Multi-individual Identity by Descent

Brian L. Browning[1,2], Sharon R. Browning[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, United States of America

In humans, homologous gene conversions occur at a higher rate than crossovers, however gene conversion tracts are small and often unobservable. As a result, estimating gene conversion rates is more difficult than estimating crossover rates. We present a method for multi-individual identity-by-descent (IBD) inference that allows for mismatches due to genotype error and gene conversion. We use the inferred IBD to detect alleles that have changed due to gene conversion in the recent past. We analyze data from the TOPMed and UK Biobank studies to estimate autosome-wide maps of gene conversion rates. For 10 kb, 100kb, and 1 Mb windows, the correlation between our TOPMed gene conversion map and the deCODE sex-averaged crossover map ranges from 0.56 to 0.67. We find that the strongest gene conversion hotspots typically die back to the baseline gene conversion rate within 1 kb. In 100 kb and 1 Mb windows, our estimated gene conversion map has higher correlation than the deCODE sex-averaged crossover map with PRDM9 binding enrichment (0.34 vs 0.29 for 100 kb windows and 0.52 vs 0.34 for 1 Mb windows), suggesting that the effect of PRDM9 is greater on gene conversion than on crossover recombination. Our TOPMed gene conversion maps are constructed from 55-fold more observed allele conversions than the recently published deCODE gene conversion maps. Our map provides sex-averaged estimates for 10 kb, 100 kb, and 1 Mb windows, whereas the deCODE gene conversion maps provide sex-specific estimates for 3 Mb windows.

16

Genetics of Time to Reach EDSS 6 in Multiple Sclerosis

Soumeen Jin[1], Klementy Shchetynsky[1], Melissa Sorosina[5], Adil Harroud[2,3,4], Ali Manouchehrinia[1], Tomas Olsson[1], Fredrik Piehl[1], Federica Esposito[5,6], Pernilla Stridh[1], Ingrid Kockum[1]

[1]Department of Clinical Neuroscience, Karolinska Institute, Sweden; [2]The Neuro (Montreal Neurological Institute-Hospital), Montréal, Quebec, Canada; [3]Department of Neurology and Neurosurgery, McGill University, Montréal, Quebec, Canada; [4]Department of Human Genetics, McGill University, Montréal, Quebec, Canada; [5]Human Genetics of Neurological Disorders, Division of Neuroscience, San Raffaele Scientific Institute, Milan, Italy; [6]Neurology Unit and MS Center, IRCCS San Raffaele Hospital, Milan, Italy

Multiple sclerosis is a complex autoimmune condition characterized by demyelination and axonal damage, disrupting neural signaling. It has a very low population prevalence (about 0.4%) in Caucasian population, and lower in other ethnicities [1]. While genetic factors underlying MS susceptibility have been extensively studied [2], determinants of disease severity remain largely unknown[3]. Recent evidence suggests that genetic variants of MS severity and progression can have different effects in early and late stages of disease[4], indicating the need for diverse phenotypes and time-points to capture the genetic architecture.

MS disability is assessed by the Expanded Disability Status Scale (EDSS), ranging from 0 to 10. This study conducted the first time-to-event genome-wide association study of two phenotypes: disease duration to EDSS6 and age at EDSS6. EDSS6 is a milestone reflecting the need for walking aids. After stringent quality control and imputation, we analyzed ~10 million autosomal variants in 7,109 Swedish and 1,353 Italian patients, adjusting for demographic and population structure covariates. The proportional hazards assumption was satisfied by adding interaction term of the covariates.

Our findings identified one replicated suggestive locus ($p<1x10-5$) on chromosome 11, associated with disease duration at observed EDSS ≥6, and two replicated suggestive loci on chromosomes 2 and 5, associated with age at observed EDSS ≥6. The mapped genes show significant brain tissue expression and associations with central nervous system-related traits, supporting the role for CNS resilience as a determinant of MS severity [3]. The identified genes may serve as potential novel targets to slow progression.

References

[1] M. Hittle et al., "Population-Based Estimates for the Prevalence of Multiple Sclerosis in the United States by Race, Ethnicity, Age, Sex, and Geographic Region," JAMA Neurology, vol. 80, no. 7, p. 693, May 2023,

doi: 10.1001/jamaneurol.2023.1135.
[2] N. A. Patsopoulos et al., "Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility," Science, vol. 365, no. 6460, Sep. 2019, doi: 10.1126/science.aav7188.
[3] A. Harroud et al., "Locus for severity implicates CNS resilience in progression of multiple sclerosis," Nature, vol. 619, no. 7969, p. 323, Jun. 2023, doi: 10.1038/s41586-023-06250-x.
[4] N. Sahi et al., "Evaluating multiple sclerosis severity loci 30 years after a clinically isolated syndrome," Brain Communications, vol. 6, no. 6, Jan. 2024, doi: 10.1093/braincomms/fcae443.

17

Formal Statistical Replication Analysis in Lung Cancer Genome-Wide Association Studies

Yung-Han Chang*[1], Jinyoung Byun[2,3,4], Bryan Gorman[5], Rayjean J. Hung[6,7], James McKay[8], Christopher I. Amos[2,3,9], Saiju Pyarajan[5,10], Arjun Bhattacharya[11], Ryan Sun[1]
[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Graduate School of Biomedical Sciences, Houston, Texas, United States of America; [2]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, Texas, United States of America; [3]Department of Medicine, Section of Epidemiology and Population Sciences, Baylor College of Medicine, Houston, Texas, United States of America; [4]University of New Mexico Comprehensive Cancer Center, Albuquerque, New Mexico, United States of America; [5]Center for Data and Computational Sciences (C-DACS), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, Massachusetts, United States of America; [6]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [7]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada; [8]Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France; [9]Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America; [10]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America; [11]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

Lung cancer is the most common cause of cancer death worldwide. Although genome-wide association studies (GWAS) have identified numerous genetic variants associated with lung cancer risk, it remains unclear how to best translate these findings to clinical applications such as new therapies or risk prevention strategies. A major challenge in GWAS is the lack of rigorous statistical methods for testing the replicated SNPs across different cohorts, as many ad-hoc approaches rely on *P* value thresholds or meta-analysis as a substitute. The absence of formal replication analysis, which requires testing a composite null hypothesis, increases the risk of false positives and undermines the reliability of SNP-phenotype associations.

Here, we robustly test the replication composite null hypothesis in lung cancer GWAS using an empirical Bayes approach. Our work relies on a conditionally symmetric multidimensional Gaussian mixture model, which improves interpretability of results while controlling false discoveries and maintaining statistical power in simulation.

We apply our method to three Caucasian lung cancer GWAS cohorts, identifying 428 replicated variants at a false discovery rate of 0.1. Meta-analysis detects 1,271 variants, but 57.0% show a large *P* value ($P > 0.01$) in one or more cohorts, compared to only 6.5% in replication analysis. Functional annotation of replication-identified variants reveals large numbers in evolutionarily conserved and epigenetically active areas. We further analyze the squamous cell carcinoma and adenocarcinoma subtypes. These results emphasize the importance of formal statistical testing for replication effects.

18

Influence of Heritable Covariates on Genetic Studies of The Human Gut Microbiome and Consequences for Mendelian Randomization Analyses

Alec J. McKinlay*[1,2], Nicholas J. Timpson[1,2], Eleanor C.M. Sanderson[1,2], Timothy M. Robinson[1,2], David A. Hughes[1,2,3], Jeroen Raes[4,5], Kaitlin H. Wade[1,2]
[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom Medical Research Council; [2]Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, United Kingdom; [3]Pennington Biomedical Research Center Baton Rouge Louisiana, United States of America; [4]Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium; [5]Center for Microbiology, Vlaams Instituut voor Biotechnologie, Leuven, Belgium

Whilst previous research efforts have struggled to resolve the interplay between the host genome and human gut microbiome, there is now evidence of heritable contribution to gut flora variation. Despite the characterisation of this contribution being incomplete and the likely complexity of current GWAS signals, there has been an exponential increase in Mendelian randomisation (MR) studies attempting to use host genetic signals to aid causal inference around associations between gut microbiome traits (GMTs) and host health. We hypothesise that human SNPs associated with the gut microbiome are likely derived from complex signals or chance and – where signals are reliable – may likely encapsulate aspects of the human phenome and exposome (e.g., as for the only strongly replicated host/microbiome signal to date – LCT/MCM6-*Bifidobacterium*). To investigate this, we used characterised genetic contributions to microbiome-related covariates, individual-level genetic data from the Flemish Gut Flora Project (FGFP; n~2500) and published summary-level data from the MiBioGen consortium (n~18,000). We identified human phenotypes (including body mass index, triglycerides and uric acid levels) associated with specific GMTs and showed that GWAS signals for associated GMTs were enriched for SNPs associated with those phenotypes. These analyses suggest that microbiome-related SNPs used in MR reflect phenotypic variation that is representative of other complex human phenotypes, which may lead to biased MR estimates from heritable confounding and/or misspecification of the primary phenotype. Lastly, we present a framework to evaluate the likelihood for this bias in

two-sample MR framework assessing the causal role of the gut microbiome on cancer as an exemplar.

19

The Role of the *ADRB2* Thr164Ile Variant in Lung Function Determination, Plasma Proteome Variability and Other Phenotypes in UK Biobank

Katherine A Fawcett*[1], Robert J Hall[2], Richard Packer[1,3], Kayesha Coley[1], Nick Shrine[1], Louise V Wain[1,3], Martin D Tobin[1,3], Ian P Hall[2]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Division of Respiratory Medicine and NIHR Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom; [3]Leicester National Institute for Health and Care Research, Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom

Introduction: Beta-2 adrenergic receptor gene (*ADRB2*) polymorphisms have been associated with multiple conditions including asthma, COPD and response to asthma treatment (long-acting beta agonists, LABA). However, large, systematic studies of the low-frequency Thr164Ile variant (rs1800888) and rare coding variation within the gene are lacking.

Methods: To identify pleiotropic effects of Thr164Ile and other coding variants, we performed respiratory-focused and phenome-wide association studies in UK Biobank, including gene-based tests of rare variants. In addition, we used Olink proteomic data to characterise enriched pathways and upstream regulators of proteins associated with *ADRB2* polymorphisms.

Results: The minor allele of Thr164Ile was associated with reduced lung function (FEV1/FVC, PEF and FEV1), increased eosinophil counts and blood lipid measurements, including increased cholesterol, reduced triglycerides and reduced apolipoprotein A, but not COPD or asthma. There was no association with asthma exacerbation risk in those self-reporting taking LABA. Proteins associated with Thr164Ile (P-value≤0.01) were enriched for various pathways, with the eosinophil-raising allele associated with reduced neutrophil degranulation, immunoregulatory interactions between lymphoid and non-lymphoid cells, TNF binding and DAP12 interactions, as well as activation of lipid metabolism pathways, including FXR/RXR activation and LXR/RXR activation. A gene-based analysis of other non-synonymous *ADRB2* variants identified a novel association with non-rheumatic pulmonary valve disorders, but no association with lung function.

Conclusion: Thr164Ile is associated with traits and proteins indicative of a role in immune and lipid metabolism pathways, identifying potential targets for therapeutic intervention.

Funding: Wellcome Trust Award WT225221/Z/22/Z, NIHR Leicester BRC and NIHR Senior Investigator Awards to M.D.T. and I.P.H.

Keywords: beta-2 adrenergic receptor, phenome-wide association studies, lung function, protein quantitative trait loci, pathway analysis

20

Large-Scale Association Analysis Identified New Susceptibility Risk Loci for Differentiated Thyroid Carcinoma by Integrating the Transcriptome and Proteome

See Hyun Park[1], Yazdan Asgari[1], Pierre-Emmanuel Sugier[1], Mojgan Karimi[1], EPITHYR consortium, EPIC consortium, Stefano Landi[2], Hauke Thomsen[3], Asta Försti[4,5,6], Thérèse Truong[1]

[1]Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France; [2]Department of Management, Università Ca' Foscari, Venezia, Italy; [3]MSB Medical School Berlin, D-14197 Berlin, Germany; [4]Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [5]Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany; [6]Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany

Background: Differentiated thyroid cancer (DTC) is a prevalent malignancy with increasing global incidence, yet its genetic susceptibility remains unclear. Although previous genome-wide association studies (GWAS) have identified several susceptibility loci, the genetic, transcriptomic, and proteomic factors influencing DTC risk are not fully understood.

Methods: We conducted a large-scale GWAS analysis of DTC from 7,681 cases and 963,550 controls of European ancestry. Transcriptome-wide association studies (TWAS) used the joint tissue imputation (JTI) model across multiple tissues (thyroid, pituitary, blood, and hypothalamus). Proteome-wide association studies (PWAS) integrated brain and plasma proteomic data to identify proteins influencing DTC risk. Mendelian randomization (MR) and Bayesian colocalization were conducted to infer causality.

Results: GWAS identified 18 genome-wide significant loci, including four previously suggested and now confirmed. TWAS identified 29 significant genes, including five genes (*LRRC34, NRG1, HEMGN, PTCSC3*, and *SMAD3*) located within known DTC risk loci, now confirmed as causal. Additional three novel genes (*SAMD4A, RAD51-AS1*, and *MPHOSPH6*) were validated as causal through MR and Bayesian colocalization. PWAS identified seven significant proteins, including the known *NANS* gene. Among the remaining six proteins, three (*MTHFR, KDELC2,* and *SAMD4A*) were validated as causal, highlighting 15q15.1 as a novel risk locus consistently emerging across all omics layers.

Conclusion: Our integrated multi-omics approach reveals novel genetic risk factors for DTC, emphasizing the relationship between genetic variation, gene expression, and protein abundance. These findings enhance new insights into the molecular etiology of thyroid cancer.

21

covImpute: Leveraging Genetic Correlations to Impute Missing EHR Phenotypes

Cue Hyunkyu Lee[1,†], Hanqing Wu[2,†], Najmeh Abiri[3], Iuliana Ionita-Laza*[1,2]

[1]Department of Biostatistics, Columbia University, New York, New York, United States of America; [2]Department of

Statistics, Lund University, Lund, Sweden; [3]School of Information Technology, Halmstad University, Halmstad, Sweden
[†]These authors contributed equally.
**Corresponding author*

Electronic Health Records contain abundant phenotypic information yet often miss or inconsistently record many binary and quantitative traits. These gaps weaken genome wide association studies, skew fine mapping, bias estimates of genetic correlation and heritability, and distort polygenic risk scores—pillars of modern genomic practice. Imputation methods that depend on observed phenotype correlations and treat every trait as continuous perform poorly when data are missing not at random, a frequent clinic driven pattern in which tests focus on suspected cases. Social and access inequalities intensify this bias, and variation within diagnostic groups further limits recovery.

We present **covImpute**, a hierarchical Bayesian method that brings genetic architecture into a liability threshold framework. Each trait receives a latent normal score; binary traits are fitted with a probit likelihood, quantitative traits with squared error, and the model penalizes any gap between the latent correlation matrix and a genetic reference. The resulting convex objective respects trait scales, separates genetic from environmental effects, and converges quickly; cross validation sets all tuning parameters.

Simulations and analyses of eight UK Biobank diseases show that covImpute lifts genome wide association power by as much as five-fold under missing not at random designs and discovers extra risk loci for diabetes, stroke, and hypertension, outperforming established methods such as LTPI, AutoComplete, and SoftImpute. By modeling shared liabilities instead of noisy observed correlations, covImpute boosts statistical power and more accurately recovers latent disease liability, offering a scalable imputation strategy for Electronic Health Record linked biobank studies.

22

Multi-trait GWAS Across 52 Infectious Diseases: Mapping Common Variants Associated with the Immune Response

Hanna Julienne[1], Gaspard Kerner[2], Jhonatan Ramos[1], Andreea Patarlageanu[1], Lluis Quintana-Murci[3], Etienne Patin[3]
[1]Institut Pasteur, Université Paris Cité, Department of Computational Biology, Paris, France; [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [3]Human Evolutionary Genetics Unit, Institut Pasteur, Université Paris Cité, Paris, France

Response and susceptibility to infectious diseases vary from person to person, owing to a combination of genetic and non-genetic factors. Recent Genome Wide Association Studies (GWAS) on COVID-19 susceptibility have highlighted the contribution of common genetic variants to this variability. However, GWAS data for other infectious diseases often consist of studies with limited sample size.

In this context, multi-trait GWAS may be helpful to capture genetic evidence scattered across underpowered studies. Here, we leveraged the Joint Analysis of Summary Statistics (JASS) pipeline to analyze 91 GWAS summary statistics obtained from the GWAS catalog, and *23 and me.* We i) performed a meta-analysis to regroup summary statistics by trait, resulting in 52 infectious traits (e.g. COVID-19, pneumonia, chronic sinus infection, ...), ii) conducted a multi-trait GWAS including all traits, and ii) detected 69 novel associations. The nearest genes of significant variants mapped to immune pathways such as: antigen processing and presentation (p-val=4.42e-17); positive regulation of T-cell activation (p-val=1.63e-14); and positive regulation of immune response (p-val=5.27e-13).

This compendium of signals across infectious traits provides an opportunity to understand how genetic variants shape the immune response across a range of immune stimuli. By calculating the genome-wide genetic correlations across the 52 traits, we observed overall diffuse positive correlations (median correlation = 0.17, and 92% of positive correlations). To study pleiotropic effects across infectious diseases, we will search for patterns of multi-trait genetic effects using a range of dimensionality reduction techniques such as FactorGo, and Genetic Factor Analysis.

23

Mapping the Epigenetic Mechanisms Underlying Musculoskeletal Disease Across the Lifecourse

Sarah E. Orr*[1], Euan McDonnell[2], Abby Brumwell[1], David J. Deehan[3], Antony K. Sorial[1], Jamie Soul2, David A. Young[1], Sarah J. Rice[1]
[1]Newcastle University Biosciences Institute, Newcastle upon Tyne, United Kingdom; [2]Liverpool University Computational Biology Facility, Liverpool, United Kingdom; [3]Newcastle University Teaching Hospitals NHS Trust, Freeman Hospital, High Heaton, NE1 7DN

Over 500 SNPs have been associated with genetic risk of musculoskeletal (MSK) phenotypes, including osteoarthritis and fracture. The challenge now lies in the biological interpretation of identified variants and identification of the regulatory mechanisms of disease effector genes (eGenes). Here, we address this via statistical finemapping to identify methylation quantitative trait loci (mQTLs) which colocalize with risk SNPs.

We hypothesise that MSK risk SNPs can exert temporal effects (acting either during development or ageing) and spatial effects (operating in different tissues).

We have quantified the methylome of human developmental, osteoarthritis and fracture hip articular cartilage and trabecular bone (n=312). We identified over 100,000 mQTLs in each tissue (Bonferroni-adjusted $P<0.05$). Across the investigated skeletal tissues, between 33% (fracture bone and cartilage) and 59% (foetal bone and cartilage) mQTLs were shared. We further overlapped our data with previously published whole blood mQTLs, identifying 55% (fracture cartilage) to 84% (fracture bone) of shared effects. Conversely, only 8% (fracture bone) to 25% (osteoarthritis cartilage) were shared with human brain mQTLs, highlighting the tissue specificity of the methylome.

Analysis of the mQTLs alongside OA and OP GWAS signals identified 638 colocalizations (*PP.H4*>0.8; 628 CpGs, 246 SNPs). Of the colocalizing signals, 17% were unique to development and 65% unique to adult tissues. This highlights that epigenetic plasticity can confer risk of

complex disease in a spatiotemporal manner. Ongoing work involves the generation and integration of additional datasets (ATAC-seq and Capture Hi-C) to further finemap these signals in relevant tissues and reveal molecular mechanisms underpinning MSK disease.

Keywords: colocalisation, mQTL, musculoskeletal, GWAS

25

Cross-Ancestry Genome-Wide Association Study of Creatine Kinase Reveals Novel Genetic Loci and Insights into Muscle Injury

Gang Chen*[1], Sizheng S. Zhao[2], Hector Chinoy[2,3,4], Andrew P. Morris[2,3], Janine A. Lamb[1]

[1]Epidemiology and Public Health Group, School of Health Sciences, University of Manchester, Manchester, United Kingdom; [2]Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom; [3]NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, United Kingdom; [4]Department of Rheumatology, Salford Royal Hospital, Northern Care Alliance NHS Foundation Trust, Manchester Academic Health Science Centre, Salford, United Kingdom

* Presenting author

Background: Creatine kinase (CK) is an enzyme predominantly expressed in cardiac and skeletal muscle, and its serum concentration is a well-established biomarker of muscle damage. To identify novel genetic loci and biological processes through which they lead to muscle injury, we conducted a cross-ancestry genome-wide association study (GWAS) meta-analysis of CK levels.

Method: We performed a fixed effects meta-analysis using publicly available CK GWAS summary statistics in 237,254 individuals from diverse populations. Comprehensive functional analyses were applied, and potential effector genes were evaluated using expression quantitative trait loci (eQTL) data. Genetic correlations with related traits were estimated using Linkage Disequilibrium Score Regression, followed by colocalization analyses using HyPrColoc. We also assessed enrichment of Mendelian muscle disease genes at CK associated loci.

Results: Cross-ancestry meta-analysis revealed 109 genome-wide significant loci for CK levels ($P < 5\times10^{-8}$), including 38 loci not previously reported. Functional analyses highlighted several genes involved in muscle structure and function, including *BAG3, MYPN, TTN, CAV3, CACNA1S,* and *FHOD3*. Notably, loci near BAG3, MYPN, and FHOD3 exhibited overlapping muscle specific eQTL signals. Significant genetic correlations between CK levels and traits, including hand grip strength and aspartate aminotransferase levels, were observed and supported by colocalization evidence. Furthermore, we observed significant enrichment of Mendelian muscle disorder genes among CK associated loci, suggesting a shared genetic architecture.

Conclusion: Our findings advance understanding of the genetic architecture underlying CK levels and provide new insights into the genetic basis of muscle injury in the general population.

26

Parent-of-Origin Inference and Its Role in the Genetic Architecture of Complex Traits: Evidence from ~265,000 Individuals

Robin J. Hofmeister[1,2,3,4], Stefan Johansson[5,6], Lili Milani[4,7], Olivier Delaneau[8], and Zoltan Kutalik[1,2,3]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; [2]University Center for Primary Care and Public Health, Lausanne, Switzerland,; [3]Swiss Institute of Bioinformatic (SIB), University of Lausanne, Lausanne, Switzerland; [4]Estonian Genome Centre, Institute of Genomics, University of Tartu, Estonia; [5]Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen, Norway, [6]Department of Pediatrics, Haukeland University Hospital, Bergen, Norway, [7]Estonian Biobank, Institute of Genomics, University of Tartu, Estonia, [8]Regeneron Genetics Center, Tarrytown, New York, United States of America

Parent-of-origin effects (POEs) occur when the impact of a genetic variant depends on its parental origin. Traditionally linked to genomic imprinting, these effects are believed to have evolved from parental conflict over resource allocation to offspring, which results in opposing parental genetic influences. Despite their potential importance, POEs remain heavily understudied in complex traits, largely due to the lack of parental genomes.

Here, we present a multi-step approach to infer the parent-of-origin of alleles without parental genomes, leveraging inter-chromosomal phasing, mitochondrial and chromosome X data, and sibling-based crossover inference. Applied to the UK (discovery) and Estonian (replication) Biobank, we inferred the parent-of-origin for up to 221,062 individuals, representing the largest dataset of its kind.

GWAS scans in the UK Biobank for more than 60 complex traits and over 2,400 protein levels contrasting maternal and paternal effects identified over 30 novel POEs and confirmed more than 50% of testable known associations. Notably, approximately half of our POEs exhibited a bi-polar pattern, where maternal and paternal alleles exert conflicting effects. These effects were particularly prevalent for traits related to growth (e.g., IGF-1, height, fat-free mass) and metabolism (e.g., type 2 diabetes, triglycerides, glucose). Replication in the Estonian Biobank and in 45,402 offspring from the Norwegian Mother, Father and Child Cohort Study validated over 75% of testable associations.

Overall, our findings shed new light on the influence of POEs on diverse complex traits and align with the parental conflict hypothesis, providing compelling evidence for this understudied evolutionary phenomenon.

27

Why Meta-Analysis Fine-Mapping Can Be Confidently Wrong and How to Fix It

Wei-Yu Lin[1], Jeffrey Pullin[1], Yurii Aulchenko[2], Toby Johnson[2], Chris Wallace[1,2,3]

[1]MRC Biostatistics Unit, University of Cambridge, United Kingdom; [2]GSK, Stevenage, United Kingdom; [3]Cambridge Institute for Therapeutic Immunology and Infectious Disease, University of Cambridge, United Kingdom

Fine mapping is a major component of post-GWAS analyses. However, recent work has demonstrated that fine mapping procedures can produce high confidence credible sets which do not contain any causal variant in larger GWAS, particularly meta-analysis datasets. We demonstrate that variable sample size across SNPs, a common characteristic of such datasets, affects both the expected distribution of Z scores which are the common inputs to fine mapping procedures, and the expected between SNP correlation matrix. This creates additional noise in the input data that standard fine mapping approaches do not allow for.

We propose a new method NIFTY that overcomes these challenges, first by estimating the between SNP correlations as a function of LD matrix and sample size. Second, we use a novel factorisation of the fine mapping likelihood that only uses data at SNPs with near-complete sample coverage for fine mapping, yet allows probabilities of causal association at any SNP in the LD reference panel to be accurately predicted. This factorisation also removes the need for imputation. We benchmark our approach against FINEMAP, SuSiE and finimom in simulated and real datasets. In these comparisons, FINEMAP and SuSiE tended to fit additional, fictitious signals to compensate for noise in the input data. Finimom avoided fictitious signals, but at the cost of detecting true causal variants less often. In contrast, our approach avoids fictitious signals whilst maintaining similar or better sensitivity to FINEMAP and SuSiE for detecting the true causal variants. By improving fine mapping accuracy, our approach can propagate advantages to other post GWAS analyses that depend on fine mapping posterior probabilities, including colocalisation, enrichment, and gene prioritisation, fully realising the potential of modern large, and meta-analysed, GWAS.

28

Bayesian Inference Model to Prioritise Rare Variants From Family-Based Whole Genome Sequencing Data

Cathal M. Ormond*[1], Mathieu Cap[1], Niamh M. Ryan[1], Carol A. Mathews[2,3], Aiden P. Corvin[1], Elizabeth A. Heron[1]

[1]Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity Centre for Health Sciences, Trinity College Dublin, St James's Hospital, Dublin, Ireland; [2]Department of Psychiatry, Mc Knight Brain Institute, Center for OCD, Anxiety, and Related Disorders, University of Florida, Gainesville, Florida, United States of America; [3]University of Florida Genetics Institute, University of Florida, Gainesville, Florida, United States of America

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 are 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. This tool quantifies the co-segregation pattern of single nucleotide variants (SNVs) with a binary trait and aggregates this with independent variant information such as allele frequency and deleteriousness. The quantitative output metrics of BICEP allow users to rank variants based on their overall evidence for causality (as well as examining evidence for co-segregation and pathogenicity), which is not possible using existing approaches. Here, we have extended the underlying model to also evaluate indels and copy number variants (CNVs).

We applied BICEP to whole genome sequencing data for a five-generational pedigree containing 15 individuals diagnosed with obsessive compulsive disorder (OCD). SNVs and indels were called using standard best practises, and CNVs were called using a family-aware consensus approach (PECAN). BICEP did not identify any rare SNVs, indels or CNVs that perfectly co-segregated with OCD, consistent with the known complex genetic architecture of psychiatric conditions. However, it did identify several SNVs carried by the majority of cases, which could be explained by a single-hit model with phenocopies or by a multi-hit model. This work showcases a novel approach to analysing pedigree data, furthering our understanding of OCD.

31

Polygenic Risk Score Convex Combinations Can Inform Two-Phase Re-Sequencing Study Design and Analysis

Chenyang Li[1], Osvaldo Espin-Garcia[1-3]

[1]Department of Epidemiology and Biostatistics, University of Western Ontario, London, Canada; [2]Department of Biostatistics, University Health Network, Toronto, Canada

[3]Dalla Lana School of Public Health and Department of Statistical Sciences, University of Toronto, Toronto, Canada

The genetic architecture of complex traits makes polygenic risk score (PRS) construction challenging given that a single PRS method might not comprehensively summarize the trait genomic susceptibility. For instance, LDpred-inf and LASSOsum, posit contrasting assumptions with respect to the trait underlying genetic architecture, i.e., infinitesimal vs. sparse. Recent work in two-phase re-sequencing study design, where only informative subsamples are selected for cost-effective targeted sequencing data collection, reduces expenses while preserves the identification of key genetic-trait links. We propose an approach that integrates multiple PRS methods for two-phase re-sequencing study design. The proposal solves a convex combination problem aiming to identify the PRS mix that minimizes the mean squared error. Under a linear regression model with non-edge solutions, the resulting combination matches a residual dependent sampling (RDS) with all PRS as covariates. The main advantage of the convex optimization approach is that the resulting PRS mix can serve as sole auxiliary covariate in maximum likelihood (ML) methods. The proposed optimization approach is evaluated against alternative RDS designs with single or all PRS as covariates under ML and sieves ML (SML) methods for linear and logistic models via simulations and real data. Preliminary results under the null hypothesis show differential bias between PRS strategies, linear/logistic models and ML/SML methods. Briefly, under a linear model, bias become apparent only when a single PRS is used. In contrast, logistic models appear unbiased only when the PRS mix is used in tandem with SML.

Keywords: Two-phase studies, Re-sequencing, PRS

32

Multiallelic Genetic Architecture Underlying AD and ADRD Proteinopathy

Yuriko Katsumata*[1,4], Inori Tsuchiya[1,4], Khine Zin Aung[1,4], Xian Wu[1,4], Lincoln M. Shade[1,5], Erin L. Abner2,4, Peter T. Nelson[3,4], David W. Fardo[1,4]

[1]Department of Biostatistics, University of Kentucky, Lexington, Kentucky, United States of America; [2]Department of Epidemiology and Environmental Health, University of Kentucky, Lexington, Kentucky, United States of America; [3]Department of Pathology, University of Kentucky, Lexington, Kentucky, United States of America; [4]Sanders-Brown Center on Aging, University of Kentucky, Lexington, Kentucky, United States of America; [5]College of Medicine, University of Kentucky, Lexington, Kentucky, United States of America

We recently reported genetic associations with burden of dementia-related proteinopathies measured by constructing three continuous latent endophenotype scores—reflecting TDP-43; Aβ and Tau; and α-synuclein pathology—using multidimensional generalized partial credit modeling. These scores were derived from harmonized neuropathological data from the National Alzheimer's Coordinating Center (NACC), Alzheimer's Disease Neuroimaging Initiative (ADNI), and the Religious Orders Study/Memory and Aging Project (ROSMAP). However, prior analyses excluded multiallelic variants, which are not compatible with standard biallelic genetic coding schemes. In this study, we conducted a genome-wide analysis of multiallelic variants using whole genome sequencing data (WGS) from the Alzheimer's Disease Sequencing Project (ADSP) and our endophenotype scores. After quality filtering, 672,004 multiallelic variants on autosomes were analyzed across 447 NACC/ADNI and 519 ROSMAP participants. Score tests under a generalized linear model framework were applied, adjusting for sex, age at death, top three genetic ancestry principal components, and the other two latent scores. Meta-analyses combined results across cohorts. Notable associations included: Chromosome 13 with the TDP-43 and the α-synuclein scores, and Chromosome 19 with the Aβ and Tau score. A multiallelic variant downstream of *APOE* was identified in significant association with AD-related proteinopathy. These findings underscore the potential of multiallelic variant analysis to uncover novel genetic contributors to neurodegenerative disease. Replication in independent datasets and functional validation are ongoing.

33

Expanding Insights into the Genetic Architecture of Coronary Artery Disease: A Multi-Ancestry and Multi-Trait Genome-wide Meta-Analysis among ~2 Million Individuals

Federico Murgia[1], Pik Fang Kho[2,3], Anu Nyberg[4], Anuj Goel[5], Xiang Zhu[6], Satoshi Koyama[7], Panagiotis Deloukas[8], Minna Kaikkonen-Määttä[4], Themistocles L Assimes[2,3], Jemma C Hopewell[1] on behalf of the CARDIoGRAMplusC4D Consortium

[1]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom; [2]Department of Medicine, Stanford University School of Medicine, Stanford, California, United States of America; [3] VA Palo Alto Health Care System, Palo Alto, California, United States of America; [4] A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland; [5] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom; [6]Department of Statistics and Huck Institutes of the Life Sciences, Penn State University, Pennsylvania, United States of America; [7]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America; [8]Barts and The London School of Medicine and Dentistry, William Harvey Research Institute, Queen Mary University of London, London, United Kingdom

Background: Genome-wide association studies (GWAS) have identified over 393 distinct loci associated with coronary artery disease (CAD). However, large-scale multi-ancestry efforts and data integration across cardiovascular traits remain limited.

Material and Methods: We conducted a multi-ancestry genome-wide meta-analysis among 329,572 CAD cases and 1,629,888 controls of European, African American, Hispanic, and East Asian ancestry. We performed multi-trait analyses across cardiovascular diseases, including peripheral artery disease and stroke, followed by a comprehensive downstream functional and bioinformatic evaluation.

Results: We identified 44 new loci, increasing the number of CAD-associated loci by 11%. Over half of these colocalized with eQTLs in CAD-relevant tissues, suggesting that some genetic variants may influence CAD risk by regulating gene expression in these tissues. Functional analysis revealed an enrichment of CAD-associated variants in endothelial and immune cells, consistent with their roles in vascular homeostasis and inflammation. Significant enrichment was also observed in pericytes, precursors to vascular smooth muscle cells. STARR-Seq was used to map regulatory variants in selected novel loci across various disease-relevant cell types and conditions, and prioritized causal variants at six loci, including *COBLL1* and *SEMA5A*. Integration of functional fine-mapping, eQTL colocalization, and gene-based approaches prioritized novel genes such as *ADAMTS9* and ENG. Multi-omic data analyses and polygenic scores also provided further insights.

Conclusion: Multi-ancestry and multi-trait CAD GWAS considerably expands our understanding of the spectrum of allelic variation underlying CAD and its translational potential. Our study also provides a valuable community resource for further studies of the causes and consequences of CAD.

Keywords: Coronary Artery Disease, Multi-Ancestry Meta-Analysis, Functional Analysis

34

Genetic Comparative Analyses of Atopic Dermatitis and Psoriasis

Katie Watts*[1], Nick Dand[2], Matthew Boyton[1], Andrew Elmore[1], Catherine Smith[3], Sara J. Brown[4], Lavinia Paternoster[1]

[1]Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, United Kingdom; [2]Department of Medical and Molecular Genetics, School of Basic & Medical Biosciences, Faculty of Life

Sciences & Medicine, King's College London, London, United Kingdom; [3]St John's Institute of Dermatology, School of Basic & Medical Biosciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom; [4]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom

Atopic dermatitis (AD) and psoriasis are two common chronic inflammatory skin diseases that rarely coexist in the clinical context. Previous comparative analyses have also found little evidence of shared loci with consistent allelic effects (operating in the same direction on both diseases). However, treatment with biologics can switch the phenotype between the two diseases for reasons that are poorly understood. By utilizing recently released large genome wide association (GWAS) meta-analyses of AD and psoriasis susceptibility, we have investigated this through genetic comparative analyses. We performed a compare and contrast meta-analysis (CCMA) and Local Analysis of [co]Variant Association (LAVA) analysis to identify variants/loci shared between the two diseases. The CCMA method identified 52 independent significant loci where allelic effect directions were consistent between the two diseases and 24 independent significant loci where allelic effect directions were opposing. Colocalisation of the two GWAS showed that of these, 22 loci with consistent allelic effects and five loci with opposing allelic effects had strong evidence of a shared causal variant (posterior probability H4 > 0.8). LAVA analysis confirmed genetic overlap between AD and psoriasis, identifying 115 regions with significant genetic correlation (r_g); 60 with a positive r_g and 55 a negative r_g. The loci we have identified may help inform drug repositioning opportunities. For example, an IL31 pathway inhibitor nemolizumab, is approved for AD but not psoriasis. However, CCMA evidence shows consistent effects at IL31 for both diseases (*P* value = $4.5x10^{-17}$, H4 = 0.95) suggesting potential treatment efficacy for psoriasis.

35

Body Size and Non-Muscle Invasive Bladder Cancer Outcome: What Do the Genes Say?

Tessel E. Galesloot*[1], S. Burgess[2,3], Jasper P. Hof[1], Katja K.H. Aben[1,4], Richard T. Bryan[5,6], James W.F. Catto[7], Neil E. Fleshner[8], Lourdes Mengual[9], Alina Vrieling[1], Sita H. Vermeulen[1]

[1]Radboud University Medical Center, IQ Health Science Department, Nijmegen, The Netherlands; [2]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [3]Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [4]Netherlands Comprehensive Cancer Organization, Department of Research and Development, Nijmegen, The Netherlands; [5]Institute of Cancer & Genomic Sciences, University of Birmingham, Birmingham, United Kingdom; [6]Bladder Cancer Research Centre, University of Birmingham, Birmingham, United Kingdom; [7]Academic Urology Unit, University of Sheffield, Sheffield, United Kingdom; [8]Department of Urology, Princess Margaret Cancer Centre, Toronto, Canada; [9]Department and Laboratory of Urology, Hospital Clínic, IDIBAPS, Universitat de Barcelona, Barcelona, Spain

Background: Patients with non-muscle invasive bladder cancer (NMIBC) have a high risk of recurrence and substantial risk of progression to muscle-invasive disease. Current clinicopathologic-based prediction of NMIBC outcome is imprecise and understanding of biological mechanisms limited. Results of observational studies into body mass index (BMI) and NMIBC outcome remain inconclusive. Here, we investigated if body size is causally associated with NMIBC recurrence and progression using Mendelian randomization analysis and also evaluated potential mediators (i.e. insulin, adipokines, and C-reactive protein).

Methods: We used data from our genome-wide association study (GWAS) consortium for NMIBC outcome (N 4,900). Causal associations between BMI, waist-hip ratio (WHR) and WHR adjusted for BMI (WHRadjBMI) and recurrence and progression were studied using 1) genetic risk scores (GRS) for the body size variables and Cox proportional hazards regression models, and 2) a 2-sample MR design.

Results: No evidence of a causal association for the body size variables was found for the total group (e.g. effect for BMI on progression: beta -0.04 per 1 SD increase in GRS (95% CI -0.12; 0.03), *P* value 0.24) and for men and women separately. Results of the 2-sample MR confirmed the findings from the GRS analyses. Sensitivity analyses showed that our findings and conclusions were robust.

Conclusion: Our results do not support a large causal effect of lifelong predisposition to higher BMI, WHR or WHRadjBMI on NMIBC recurrence and progression. Currently, we are evaluating direct effects of the proposed mediators on NMIBC recurrence and progression, and performing sensitivity analyses on collider bias.

36

Age-Informative Polygenic Score for Quantitative Traits: An Approach and Its Chances and Challenges to Predict Kidney Function and Kidney Function Decline

Janina M. Herold[1], Simon Wiegrebe[1,2], Barbara Thorand[3,5,7], Thomas W. Winkler[1], Christian Gieger[3,4,5], Florian Hartig[8], Annette Peters[3,4,5,8], Helmut Küchenhoff[2], Iris M. Heid*[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Munich, Germany; [3]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [4]German Center for Cardiovascular Disease Research (DZHK), Munich Heart Alliance, Munich, Germany; [5]Research Unit Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [6]German Center for Diabetes Research (DZD), Munich-Neuherberg, Neuherberg, Germany; [7]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilians-Universität, Pettenkofer School of Public Health, Munich, Germany; [8]Theoretical Ecology, University of Regensburg, Regensburg, Germany; [9]Chair of

Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Munich, Germany

Polygenic scores (PGSs) for quantitative traits are widely used to identify individuals at high genetic risk. It is increasingly recognized that the effect of such genetic risk factors can be age-dependent, but the integration of such interactions into PGSs remains underexplored. A particular prominent example for this phenomenon is kidney function, assessed by estimated glomerular filtration rate (eGFR), which is known for strong age-related genetic effects.

We develop an age-informative PGS for quantitative traits by generating age-specific weights via main and interaction effects and compare it to a conventional age-agnostic PGS in theory and real eGFR data.

Our data comprises 282 kidney function SNPs in the cross-sectional and longitudinal UK Biobank data (n=348,275, m=1,520,382) and independent data of population-based individuals aged 25 to 98 years (KORA&AugUR; n=9,057, m=16,804).

Our results demonstrate that the age-informative PGS outperforms age-agnostic PGS in young and old individuals (KORA&AugUR: 6.3% versus 5.8% of eGFR variance in 25- to 45-year-old, 2.3% versus 1.8% in 75- to 98-year-old). Moreover, age-informative PGS explained more of the eGFR-decline variability. Decline per year of age was -0.4 versus -1.1 ml/min/1.73m² in low- versus high-risk group. However, genetic risk strata had a modest impact on predicting eGFR or its decline, comparable to diabetes, obesity, or albuminuria status.

In conclusion, we present a simple approach to conduct age-informative PGS for quantitative traits. Although the impact in our study on eGFR genetics was modest, the small additional effort required for an age-informative PGS supports its use when age-dependent genetic effects are suspected.

Keywords: Polygenic Scores, age-dependent genetic effects, kidney function, prediction, age-informative PGS

37

Design and Development of a Results Relational Database for Cardiovascular Phenotypes

Cristian Riccio[1], Linlin Guo[2,3,4], Georgios Koliopanos[1], Amra Dhabalia Ashok[1], Felicia Sandberg[1], Raphael Oliver Betschart[1,5], Tanja Zeller[3,5], Raphael Twerenbold[2,3,4], Andreas Ziegler[1,2,3,6]

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

Introduction: Advances in multi-omics technologies have promoted multimodal research of complex cardiovascular conditions. Integrating and querying results across large public and in-house datasets can be challenging for diverse stakeholders.

Methods: To overcome this challenge we turned towards relational databases. We evaluated commercial solutions that could help develop a database to house the GWAS catalog, FAVOR, and association results. Product specifications were defined in close collaboration with biologists, clinicians, data scientists and bioinformaticians.
Results: This process led to the identification of key product requirements and a vendor to implement them. These requirements include data upload and maintenance, graphical user interface, identity and access management, and data protection. The resulting system enabled efficient querying and interactive visualizations of results with allowing users to easily access and explore results. Thus, a safe, scalable, informative, and user-friendly platform was deployed.

Conclusion: In conclusion, our relational database offers a collaborative and multimodal exploration of cardiovascular phenotypes.

38

A Multi-Trait GWAS to Disentangle Kidney Trait Genetics

Hannah C. de Hesselle[1], Alexander Teumer[2,3], Klaus J Stark[1], Cristian Pattaro[4], Richard Warth[5], Iris M. Heid[1], Thomas W. Winkler[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany; [3]German Center for Cardiovascular Research (DZHK), Partner Site Greifswald, Greifswald, Germany; [4]Eurac Research, Institute for Biomedicine, Bolzano, Italy; [5]Medical Cell Biology, University of Regensburg, Regensburg, Germany

To identify the genetic basis of chronic kidney disease, genome-wide association studies (GWAS) were conducted on glomerular filtration rate estimated based on serum creatinine (eGFRcrea) or serum cystatin C (eGFRcys) and blood-urea-nitrogen (BUN), which are markers of kidney filtration, as well as urinary-albumin-to-creatinine-ratio (UACR), a marker of kidney damage. However, single trait GWAS cannot distinguish between kidney function and metabolisms loci. Application of multi-trait methods to kidney traits is lacking. We applied multi-trait GWAS (C-GWAS) and fine-mapping (flashfm) to four kidney traits (eGFRcrea, eGFRcys, BUN, UACR; up to 1.4M individuals from UK-Biobank and CKDGen), compared results to single-trait approaches and classified loci regarding kidney function relevance. Our multi-trait GWAS identified 812 independent association signals. Multi-trait fine-mapping substantially sharpened credible-sets (~26% set size reduction) and identified novel likely causal variants (PIP>50%, including a POR missense variant). Overall, 333 signals were associated with filtration function (eGFRcrea and eGFRcys, consistent effects) or UACR (with effect on urinary albumin), but only 11 signals overlapped. Interestingly, overlapping signals demonstrated detrimental effects on filtration function but beneficial effects on structural kidney damage (including kidney eQTLs near

MUC1, SHROOM3). Finally, we observed a significant (Penrich=9x10-27) enrichment of fluid intake associations among filtration function variants (including the POR missense allele decreasing filtration function and fluid intake). Multi-trait fine-mapping sharpened identification of likely causal variants for kidney traits, demonstrated a distinction between filtration function and structural damage genetics, and highlighted an important novel link between filtration function and fluid intake genetics.

Keywords: multi-trait, fine-mapping, kidney, GWAS

39

Genomic Prediction of Actinic Keratosis Risk Using GWAS-Derived Polygenic Scores and Machine Learning Approaches

Yu-Ming Lee*[1], Amrita Chattopadhyay[2], Eric Y. Chuang[1]
[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan; [2]Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

Actinic keratosis (AK) is a common precancerous skin condition linked to ultraviolet (UV) exposure and genetic susceptibility. Leveraging the UK Biobank cohort, we conducted a genome-wide association study (GWAS) comprising 5,316 AK cases and 53,160 controls, using propensity-score matching by age and sex. Multiple genome-wide significant loci were identified, including variants in IRF4 and SPATA33. Based on these results, we constructed polygenic risk scores (PRS) using PRSice-2 and evaluated their predictive performance in logistic regression models. In addition, we integrated PRS with environmental and lifestyle covariates—such as skin color, tanning ability, and sunscreen use—into machine learning models to assess improvements in prediction. The preliminary models yielded modest predictive performance, with an area under the curve (AUC) reaching approximately 0.61. Functional annotation and expression quantitative trait loci (eQTL) analysis suggested enrichment in UV response and skin-related pathways. Our findings highlight the utility of PRS and covariate integration in risk stratification of AK and provide insight into the underlying genetic architecture of the disease.

Keywords: actinic keratosis, GWAS, polygenic risk score, UK Biobank, machine learning

40

Incorporating Large-scale Genetic Association Results Into Clinical Genetics Consortium (ClinGen) Sequence Variant Classification

Elika Garg[1], Andrew D. Paterson*[1,2]
[1]Program in Genetics & Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; [2]Biostatistics and Epidemiology Divisions, Dalla Lana School of Public Health, University of Toronto, Ontario, Canada

Accurate rare variant interpretation is challenging: each variant is typically observed only in a small number of individuals. Over the last decade, Clinical Genetics Consortium (ClinGen) used disease-gene specific expert curation groups who follow and develop criteria to interpret sequence variants, providing classifications for 9.9k variant-disease pairs across 138 genes. However, upscaling this activity to the whole genome is challenging. To assist this activity, we examined whether single-variant association results from exome sequencing (WES) in UK Biobank (genebass.org) differ between ClinGen variant classes. Using phewas for predicted loss of function and missense variants within each gene, we first determined which of 1,493 traits (with case n >1k and total n>100k) revealed the most significant association for that gene using SKAT-O. We then used that trait for the association results of all ClinGen variants in that gene. Next, we examined whether there were significant differences in the variant effect sizes (absolute beta) by ClinGen classes across genes. 31% of unique ClinGen variants (n=2,624) were found in UKB WES in 101 genes. ClinGen variants classified as pathogenic had significantly larger effect sizes (mean±SE=1.19±0.06, n=537) than those classified as benign (0.57±0.34, n=517, p=E-22), likely benign (0.70±0.37, n=442, p=E-7), and uncertain significance (0.71±0.02, n=823, p=E-7). Those classified as likely-pathogenic (0.92±0.06, n=305, p=0.1) were not significantly different than pathogenic. This indicates that statistical association with traits from large population-based cohorts could be added to interpretation criteria for rare variants for clinical genetics purposes. Combining association results across multiple phenotypes and cohorts could further improve this approach.

41

A Multivariate Approach to Identify CpGs in Peripheral Blood Where DNA Methylation is Associated with Cerebrospinal Fluid Biomarkers of Alzheimer Disease

Bowei Xiao*[1], Yixiao Zeng[1,2], Kathleen O. Klein[2], Bianca Granato[1], Mathieu Blanchette[3], Xiaojian Shao[4,5], Celia M.T. Greenwood[2,6,7], *Alzheimer's Disease Neuroimaging Initiative*
* Presenting author
[1]Quantitative Life Sciences, McGill University, Montreal, Quebec, Canada; [2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada; [3]School of Computer Science, McGill University, Montreal, QC, Canada; [4]Digital Technologies Research Centre, National Research Council Canada, Ottawa, Ontario, Canada; [5]Ottawa Institute of Systems Biology, Department of Biochemistry Microbiology and Immunology, University of Ottawa, Ottawa, Ontario, Canada; [6]Gerald Bronfman Department of Oncology, McGill University, Montreal, Quebec, Canada; [7]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

DNA methylation patterns are known to show associations with many diseases, including Alzheimer's disease (AD). Although many studies have correlated DNA methylation in blood samples with risk of clinical AD diagnosis, fewer studies have examined links with AD neuropathology. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, we investigate the associations between peripheral blood DNA methylation and three biomarkers in cerebrospinal fluid: amyloid-β (A), tau pathology (T), and neurodegeneration (N) using an innovative multivariate penalized approach. When analyzing multiple correlated phenotypes and analyzing gene-adjacent regions of methylation probes, our method is capable of not only detecting strong signals that can also be identified

by conventional association tests between a single phenotype and a single probe (EWAS) but also identifying weaker signals that would have been missed by conventional methods. We validated our method both in simulations and in the ADNI dataset by examining jointly all CpG probes annotated to a gene. We then compared our findings from the ADNI dataset with both conventional EWAS and with another multivariate method, *cglasso*. Our finding shows that when we identified associations between biomarkers and gene-adjacent methylation patterns, these genes had often been previously reported to show associations with AD-related phenotypes. Our multivariate strategy has the potential to increase sensitivity in epigenetic studies while also improving the selection accuracy among correlated predictors.

42

Epigenetic Factors Predict Incident Heart Failure in Multi-Ancestral Populations

Yan V. Sun[1,2,3], Gregorio V. Linchangco[1,2], Qin Hui[1,2], Maryam Rahafrooz[4], J. Michael Gaziano[5,6], Peter W.F. Wilson[1,3], The Million Veteran Program, Lawrence S. Phillips[1,3], Jacob Joseph[4,7]

[1]Atlanta VA Healthcare System, Decatur, Georgia, United States of America; [2]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, Georgia, United States of America; [3]Department of Medicine, Emory University School of Medicine, Atlanta, Georgia, United States of America; [4]VA Providence Healthcare System, Providence, Rhode Island, United States of America; [5]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, Massachusetts, United States of America; [6]Division of Aging, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America; [7]The Warren Alpert Medical School of Brown University, Providence, Rhode Island, United States of America

Background: Heart failure (HF) is a life-threatening aging-related syndrome with growing impact on the global population. HF progression can be affected by both gene and environment via different molecular pathways, which can be represented by epigenetic factors measured by DNA methylation (DNAm).

Methods: After excluding samples with low quality or discordant sex, and merging with the phenotypic data, the multi-ancestry (~70%, 24%, 5.4%, and 0.6% are European, African, Hispanic, and Asian Americans, respectively) epigenomic study in the Million Veteran Program (MVP) included 3,525 with incident HF and 31,329 participants without any record of HF. Incident HF was determined by the first post-enrollment HF diagnosis with a measured left ventricle ejection fraction within 90 days. For each DNAm site, the association with HF was examined in a multiple regression model adjusted for age, sex, race/ethnicity, calculated blood cell proportions, and potential batch effects. Epigenome-wide significance (EWS) threshold ($p<6.6x10^{-8}$) was corrected for multiple testing.

Results: We identified 107 EWS DNAm sites associated with incident HF, including the most significant DNAm in chromosomes 17 (17:76354934, *SOCS3*, $p=7.40x10^{-15}$), 11 (11:315102, IFITM1, $p=3.77x10^{-13}$), 19 (19:1126342, *SBNO2*, $p=3.00x10^{-12}$), and 2 (2:20232577, *LAPTM4A*, $p=4.97x10^{-12}$). The identified DNAm sites showed mostly negative association (86%), and were enriched for inflammatory response pathways.

Conclusion: The significant epigenetic associations with incident HF revealed the genetic pathways underlying disease progression. These findings may lead to novel targets for treatment and prevention which could reduce the burden of HF among the growing aging population.

Keywords: DNA methylation, EWAS, heart failure, aging, incidence

43

Phenome-Wide Analysis Affirms Utility of Combining Polygenic Scores with Proteomics in Risk Prediction of Incident Disease

Jakob Woerner*, Thomas Westbrook, Jaehyun Joo, Yonghyun Nam, Dokyoon Kim

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Polygenic risk scores (PRS) and proteomic risk scores (ProRS) have independently demonstrated predictive utility for complex human diseases, but their combined value across a broad disease spectrum remains underexplored. Integrating genetic and 2,920 plasma protein measurements in 39,843 UK Biobank participants, we computed PRS and ProRS for 301 diverse disease phenotypes. Models trained on prevalent cases and controls were evaluated for their ability to predict incident disease.

ProRS improved prediction of future disease onset over PRS in 95% of traits, especially for conditions with lower heritability. PRS remained important in predicting more heritable diseases, particularly autoimmune and cancer phenotypes. Combined modalities (PRS+ProRS) achieved the highest predictive accuracy and improved risk stratification, especially in those at the highest risk. In proportional hazards models, PRS explained >20% of the relative variation in eight autoimmune diseases. While ProRS provided limited added value for most cancer outcomes, genetics were significant in predicting incidence. Temporal analyses revealed that although ProRS performance wanes as time since blood draw increases, it can still identify individuals at risk up to a decade before disease onset.

Our findings support leveraging both static genetic and dynamic proteomic markers for the care and screening of complex diseases. While previous studies have questioned whether PRS provides any additional information beyond protein risk scores, we showed that PRS further stratifies risk even among individuals in the highest ProRS strata. These results reaffirm the utility of the complementary information inferred from genetics and proteomics to more precisely identify individuals at elevated disease risk.

Keywords: plasma proteomics, polygenic risk score, multi-omics, phenome, complex traits

45

Framework for Allelic Effect Heterogeneity Assessment in Genome-Wide Association Study Meta-Analyses

Chuan Fu Yap & Andrew P. Morris
Centre for Genetics and Genomics Versus Arthritis, The University of Manchester, Manchester, United Kingdom

Multi-ancestry genome-wide association studies (GWAS) of disease have reported heterogenous allelic effects between diverse populations. The standard approach to capturing such heterogeneity is to stratify participants based on continental labels (e.g. European and African) and aggregate through multi-ancestry meta-analysis. However, this approach can lead to loss of genetic representation as not all participants can be assigned to a label. To maintain all samples in a single GWAS without stratification, we propose a continuous and multi-dimensional representation of genetic diversity by first projecting participants onto genetic principal components (PCs) derived from the Human Genome Diversity Project. The resulting PCs are used to assess allelic effect heterogeneity via SNP-by-PC interaction terms (e.g., SNP×PC1–3), by extending the whole-genome regression framework to support multiple interaction testing. To facilitate the meta-analysis of multiple interaction terms across studies, we implemented a multivariate generalized least squares approach to synthesis the regression slopes. We applied the new interaction meta-analysis to GWAS on type 2 diabetes (T2D) in up to 55,090 cases and 503,503 controls of diverse ancestry. Interaction tests were performed for lead SNPs at 170 loci attaining genome-wide significant evidence of association($P<5x10^{-8}$). Of these, there was significant evidence of allelic effect heterogeneity (P<0.05) at 14 loci. Simulations indicated that the power of this method to detect heterogenous association signals depends on the interaction effect size but not the marginal allelic effect size. These findings highlight the value of modelling continuous genetic structure to uncover allelic effect heterogeneity and improve inclusivity in GWAS.

46

Comparison of Classic Polygenic Scores with Machine Learning Algorithms to Predict Hypertension

Tanja K. Rausch*, Silke Szymczak, Inke R. König
Institute of Medical Biometry and Statistics, University of Lübeck, University Hospital of Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Hypertension is the leading risk factor for cardiovascular disease and blood pressure is a frequently measured parameter. Given the polygenic heritability of complex traits like hypertension, polygenic scores (PGS) are increasingly used to stratify individuals by genetic susceptibility for targeted prevention, therapy, or prognosis. However, classic PGS use a simple sum of genotypes, weighted by effect sizes from single variant genome-wide association studies (GWAS), ignoring multivariable and non-linear effects. Because classic statistical methods reach their limits with many independent variables, machine learning (ML) algorithms offer an alternative for score construction.

ML algorithms have not yet been applied to construct PGS for predicting hypertension. It remains unclear whether more complex algorithms improve performance. This study evaluates ML algorithms e.g., random forest and k-nearest neighbor, using UK Biobank data. Hypertension is defined as antihypertensive medication use, diastolic blood pressure >90 mmHg, or systolic >140 mmHg at baseline. The dataset is repeatedly and randomly split into variable selection, training, and test datasets. Important variables are identified in the selection set. The training set is then used to generate a PGS via GWAS or train ML models. Hyperparameter tuning will be performed. Prediction performance is assessed on the test set by the area under the receiver operating curve and the Brier score.

Preliminary results indicate that PGS is sufficient to predict hypertension. This study provides insight into whether genetic information compressed by complex ML algorithms outperforms classic PGS and which model is best suited. Additionally, results contribute to understanding the genetic structure of hypertension.

Keywords: polygenic score, machine learning, comparison, hypertension, UK biobank

47

Evaluating the Consequences of Childhood Adiposity on the Human Plasma Proteome at Three Timepoints Across the Lifecourse

Phoebe H. Dickson*[1,2], Grace M. Power[1,2,3], Tom R. Gaunt[1,2], George Davey Smith[1,2†], Tom G. Richardson[4†]
[1]MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia; [4]Human Genetics and Genomics, GSK Research and Development, Stevenage, United Kingdom
*Presenting author
†These authors contributed equally.

Adiposity has a profound influence on the human proteome, although determining whether this is due to a legacy effect of being overweight in childhood is challenging due to confounding factors throughout the lifecourse. In this study, we aimed to address this question by applying lifecourse Mendelian randomization to circulating protein data measured at three key life stages: childhood (mean age 9.9 years) and early adulthood (mean age 24.5 years) using data from the ALSPAC cohort, and midlife (mean age 55.9 years) using data from up to n=34,557 UK Biobank participants.

Our analyses highlight direct effects of early life adiposity on eight circulating proteins during childhood, such as Osteoprotegerin and TNFSF11. We further demonstrate that TNFSF11 confers risk of childhood-onset asthma (colocalization posterior probability (PPA)= 94.58%) but not adult-onset asthma (PPA=0.44%), highlighting its putative role as a causal intermediate between adiposity and asthma risk specific to early life.

By midlife, we found that genetically predicted childhood adiposity had a robust effect on 139 circulating proteins. Evidence of an independent effect persisted for 20 of these proteins after accounting for adulthood adiposity. Validation analyses conducted in n=35,559 Icelanders from the

deCODE study found corroborating evidence for an independent effect of childhood adiposity on proteins such as NOTCH3 which has been found to play a protective role in breast cancer susceptibility.

Our findings provide considered insight into the time-varying influence of adiposity on the circulating proteome and highlight key proteins warranting prioritisation by future work to unravel their mechanistic roles in disease aetiology.

48

Neurodevelopmental Disorders Copy Number Variants and Risk of Internalising and Cardiometabolic Disorders

Lam O. Huang*[1,2,3] , Simone Montalbano[1,2], Seyedmorteza Vaez[1,2], Dorte H. Mikkelsen[1,2], Thomas Werge[1,2,4], Andrés Ingason[1,2,3]

[1]Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. [2]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen, Denmark. [3]Lifespan Multimorbidity Research Collaborative (LINC), United Kingdom. [4]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

[**Background**] In a population-based sample of young Danes we found no association between carriage of neurodevelopmental copy number variations (NDD-CNVs) and increased risk of depression. However, some recent evidence indicates that NDD-CNV-associated risk of internalising disorders in the UK Biobank is higher among older female subjects with comorbid cardiometabolic disorders (CMD) compared to those without CMD.

[**Hypothesis**] We want to investigate whether risk of depression associated with NDD-CNVs is mediated through lifestyle-associated complications across lifespan, in a sex-dependent manner.

[**Data sources**] We leverage on two existing Danish biobanks, with Copenhagen Hospital Biobank (CHB; N=227,672) used as the case population, and the Danish Blood Donor Study (DBDS; N=100,872) as the control population. CHB includes samples from patients admitted to hospitals for diagnostic and treatment purposes, and DBDS includes healthy volunteer blood donors. Both biobanks are linked with a comprehensive collection of nation-wide registers, which includes hospital diagnoses, drug prescriptions, education, and income. This grants us unprecedented opportunity to study a wide range of disease outcomes and ancillary variables in association with genetic exposures such as NDD-CNVs. [Methods] We conduct a sex-stratified analysis of the association between NDD-CNVs and risk of internalising disorders in the elderly, with and without comorbid CMD. To address selection bias, we use propensity score weighting to adjust for disparity in age, gender, and overall health status between CHB and DBDS. Generalized estimating equation (GEE) is used to account for multimorbidity within individuals. Multistate modelling is also employed to investigate the progression through different disease states across lifespan.

Keywords: register-based biobank, comorbidity, propensity score weighting, generalized estimating equation, multistate model

49

Recurrent Copy Number Variants and Polygenic Scores Jointly Influence the Risk of Psychiatric Disorders in the iPSYCH2015 Case-Cohort Sample

Morteza Vaez[1,2], Simone Montalbano[1,2], Ryan Waples[1,2], Morten Dybdahl Krebs[1,2], Kajsa-Lotta Georgii Hellberg[1,2], Jesper Gådin[1,2], Dorte Helenius[1,2], Thomas Werge[1-4], Andrew J. Schork[1-3,5], Andrés Ingason[1-3]

[1]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Roskilde, Denmark; [2]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen and Aarhus, Denmark; [3]Lundbeck Foundation Center for GeoGenetics, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark; [4]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [5]Neurogenomics Division, The Translational Genomics Research Institute (TGEN), Phoenix, Arizona, United States of America

Background: While both recurrent copy number variants (rCNVs) and common SNP variants are known to influence the risk of psychiatric disorders, their joint effect remains underexplored. Using the population-based iPSYCH2015 case-cohort, we investigated the combined effect of rCNVs and polygenic scores (PGS) on major psychiatric disorders.

Methods: The iPSYCH2015 case-cohort includes all individuals diagnosed with major psychiatric disorders (n=82,626) and a random population-based sample (n=41,346) from a Danish birth cohort (1981 to 2008). We identified rCNVs at 27 genomic loci using PennCNV and grouped them according to their gene constraint. PGSs were derived from external genome-wide association studies, with SNP effect sizes rescaled using SBayesR. Survival models, incorporating inverse probability weights, were used to estimate absolute risks, while generalized linear models were employed to evaluate additive and interactive effects between rCNVs and PGSs, as well as differences in PGS distributions between rCNV carriers and non-carriers.

Results: Higher PGS and gene-constrained rCNVs were associated with increased absolute risk for autism, ADHD, and schizophrenia, but not major depressive disorder. For ADHD and Schizophrenia, more individuals at similar risk levels attributed to rCNV groups were identified through PGSs. We observed additive, but not multiplicative, effects of rCNVs and PGSs on ADHD, ASD, and schizophrenia. PGS profiles for psychiatric and several non-psychiatric traits did not differ significantly between rCNV carriers and non-carriers.

Conclusion: Joint assessment of rCNVs and PGSs in a population-based iPSYCH2015 case-cohort highlights their additive contributions to psychiatric risk and supports integrated genetic profiling as a tool for precision psychiatry.

50

Genetic and Geographic Influence on Phenotypic Variation in European Sarcoidosis Patients

Sandra Freitag-Wolf[1], Astrid Dempfle[1], Joachim Müller-Quernheim[2]

[1]Institute of Medical Informatics and Statistics, Kiel University,

Kiel, Germany; [2]Department of Pneumology, University Medical Centre, Faculty of Medicine, Freiburg, Germany

Sarcoidosis is a highly variable disease and associations of genetic polymorphisms with sarcoidosis phenotypes have been observed and suggest genetic signatures. The Genetic-Phenotype Relationship in Sarcoidosis (GenPhenReSa) cohort consists of 1909 deeply phenotyped patients recruited from 31 centers in 12 European countries with 116 potentially disease-relevant single-nucleotide polymorphisms (SNPs). We investigated the association of relevant phenotypes (acute vs. sub-acute onset, phenotypes of organ involvement, specific organ involvements, and specific symptoms) with genetic markers and build subgroups on the basis of geographical, clinical and hospital provision considerations. In the meta-analysis of the full cohort, there was no significant genetic association with any considered phenotype after correcting for multiple testing. In the largest sub-cohort (Serbia), we confirmed the known association of acute onset with TNF and reported a new association of acute onset an HLA polymorphism. Multi-locus models with sets of three SNPs in different genes showed strong associations with the acute onset phenotype in Serbia and Lublin (Poland) demonstrating potential region-specific genetic links with clinical features, including recently described phenotypes of organ involvement.

The observed associations between genetic variants and sarcoidosis phenotypes suggest that gene-environment-interactions may influence the clinical phenotype. In addition, we show that two different sets of genetic variants are permissive for the same phenotype of acute disease only in two geographic subcohorts pointing to interactions of genetic signatures with different local environmental factors.

51

Comparison of Meta-Learners for Late-Stage Prediction Modeling for Multi-Omics Data

Marina Bleskina*, Cesaire J. K. Fouodo, Silke Szymczak
Institute of Medical Biometry and Statistics, University of Lübeck, University Hospital of Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Advances in modern technologies have made it feasible to collect diverse data modalities, such as multi-omics data, from the same individuals. Their unique characteristics make it challenging to integrate them for predictive purposes.

A promising approach involves training modality-specific prediction models separately. Their predictions are used as input for a meta-model that delivers the final predictions. However, it is currently unclear which of the many proposed meta-learners should be applied for a specific research question and combination of data modalities.

We systematically evaluated the prediction performance of different meta-learners for multi-omics data in a simulation study. We investigated settings reflecting both independent and correlated effects between modalities, as well as varying effect sizes. The evaluated meta-learners included weighted average, best modality-specific learner, logistic regression, least absolute shrinkage and selection operator (Lasso), the combined regression alternative (COBRA), and random forest. Our results demonstrate that complex meta-learners, including logistic regression, Lasso, random forest, and COBRA, consistently outperform simpler approaches (weighted average and best modality-specific learner), particularly in settings with stronger effects. Interestingly, selecting the best modality-specific learner often yields better prediction performance than the weighted average.

Keywords: multi-omics, machine learning, integrative predictive modeling

52

Re-Evaluating the Association Between 22q11 Deletion Syndrome and Schizophrenia

Andrés Ingason[1,2], Morteza Vaez[1,2], Simone Montalbano[1,2], Lam Opal Huang[1,2], Soha Sabtiy[1,2], Wesley K. Thompson[2,3], Armin Raznahan[4], Dorte Helenius[1,2], Thomas Werge[1,2,5]
[1]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Roskilde, Denmark; [2]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen and Aarhus, Denmark; [3]Laureate Institute for Brain Research, Tulsa, Oklahoma, United States of America; [4]Section on Developmental Neurogenomics, Human Genetics Branch, National Institute of Mental Health Intramural Research Program, Bethesda, Maryland, United States of America; [5]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

Background: For a long time 22q11 deletion syndrome (22q11DS) has been considered a major risk factor of schizophrenia and related psychoses, with reported ORs as high as 80-fold and penetrance estimates exceeding 50%. Our recent research based on nation-wide Danish registers and biobanks indicates that the true increase in risk of schizophrenia associated with 22q11DS is much lower than previously held. In our most recent study, we found only a threefold increased risk of schizophrenia spectrum disorder (SSD) associated with 22q11DS in the population-based iPSYCH2015 case-cohort sample. The stark contrast with previous estimates has sparked skepticism as to the suitability of the iPSYCH2015 sample to study schizophrenia and related psychoses, given the young age of participants and the reliance on diagnoses from public hospital registries.

Aim: To address this critique we compared the prevalence and characteristics of SSD in iPSYCH2015 with those reported in other studies, and performed simulations to study how risk estimates for 22q11DS and other schizophrenia-associated copy number variants (CNVs) change with different length of follow-up and with screening of controls.

Finding: Our results show that the characteristics of the iPSYCH2015 SSD sample are comparable to those of other studies and that risk estimates do not increase with longer follow-up, whereas removing individuals with other psychiatric disorders or a family history thereof from the comparison group leads to inflated risk estimates, suggesting that the use of screened controls is a major factor contributing to the high ORs reported for 22q11DS and some other CNVs in case-control studies.

53

The Neurodevelopmental Risk Associated With Congenital Heart Disease and Recurrent Copy Number Variants

Soha H. Sabtiy*[1,2], Sara H. Lau-Jensen[3,4], Morteza Vaez[1,2], Simone Montalbano[1,2], Lars Allan Larsen[5], Thomas Werge[1,2,4], Vibeke Hjortdal[3,4], Andrés Ingason[1,2], and Dorte Helenius[1,2]

[1]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Roskilde, Denmark; [2]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen and Aarhus, Denmark; [3]Department of Cardiothoracic Surgery, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; [4]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [5]Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

Background: Advancements in diagnostics and treatments have allowed more individuals with congenital heart disease (CHD) to survive into adulthood. As a result, there is an increased awareness of comorbid neurodevelopmental disorders (NDDs) in CHD. However, the underlying mechanisms linking CHD to NDDs are still poorly understood. We aimed to investigate the influence of rare recurrent copy number variants (rCNVs) on the associations between CHD and five NDDs: attention-deficit/hyperactivity disorder (ADHD), affective disorder (AFF), autism spectrum disorder (ASD), bipolar disorder (BDP), and schizophrenia spectrum disorder (SCZ).

Method: We utilized the iPSYCH2015 case-cohort study sample, which includes 141,265 individuals born in Denmark in 1980-2008 and followed until 2015. We assessed the risk of NDDs based on CHD and rCNV carrier status using a weighted Cox proportional hazards model.

Findings: CHD significantly increased the risk of ADHD (HR = 1.50) and ASD (HR = 1.60). Those under five are nearly four times more likely to develop ADHD (HR = 3.90) and over twice as likely for ASD (HR = 2.60), with risk decreasing as they age. While rCNVs were associated with CHD, ADHD, and ASD, adjusting for these did not moderate the effect of CHD on either NDDs.

Conclusion: These findings suggest that CHD and the studied rCNVs influence ADHD and ASD independently. Additionally, individuals with CHD are diagnosed with ADHD or ASD at an earlier age than those without CHD, highlighting a potential common factor that may contribute to and accelerate this association.

54

Large-Scale Genotype-Phenotype Simulations for Genomic Studies

Victor Vera Frazao[1], Sebastian Sendel[2,3], Amke Caliebe[2,3], Michael Nothnagel[1,4]

[1]Cologne Center for Genomics, University of Cologne, Cologne, Germany; [2]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [3]University Medical Centre Schleswig-Holstein, Kiel, Germany; [4]University Hospital Cologne, Medical Faculty, University of Cologne, Cologne, Germany

Large-scale genotype and phenotype simulations are frequently used to integrate population genetics approaches into genetic epidemiological research, highlighting evolutionary influences in human genetic variation. In genome-wide association studies (GWAS), synthetic datasets mirroring real-life populations can help improve the methodology of polygenic scores (PGS), particularly in terms of their accuracy across different ancestries. Although numerous simulation software packages exists, they frequently require extensive setup before large-scale genotype and phenotype simulations can be performed effectively. HAPNEST is a promising toolkit for simulating large, representative datasets, offering algorithms for both single- and cross-population genotype and phenotype simulation, along with validation tools. However, in our experience, it demands extensive data preparation to mitigate runtime and disk space challenges when constructing large datasets. Furthermore, while HAPNEST provides a function to run GWAS, no PGS method are incorporated yet and this requires further and method-dependent data preparation. Here, we present an extensive modular pipeline designed to streamline data preparation, simulation and validation with HAPNEST, minimizing the effort to to provide necessary input. The pipeline offers flexible filtering options, including the ability to retain pre-selected sets of variants, applying allele frequency threshold and including or excluding (inter-)genic variants. It also automates variant mapping and annotation as well as re-formatting the input. Additionally, the pipeline supports multiple PGS construction methods, enabling seamless PGS generation after the simulation and GWAS steps. This framework facilitates benchmarking of existing PGS construction methods and the comparison of their performance across different ancestries using simulated data.

Keywords: genotype simulation, phenotype simulation, PGS, pipeline

55

Collection of Multi-Omics Data for the Investigation of Long-Term Conditions in EXCEED (Extended Cohort for E-health, Environment and DNA)

Catherine John*[1,2], Chiara Batini[1,2], Brandon Lim[1], Gerald Sze[1,2], Nick Shrine[1], David J. Shepherd[1], Hiten D. Mistry[1], Laura Venn[1], Martin D. Tobin[1]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]University Hospitals of Leicester NHS Trust, Leicester, United Kingdom

The Extended Cohort for E-health, Environment and DNA (EXCEED) is a longitudinal population study based in the Midlands (UK), which aims to facilitate the study of genetic and environmental determinants of health and disease. Over 11,000 participants, primarily older adults, have been recruited since 2013, by completing baseline questionnaires and providing biological samples. Consent for linkage to electronic health records permits longitudinal follow-up of a broad range of long-term conditions and health-related traits.

Samples were genotyped using the UK Biobank Axiom array and imputed to the TOPMed reference panel. Ancestry groups were defined using k-means clustering on principal components, with 1000 Genome Project Phase 3 (1KG) populations as reference. Measurement of 319 proteins and

7750 metabolites was undertaken for a subset of participants using mass spectrometry, and 39 glycoproteins were measured using liquid chromatography. We tested associations between genetic variants and protein levels using regenie in 1KG-EUR-like individuals.

After quality control, genomic data was available for 8499 individuals, proteomic for 1979, metabolomic for 1524 and glycomic for 900. There were 541 participants with quality-controlled data from all four sources, whilst 1551 had both genomic and proteomic data. We detected cis-protein quantitative trait loci for 114 proteins; 59 of these proteins are not available in UK Biobank proteomic data.

We present a multi-omics resource with potential for novel insights into the biology of long-term conditions. Next steps will include exploration of associations between protein and metabolite levels and long-term conditions, through genome-wide association and Mendelian randomisation analyses.

56

The Relationship between Maternal Migraine During Pregnancy and Offspring ADHD Traits

Yaxin Luo*[1,2], Christina Dardani[1,2 3,4], Robyn E Wootton[2,3,4,5], Evie Stergiakouli[1,2]

[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [2]Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Research Department, Lovisenberg Diaconal Hospital, Oslo, Norway; [4]PsychGen Centre for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway; [5]School of Psychological Science, University of Bristol, Bristol, United Kingdom

Background: Observational studies suggest associations between maternal migraine and offspring ADHD. It is unclear whether these associations indicate causal relationships. Towards this, it is important to utilise a causal triangulation framework to assess whether maternal migraine in pregnancy might causally influence ADHD in the offspring.

Methods: Using Avon Longitudinal Study of Parents and Children (ALSPAC) data, we examined associations between maternal migraine during pregnancy and offspring ADHD traits at age 7. We conducted negative control analyses using paternal migraine to disentangle intrauterine effects from familial confounding. Maternal non-transmitted and transmitted Polygenic Risk Scores (PRSs) for migraine were utilized as instrumental variables in one-sample Mendelian Randomization (MR) to investigate in utero influences and genetic confounding on offspring ADHD, separately.

Result: A total of 4171 family trios were included in the analysis. Maternal migraine during the first trimester was observationally associated with elevated ADHD traits in offspring (OR = 1.59 [1.22, 2.06]). There was little evidence supporting a similar association for partner's migraine (OR = 1.31 [0.95, 1.82]). Point estimates from one-sample MR suggested possible transmission (OR = 1.06 [0.94, 1.18]) exceeding potential in utero effects (OR = 1.02 [0.87, 1.2]), though wide confidence intervals precluded definitive conclusions.

Conclusion: There was evidence of associations between maternal migraine in pregnancy and ADHD symptoms in the offspring. Although pregnancy-specific effects cannot be entirely excluded, genomic evidence suggests shared genetic effects—rather than causal in utero mechanisms—underlie this multigenerational neurodevelopmental relationship.

Keywords: Mendelian Randomization, Cross-generation association, genetic confounding

57

Multi-Ancestral, Trans-generational Genome-Wide Association Meta-Analysis of Gestational Diabetes Mellitus and Glycemic Traits During Pregnancy

Valentina Rukins*[1], Caroline Brito Nunes[2], Nancy McBride[3,4], Aminata H. Cissé[5], Frédérique White[6], Gunn-Helen Moen[2,7,8], and Reedik Mägi[1] on behalf of the GENetics of Diabetes In Pregnancy (GenDiP) Consortium

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Institute for Molecular Bioscience, The University of Queensland, St Lucia, Queensland, Australia; [3]MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom; [4]Population Health Science, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [5]Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, United Kingdom; [6]Département de Biologie, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada; [7]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway; [8]University of Queensland Frazer Institute, University of Queensland, Woolloongabba, Australia

Gestational diabetes mellitus (GDM) is a metabolic complication affecting ~14.0% pregnancies worldwide. It is associated with adverse maternal and fetal health outcomes and notably increases the risk of maternal type 2 diabetes (T2DM) later in life. We investigated the contributions of maternal and fetal genetic variants to GDM and glycemic traits during pregnancy by performing the largest multi-ancestral genome-wide association study (GWAS) meta-analyses of these traits to date.

In this study, we leveraged data from 30 cohorts representing African, East Asian, South Asian, European, and Hispanic ancestral groups. We conducted GWAS meta-analyses of: i) GDM diagnosis (Nmaternal=38,375 cases, 776,075 controls; Nfetal=3,126 cases, 90,877 controls), ii) Fasting glucose ($N_{maternal}$=55,371; N_{fetal}=15,855), iii) 1-hour glucose post-oral glucose tolerance test (OGTT) ($N_{maternal}$=38,439; N_{fetal}=9,365), iv) 2-hour glucose post-OGTT ($N_{maternal}$=46,401; N_{fetal}=15,046) and v) HbA1c ($N_{maternal}$=9,724; N_{fetal}=7,035). Downstream analyses included conditional analyses to partition maternal and fetal genetic effects, and shared variant analyses to distinguish GDM or T2DM-predominant effects at individual loci.

We identified 37 loci for GDM (29 novel) and 16 novel trait-variant associations for glycemic traits at genome-wide significance in the maternal meta-analyses. Conditional analyses suggested that associations detected in the fetal GWAS were likely due to correlations with maternal genotypes, not fetal genetic effects. Several genetic variants discovered were known to associate with glycemic traits

outside of pregnancy. Likewise, LD score regression analyses supported substantial genetic overlap between GDM and T2DM (r_G=0.78). However, 11 variants showed stronger associations with GDM than with T2DM, including novel variants at *TENT5C, GCK, HKDC1, RMST, FOXA2* and *NFATC2*.

Keywords: Pregnancy, GWAS, Glycemic Traits, Gestational Diabetes.

58

Characterisation of Diverse Global Ancestries within Participants of the UK Biobank

Fiona Pantring*[1,2,3], Gianpiero L. Cavalleri[1,2,3], Edmund Gilbert[1,2]
[1]School of Pharmacy and Biomedical Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland; [2]The FutureNeuro Research Centre, Dublin, Ireland; [3]Research Ireland Centre for Research Training in Genomics Data Science, Ireland

The UK Biobank (UKB) is a large dataset containing in-depth phenotype and genotype data of 500,000 UK-based participants. To control for cryptic population genetic confounders, studies leveraging the UKB are typically restricted to a subset of the participants with homogenous European ancestry. By analysing the 78,573 UKB participants with non-UK-like ancestries using population genetic approaches, there is an opportunity to better understand the global genetic diversity in the UKB and empower their inclusion in disease association studies.

Here we characterise these diverse ancestries by identifying eight primary continental-like ancestry clusters and 293 fine-scale communities. Individuals were assigned to continental-like ancestry clusters using the machine learning algorithm XGBoost and each cluster was further divided by applying community detection to a network of Identity-By-Descent sharing.

We find that the UKB is a repository of diverse ancestries primarily of European-, African-, and South Asian-like descent. Whilst capturing worldwide diversity, the 293 communities appear to reflect the immigration history of Great Britain and its Commonwealth in the 20th century and are likely less represented in other large global biobanks.

Our communities facilitate novel findings of community-specific genetic risk factors, such as one of the highest worldwide frequencies of idiopathic pulmonary fibrosis risk variant rs35705950 in individuals of Maltese-like ancestry. Analysis of fine-scale ancestry communities enables further analysis into the distribution of genetic risk factors in the general population as well as into the demographic history. Thus, the work provides a framework for future studies of health-related genetic variation specific to otherwise understudied genetic communities.

59

A Genome-Wide Association Study Using the Bacillus Calmette-Guérin Scar to Elucidate the Genetic Basis of Scarring in a European Population

Iona Collins[1,2], Marisa Cañadas-Garre[1,2], Oscar A Peña[3], Vanessa Tan[1,2], Leila Thuma[4], Paul Martin[3], Nicholas J. Timpson[1,2]
[1]Bristol Medical School, MRC Integrative Epidemiology Unit, University of Bristol; [2]Population Health Science, Bristol Medical School, University of Bristol; [3]School of Biochemistry, University of Bristol; [4]School of Medicine, University of Nottingham

Background: Scarring is a complex disease, whose severity is impacted by environmental and genetic factors. Systematically collecting human wound healing information has been challenging. Previous genetic research has focused on fibroproliferative lesions including keloids and hypertrophic scarring rather than scarring severity and wound repair variation. The aim of our study was to identify variants associated with scarring severity within genes that might influence the mechanisms of wound healing and scarring.

Methods: We performed a genome-wide association study (GWAS) in 749 women from the European Avon and Longitudinal Study of Parents and Children (ALSPAC) population, using the size of the Bacillus Calmette Guérin (BCG) scar as a proxy for scarring severity. After quality control, 11,237,058 variants imputed from TOPMed were analysed. Post GWAS analyses were performed to uncover the biological context and functional relevance of our findings.

Results: We identified 181 independent variants with a *P* value < 5e-5 associated with the BCG scar size. The most significant association was observed at rs61887838 in leucine rich repeat containing G protein-coupled receptor 4 (*LGR4*), where the G allele related to a smaller sized BCG scar (beta: -1.472; *P* value:1.03e-7). *LGR4* is a gene which influences the Wnt/β catenin pathway, involved in wound healing. Other genes were identified *(PTPRD, FGF1, CAST, COL22A1, PIEZO1*), implicated in pathways associated with wound healing.

Conclusion: Several variants within genes involved in pathways associated with wound healing and scarring were associated with the size of the BCG scar. Experimental studies will validate the functional relevance of these findings.

60

Penalized Generalized Linear Mixed Models for Longitudinal Outcomes in Genetic Association Studies

Julien St-Pierre*[1], Sahir Rai Bhatnagar[1], Massimiliano Orri[1,2,3], Michel Boivin[4,5], Josée Dupuis[1], Karim Oualkacha[6]
[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada; [2]McGill Group for Suicide Studies, Douglas Mental Health University Institute; [3]Department of Psychiatry, McGill University, Montreal, Canada; [4]Research Unit on Children's Psychosocial Maladjustment, University of Montreal, Montreal, Canada; [5]Department of Psychology, Laval University, Quebec, Canada; [6]Département de Mathématiques, Université du Québec à Montréal, Montreal, Canada
*corresponding author

This work is motivated by analyses of longitudinal data collected from participants in the Quebec Longitudinal Study of Child Development (QLSCD) and the Quebec Newborn Twin Study (QNTS) to identify important genetic predictors for emotional and behavioral difficulties in childhood and adolescence. We propose a lasso penalized mixed model for continuous and binary longitudinal traits that allows the inclusion of multiple random effects to account for random individual effects not attributable to the genetic similarity

between individuals. Through simulation studies, we show that replacing the estimated genetic relatedness matrix (GRM) by a sparse matrix introduces bias in the variance components estimates, but that the obtained computational gain is major while the impact on the performance of the penalized model to retrieve important predictors is negligible. We compare the performance of the proposed penalized mixed model to a standard lasso and to a univariate mixed model association test and show that the proposed model always identifies causal predictors with greater precision. Finally, we show an application of the proposed methodology to predict three externalizing behavorial scores in the combined QLSCD and QNTS longitudinal cohorts.

62

Identifying Shared Genetic Associations in Fibrosis: A Multi-Organ Rare Variant Analysis

Dominic P. Sayers*[1,2], Ebrima Joof[1,2], Nick R. J. Shrine[1,2], Georgie M. Massen[3], Jennifer K. Quint[3], Hilary Longhurst[5], R. Gisli Jenkins[4], Louise V. Wain[1,2], Katherine A. Fawcett[1,2], Richard J. Allen[1,2], DEMISTIFI Consortium

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, United Kingdom; [3]School of Public Health, Imperial College London, London, United Kingdom; [4]National Heart & Lung Institute, Imperial College London, London, United Kingdom; [5]DC Action, London, United Kingdom

Introduction: 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.

Methods: A Delphi survey classified fibrotic diseases across 12 organ-systems: 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 per case matched by age and sex. Rare variant genome-wide gene-based analyses, incorporating age and sex as covariates, 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. Significant genes were defined as those reaching Bonferroni correction. Significant genes were investigated in the other organ groups with suggestive significance defined as those reaching a Bonferroni corrected threshold for the number of genes followed-up.

Results: In total, 89 unique significant genes were identified across all phenotypes. *PKD1* was significantly associated with liver and renal fibrosis and suggestively significant with diabetes. *PKD2* and *OR6C70* were suggestively significant across liver and renal.

Discussion: The identification of *PKD1, PKD2,* and *OR6C70* in multiple fibrotic phenotypes suggests potential shared genetic pathways with fibrosis across different organ systems. These genes may represent promising targets for future therapeutic interventions in fibrotic diseases. Replication analyses are currently being performed using the All of Us biobank.

Keywords: Fibrosis, multi-organ, rare variant

63

Back-in-Time Reconstruction of Population Structure Using Reconstructed Haplotypes

Jin Du[1], Stefan Böhringer*[1]

[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The 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.

64

Single Variant and Gene-Based Collapsing Association of Rare Variants With Lung Function to Refine Mapping of Causal Genes and Biological Pathways

Nick Shrine[1,3], Kayesha Coley[1,3], Alex Williams[1,3], Anna L. Guyatt[1,3], Samuel T. Moss[1,3], SpiroMeta Consortium, TOPMed Consortium, Louise V. Wain[1,3], Ian P. Hall[2] and Martin D. Tobin[1,3]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Division of Respiratory Medicine and NIHR Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom; [3]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, University of Leicester, Leicester, United Kingdom

Genome-wide association studies (GWAS) of quantitative lung function have increased power to detect risk loci for COPD over case/control studies and have implicated over 500 putative causal genes. Whole-exome/genome

sequencing (WES/WGS) is better suited to study rarer protein-coding variants that may not be well measured in GWAS.

We meta-analysed single rare variant (MAF <1%) associations with lung function traits forced expiratory volume in 1 second (FEV_1), forced vital capacity (FVC), FEV_1/FVC and peak expiratory flow (PEF) from 13 multi-ancestry cohorts (11 WES, 2 WGS, total N=433,869). We also ran gene-based collapsing analysis in UK Biobank European WES data (N= 356,860).

Focussing on exonic regions covered in both WES and WGS data we report 15 significant ($P<5x10^{-9}$) single variant signals; 10 were independent of previous GWAS signals. From collapsing analyses, we report 16 significant genes ($P<2.7x10^{-6}$; Bonferroni correction for 18,600 genes). In total we report associations at 25 genes, 7 not previously implicated for lung function. Using 7 variant-to-gene (V2G) criteria (nearest gene, eQTL, pQTL, rare disease, mouse knockout, polygenic priority score, functional deleterious) previously prioritised 596 genes with ≥2 lines of V2G evidence. Adding rare variant associations passing thresholds of $P<5x10^{-6}$ and $P<10^{-4}$ for single variant and collapsing signals respectively, increased this to 693 genes. The number of biological pathways enriched for lung function signals at FDR<5% increases from 299 to 361 by incorporating rare variant evidence.

In summary, rare variant associations elucidate potential mechanisms of genetic effects on lung function which could help target therapeutic intervention for COPD.

65

Multi-Ancestry Polygenic Risk Scores for Chronic Obstructive Pulmonary Disease Improve Transferability Across Diverse Populations

Matthew Moll[1,†], Jing Chen[*2,†,] Cynthia Zhang[1], Emmanuel Adonyo[2], Anna Guyatt[2], Nick Shrine[2], Lisa Micklesfield[3], Michele Ramsay[4], H3Africa AWI-Gen Study, Coronary Artery Risk Development in Young Adults (CARDIA) Study, Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) Study, SPIROMICS study, GenKOLS, KARE, MGB Biobank, Martin D. Tobin[2,‡], Michael Cho[1,‡]

[1]Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; Harvard Medical School, Boston, Massachusetts, United States of America; [2]Genetic Epidemiology Group, Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [3]SAMRC/Wits Developmental Pathways for Health Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [4]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[†] Contributed equally, [‡] Contributed equally, * Presenting author

Background: Polygenic risk scores (PRS) for chronic obstructive pulmonary disease (COPD) developed in European populations have demonstrated limited transferability to individuals of non-European ancestry. We hypothesized that multi-ancestry PRS, tailored to target populations, would enhance the cross-ancestry performance and improve risk prediction across diverse populations.

Methods: We developed a multi-ancestry PRS using the PRS-CSx method, leveraging recent advances in genome-wide association studies (GWAS) of lung function traits. Our training data included 588,452 participants from diverse ancestral backgrounds. To optimize predictive performance, we tuned the PRS using the COPDGene non-Hispanic white cohort for European ancestry and the AWI-Gen cohort for African ancestry. We integrated PRSs for forced expiratory volume in 1 second (FEV1) and the FEV1/forced vital capacity (FEV1/FVC) ratio to improve COPD prediction, and evaluated the scores in independent, ancestrally diverse populations from both research cohorts and biobanks.

Results: We observed improved transferability of COPD risk prediction across diverse populations when using multi-ancestry PRS compared to European-specific PRS, with the overall OR per SD change of PRS rising from 1.38 [95% CI 1.34 – 1.42] to 1.42 [95% CI 1.38 – 1.46]. Notably, the gap in risk score performance between individuals of European and African ancestry was reduced, particularly within the research cohorts. Additionally, the multi-ancestry PRS showed associations with CT imaging phenotypes and quantitative measures of emphysema.

Conclusion: A PRS derived from multi-ancestry GWAS more accurately predicts COPD across ancestries compared to a European-specific PRS. This finding highlights the importance of ancestry-relevant scores for personalised medicine and in reducing health disparities.

66

Mind the Colocalisation Gap: Characterising GWAS Signals in Immune-Mediated Diseases Using Immune Cell eQTL Data

Guillermo Reales[1,2], Jeffrey M Pullin[3], Ichcha Manipur[1,2], Elena Vigorito[3], Chris Wallace[1,2,3]

[1]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Cambridge, United Kingdom; [2]Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, United Kingdom; [3]MRC Biostatistics Unit, University of Cambridge, School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, United Kingdom

Genome-wide association studies (GWAS) have identified genetic risk variants for thousands of complex traits. The challenge is to match those risk variants to their regulated genes. The expansion of expression quantitative trait loci (eQTL) studies in cells with increasing granularity offers an opportunity to identify the causal gene and context of GWAS variants. Here, we used colocalisation across 14 immune-mediated disease (IMD) GWAS and 101 blood cell eQTL datasets to quantify the proportion of GWAS loci explained by an eQTL, and to identify the most informative eQTL data types to inform future studies. We found that on average, 30% of GWAS loci colocalised with eQTLs, with 46.9% of loci colocalising with multiple genes. The nearest gene to a GWAS peak is usually considered the most likely causal gene, supported by our results which linked ~66.6% of

significant colocalisations to the closest gene. We show that single-cell studies and samples including patients have greater potential to annotate GWAS signals, but face challenges in achieving larger sample sizes. Surveying the disease-appropriate cell types also matters, with T1D and asthma having proportionally more CD4+ T cells colocalisations and IBD and its subtypes having more monocyte colocalisations. Altogether, we show that circulating immune cell eQTLs explain a substantial proportion of IMD GWAS loci and that specific immune cells are more informative for certain IMDs, highlighting the importance of using adequate cell granularity in future studies.

67

Roadmap to a Successful Rare Variant Association Study – A Topic Review

Vivian Link*[1], Amra Dhabalia Ashok*[1], Andreas Ziegler[1,2,3,4]
[1]Cardio-CARE, Davos, Switzerland; [2]University Heart and Vascular Center Hamburg, University Medical Center Hamburg–Eppendorf, Hamburg, Germany; [3]German Center for Cardiovascular Research (DZHK), Partner Site Hamburg–Kiel–Luebeck, Hamburg, Germany; [4]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa
*These authors contributed equally

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.

68

Transcriptome-Wide Association Analysis of Age-Related Macular Degeneration Across Two Retinal Layers

Inti A. Pagnuco*[1], Jacob Sampson[2], Jamie Ellingford[2], Andrew P. Morris[1]
[1]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Manchester, United Kingdom; [2]Division of Evolution, Infection and Genomics, The University of Manchester, Manchester, United Kingdom

Age-related macular degeneration (AMD) is a leading cause of vision loss in adults, with heritability estimated at up to 71%. Genome-wide association studies (GWAS) have identified numerous risk loci, yet the causal genes and tissues of action remain incompletely understood.

To refine GWAS findings and uncover gene-level associations, we conducted tissue-specific transcriptome-wide association studies (TWAS) in two functionally distinct retinal layers: the neurosensory retina (NSR) and the retinal pigment epithelium (RPE). Using genotype and normalized transcriptomic data from 183 NSR and 176 RPE samples from the Manchester Eye Tissue Repository, we built predictive expression models for 2,263 genes in NSR and 2,116 in RPE.

We applied S-PrediXcan to test these models against a large-scale AMD meta-GWAS (57,290 cases, 324,430 controls), followed by replication in FinnGen (12,495 cases, 461,686 controls). In NSR, we identified four significantly associated genes ($p < 1\times10^{-4}$), (*HLA-DQB2, CFH, MICA, NCOA5*), while six were implicated in RPE (including *NDUFA11, PET100, ABHD16A*). Replication confirmed *HLA-DQB2* and *MICA* (NSR), and *ABHD16A* (RPE), with consistent effect directions across datasets, underscoring the robustness of these associations.

Our tissue-specific analysis showed strong overlap between TWAS-prioritized genes and known AMD loci, including *CFH* and *HLA-DQB2* in NSR, and *NDUFA11* and *ABHD16A* in RPE. *PET100* and *FTL*, while not previously implicated in AMD, are linked to other ocular disorders and represent novel candidate genes.

These findings demonstrate the utility of TWAS for refining GWAS signals, identifying putative causal genes, and revealing potentially tissue-specific mechanisms that may contribute to AMD risk.

69

Accounting for Heterogeneity Due to Ancestry and Environment Improves the Resolution of Multi-Ancestry Fine-Mapping

Siru Wang*[1], Oyesola O. Ojewunmi[2], Fraser J. Pirie[3], Ayesha A. Motala[3], Michele Ramsay[4], Andrew P. Morris[5], Segun Fatumo[2,6], Tinashe Chikowore[7,8,9], Jennifer L. Asimit[1]
[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [2]Precision Healthcare University Research Institute, Queen Mary University of London, London, United Kingdom; [3]Department of Diabetes and Endocrinology, School of Clinical Medicine, University of KwaZulu-Natal, Durban, South Africa; [4]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the

Witwatersrand, Johannesburg, South Africa; [5]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom; [6]The African Computational Genomic (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda; [7] MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [8]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [9]Harvard Medical School, Boston, Massachusetts, United States of America

Amongst diverse population groups it is likely for there to be heterogeneity in effect sizes due to ancestry, as well as environmental exposures. This allelic heterogeneity impacts the power to detect genetic associations, and in turn, refinement of sets of potential causal variants underlying genetic associations, through statistical fine-mapping. MR-MEGA (Meta-regression of multi-ethnic genetic association) adjusts for and assesses heterogeneity due to genetic ancestry and the recently developed environment-adjusted MR-MEGA accounts for environmental exposures alongside genetic ancestry. In this work, we developed novel multi-ancestry fine-mapping methods, (env-)MR-MEGAfm, which allows for multiple causal variants in a genomic region. Employing a stepwise selection procedure, (env-)MR-MEGAfm integrates approximate conditional analyses with (env-)MR-MEGA to construct credible sets of potential causal variants. Both methods use genome-wide association study (GWAS) summary statistics and account for differing linkage disequilibrium (LD) from multiple cohorts, and env-MR-MEGAfm also accounts for summary-level environmental covariates. Additionally, (env-)MR-MEGAfm does not require that variants appear in all cohorts. Through extensive simulation studies, we showed that (env-)MR-MEGAfm had significant improvements in coverage, resolution and prioritization over current multi-ancestry approaches. In addition, env-MR-MEGAfm had improved resolution over MR-MEGAfm, producing significantly smaller credible sets (p-value < $3.29\times10^{\wedge-14}$). In summary, (env-)MR-MEGAfm accounts for cohort-level differences in genetic ancestry and environmental factors (for env-MR-MEGAfm) and allow for variants to be present in only a subset of cohorts. Finally, these methods only require summary-level data and allow for any number of cohorts, making them useful tools in consortia efforts.

Keywords: Fine-mapping; environment; heterogeneity; multi-ancestry; genome-wide association studies

70

Association Between Pace of Aging Estimated Using Blood DNA Methylation and All-Cause Mortality: The HUNT Study

Yi-Qian Sun[1,2,3], Ilona Urbarova[4], Lin Jiang[5,6], Therese Haugdahl Nøst[4,7,8,9], Xiao-Mei Mai[5]

[1]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway; [2]Department of Pathology, Clinic of Laboratory Medicine, St. Olavs Hospital, Trondheim, Norway; [3]Center for Oral Health Services and Research Mid-Norway (TkMidt), Trondheim, Norway; [4]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway, [5]Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway, [6]Clinic of Cardiology, St. Olav's Hospital, Trondheim, Norway; [7]HUNT Research Centre, Norwegian University of Science and Technology, Levanger, Norway; [8]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway; [9]HUNT Centre for Molecular and Clinical Epidemiology, Norwegian University of Science and Technology, Trondheim, Norway

Aim: We aimed to investigate the relationship between pace of aging, estimated using blood DNA methylation, and all-cause mortality in a population-based Norwegian cohort. Methods: This study included 140 cancer-free controls from a lung cancer nested case-control study within the Trøndelag Health Study (HUNT). DNA methylation was measured in blood samples for the study participants in both HUNT2 and HUNT3, 11 years apart. The pace of aging was estimated using four established measures of biological age based on blood DNA methylation developed by Levine 2018, Lu 2022, Belsky 2020, and Belsky 2022.

Results: Pearson correlation coefficients between the four measures of the pace of aging ranged from 0.31 to 0.73. There was a moderate to excellent reliability of the repeated measurements, with intraclass correlation coefficient (ICC) values ranged from 0.69 to 0.91. University education appeared to be associated with a slow pace of aging, while smoking was associated with a fast pace. A one standard deviation increase in the pace of aging was associated with a 45% to 242% higher likelihood of all-cause mortality across the four measures in HUNT2, and a 17% to 227% higher likelihood in HUNT3. Accelerated pace of aging from HUNT2 to HUNT3 was significantly associated with higher all-cause mortality for three of the four measures.

Conclusion: The pace of aging, as estimated using blood DNA methylation, is a robust and independent predictor of all-cause mortality. This measure may reflect the combined influences of genetic, lifestyle, and environmental factors on individual aging trajectories.

Keywords: Biological age, DNA methylation, epigenetic clocks, HUNT, mortality, pace of aging

71

The Effect of Genetic Profiles on Physical Activity and Sedentary Behavior in Children - The GECKO Drenthe Cohort

Yeliz Eski[1], Lu Yang[2], Harold Snieder[1], Eva Corpeleijn[1], and Ilja M. Nolte[1]

[1]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2]Department of Human Movement Sciences, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

Purpose: Increasing physical activity (PA) and reducing sedentary behaviour (SB) benefit childhood well-being and long-term health. This study aims to investigate how much variance in PA and SB can already be explained by polygenic risk scores (PRS) in early childhood.

Methods: We analysed data from 536 children (52.6% male, mean age [SD]=5.65 [0.8]) in the GECKO Drenthe

birth cohort. Device-based PA and SB traits were measured using hip-worn accelerometers (ActiGraph GT3X; ≥600 min/day; ≥3 days). Ten PRSs for various activity outcomes were computed using SBayesRC, a method that incorporates functional annotations to enhance prediction accuracy. Linear regression analyses with and without PRS were performed adjusted for relevant covariates to assess the contributions of PRSs to PA and SB outcomes (ΔR^2).

Results: Several activity PRSs showed significant associations with device-measured PA and SB outcomes in expected directions. The strongest effects were observed for PRSs related to moderate-to-vigorous PA (MVPA) and metabolic equivalent of task values, both explaining 2.1% of the variance in vigorous PA ($p<0.001$). Children in the highest PRS quartile engaged in 10.5 more minutes of MVPA per day compared to those in the lowest quartile.

Conclusion: Activity PRSs derived from adult data were associated with objectively measured PA in early childhood. These findings suggest that genetic predispositions for PA are already observable at a young age. PRSs may serve as early indicators to identify children at risk for low PA, enabling timely and targeted interventions to promote lifelong health and prevent inactivity-related conditions.

72

Evaluating Transferable Polygenic Risk Scores for Internalising and Cardiometabolic Multimorbidity

Daniel Stow*[1], Ruby Tsang[2], Ioanna Katzourou[3], The Lifespan Multimorbidity Research Collaborative (LINC), Peter Holmans[3], Inês Barroso[5], Hilary Martin[4], Marianne B.M. van den Bree[3,6], Sarah Finer[1], Nic Timpson[2,7]

[1]Wolfson Institute of Population Health, Queen Mary University of London, United Kingdom; [2]MRC Integrative Epidemiology Unit, University of Bristol, United Kingdom; [3]Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, United Kingdom; [4]Wellcome Sanger Institute, Cambridge, United Kingdom; [5]Medical School, University of Exeter, United Kingdom; [6]Mental Health Innovation Institute, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, United Kingdom; [7]Population Health Sciences, Bristol Medical School, University of Bristol

Internalising and cardiometabolic multimorbidity (ICM-MM) is a common type of physical and mental health multimorbidity in later-life. Our understanding of early ICM-MM manifestations, and development of timely prevention strategies would benefit from identification of young people at highest risk. We aimed to evaluate polygenic risk scores (PRSs) to predict ICM-MM and create a transferable score to identify precursor risk-stages of ICM-MM in studies across the lifespan.

We included 45,492 UK BioBank participants with primary care data and Genomics PLC PRSs (PRS_{GPLC}), of whom 8,758 (19.3%) had ICM-MM (lifetime occurrence of: >one internalising condition: depression, anxiety, somatoform disorder; AND >one cardiometabolic risk-factor condition: type 2 diabetes, obesity, hypertension, dyslipidemia, chronic kidney disease). We used PRS_{GPLC} covering 51 traits and generated PRS_{trait} for seven ICM-MM traits using PRS-CS and summary statistics from GWAS excluding UKB. We modelled associations between PRSs and ICM-MM using receiver operating characteristic analysis adjusted for sex and ten genetic principal components, and generated two novel PRSs using PRSMix (Elastic Net) on PRS_{trait} (ICM-MM-PRS_{trait}) and PRS_{GPLC} (ICM-MM-PRS_{GPLC}) separately.

Hypertension and ischemic stroke $PRSs_{GPLC}$ best predicted ICM-MM (both area under the curve (AUC) =0.56, 95%CI:0.55-0.56). ICM-MM-PRS_{trait} retained five PRSs and had similar predictive performance to single PRSs (AUC=0.56, 95%CI:0.55-0.57). ICM-MM-PRS_{GPLC} retained 13 PRSs and improved predictive performance (AUC=0.59, 95%CI:0.58-0.60).

Single trait PRSs are associated with ICM-MM, but offer limited formal predictive value, which is slightly improved by ICM-MM-PRSs. In this context, we report on differences in developmental trajectories by ICM-MM-PRS in external birth cohorts, capturing early signals of ICM-MM risk.

74

Genome-Wide Association Study of Eye Protrusion

Alexander T. Williams*[1], Callum Hunt[2], Alvin Lirio[2], Ha-Jun Yoon[2], Martin D. Tobin[1,3], Mervyn G. Thomas[2], Catherine John[1,3]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Ulverscroft Eye Unit, College of Life Sciences, School of Psychology and Vision Sciences, University of Leicester, Leicester, United Kingdom; [3]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, University of Leicester, Leicester, United Kingdom

Motivation: Protrusion of the eyes (proptosis or exophthalmos) is seen in thyroid eye disease (TED), an autoimmune condition that causes inflammation and remodelling of tissues around the eye, including extraocular muscles and orbital fat. Graves' disease and other autoimmune thyroid disease commonly cause TED. However, the genetic determinants remain poorly understood.

Methods: We derived an imaging-based quantitative trait for eye protrusion using axial T1-weighted MRI scans from UK Biobank. The maximum distance between the lateral orbital rim (approximated by the zygomatic bone) and the posterior surface of the globe was computed for each individual. We performed genome-wide association studies (GWAS) of this trait across four ancestry groups in UK Biobank. We meta-analysed ancestry-specific GWAS results using METAL. We defined primary sentinel variants by *P* value $<5\times10^{-8}$ in the overall meta-analysis and then used GCTA-COJO to identify conditionally independent sentinel variants.

Results: We analysed 79,528,398 genetic variants across 55,227 individuals (52,916 (95.8%) European). We report 48 significant sentinel variants, many in or near to genes implicated in musculoskeletal abnormalities. We highlight associations in the *RARB* (retinoic acid receptor beta) gene which has been implicated in microphthalmos, a developmental disorder resulting in abnormally small eyes. Despite small numbers of non-European participants, as many as six of our sentinel variants were at least nominally significant (*P* value <0.05) in other ancestry groups in UK Biobank.

Conclusion: We conducted the first GWAS of an imaging-derived phenotype for eye protrusion. Extensive variant-to-gene mapping is underway to identify putative causal genes..

Keywords: genome-wide association study, thyroid, common disease genetics

75

Genetic Prediction of Blood Glucose Variability in Preterm Infants: A Polygenic Score Approach

Lisa-Marie Nuxoll*[1], Franziska Hanßmann[2], Wolfgang Göpel[2], Inke R. König[1]

[1]Institute of Medical Biometry and Statistics, University of Lübeck, University Hospital Schleswig-Holstein, Campus Lübeck, Lübeck, Germany, [2]Department of Paediatrics and Adolescent Medicine, University of Lübeck, University Hospital Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Maintaining stable blood glucose levels in preterm infants is a clinical challenge, with implications for the further development of the child. While genetic factors influencing glucose metabolism are increasingly understood in adults, their relevance in neonates, especially in preterm infants, remains largely unexplored. We investigate the applicability of polygenic scores (PGS) for blood glucose phenotypes in predicting blood glucose variability during the early life of preterm infants.

We utilize data from the German Neonatal Network (GNN). This is a German cohort study of preterm infants, in which longitudinal clinical data, specifically blood glucose measurements recorded over the first 70 days of life, were combined with genetic information. Established polygenic scores, developed mostly in adult populations, are computed for each infant to assess their association with blood glucose measures. Our analysis focuses on the temporal dynamics of blood glucose measurements and explores the predictive value of genetic predisposition in this critical developmental phase.

This presentation aims to provide insights into the genetic contribution to glucose regulation in preterm infants and inform future development of age-specific genomic tools for neonatal care.

Keywords: Polygenic Scores, Preterm infants, Blood glucose

76

Using Mendelian Randomization to Identify Mass Spectrometry Quantified Proteins Causally Associated with Lung Function

Brandon E. M. Lim*[1], Nick Shrine[1], Jing Chen[1], Colleen Maxwell[2], Don Jones[2], Catherine John[1], Chiara Batini[1], Martin D. Tobin[1], Anna L. Guyatt[1]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]The Leicester van Geest MultiOmics Facility, Hodgkin Building, University of Leicester, Leicester, United Kingdom

Introduction: Chronic obstructive pulmonary disease (COPD) is a leading cause of death worldwide, characterised by poor lung function. Identifying proteins causally associated with lung function may identify new drug targets. We used cis protein quantitative trait loci (pQTLs) in a two sample Mendelian randomization (2SMR) to find proteins causally associated with lung function.

Methods: The UK-based EXCEED study contains 1,471 individuals with genetic data and mass spectrometry (MS) measured protein quantities, including proteins not available in UK Biobank.

pQTLs were identified in individual GWAS of 319 protein levels, using REGENIE v3.4.1. Proteins were residualised on age, sex and MS batch, before rank inverse normal transformation. Covariates were genotyping batch and the first 10 ancestry principal components.

Sentinel SNPs within 1Mb of the cognate gene's transcription start site (sample 1) were used for SNP-exposure data. SNP-outcome data (sample 2) were from a meta-analysis of forced expiratory volume in 1 second (FEV_1), forced vital capacity (FVC) and FEV_1/FVC, excluding EXCEED.

2SMR was conducted using the Wald ratio test in the TwoSampleMR package (R v4.2.3). Bonferroni corrected $P<0.05$ associations were considered significant.

Results: We identified 114/319 proteins as having cis pQTLs. Of these, 12 were associated with lung function ($P<4.39\text{x}10^{-4}$, corrected for 114 tests). Notably increased C4A was associated with decreased FEV_1, FVC and FEV_1/FVC.

Discussion: We identified 12 proteins associated with lung function. Ongoing sensitivity analyses and phenome-wide association studies will clarify their potential as therapeutic targets for COPD.

77

Revisiting the Association Between BMI and Depression Using Phenome-Wide Association Clustering of MR Instruments (PWC-MR)

Stephanie Sheir[1], Neil Davies[2], Jean-Baptiste Pingault[1]

[1]Department of Clinical Educational and Health Psychology, University College London (UCL), London, United Kingdom; [2]Division of Psychiatry, Department of Statistical Science, University College London (UCL), London, United Kingdom

Higher body mass index (BMI) is associated with depression at both phenotypic and genotypic levels, yet the extent of causality and potential underlying mechanisms remain unclear. We revisit this relationship using Phenome-Wide Clustering of Mendelian Randomization instruments (PWC-MR), an approach developed in Darrous et al. 2024 that accounts for heterogeneity in both the exposure and outcome.

We conducted a PheWAS on 324 BMI-associated SNPs across 407 UK Biobank traits, identifying six genetic clusters. Two clusters—one characterised by adiposity-related traits (e.g. fat mass) and another by socioeconomic traits (e.g. occupation)—were the focus of our analysis. We then estimated the causal effects of each cluster on nine depressive symptoms measured in the UK Biobank.

BMI exhibited differential effects across symptoms and clusters not captured in conventional MR, which estimates the effect of single exposure on a single outcome. The most consistent effects across clusters were on somatic symptoms such as appetite changes and tiredness. For the adiposity cluster, appetite changes showed the strongest association ($\beta = 0.26$, 95% CI [0.22, 0.31]), with moderate effects on tiredness ($\beta = 0.15$) and anhedonia ($\beta = 0.10$). Weaker or non-

significant associations were observed for affective symptoms, such as mood and suicidal ideation.

These differential effects across symptoms highlight differences in how BMI relates to non-specific somatic symptoms and psychological symptoms thought to be core to depression. Ultimately, we demonstrate symptom-specific heterogeneity in the psychiatric outcomes of BMI and the utility of clustering MR instruments for polygenic traits.

Keywords: Mendelian randomization, body mass index, depression, symptom-level, psychiatric genetics

78

Genetic Correlations Between Asthma Subtypes and Neuropsychiatric Disorders

Haibo Huang*[1], Raphaël Vernet[1], Lucie Troubat[1], Florence Demenais[1], Christophe Linhard[1], Yuka Suzuki[2], Hanna Julienne[2], Emmanuelle Bouzigon[1]

[1]Université Paris Cité, Inserm, HealthFex, group of Genomic Epidemiology of Multifactorial diSeases, Paris, France; [2]Institut Pasteur, Université Paris Cité, Department of Computational Biology, 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 (rg) 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 rg (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, rg=0.26-0.37), post traumatic stress (rg=0.28-0.40), bipolar (BIP, rg=0.10), and attention deficit hyperactivity disorder (rg=0.26-0.36). No significant rg 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 rg, 49 trait pairs had non significant global rg. Notably, asthma and schizophrenia pair exhibited 11 significant local correlations (global rg=0.06), and childhood asthma and BIP pair exhibited 4 significant local correlations (global rg=0.07). Across these trait pairs, the proportion of positive and negative local correlations was balanced, which may explain the absence of significant global rg.

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.

Funding: ANR-20-CE36-0009, CSC scholarship

Keywords: Asthma, Neuropsychiatric Disorder, Genetic Correlation, Heterogeneity

79

Fine-Scale Pharmacogenetic Diversity in Europe: The Example of France

Marc Gros-La-Faige*[1], The POPGEN Study Group[1], Emmanuelle Génin[1,2], Anthony F. Herzig[1]

[1]Inserm, University of Brest, Brest, France; [2]Centre Hospitalier Régional Universitaire, Brest, Brest, France

*Corresponding Author

Pharmacogenetics is the study of genetic variants responsible for variable response to medication. These variants can explain alternate drug responses and understanding their effects thus represents a key public health issue. Previous studies have shown that some of these variants have frequencies that are stratified across human populations but little is known about their distribution at fine geographic scale within a country such as France.

To study the diversity of pharmacogenes of interest in different regions of France, we used SNP-chip genotyping data on 9598 French individuals and associated spatial co-ordinates derived from the birthplaces of their ancestors collected as part of the POPGEN project.

We derived different statistics commonly used in population genetics to identify pharmacogenetic variants with a heterogeneous frequency distribution and detected variant stratifications, such as gradients from north to south or east to west. We also found clusters of variants within specific sub-populations. We studied how these patterns could be explained by selective constraints by comparing their gene constraint metrics against those of other genes with similar sizes and we observed that certain pharmacogenes are significantly less constrained, which may explain their observed high levels of genotypes and phenotypes diversity. Overall, we identified some important pharmacogenes, like CYP2D6 or ABCG2, with fine-scale geographic specificities that have phenotype consequences for drug with prescribing recommendations.

Exploring genetic diversity in pharmacogenes at finer geographic scales than previously done will improve our understanding of drug-gene interactions, while also informing potential benefits of personalized treatment based on pharmacogenetic variant data.

Keywords: Pharmacogenetics, Population genetics, bioinformatics

80

Knowledge Graph to Dissect Genotype-Phenotype Associations

Florence Ghestem*[1], Gaëlle Marenne [2], Emmanuelle Génin[2], Anne-Sophie Jannot [3,4], Anaïs Baudot [5,6], Anne-Louise Leutenegger [1,7]

[1]Université Paris-Saclay, UVSQ, Inserm, Villejuif, France; [2]University of Brest, Inserm, Brest, France; [3]BNDMR–AP–HP–Campus Picpus Département I&D, Paris, France; [4]Université Paris Cité, HeKA INSERM, INRIA Paris, centre de recherche des cordeliers, Paris, France; [5]Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, CNRS, Turing Center for Living Systems, Marseille, France; [6]Barcelona Supercomputing Center (BSC), Barcelona, Spain; 7 Université Paris Cité, Inserm, Paris, France

Defining the mechanisms underlying complex traits and

phenotypes requires understanding their genetic basis and how genes interact with environmental and lifestyle factors. Population-based prospective cohorts provide valuable resources for such research, collecting extensive phenotypic and omics data. However, the wide range of phenotypes in these cohorts makes it difficult to define homogeneous and/or clinically meaningful subgroups, limiting traditional genome-wide association studies (GWAS).

We aim to develop a novel graph-based methodology to identify genotype phenotype associations. Our method represents data as a graph, where nodes correspond to variables (for example, participants, phenotypes), and edges represent relationships between them. These edges can connect the same type of nodes (for example, participant to participant interactions) or different types (for example, bipartite interactions connecting participant to SNP).

We applied our approach to the GOLD project, which includes comprehensive information on medical conditions, drug consumption, demographics, and genotypes of 10,000 participants. The graph contains four node types (participant, drug, SNP, phenotype) and 10 edge types, with attributes such as drug reimbursements, genotypes, and participant similarity.

This knowledge graph representation allows us to use graph theory tools, including clustering algorithms, random walks, and deep learning-based graph representation methods. We expect to detect weak genotype phenotype association signals that could not be detected by GWAS.

Acknowledgment: Inserm cross-cutting program GOLD.

81

Assessing the Role of Insomnia in Breast Cancer Risk Across Multiple Ancestries Within the "All of Us" Research Program

Bryony L. Hayes*[1,2], Alok Singh[3], Marina Vabistsevits[3], Katherine S. Ruth[3], Anna Murray[3], Michael N. Weedon[3], Rebecca C. Richmond[1,2,4]

[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [2]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [3]Department of Clinical and Biomedical Science, University of Exeter Faculty of Health and Life Sciences, Exeter, United Kingdom; [4]National Institute for Health and Care Research, Oxford Health Biomedical Research Centre, University of Oxford, Oxford, United Kingdom

Poor sleep may be a risk factor for breast cancer, but evidence is inconsistent, and few studies have explored this relationship in non-European populations. The "All of Us" Research Program is a diverse longitudinal US cohort study, with genomic data linked to electronic health records. In this study, we aimed to use both clinical diagnosis (N=30,380) and medication records (N=35,840) to explore the relationship between clinical insomnia and breast cancer risk (Ncases=10,118) across multiple ancestral populations within a single cohort.

In observational analysis, an inverse relationship was found between clinical insomnia and breast cancer for European (OR=0.43, 95%CI=0.39,0.47), African (OR=0.58, 95%CI=0.47,0.71), Hispanic (OR=0.63, 95%CI=0.48,0.81), and Asian individuals (OR=0.67, 95%CI=0.35,1.27). These results were consistent when adjusted for age, body mass index, income, education, menopause, smoking, and alcohol intake. Conversely, a positive association was observed across all groups when medication use was used to define insomnia.

We conducted genome-wide association studies in each population, and found novel variants associated with insomnia. Notably, a genome-wide significant variant within *WSB1 (17:26866905;CAAATG)* was identified in White (*P* value=$3.98x10^{-43}$ and $5.05x10^{-41}$) and African (*P* value=$3.93x10^{-74}$ and $2.92x10^{-81}$) populations using both insomnia definitions, and Hispanic populations (*P* value=$2.17x10^{-11}$) only when medication use was included, suggesting evidence for conservation. *WSB1* is a regulator of metastatic disease in hormone receptor-negative breast cancer and may serve as a therapeutic target. Further investigation and pathway analysis is ongoing to better understand the role of this novel variant in insomnia and breast cancer, and to explore the causality of the associations observed.

Keywords: Insomnia, Breast Cancer, Ancestry, Epidemiology

82

Mapping French Genetic Diversity: From Regions to Europe for Genomic Medicine

Anthony Herzig[1], Gaëlle Le Folgoc[1], Hélène Blanché[2], Gaëlle Marenne[1], Aude Saint Pierre[1], Marie Zins[3], Frédérique Nowak4, Christian Dina[5], Richard Redon[5], Jean-François Deleuze[2,6], Emmanuelle Génin[1], POPGEN Study group and FranceGenRef Consortium

[1]University of Brest, Inserm, Brest, France; [2]Fondation Jean Dausset—CEPH, Paris, France; [3]Inserm-Paris Saclay University, University of Paris, Villejuif, France; [4]Institut Thématique Technologies pour la Santé, Inserm, France [5]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, Nantes, France [6]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), Evry, France

With the decreasing cost of whole-genome sequencing (WGS), several countries have begun developing national reference panels reflecting the genetic diversity of their local populations. In Europe, such efforts align with the Genome of Europe (GoE) project, which aims to provide access to >100,000 WGS representative of European populations. Prior to the GoE launch, two pilot projects were conducted in France to explore fine-scale genetic diversity within and between regions. Both projects prioritized individuals with all four grandparents born within a geographic area of less than 100 km.

The first pilot, FranceGenRef, involved WGS of 856 individuals sampled from existing biobanks, focusing on limited French regions. To achieve broader geographic coverage, the second pilot, POPGEN, leveraged the Constances cohort. A total of 15,000 volunteers received salivary DNA extraction kits, with 10,250 successfully included and 9,772 genotyped using Illumina GSA. Based on genotyping data and ascendants' birthplaces, 4,000 individuals were selected for WGS. These pilot studies,

along with new samples including minority populations via the France Genomic Medicine Initiative, will constitute the 17,000 WGS contribution to GoE.

Here, we evaluate the effectiveness of this sampling design in capturing local genetic variation compared to publicly available databases. Furthermore, with support of exploratory analyses of haplotypes sharing, we discuss how such sampling approaches could aid in distinguishing local neutral variants from pathogenic variants, improving the diagnosis of rare diseases. These findings could provide valuable insights for the GoE project, refining sampling strategies and serving as a model for implementing similar pilots across Europe.

83

Evaluation of Star Allele Annotation Tools and the Influence of Imputation and Population Origin

Marc Gros-La-Faige*[1], FranceGenRef Consortium[2], Emmanuelle Génin[1,3], Anthony F. Herzig[1]

[1]Inserm, University of Brest, Brest, France; [2]LABEX GENMED, Centre National de Recherche en Génomique Humaine, Evry, Paris; [3]Centre hospitalier universitaire de Brest (CHRU), Brest, France

*Corresponding Author

Pharmacogenetics is the study of genetic variants responsible for variable response to medication. These variants can explain adverse drug reactions or lack of drug response and understanding their effects thus represents a key public health issue. Knowledge of these variants allows for adaptive patient care.

To standardize pharmacogenetics studies, a specific nomenclature has been put in place: the star-alleles, which associates haplotypes with the activity of a given Pharmacogene. As haplotype calling is influenced by population's origin, genomic region's variability and dataset used, many star-allele callers have been developed with little hindsight on their performance.

The objective is to determine the performances of the different tools available. To do this, we used sequencing and genotyping data from the 1000 Genomes and FranceGenRef projects.

To instruct analysis, we have benchmarked different pipelines of star allele annotation, as they all use different methods and no gold standard has yet emerged. We have observed discrepancies explained by differences in the reference database used, the method employed and the type of input data. We also evaluate the impact of imputation on star-allele calling, and we notice that imputation improves haplotype accuracy, especially when it is performed using a local reference panel. Finally, we discern differences between tool concordance depending on the individual's origin, with the African continent showing the lowest concordance between tools.

This study will enable us to better understand the strengths and weaknesses of the different star-allele callers, and provide insights on necessary improvements for creating a complete and accurate tool.

Keywords: Pharmacogenetics, bioinformatics, Population genetics

84

Association of Polygenic Scores for ACE Inhibitor-Induced Cough Across Cohorts, Ancestries and Phenotypes

Kayesha Coley[1,2], Catherine John[1,2], Jonas Ghouse[3,4], David J. Shepherd[1], Nick Shrine[1], Stavroula Kanoni[5], Emma F. Magavern[5], Louise V. Wain[1,2], Martin D. Tobin[1,2], Chiara Batini[1,2]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]University Hospitals of Leicester NHS Trust, Groby Road, Leicester, United Kingdom; [3]Laboratory for Molecular Cardiology, Department of Cardiology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; [4]Laboratory for Molecular Cardiology, Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark; [5]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom

Chronic cough can be a symptom of common lung conditions, an adverse reaction to ACE inhibitors (ACEis), or be unexplained. Chronic dry cough and ACEi-induced cough share similar clinical manifestations, and we have shown they are genetically correlated.

After defining ACEi-induced cough in UK Biobank, EXCEED and Copenhagen Hospital Biobank, and chronic dry cough in UK Biobank, we performed GWAS of each phenotype including individuals of European ancestries. We calculated polygenic scores (PGS) weighted by variant effect sizes from the ACEi-induced cough GWAS, and tested association with ACEi-induced cough defined in All of Us (European and African ancestries), UK Biobank (African ancestries) and Genes & Health (South Asian ancestries). We also performed a phenome-wide association study of the PGS and estimated genetic correlations between significantly associated clinical traits and each cough trait.

The PGS was significantly associated with ACEi-induced cough in individuals of European ancestries in All of Us (OR=1.27 per SD unit increase in the PGS, $P=7.45\times10^{-23}$), individuals of South Asian ancestries (OR=1.23, $P=4.31\times10^{-8}$) and individuals of African ancestries (OR=1.09, $P=0.045$). The PGS was also associated with increased risk of multi-site chronic pain (MSCP), diabetes and asthma (false discovery rate <1%). ACEi-induced cough was significantly genetically corelated with MSCP, diabetes and asthma, and chronic dry cough also significantly genetically correlated with MSCP and asthma, although with stronger correlations.

We have utilised PGS to assess the transferability findings across cohorts and ancestral groups, and provided insights into the potential consequences modulating pathways involved in chronic cough.

85

Uncovering Genetic Pathways in Type 2 Diabetes Using Genomic Structural Equation Modeling

Merli Koitmäe[1,2], Märt Möls[1], Kristi Läll[2], Krista Fischer[1,2], Reedik Mägi[2]

[1]Institute of Mathematics and Statistics, University of Tartu, Estonia; [2]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

Given the heterogeneous nature of type 2 diabetes (T2D), its pathophysiology, disease trajectory, and treatment responses vary substantially across individuals. Understanding the biological mechanisms leading to a T2D diagnosis is crucial for improving prevention and personalized strategies.

To unravel these complex pathways, we propose leveraging genomic structural equation modeling (genomic SEM) to identify latent genetic factors contributing to disease onset. Genomic SEM integrates genome-wide association study summary statistics from multiple phenotypes to uncover shared and independent genetic architectures of the traits. By applying this approach, we aim to characterize biologically distinct pathways leading to T2D and subsequently construct polygenic risk scores (PRS) for each of the pathways.

Using genomic SEM on the risk factors of T2D, we found three distinct genetic pathways contributing to disease development: glucose regulation, insulin resistance & cardiometabolic and obesity and lifestyle-related pathways. We evaluated the predictive utility of the pathway-specific PRSs in the Estonian Biobank dataset to determine disease progression and comorbidity profiles. Phenome-wide association study on the PRSs uncovered that glucose regulation pathway contributes to far fewer comorbidities compared to obesity and lifestyle-related pathways. Moreover, 75% of individuals with a high overall T2D PRS also had at least one elevated pathway-specific PRS.

In summary, our study aims to uncover distinct genetic pathways in T2D, enabling better risk prediction and treatment stratification. These insights could pave the way for more effective and personalized interventions, highlighting the potential of pathway-specific scores to explain individual variation in genetic risk and provide insight into the underlying biology.

86

TrACES of Time: A Targeted mRNA Sequencing Approach for Estimating Time-of-Day of Bloodstain Deposition in Forensic Casework

Annica Gosch[1], Sebastian Sendel*[2], Amke Caliebe[2], Cornelius Courts[1]

[1]Institute of Legal Medicine, University Hospital Cologne; [2]Institute of Medical Informatics and Statistics, Kiel University and University-Hospital Schleswig-Holstein, Kiel, Germany

In forensic casework, identifying the donor of a biological trace often requires contextual information to reconstruct the sequence of events. One such contextual parameter is the time-of-day of trace deposition (ToD). Candidate ToD mRNA markers exhibiting diurnally rhyhtmic expression patterns have previously been identified through whole transcriptome sequencing.

Here, we developed a statistical model for estimating the time-of-day of bloodstain deposition based on gene expression quantification by targeted sequencing of selected mRNA markers.

We quantified the expression of 69 candidate ToD markers in 408 blood samples collected from 51 individuals at eight time points over a 24-hour period. To estimate ToD by classification and regression, we evaluated statistical approaches based on penalized regression, support vector machines and random forest within a five-fold cross-validation framework. For classification models, ToD was treated as a categorical variable. For regression models, ToD was represented as an angle on a 24-hour clock and transformed into sine and cosine components. Separate regression models were trained for each component, and predicted values were re-transformed to yield a final continuous ToD estimate.

Based on root mean squared error, the best-performing model was penalized regression, which achieved an error of 3 hours and 44 minutes, with 78% of predictions falling within ±4 hours of the actual deposition time. The results revealed substantial inter-individual variation in expression rhythms. While the current accuracy may not yet support use in the evaluative phase of criminal proceedings, the method shows promise for providing valuable temporal intelligence during the investigative phase.

Keywords: statistical modeling, gene expression, forensics

87

Multi-Population GWAS of Waist-To-Hip Ratio Reveals Heterogeneous Pathways to Central Obesity

Anne E. Justice*[1], Emmaleigh Wilson[2], Daeeun Kim[3], Navya Shilpa Josyula[1], Shreyash Gupta[1], Qianqian Liang[1], Virginia Diez-Obrero[4], Eric Bartell[5,6,7], on behalf of the GIANT Waist Traits Working Group

[1]Department of Population Health Sciences, Geisinger, Danville, Pennsylvania, United States of America; [2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [3]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [4]Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark; [5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [6]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, United States of America; [7]Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

To understand the genetic architecture of central obesity and body shape, we conducted an extensive genome-wide association study (GWAS) for waist-to-hip ratio adjusted for BMI (WHR) in ~1.2 million individuals from diverse populations. We employed gene prioritization, gene-set enrichment, colocalization with genetic association with expression in eight tissues, and systematic comparisons of published GWAS to explore underlying biological mechanisms and shared genetic etiologies. Across sex- and population-stratified and pooled meta-analyses, we identified 1009 association signals, 49% showing sex effect heterogeneity. We identified 1360 significant genes across our gene prioritization and colocalization analyses. As in previous GWAS of WHR, we found enrichment for genes in adipose tissues; however, we also observed significant enrichment in those related to musculoskeletal/connective, cardiovascular, reproductive, digestive, and endocrine

systems. Gene-set enrichments were concordant with previous findings by identifying gene-sets for the abundance or growth of metabolically active tissues, especially adipose. Yet, we also found gene-sets for morphology of the skeletal and vascular systems. Of 1009 signals, 706 have previously been associated with another GWAS trait with 5531 unique trait-index signal associations. Anthropometric traits are the most common; with triglycerides being the only non-anthropometric trait associated with >100 overlying association signals. Also, high-density lipoprotein cholesterol, sex hormone-binding globulin levels, type II diabetes, and estimated glomerular filtration rate all share >50 signals. This study supports previous findings for the relevance of adipose tissue biology in the genetic predisposition to elevated WHR, and provides robust evidence for other potential tissues and pathways leading to changes in body morphology.

88

VACANT-M: An Annotation-Enhanced Variant-Set Association Test Accounting for Genetic Relatedness and Population Structure

Shu-Hsien Cho*[1,2], Yu Yao[2], Ryan Bohlender[2], Chad Huff[2]
[1]University of Texas Health Graduate School of Biomedical Sciences, Houston, Texas, United States of America; [2]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

Variant-set tests boost power in rare-variant association studies by aggregating multiple rare genetic variants within predefined genomic regions. However, existing set-based methods frequently overlook quantitative functional annotations that prioritize deleterious variants, potentially reducing power and interpretability. Moreover, large-scale sequencing studies often include related individuals and participants from diverse ancestral backgrounds, which may confound association signals if genetic relatedness and population structure are not appropriately modeled.

We recently developed the Variant Annotation Clustering AssociatioN Test (VACANT), a gene-based method clustering rare variants by predicted severity to reflect functional heterogeneity of variants and tests each cluster against a binary trait using a Firth-penalized generalized linear model to accommodate imbalanced case-control ratios in biobank-scale data. We develop VACANT-M, a VACANT-extended mixed-model framework, which incorporates a genetic relatedness matrix as a random effect and principal components to account for stratification. We derive a penalized quasi-likelihood via Laplace approximation combined with Jeffrey's invariant prior, yielding closed-form score equations for fixed effects and the genetic variance component solved iteratively to approximate the intractable marginal likelihood.

Applying VACANT-M to biobank-scale whole-exome sequencing data from Down's Syndrome and pediatric Acute Lymphoblastic Leukemia cohorts, we demonstrate stringent type-I error control at genome-wide significance thresholds and superior empirical power relative to alternative set-based methods. VACANT-M integrates functional annotation, genetic relatedness, and population structure into a scalable framework for robust rare variant association analysis in complex, large-scale sequencing studies.

89

Heritability-Informed MR Reveals Divergent Roles of IL6 and TNF in Asthma and Lung Cancer

Matthew Boyton*, Tom Gaunt, Venexia Walker, Tom Richardson
MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom

Introduction: Asthma and lung cancer are distinct respiratory diseases with emerging evidence for shared and divergent genetic and immunological mechanisms. We aimed to clarify the role of circulating plasma proteins and their relationship to common disease heritability using an integrative multi-omic analysis method.

Methods: We first performed local genetic correlation analysis across both disorders to identify regions of robust shared heritability. We then applied shared causal variant mapping via SuSiE colocalization before identifying cis-pQTL instruments in the UK Biobank Pharma Proteomics Project cohort associated with common variants. High-confidence protein-protein interaction networks were then constructed for 2-sample Mendelian Randomization (MR) analysis to identify causal associations with disease risk.

Results: Six MHC region loci showed strong evidence for shared heritability, prioritising variants near key immunomodulators such as TNF. Network-informed MR identified IL6 as a disease-specific protective marker for lung cancer (Z = -4.85), whereas TNF showed a risk-increasing effect in asthma (Z = +3.51) and a weaker inverse association with lung cancer. Additional hits suggested a divergent immune axis among both traits, with Th2 and Th17-related effects appearing disease-specific. AZI2, a less-characterized immune regulator, showed a notable bidirectional association, potentially representing a novel marker. Further analysis via gene set enrichment underscored the role of Th2 polarization in asthma and a potential protective role for IL6 in lung cancer, consistent with MR findings.

Conclusions: These findings highlight TNF and IL6 as context-specific mediators of respiratory disease, and demonstrate the utility of heritability-informed MR to offer insights into divergent immunopathology relevant for therapeutic targeting.

90

Alzheimer's Disease Genetic Risk Variants Influence Age Trajectories of Circulating Peripheral Proteins

Anna Lorenz[1,2], Heather M. Highland[3], Misa Graff[3], Derek B. Archer[1], Jennifer E. Below[2]
[1]Vanderbilt Memory and Alzheimer's Center, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America ; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America ; [3]Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America

Background: Late-onset Alzheimer's disease (LOAD), the most common form of dementia, is highly heritable and

influenced by numerous genetic variants.[1–6] While the functional consequences of many LOAD-risk variants are unclear, the clinical symptoms of LOAD typically manifest late in life making aging the strongest risk factor. Aging is associated with systemic physiological changes which are reflected in the peripheral blood circulation by quantifiable changes in protein abdunance.[7,8] Some of these changes may be influenced by LOAD-risk variants and contribute to or reflect deleterious brain aging. However, LOAD-specific peripheral proteomic alterations associated with aging are largely understudied.

Methods: To examine how LOAD-risk variants modulate age-related changes in the peripheral proteome, we analyzed 2,920 plasma proteins using cross-sectional and cross-pooled linear models. Models included age, LOAD-risk variant, and their interaction, adjusting for sex, menopause, and population substructure. Data were leveraged from 261 participants (66% female, aged 18.3–87.4 years, mean age = 52.9) in the Cameron County Hispanic Cohort (CCHC). From 177 known LOAD-risk variants,[1–6] we analyzed 152 with a minor allele frequency ≥1%.

Results: Ten proteins showed significant age-by-genotype interactions ($p_{FDR} < 0.05$): CACNA1C, NCAM2, TIMP2, SLC39A14, CEACAM1, PRTG, KHDC3L, KRT17, KLRF1 and BTN1A1. These proteins are involved in neurogenesis, neural patterning, calcium signaling, metal transport, immune regulation, and blood-brain barrier integrity.

Conclusion: These finding indicate that LOAD-risk variants can influence the age trajectory of specific peripheral proteins. The enrichment of brain-relevant pathways suggests that peripheral aging signatures can provide insights into LOAD pathophysiology.

Keywords: Late-onset Alzheimer's Disease, genetic risk variants, peripheral proteome, aging.

References:

1. Bellenguez, C., Küçükali, F., Jansen, I.E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A.C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., et al. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet *54*, 412–436. https://doi.org/10.1038/s41588-022-01024-z.

2. Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet *51*, 404–413. https://doi.org/10.1038/s41588-018-0311-9.

3. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet *51*, 414–430. https://doi.org/10.1038/s41588-019-0358-2.

4. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452–1458. https://doi.org/10.1038/ng.2802.

5. Marioni, R.E., Harris, S.E., Zhang, Q., McRae, A.F., Hagenaars, S.P., Hill, W.D., Davies, G., Ritchie, C.W., Gale, C.R., Starr, J.M., et al. (2018). GWAS on family history of Alzheimer's disease. Transl Psychiatry *8*, 99. https://doi.org/10.1038/s41398-018-0150-6.

6. Wightman, D.P., Jansen, I.E., Savage, J.E., Shadrin, A.A., Bahrami, S., Holland, D., Rongve, A., Børte, S., Winsvold, B.S., Drange, O.K., et al. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet *53*, 1276–1282. https://doi.org/10.1038/s41588-021-00921-z.

7. Lehallier, B., Gate, D., Schaum, N., Nanasi, T., Lee, S.E., Yousef, H., Moran Losada, P., Berdnik, D., Keller, A., Verghese, J., et al. (2019). Undulating changes in human plasma proteome profiles across the lifespan. Nat Med *25*, 1843–1850. https://doi.org/10.1038/s41591-019-0673-2.

8. Shen, X., Wang, C., Zhou, X., Zhou, W., Hornburg, D., Wu, S., and Snyder, M.P. (2024). Nonlinear dynamics of multi-omics profiles during human aging. Nat Aging, *1–16*. https://doi.org/10.1038/s43587-024-00692-2.

91

Heritability and Shared Genetics Among Stress Exposures and With Personality: Insights From the Multigenerational Lifelines Cohort Study

Felix Reichelt[1], Rujia Wang[1,2], Rima D. Triatin[1,4], Martje Bos[3], Maryam Kavousi[5], Eco JC de Geus[6], Harold Snieder*[1]

[1]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2]Social, Genetic, and Developmental Psychiatry Centre; Institute of Psychiatry, Psychology and Neuroscience; King's College London, Denmark Hill, Camberwell, London, United Kingdom; [3]University of Groningen, University Medical Center Groningen, Department of Psychiatry, Groningen, The Netherlands. [4]Department of Biomedical Sciences, Faculty of Medicine Universitas Padjadjaran, Jawa Barat, Indonesia; [5]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands; [6]Department of Biological Psychology, Vrije University Amsterdam, Amsterdam, The Netherlands

Aim: This family-based study aims to estimate the heritabilities and shared genetics among stress exposures and with personality traits using data from the multigenerational Lifelines Cohort Study.

Methods: We analyzed self-reported data from up to 141,751 adults in the multigenerational Lifelines Cohort Study. Stress exposures included childhood trauma, loneliness, social support, stressful life events, and chronic stress. Personality traits were neuroticism, extraversion, and conscientiousness. Heritabilities (h^2), genetic (rg), shared (rc), and unique (re) environmental correlations were estimated using variance decomposition from linear mixed models in ASReml, adjusted for age, age^2, and sex.

Results: Heritability was highest for childhood trauma (50.7%) and lowest for loneliness (14.1%) and social support (17.6%). Chronic stress and life events showed moderate heritability (both 22.3%). Strong genetic correlations between childhood trauma and chronic stress (rg=0.66), and between loneliness and social support were observed (rg = -0.85). Personality traits were also heritable: 29.2% for neuroticism, 27.9% for extraversion, and 22.9% for conscientiousness. Neuroticism correlated genetically with chronic stress (rg = 0.49) and loneliness (rg = 0.61), while extraversion and conscientiousness showed negative genetic correlations with loneliness (rg = -0.62; -0.37). Shared environmental correlations were strong between chronic stress and neuroticism (rc = 0.67) and between social support and extraversion (rc = 0.72). A moderate unique environmental correlation was found between chronic stress and life events (re = 0.31), reflecting

overlapping individual experiences.

Conclusions: Stress exposures and personality traits show variable heritability and share genetic and environmental pathways with potential impact on mental and cardiometabolic health.

92

An Update on Selection Bias in Genetic Risk Estimates for Emerging Global Infectious Diseases: A Hostseq-Based Simulation Study and Application to a GWAS of COVID-19 Severe Outcomes

Ohanna C. Bezerra[1], France Gagnon[1,2], Shelley B. Bull[1,3], Jerry F. Lawless[1,4], Celia M. T. Greenwood[5,6]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Office of the Vice-Principal of Research and Innovation, University of Toronto Mississauga, Mississauga, Canada; [3]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [4]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada; [5]Lady Davis Institute, Jewish General Hospital, Montreal, Canada; [6]Department of Epidemiology, Biostatistics and Occupational Health and Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada

Elucidating the true risk factors for an emerging global infectious disease like COVID-19 is extremely challenging due to several convenience sampling complications, including those related to the rapidly changing nature of the disease and public health policies. The Host Genome Sequencing initiative (HostSeq) performed whole genome sequencing on 10,000 SARS-CoV-2-positive participants from 13 independent studies across Canada. These studies varied in design, recruitment location, and recruitment dates, leading to substantial variability in symptom definitions, disease severity rates, and virus evolution. We previously developed a simulation framework to evaluate potential selection bias in genetic association estimates for severe COVID-19 (hospitalization). Simulated data matched as closely as possible to HostSeq study characteristics and publicly available Canadian pandemic data on age, sex, and comorbidity rates (proxied by hypertension) for SARS-CoV-2 cases across six pandemic waves, and included a single associated SNP. In our new work, we contrast sampling based on age, sex, and comorbidity distributions with sampling based on severe disease distributions; explore the effects of ascertainment when the SNP acts as an effect modifier of comorbidity; and add analyses with survey weights and propensity scores to account for ascertainment. Logistic regression, whether adjusted or not for each individual study, or adjusted using propensity scores, yielded less biased genetic risk estimates than survey-weighted approaches, although biases tended to be small. Simulation results will illuminate our thromboembolism-informed hypothesis-driven GWAS of COVID-19 severe outcomes, to test whether COVID-19 morbidity is partially explained by sequenced variants known to be associated with thromboembolic outcomes.

Keywords: Selection bias, genetic association, simulation, HostSeq, COVID-19 severity

93

Enabling Metanalysis across multiple biobank studies for Phenome Wide Association Studies with DeepPheWAS

Richard Packer[1,2,3], Nick Shrine[1], William Hennah[4,5], Martin D. Tobin[1,2,3]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Leicester NIHR Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom; [3]Leicester British Heart Foundation Centre of Research Excellence, Leicester, United Kingdom; [4]Orion Pharma, Espoo, Finland; [5]Neuroscience Center, HiLIFE, University of Helsinki, Helsinki, Finland

Phenome wide association studies (PheWAS) are an important tool for measuring pleiotropy, and have been used to aid drug development, with results used as a proxy for drug side effects and drug repurposing. More recently, patterns in pleiotropy have also been used to gain insights into pathways of disease development. We developed and have been using DeepPheWAS as an R package for PheWAS that enables fine control over phenotype development and flexible testing, whilst utilising the maximal available data beyond ICD codes including primary care data, prescription data, bloods measurements and study questionnaires.

DeepPheWAS has seen significant progress. We integrated association testing with regenie, enabling gene-based testing and inclusion of time-to-event phenotypes in PheWAS. We integrated the package into the UK Biobank research access platform (RAP) using a flexible WDL workflow, which allows deployment across different computing environments. This development is allowing us to make DeepPheWAS operational within the All of Us Research Program. We introduced a mapping process, mapping key data fields across cohort studies to allow the use of questionnaire fields and study measurements alongside ICD-10 data. By doing this, we can dramatically increase case ascertainment, improving power for association testing and better understanding of pleiotropic effects of genetic variation. We will significantly increase representation from non-European ancestry by targeting All of Us, thus allowing better insight into the pleiotropic variation across ancestries that existing PheWAS resources have poorly served.

94

Comparison of Statistical Methods for Identifying X Chromosome Inactivation Patterns - Application in Asthma

Mozart Nerva Deneus[1], Zhonglin Li[1], Nathalie Gaudreault[1], Sébastien Thériault[1,2], Yohan Bossé[1,3], Aida Eslami*[1,4]

[1]Institut universitaire de cardiologie et de pneumologie de Québec, Université Laval, Quebec, Quebec, Canada; [2]Department of Molecular Biology, Medical Biochemistry and Pathology, Université Laval, Quebec, Quebec, Canada; [3]Department of Molecular Medicine, Université Laval, Quebec, Quebec, Canada; [4]Department of Social and Preventive Medicine, Université Laval, Quebec, Quebec, Canada

Context: Asthma is a common respiratory disease associated with both genetic and environmental risk factors. Genome-wide association studies (GWASs) have identified

approximately 200 genetic loci associated with asthma. However, most published asthma GWASs focus on autosomal variants and exclude sex chromosomes. Analyzing the X chromosome presents unique challenges, including X chromosome inactivation (XCI) in females, which silences one X chromosome to equalize gene dosage with males. XCI can be random, skewed, or escape inactivation. The degree of skewness is quantified by a parameter ranging from 0 to 2, where 1 indicates random XCI. The XCI status is not known a *priori*.

Aim: To compare existing methods in terms of their performance in identifying XCI.

Methodology: We evaluated three methods (XCMAX4, Xlink, and max-LLR) using simulated datasets reflecting various XCI patterns and allele frequencies. These methods were then applied to the Quebec City Case-Control Asthma Cohort, including 1,618 French-Canadian participants (1,089 with asthma and 529 controls), and 14,129 X-chromosome SNPs.

Results: In simulations, XCMAX4 was the most reliable for identifying true XCI status. Both XCMAX4 and max-LLR accurately detected skewness when the parameter was 0 or 2. For SNPs escaping XCI, Xlink performed comparably to XCMAX4. In real data, we assessed the results by comparing their concordance in identifying XCI status.

Conclusion: Accurate identification of XCI can improve statistical power in X-chromosome association analyses and deepen our understanding of complex diseases. XCMAX4 emerged as a robust and reliable method for detecting XCI patterns.

95

Polygenic Liability to Internalising and Cardiometabolic Multimorbidity and Health Trajectories Across Early Life

Ruby S. M. Tsang*[1,2], Daniel Stow[3], Ioanna Katzourou[4], The Lifespan Multimorbidity Research Collaborative (LINC), Peter A. Holmans[4], Marianne B. M. van den Bree[4,5], Sarah Finer[3], Golam M. Khandaker[1,2,6], Nicholas J. Timpson[1,2]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom; [4]Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, United Kingdom; [5]Mental Health Innovation Institute, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, United Kingdom; [6]NIHR Bristol Biomedical Research Centre, Bristol, United Kingdom

Co-occurrence of internalising and cardiometabolic conditions (ICM-MM) represents the most common type of physical-mental multimorbidity in older age. Examining early ICM-MM manifestations and development across time is critical to better understanding its aetiology and health factors marking premorbid presentation, potentially enabling better targeted prevention.

In earlier work, we trained an ICM-MM-specific polygenic risk score (ICM-MM-PRS) using PRSmix (elastic net) on 51 single-trait PRSs from Genomics plc in 45,492 UK Biobank participants. Leveraging longitudinal data from the Avon Longitudinal Study of Parents and Children (ALSPAC), we used natural cubic spline mixed effects models to estimate trajectories between 7 and 24 years in health-related measures in samples of 3,450-5,730 individuals. We tested the association of PRS with means and slopes of 12 measures, adjusting for sex and ten genetic principal components.

We found evidence for higher mean body mass index (BMI), fat mass, non-high-density lipoprotein cholesterol (non-HDL-c) and C-reactive protein across time with increased ICM-MM-PRS. In addition, ICM-MM-PRS was associated with accelerated increases in BMI, fat mass and lean mass, diastolic blood pressure and non-HDL-c, and a slower increase in HDL-c, with these changes most commonly observed in the transition to adulthood. No effects were found on depressive symptoms, systolic blood pressure, triglycerides, glucose or insulin.

Our findings suggest that genetic liability for ICM-MM is associated with both the overall level and accelerated development of cardiometabolic risk factors from childhood to early adulthood. This may help identify those at-risk of developing ICM-MM who will benefit from primary prevention.

96

The Role of Stress Sensitivity in the Genetics of Juvenile Myoclonic Epilepsy

Eric Sanders[1,2], Deb Pal[3,4], Lisa Strug[1,2,5,6]

[1]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [2]Program in Genetics and Genome Biology, The Hospital for Sick Children Toronto, Ontario, Canada; [3]Department of Basic & Clinical Neurosciences, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom; [4]MRC Centre for Neurodevelopmental Disorders, King's College London, London, United Kingdom; [5]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada; [6]Departments of Statistical Sciences and Computer Science, University of Toronto, Toronto, Ontario, Canada

Juvenile myoclonic epilepsy (JME) is a complex genetic disorder with known sex differences in etiology and a sex-specific relationship between stress sensitivity and anti-seizure medication resistance. We need to understand the mechanisms of sex-specificity to develop and improve treatment pathways for women.

We conducted a case-control GWAS of JME (n=3155), using cases from the BIOJUME Consortium (n=631) and controls from the Toronto Spit-for-Science initiative (n=2524). We applied our novel sex-specific colocalization technique on significant GWAS hits, and enrichment analysis using gene ontology terms. In follow-up analyses, we calculated polygenic risk scores for "depressed affect neuroticism" and "worry neuroticism" – correlates of stress-sensitivity and known risk factors for depression – using a derivation published by Liuhanen et al., testing association with JME incidence.

GWAS results implicate several regions, including loci previously associated with neuroticism, depressive disorder, and schizophrenia such as rs11191607 on chromosome 10

(p=3.84E-12). One chromosome 11 locus, rs3763912, was not previously reported in the GWAS catalog, but showed a female-specific effect on JME risk colocalizing with female-specific expression of several genes in brain tissues (min p=3.90e-07). Enrichment analysis indicated over-representation of several gene ontology terms including "neuron development", "synaptic membrane adhesion", and "positive regulation of MAPK cascade" (1.7-fold, 4.1-fold, and 1.8-fold enrichment). Depressed affect neuroticism PRS associated with JME incidence (p=0.00183), while worry neuroticism did not show significance at the 0.05 level (p=0.128).

These results provide evidence for a role of stress-response pathways in JME etiology, a potential explanation for shared comorbidity to depression, and a biological basis for sex-specificity.

97

Biological Group-Guided High-Dimensional Mediation Analysis of Omics Data

Yixin Zhang*[1], James B. Meigs[2,3,4], Ching-Ti Liu[1]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2]Division of General Internal Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [4]Program in Population and Medical Genetics, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, United States of America

Causal mediation analysis is a powerful tool for identifying omics mediators that link exposures to disease outcomes. Omics data provide unique snapshots of biological processes and physiological states, offering promising biomarkers for personalized medicine.

Due to the complex structure of omics data, mediation methods that focus only on individual mediators often overlook inherent connectivity, leading to poor interpretability and reduced robustness. To address this, we propose a novel framework that leverages prior knowledge-based groupings to enhance the identification of biologically meaningful effects. We employ sparse group lasso to select relevant mediator groups, then integrate a tailored debiasing procedure that enables valid statistical inference for individual mediator effects within a group, without requiring model refitting. Simulation studies indicated that this framework improved estimation accuracy and mediator identification, even under moderate group misspecification.

We demonstrate the utility of our framework using 925 non-diabetics in the Framingham Heart Study. Leveraging plasma metabolites, we investigated their mediating role in bridging BMI and glucose levels. Using pathway groupings based on existing databases, we identified 37 pathways among the 135 metabolites that passed sure independence screening. Notable significant mediators included glutamate, pyruvic acid, alanine, and glyoxylic acid from the alanine-aspartate-glutamate metabolism pathway, which is crucial for glucose regulation; as well as kynurenic acid, kynurenine, quinolinic acid, and xanthurenic acid from the tryptophan metabolism pathway, which influences insulin signaling. Impairment in these pathways is linked to diabetes risk, highlighting the utility of pathway-informed mediation analysis for revealing distinct biological mechanisms connecting BMI to glucose level.

Keywords: Casual mediation, High-dimensional omics, Knowledge integration, Debiased sparse group lasso, Diabetes

100

Challenges and Innovations in Constructing and Validating Individual-Specific Networks for Precision Medicine

Kristel Van Steen*[1], Fabio Stella[2]

[1]GIGA- Molecular and Computational Biology, University of Liège, Liège, Belgium; [2]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

Understanding patient-level differences is key to improving precision medicine, where the goal is to tailor prevention, diagnosis, and treatment to the individual. Individual-Specific Networks (ISNs) are a promising way to achieve this. By using network theory to represent molecular or clinical interactions unique to each person, ISNs offer a flexible method for combining omics data with population-level information to highlight individual variability.

However, building and validating these networks presents several challenges. One of the most important is choosing which features to include—deciding what parts of the system (like genes or proteins) should form the backbone of each network, as well as determining the most appropriate edge definition. New machine learning tools, such as visible neural networks and multi-view learning, can help prioritize the most relevant features or feature combinations based on disease relevance.

Another challenge is choosing how detailed or complex the networks should be. More detailed networks may capture richer information but can be harder to scale or interpret. These choices directly affect tasks like grouping similar patients, identifying subtypes of disease, or finding new uses for existing drugs.

In this work, we explore strategies to make ISNs more practical and informative for real-world use. We present ways to improve their construction, balance complexity and scalability, and better capture meaningful biological variation. Overall, ISNs represent a useful middle ground between generic models and highly complex digital twins—offering a practical path toward personalized, data-driven healthcare.

Keywords: individual-specific networks, digital twins, reverse-engineering, systems health

101

Advances in Individual-Specific Networks for Patient Subtyping in Precision Medicine

Kavya Singh*[1], Giulia Tremolada[2], Gaia Righetti[2], Zuqi Li[1], Federico Melograna[1], Dave Fardo[3], Fabio Stella[2], Kristel Van Steen[1]

[1]GIGA-Molecular and Computational Biology, University of Liège, Liège, Belgium; [2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy; [4]Department of Biostatistics, University of Kentucky,

Lexington, United States o America; [5]Sanders-Brown Center on Aging, University of Kentucky, Lexington, Kentucky, United States of America

Network-based methods provide a powerful framework for modeling biological systems as interconnected components (e.g., genes, proteins, metabolites), enabling the capture of complex interactions that underlie disease heterogeneity. Individual-Specific Networks (ISNs), reverse-engineered from population data, are subject-specific graphs whose nodes and edges reflect individual molecular or biological profiles. As such, ISNs offer a systems-level representation of patients, refining traditional approaches to uncover personalized disease mechanisms or therapeutic responses.

Advances in machine learning have led to the development of diverse network analytics techniques for identifying groups of structurally similar graphs. In this work, we evaluate a range of graph distance measures—both established and new—for their ability to detect subtle but biologically meaningful differences in ISN network topology. Inter-individual distances are processed through a consistent agglomerative hierarchical clustering pipeline. While graphlet-based metrics demonstrate strong discriminative power, they are computationally intensive. More scalable and interpretable alternatives, such as those based on curvature filtrations and persistent homology, retain edge weights—crucial for minimizing information loss during thresholding.

We furthermore show that careful selection of the variable support for ISN construction—i.e., the subset of nodes over which edges are defined—may improve the detection of clinically relevant subgroups. Moreover, integrating multiple data modalities or views, informed by the disease under investigation, enhances the ability of ISNs to uncover clusters that are disease- relevant, as demonstrated in the context of tau pathology.

Keywords: patient subtyping, individual-specific networks, representation learning

103

Improving Association Analysis of Mitochondrial DNA Heteroplasmy and Pancreatic Cancer by Hierarchical Testing

Brahim Aboulmaouahib[1,2,3], Martina Müller-Nurasyid[1,2,3,4], Antònia Flaquer[2,3], Hansi Weissensteiner[5], Peter Lichtner[6], Elvira Matthäi[7], Emily P. Slater[7], Detlef K. Bartsch[7], Konstantin Strauch[1,2,3]

[1]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany; [2]Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), , LMU Munich, Munich, Germany; [3]Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; [4]Pettenkofer School of Public Health Munich, Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Ludwig Maximilian University of Munich, Munich, Germany; [5]Institute of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria; [6]Core Facility Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; [7]Department of Visceral Thoracic and Vascular Surgery, Philipps University Marburg, Marburg, Germany

Somatic mutations in mitochondrial DNA (mtDNA) are increasingly recognized as key contributors to many human cancers and age-related diseases. In this study, we aimed to identify mitochondrial genetic variants associated with pancreatic ductal adenocarcinoma (PDAC), a malignancy with a currently poor prognosis. To achieve this, we performed a mtDNA association analysis of 10,350 single-nucleotide genetic variants (mtSNVs) in a cohort of 219 affected individuals and 2,880 population-based controls.

The study comprises sporadic cases as well as cases from the German National Case Collection of Familial Pancreatic Cancer (FaPaCa). Mitochondrial heteroplasmy was derived using our previously presented pipeline for processing of mitochondrial sequencing data. For association analysis, we developed a hierarchical analytical framework. First, we performed a stepwise comparison of dichotomized heteroplasmy between cases and controls. Second, we used a log-linear mixed model (LLMM) with the continuous heteroplasmy ratio at mtSNVs as the outcome and disease status as the primary predictor, adjusting for age, sex, sequencing coverage, and batch.

After adjusting for multiple comparisons, we identified 17 genes associated with PDAC. In 6 genes the signal could be traced down to 12 mtSNVs, while the remaining 11 genes only produced an aggregated signal. 3 mtSNV signals were also identified using LLMM, along with 1 additional mtSNV.

Our hierarchical testing strategy provides more signals on gene and mtSNV level than mtSNV LLMM alone. This improves the potential to gain insights into the role of mitochondrial DNA variants in the pathogenesis of PDAC.

Keywords: Mitochondrial DNA Heteroplasmy, Log-Linear Mixed Model, Hierarchical Testing, PDAC, Familial pancreatic cancer

104

HLA Genotype Combinations Impact Allele Association With Multiple Sclerosis Risk

François Cornélis*†[1,2], Igor Faddeenkov[1], Stanislas Demuth[1], Sonia Bourguiba-Hachemi[1], Pierre-Antoine Gourraud1,3, Nicolas Vince[1], and Anna Serova-Erard ‡[1,2,4]

[1]Team 5: Neuroinflammation, Mechanisms, Therapeutic Options (NEMO), Centre Hospitalier Universitaire, Nantes, Nantes Université, INSERM, Centre de Recherche Translationnelle en Transplantation et Immunologie (CR2TI), Nantes, France; [2]Genetique Oncogenetique Adulte Prevention (GENOAP), Centre Hospitalier Universitaire, Clermont-Ferrand, France; [3]Pôle Hospitalo-Universitaire, Santé Publique, Clinique des données, Nantes Université, Centre Hospitalier Universitaire, Clermont-Ferrand, France; [4]UFR de Médecine et des professions paramédicales, Université Clermont Auvergne, France

Associations between Multiple sclerosis (MS) and HLA involve 7 predisposing and 6 protective alleles (HLA-DRB1*03:01/*08:01/*13:03/*15:01, HLA-DQB1*03:02, HLA-DPB1*03:01, LTA-H51P and HLA-A*02:01, HLA-B*38:01/*44:02/*55:01, HLA-DQA1*01:01, HLA-DQB1*03:01,respectively). We investigated HLA-wide

genotype combinations in MS associations. We analysed WTCCC HLA data for 11,376 MS-cases (2005 diagnosis criteria) and 18,872 controls. We performed principal component analysis for European ancestry selection. HLA alleles were imputed with HIBAG R package or inferred with proxy SNPs (rs2229092, rs9273912 and rs9277565) and recoded to consider only the 13 MS-associated alleles. Dataset was divided: 20% to search for MS-associated genotype combinations and 80% to test them for replication. Replicated combinations were assessed on the whole dataset. We retained 9,024 MS and 13,923 controls from European Ancestry. In the 20% sub-sample, we observed 41 nominally MS-associated genotype-combinations ($P < 0.05$) (out of 776). In the 80% subsample, 22 combinations were replicated (P corrected < 0.05/41). In the full dataset, they accounted for 23.61% of MS-cases with 14 predisposing combinations (OR 1.83-6.75) and 4.29% of MS-cases with 8 protective combinations (OR 0.30-0.57). Surprisingly, some predisposing combinations carried "protective" alleles and vice versa: HLA-allele MS-association depends on HLA-wide-genotypes. Strikingly, 4.29% of patients, diagnosed with MS *prior* to 2005, carried protective combinations: those combinations are to be investigated in patients whose former MS diagnosis would in 2025 be changed to NeuroMyelitis Optica Spectrum Disorder or Myelin Oligodendrocyte Glycoprotein Antibody-associated disease. Those HLA-combinations could help clarifying disease heterogeneity and improve diagnosis.

105

Uncovering the Genetic Nexus of Obesity and Addiction: Pleiotropic Loci from Large-Scale GWAS Reveal Shared Neurobehavioral Risk

Baiyu Qi[1], Heather M. Highland[1], Mariaelisa Graff[1], Cynthia M. Bulik[2,3,4], Kristin L. Young[1], Daeeun Kim[5], Melissa A. Munn-Chernoff[6], Kari E. North[1]

[1]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina , United States of America; [2]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina , United States of America; [3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [4]Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [5]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [6]Department of Community, Family, and Addiction Sciences, Texas Tech University, Lubbock, Texas, United States of America

Obesity has increasingly been conceptualized as a neurobehavioral disorder with overlapping features with substance-related addictions. Evidence suggests a shared genetic architecture between body mass index (BMI) and addictive behaviors such as binge eating (BE), smoking, and alcohol use, but specific pleiotropic variants remain largely unknown. We analyzed genome-wide association study (GWAS) summary statistics for BE (39,279 cases; 1,227,436 controls), cigarettes per day (CigDay; N = 618,489), drinks per week (DrnkWk; N = 2,428,851), and BMI (N = 1,924,645) in individuals of European ancestry. Using Genomic Structural Equation Modeling (Genomic SEM), we identified a latent genetic factor capturing shared liability across these traits and estimated its genetic correlations (rg) with neuropsychiatric and metabolic outcomes. We applied gene prioritization, gene-set enrichment, gene expression colocalization, and METASOFT to identify pleiotropic loci.

The latent factor explained 31% of shared genetic variance and was positively correlated with ADHD (rg = 0.26), PTSD (rg = 0.18), MDD (rg = 0.12), type 2 diabetes (rg = 0.49), and HbA1c (rg = 0.15), and negatively correlated with schizophrenia (rg = –0.09), OCD (rg = –0.24), and HDL cholesterol (rg = –0.35). Of 690 genome-wide significant loci ($P < 5 \times 10^{-9}$), 12 were pleiotropic across all four phenotypes, 55 across three, and 317 across two. Notably, rs12739892 in PDE4B was associated with all traits; *PDE4B* has been linked to substance use, and *PDE4* inhibitors reduce alcohol intake in rodents. Additionally, rs8074078 (*BPTF*) and rs6567315 (*PHLPP1*) influenced BMI via binge eating. These findings highlight shared genetic mechanisms linking obesity and addiction, with potential therapeutic relevance.

Keywords: obesity, addiction, GWAS

106

Proprotein Convertase Subtilisin/Kexin Type 9 and Breast Cancer Survival: a Mendelian Randomization Study

Janne Pott*[1], Amy Mason[2,3], Stephen Burgess[1,2]

[1]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [2]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [3]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, United Kingdom

Proprotein convertase subtilisin/kexin type 9 (PCSK9) is well known for its causal effects on the lipid metabolism. A recent study of 553 breast cancer patients identified an association between missense mutation rs562556 within *PCSK9* and breast cancer survival (BCS). It was suggested that PCSK9 inhibition might allow for early intervention strategies to prevent metastasizing breast cancer. Here, we explored this link between PCSK9 and BCS using data from publicly available genome-wide association study (GWAS) summary statistics.

Using PCSK9 protein levels as exposure, we applied three 2-sample Mendelian Randomization (MR) approaches on BCS. Via inverse variance weighted (IVW) MR, we observed negative, but insignificant estimates (logHR=-0.199, *P*=0.274). To correct for indirect effects via lipid metabolism, we applied multivariable MR-IVW including low density lipoprotein as exposure. The causal estimate changed effect direction, but remained insignificant (logHR=0.328, *P*=0.264). Finally, we used only rs562556 on both data from the original study and a large meta-GWAS (N=99,217). The MR-ratio estimates were both positive (logHR=12.6 and logHR=1.66, respectively), but only significant when using the small study ($P=4.25\times10^{-4}$ and $P=7.56\times10^{-2}$, respectively).

The different scaling and significance in the MR-ratio estimates could be explained by different genetic models,

sample selection, and the time-variability of HR estimates. The difference between the ratio and IVW approaches might reflect different pathways of PCSK9 action on BCS. In conclusion, a significant positive effect of PCSK9 on BCS could only be reproduced when using the exact same outcome data as in the original study, but not when using an independent, larger meta-GWAS.

107

High-Resolution Identity-by-Descent Mapping Test Identifies Genetic Associations for Deep-Learning-Derived Brain Imaging Phenotypes

Han Chen[1], Bohong Guo[2], Ziqian Xie[3], Wei He[3], Ardalan Naseri[3], Degui Zhi[3]

[1]Human Genetics Center, Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [2]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [3]Department of Bioinformatics and Systems Medicine, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America

In the past two decades, genome-wide association studies (GWAS) have identified numerous genetic loci associated with complex diseases and quantitative traits. Most GWAS have focused on testing the associations with common and rare variants, but ignored the phased haplotype information. Therefore, little is known about the roles of mid-range and long-range haplotypes on the genetic architecture of complex traits. Recently, efficient identity-by-descent (IBD) detection algorithms have facilitated IBD segment calling and IBD-based association mapping in biobanks and large-scale populations. Existing IBD mapping methods typically require chunking the whole genome into short segments and testing each segment to reduce the computational burden, but one major limitation is their low resolution in association mapping. Here we present an efficient IBD mapping test that takes snapshots of IBD segments at the single nucleotide polymorphism (SNP) level resolution. Our benchmark study with an existing fast IBD mapping test (FiMAP) shows that the new algorithm can finish genome-wide IBD mapping at 640,899 SNPs in 50 CPU days, compared to 90 CPU hours for 3,403 1-centiMorgan windows from FiMAP, in the UK Biobank with 407,827 individuals. The new algorithm identified all association peaks previously reported for 6 anthropometric traits, but at a much higher single-SNP level resolution. We then applied the high-resolution IBD mapping test to unsupervised deep-learning-derived brain magnetic resonance imaging phenotypes from over 35,000 individuals in the UK Biobank, and identified 7 novel associations that had not previously been identified using a GWAS approach, illustrating its potential in providing complementary evidence in association mapping.

Keywords: identity-by-descent, association mapping, imaging genetics, deep learning, biobank

108

Comparison of Linear and Non-Linear Approaches to Ancestry Estimation in Highly-Admixed Populations

Mark H. Lamin*[1], Saonli Basu[1,2], Michael J. Anderson[2], Kody A. DeGolier[2], Andrew R. Raduski[3], Logan G. Spector[3]

[1]Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, Minnesota, United States of America; [3]Division of Pediatric Epidemiology & Clinical Research, University of Minnesota, Minneapolis, Minnesota, United States of America

Ancestry estimation in genome-wide association studies (GWAS) is critical for understanding genetic diversity and its relationship to diseases and traits across different populations. Linear approaches, such as principal component analysis (PCA), capture population structure by projecting genetic data onto a set of orthogonal axes representing ancestral variation. These methods are computationally efficient and have existing frameworks for interpreting results, making them popular for controlling confounding factors in GWAS. However, linear methods may struggle to capture highly-admixed population structures, potentially leading to oversimplification of ancestry. Nonlinear approaches, such as Uniform Manifold Approximation Projection (UMAP), offer greater flexibility in modeling intricate ancestral structures. We studied several linear and non-linear approaches for ancestry estimation and illustrated them through simulation studies and real data analysis. In summary, linear methods are efficient and effective for simpler ancestry estimation while nonlinear methods provide more accurate estimates of admixed population structures at the cost of increased computational complexity.

109

High Consanguinity, Underrepresented Populations, and Whole Genome Sequencing: Ingredients for Novel Rare Variant Discoveries?

Abdullah Shaar[1], Areeba Irfan[2], Khalid Kunji[1], Mohamad Saad*[1]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [2]College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar

*Presenting author

Background: Rare variants have the potential to explain part of the missing heritability of complex diseases. For underrepresented Middle Eastern populations that exhibits high levels of relatedness and consanguinity, novel rare variant discovery could be made due to enrichment of rare variants.

Methods: Using the Qatar Precision Health Institute of 13,994 whole genome sequencing individuals, we performed gene-based burden test for 21 cardiometabolic traits. The impact of variant inclusion criteria such as the minor allele frequency cutoff and the functional annotation of variants. Conditional analysis of the obtained genes with respect to common variants was performed. Burden of pathogenic rare variants within known cardiometabolic genes with monogenic mode of inheritance was quantified.

Results: We showed a few pathogenic variants that are substantially more prevalent in our dataset (e.g.,

chr19:11102742:A:G in *LDLR*, associated with familial hypercholesteremia, 50-fold increase compared to gnomAD v3). Our results showed 70 gene-trait associations with $P<5\times10^{-8}$. The number of obtained associations increased for higher minor allele frequency cutoffs. The identified genes were not restricted to one functional variant annotation approach. Thirty one percent of the associations were already present in the GWAS catalog. Eight novel genes were prioritized based on their plausible biological function related to the associated traits (e.g., *FASN* for LDL; *OTUD3* for hemoglobin A1C; *KCNK9* for electrocardiography JT interval. Pathogenicity rate within known cardiometabolic genes was higher than in the UK Biobank (5.5% vs 2.4%). Of the total significant burden test associations, 24% could not be discovered by single marker analysis. Finally, 79% of the identified genes remained significant at $P < 5\times10^{-8}$ conditional on the best common variants.

Keywords: Rare Variant Analysis, Consanguinity, Middle East

110

WGS-based HLA Allele Imputation Quantifies Pleiotropic Associations With Disease-Related Traits

Peyton McClelland*[1], Ken Sin Lo[2], Guillaume Lettre[2,3], Simon Gravel[1], Claude Bhérer[1], Daniel Taliun[1]

[1]Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada; [2]Montreal Heart Institute, Montreal, Quebec, Canada; [3]Department of Medicine, Université de Montréal, Montreal, Quebec, Canada

Human leukocyte antigen (HLA) alleles are population-specific and associated with autoimmune diseases affecting millions of people. We assembled a novel HLA haplotype reference panel comprising individuals of French-Canadian, Haitian and Moroccan ancestries, and applied phenome-wide association analysis (PheWAS) to elucidate HLA allele-phenotype relationships.

The CARTaGENE (CaG) population-based biobank in Quebec, Canada, contains high-depth whole-genome sequencing (WGS) (N=2,180) and genotyping (N=27,239) data. We called 307 high-quality HLA alleles from WGS and constructed a reference panel for HLA imputation. We found 42 (14%) alleles significantly enriched in frequency compared to a global panel, including HLA-B*51:01 (7% vs 4%, $p=4.0\times10^{-13}$) linked to Behçet's disease, most often reported in populations along the Silk Road. After validating the panel in an independent dataset, we imputed HLA alleles into genotyping data and performed PheWAS of 137 imputed alleles with 42 phenotypes. We detected 47 statistically significant associations ($p<3.7\times10^{-4}$) in all participants (N=27,239); 38 replicated in Europeans (N=25,210); 12 alleles were associated with ≥2 phenotypes. HLA-B*08:01 was associated with decreased leukocyte ($p=8.4\times10^{-8}$, β=-0.20), monocyte ($p=4.9\times10^{-10}$, β=-0.02), lymphocyte ($p=2.4\times10^{-5}$, β=-0.06) counts and increased thyroxine levels ($p=3.6\times10^{-5}$, β=0.15). HLA-B*08:01 is expressed on the surface of lymphocytes/monocytes, which play a role in thyroid disorders, and was associated with thyroid status in the UK Biobank.

We constructed and validated a new HLA reference panel which included enriched known disease-risk alleles such as HLA-B*51:01. It helped us identify 47 associations with 20 phenotypes and quantify pleiotropic effects of HLA-B*08:01 with thyroxine and blood cell levels, providing valuable insights into thyroid molecular mechanisms.

111

Inflammation, Brain Structure, and Cognitive Decline: An Integrated Analysis in the Rotterdam Study

Midas M. Kuilman, Eline J. Vinke, Costanza L. Vallerga, Joyce van Meurs, Meike W. Vernooij, M. Arfan Ikram, Frank J. Wolters, Mohsen Ghanbari

Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands

Background: Inflammation plays an important role in neurodegenerative processes, but its role in brain structural changes and cognitive decline remains unclear. We investigated the relationship between inflammatory proteins, MRI-derived brain markers, and cognitive performance in the Rotterdam Study and applied Mendelian Randomization and enrichment analysis to gain mechanistic insight.

Methods: We included 3,456 participants from the third Rotterdam Study cohort with proteomics, cognition, and MRI data. Inflammatory proteins were profiled using Olink's Inflammation panel. A general cognitive factor was derived from multiple cognitive tests. Brain volumetric markers were obtained via MRI. Cross-sectional and longitudinal associations were analyzed using linear regression and mixed models. Mendelian randomization was performed using genetic instruments for selected inflammatory proteins to assess causality with cognitive function. Gene-set enrichment analysis was performed on proteins associated with cognition or MRI metrics to explore biological pathways.

Results: This exploratory analysis identified several inflammatory proteins associated with reduced cognitive performance and brain volume. MR findings suggested possible causal effects for specific proteins on cognitive decline.

Conclusion: Our findings highlight the complex interplay between systemic inflammation and cognitive aging. While exploratory, this integrative approach combining proteomics, neuroimaging, and cognitive tests offers insight for future research into mechanisms underlying cognitive decline.

Keywords: Cognitive decline, inflammation, brain MRI, Mendelian Randomization

112

DrFARM: Identification of Pleiotropic Genetic Variants in Genome-Wide Association Studies

Lap Sum Chan[1,2], Gen Li[1], Eric B. Fauman[3], Xianyong Yin[4], Markku Laakso[5], Michael Boehnke[1] and Peter X.K. Song[1]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America; [2]Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [3]Internal Medicine Research Unit, Pfizer Worldwide Research, Development and Medical, Cambridge, Massachusetts,

United States of America; [4]Department of Epidemiology, Nanjing Medical University, Nanjing, Jiangsu, China; [5]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland

In a standard analysis, pleiotropic variants are identified by running separate genome-wide association studies (GWAS) and combining results across traits. But such statistical approach based on marginal summary statistics may lead to spurious results. We propose a new statistical approach, Debiased-regularized Factor Analysis Regression Model (DrFARM), through a joint regression model for simultaneous analysis of high-dimensional genetic variants and multilevel dependencies. This joint modeling strategy controls overall error to permit universal false discovery rate (FDR) control. DrFARM uses the strengths of the debiasing technique and the Cauchy combination test, both being theoretically justified, to establish a valid post selection inference on pleiotropic variants. Through extensive simulations, we show that DrFARM appropriately controls overall FDR. Applying DrFARM to data on 1,031 metabolites measured on 6,135 men from the Metabolic Syndrome in Men (METSIM) study, we identify five first-time reported putative causal genes, none of which had been implicated in any prior metabolite GWAS (including the prior METSIM analysis).

Keywords: Pleiotropy, debiasing, metabolomics, FDR control, post-selection inference

113

Leveraging a Multi-Population Likelihood Framework for Bayesian Model Uncertainty in PRS Construction

Gillian King*[1], Jiayi Shen[1], David V. Conti[1,2,3]

[1]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [2]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [3]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America

Bayesian Model Uncertainty (BMU) provides a framework for exploring model spaces by averaging over multiple plausible models, which in the context of SNP association studies could enable robust variant selection and effect estimation. We adapted an existing BMU approach, originally designed for individual -level data, to instead leverage Genomic Summary Results (GSR) within a flexible likelihood framework that enhances variant detection across diverse populations. Specifically, we define the model likelihood using sufficient statistics from the Joint Analysis of Marginal SNP Effects (mJAM) and explore models via Markov Chain Monte Carlo (MCMC). Our method also allows for the incorporation of informative priors, which could further refine variant prioritization and improve the precision of signal localization.

Models are ranked by posterior probability, and SNP-specific quantities, including marginal inclusion probabilities and Bayes factors, are computed. Our method tends to highlight key subregions, often aligning with regions identified by fine-mapping methods such as mJAM Forward and mJAM Sum of Single Effects (SuSiE). Rather than pinpointing specific index SNPs, our method captures broader signals and provides averaged associations across regions, which may better support PRS construction by aggregating effects across many variants. Through simulations and applications to prostate cancer risk regions on chromosomes 10, 11, and 12, we demonstrate that our method not only recovers the subregions identified by established fine-mapping approaches, but also extends the analysis by producing posterior quantities useful for PRS construction.

Keywords: PRS Construction, Bayesian Model Uncertainty, Prostate Cancer, Summary Statistics

114

Fantasio: A Case-Control Approach To Detect Rare Recessive Rariants In Multifactorial Diseases

Sidonie Foulon[1,2], Thérèse Truong[1], Anne-Louise Leutenegger[2], Hervé Perdry[1]

[1]Université Paris-Saclay, Villejuif, France; [2]NeuroDiderot, Inserm, Université Paris Cité, Paris, France

Genome-wide association studies (GWAS) aim to detect associations between genetic variants and multifactorial traits. They mainly use common variants and study them according to the additive genetic model. However, the genetic component of most multifactorial diseases is not yet fully elucidated. This could be partly due to the contribution of rare variants with recessive effects, which are difficult to identify in GWAS. We propose Fantasio, a method based on an excess of Homozygous-by-Descent (HBD) segments shared among cases compared to what is expected among controls. HBD segments, found in consanguineous individuals, are regions where rare recessive variants are more likely to be found.

We present a simulation framework to assess the type I error and power of Fantasio, and the results. In these simulations, haplotypes from 1000 Genomes are shuffled to create new 'mosaic' haplotypes, allowing to control the consanguinity coefficient of simulated individuals. Some consanguineous cases are selected to carry rare recessive variants in a specific genomic region, while other cases and controls have varying degrees of consanguinity. The sample size, the percentage of cases linked to the rare recessive variants and the types of consanguinity are varied.

Our results show that the type I error is well controlled. For some genetic models (e.g. rare disease with 10% of cases sharing deleterious alleles), Fantasio achieves high power starting from a case-control sample size of 1000 individuals. Our method is promising for studying rare recessive variants in real data and with more common multifactorial diseases, particularly with large sample sizes.

Keywords: Rare recessive variants, Consanguinity, Multifactorial diseases, Statistical power, GWAS

115

Robust Rare Variant Association Tests for Machine-Learning Derived Phenotypes

Andrey Ziyatdinov[1], Anjali Das[1,2,3], Joseph Herman[1,] Benjamin Geraghty[1], Karl Landheer[1], Joelle Mbatchou[1], Olivier Delaneau[1], Jonathan Marchini[1]

[1]Regeneron Genetics Center, Tarrytown, New York, United States of America; [2]New York Genome Center, New York, New York, United States of America; [3]Computer Science, Columbia University, New York, New York, United States of America

In the UK Biobank many anthropometric and biomarker measurements are available on almost all of the 500,000 participants, but only a subset (~70,000) participants have imaging-based body fat and muscle composition measures. Phenotype imputation approaches can potentially be applied to impute the missing imaging-based measures from the fully observed measures and lead to a boost in power to find genetic associations.

Recently a method called POP-GWAS was introduced that takes into account the phenotype imputation uncertainty when testing common variants in GWAS. The approach works by combining together three association statistics calculated using observed and imputed phenotypes in the subsets of individuals with and without missing data.

However, in exome wide association studies (ExWAS) where rare variants and genes are tested we observed that the POP-GWAS approach is not well calibrated. Using real and simulated data we show why this approach leads to inflated test statistics at rare variants, and propose a new method that involves constructing an intermediate composite phenotype in those individuals with observed phenotype values.

We analyzed 34 whole-body MRI traits (N ~ 68,000) and 24 broadly measured traits (N ~ 490,000) with genotype imputed and exome sequencing data in UK Biobank. We used an autoencoder to predict missing phenotypic data in the MRI traits and obtained prediction accuracy r^2 = 0.31-0.82. POP-GWAS and our method produce near identical results at common variants, but POP-GWAS exhibits clear evidence of inflation at rare variants. Our results are well calibrated in simulations and the real MRI traits and suggest novel associations for MRI-derived total muscle water volume.

116

Glycemic Traits are Associated with Insulin Signaling and Hormone Metabolism in a High-Risk Hispanic/Latino Population on the US/Mexico Border

Heather M. Highland[1], Elizabeth G. Frankel[2], Wanying Zhu[2,] Xinruo Zhang1, Rashedeh Roshani[2], Mohammad Yaser Anwar[1], Kristin L. Young[1], Joseph B. McCormick[3], Susan P. Fisher-Hoch[3], Jennifer E. Below[2]

[1]University of North Carolina Gillings School of Global Public Health. Department of Epidemiology, Chapel Hill, North Carolina, United States of America; [2]Vanderbilt University, Department of Medicine, Nashville, Tennessee, United State of America; [3]The University of Texas Health Science Center at Houston, School of Public Health, Brownsville, Texas, United States of America

Insulin resistance, impacting glucose homeostasis, precedes the development of Type 2 Diabetes (T2D). T2D disproportionately impacts Hispanic/Latino populations (H/L) who are often underrepresented in scientific studies. Studies of the circulating proteome are crucial for a better understanding of disease pathogenesis. Here we assess proteins associated with fasting glucose (FG), glycated hemoglobin (A1c), and Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) in H/L individuals without T2D from the Cameron County Hispanic Cohort (CCHC) to reveal insights on biological processes, potential biomarkers for diagnosis and prognosis, and pathways for drug development.

We leveraged plasma proteomic data from 313 CCHC study participants without T2D using the Olink ExploreHT panel to assess 5,420 proteins. Normalized protein values were log_2 transformed for normality. We used linear regression models, adjusting for age, sex, BMI, and ancestral principal components to identify 8, 15, and 239 proteins associated ($P<1x10^{-5}$) with FG, HOMA-IR, and A1c respectively. Notable findings included a positive association of INS-CPEPTIDE ($P_{HOMA-IR}=2.0x10^{-14}$), CHUK ($P_{A1c}=8.1x10^{-14}$), INPP5D ($P_{A1c}=3.5x10^{-13}$), TMPRSS15 ($P_{HOMA-IR}=6.9x10^{-10}$), IGSF9 ($P_{HOMA-IR}=9.1x10^{-9}$. $P_{FG}=1.2x10^{-7}$), IGSF3 ($P_{FG}=1.3x10^{-8}$), and GSTA3 ($P_{FG}=6.3x10^{-7}$) and a negative association of IGFBP1 ($P_{HOMA-IR}=1.5x10^{-8}$), CKB (PHOMA-IR=$3.0x10^{-7}$), and IGFBP2 ($P_{HOMA-IR}=4.7x10^{-6}$). Many of these proteins are well characterized in the pathogenesis of T2D (e.g., insulin related genes and insulin-like growth factor binding proteins) whereas other findings appear to support more emerging biology for T2D, for example AKT signaling (CKB, CHUK, and INPP5D), and steroid hormone metabolism (GSTA3). In summary, our findings provide important insights into the biological pathways driving the development of T2D.

Keywords: proteomics, diabetes, insulin resistance

118

Mendelian Randomization Provides No Evidence for the Associations of Genetically Predicted Heart Rate Variability with Seven Psychiatric Diseases

Wenyu Huang[1], Zekai Chen[1], Catharina A Hartman[2], Harold Snieder[1]

[1]Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [2]Interdisciplinary Center Psychopathology and Emotion Regulation, Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Background: Observational studies have shown that cardiac autonomic function, as indicated by heart rate variability (HRV), is associated with psychiatric disorders. While a causal role is often suggested in the literature, this is currently unknown. We studied potentially causal effects of HRV-traits on multiple psychiatric disorders, including major-depressive-disorder (MDD), anxiety-disorders (ADs), attention-deficit-hyperactivity-disorder (ADHD), post-traumatic-stress-disorder (PTSD), bipolar-disorder (BIP), schizophrenia (SCZ), and alzheimer's disease (ALZ), using two-sample Mendelian randomization (MR) methods.

Methods: The primary exposure was the root mean square of the successive differences (RMSSD). We additionally studied related HRV indices SDNN, HF and those corrected(c) for heart rate: RMSSDc, SDNNc, HFc. These six HRV-related phenotypes were derived from the largest, up-to-date GWAS of cardiac autonomic function. As outcomes, we likewise used summary-level statistics from the largest, most up-to-date GWAS of psychiatric disorders. Multiple complementary MR methods, with the inverse-variance-weighted method as main analysis, were

performed to evaluate causality..

Results: The IVW-MR analyses indicated that genetic liability to RMSSD was not significantly associated with MDD (OR, 1.08, 95%CI, 0.98 - 1.20; *P* =0.14), ADs (1.05, 0.93 - 1.20; 0.43), ADHD (1.02, 0.96 - 1.08; 0.54), PTSD (1.03, 0.93 - 1.13; 0.61), BIP (1.10, 0.96 - 1.26; 0.18), SCZ (0.97, 0.82 - 1.16; 0.75) and ALZ (1.06, 0.88 - 1.27; 0.53), respectively. Similarly, there were no significant associations of the other five HRV with seven psychiatric diseases (all *Ps* > 0.05). Complementary MR methods yielded consistent findings.

Conclusions: Our study does not support that cardiac autonomic function, as indicated by HRV traits, has a causal role in these seven psychiatric diseases.

Keywords: Heart rate variability traits, Psychiatric disorders, Mendelian Randomization, Causal Inference

119

Genetic Dysregulation of Protein Expression in Aging and Neurodegeneration

Mykhaylo M. Malakhov*, Wei Pan

Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America

Genetic variation influences complex traits in part by regulating gene and protein expression. Precise coordination of expression is necessary for the maintenance of healthy cellular and physiological functions, but aging has been shown to disrupt transcriptional relationships between genes. Here we explore how aging impacts the genetic regulation of protein expression and quantify the downstream effects of genetic dysregulation on Alzheimer's disease and other neurodegenerative disorders. We considered UK Biobank Pharma Proteomics Project (UKB-PPP) participants of White British ancestry with data available at the baseline visit (N = 36k) and stratified them into non-overlapping groups by age. We then performed age-stratified protein quantitative trait locus (pQTL) mapping and trained elastic net models within each stratum to predict plasma protein concentrations from genotype data. Our results demonstrate that although most genetic effects on proteomic levels remain stable with age, aging attenuates the overall heritability of protein expression. Moreover, we conducted proteome-wide association studies (PWAS) and co-expression-wide association studies (COWAS) to identify age-specific associations between genetically predicted protein concentrations and Alzheimer's disease, highlighting disease-relevant molecular pathways that change over time. Together, our findings provide insight into the mechanisms underlying aging-related changes in protein expression and disease.

Keywords: Alzheimer's disease, genetic regulation, plasma proteome, proteome-wide association study, quantitative trait loci

120

Effects of Trait Ascertainment and Selection on Polygenic Scores in Family-based Study Designs for Rare Variant Genetic Association Analysis

Shelley B. Bull*[1,2], Kexin Luo[1], Sumin Kim[1], Michela Panarella[2], Razvan Romanescu[3]

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada; [2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [3]Department of Community Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada

Two classic family-based designs involve ascertainment of affected family members: eg. affected sibling pairs (ASPs) with early age at onset, parent-offspring trios ascertained on childhood disease diagnosis. Because family ascertainment depends on phenotype(s) of family members, followed by genotyping or sequencing, individual genotypes are often modelled as a response conditional on phenotype. Motivated by investigations of breast cancer and autism spectrum disorder, we designed simulation studies that imitate ascertainment of families from a large population generated under an individual-level disease susceptibility model. We compare validity, efficiency and power of rare variant (RV) tests, and quantify effects of alternative ascertainment criteria. For ASP studies in which RV counts are stratified by local identical by descent (IBD) sharing in a linear model, power to detect association can decrease dramatically when disease susceptibility also depends on a common variant polygenic score (PGS); however, efficiency can be improved by excluding sibpairs with high PGS and/or by setting stricter age-at-onset criteria. For trio designs, validity and efficiency can be improved by modelling the PGS as an effect modifier in conditional logistic regression of parent-child allelic transmissions. When disease susceptibility consists of the combined effect of rare germline variants, polygenic background, and family history, the detection of RV association can be affected by ascertainment in complicated ways that have implications for both design and analysis.

121

Large Language Model-Driven Single-Cell Analysis Enhances Prediction of Breast Cancer Therapy Response via Cell-Type-Specific Markers

Victoria Truong*[1], Yiming (Emmett) Peng*[1], Yu Shi[1], Pingzhao Hu[1,2,3]

[1]Dalla Lana School of Public Health, Biostatistics Division, University of Toronto, Toronto, Canada; [2]Department of Biochemistry, Western University, London, Canada; [3]Department of Computer Science, Western University, London, Canada

*Equal contributions

Background: The cellular heterogeneity of breast cancer (BC) limits the predictive power of models based on bulk RNA sequencing. Single-cell RNA sequencing (scRNA-seq) offers higher resolution, yet current approaches struggle to identify meaningful cell-type-specific markers for therapy outcome prediction.

Objectives: We introduce a large language model (LLM)-based framework that combines scGPT, a foundation model for single-cell data, and MixedBread, a general-purpose language model. This framework generates unified scRNA-seq embeddings for cell type classification and builds cell-type-specific classifiers to predict pathologic complete response (pCR) to therapy.

Methods: Using scRNA-seq data from Bassez et al. (2021), including 157,760 cells and 25,228 genes from 29 BC patients, we generated embeddings with scGPT and MixedBread, followed by Louvain clustering. Cell-type-specific marker genes were identified and annotated using the Blueprint/ENCODE reference. These markers were used to train XGBoost and SVM-based pCR classifiers, validated on the independent I-SPY2-990 trial bulk RNA-seq dataset with 987 BC patients, 19,134 genes.

Results: The integrated embeddings revealed nine predicted cell types, which aligned well with established breast cancer cell populations. Marker genes derived from these specific cell types varied in their ability to predict pCR, with area under the curve (AUC) values ranging from 0.34 to 0.95. Notably, marker genes from macrophage cells achieved the highest predictive performance, with an AUC of 0.78 on the Bassez scRNA-seq dataset and 0.95 on the I-SPY2-990 bulk RNA expression data.

Conclusions: Our LLM-driven framework enables accurate cell type classification and improves treatment response prediction via cell-type-specific biomarkers, advancing precision oncology for breast cancer.

Keywords: Breast cancer, Single-cell RNA sequencing, Large language models, Cell-type classification, Pathologic complete response prediction

122

Improving Epigenetic Age Estimation by Combining Epigenetic Clocks

Denitsa Vasileva*[1], Celia M. T. Greenwood[2], Denise Daley[1]
[1]Center for Heart Lung Innovation, Faculty of Medicine, University of British Columbia, Vancouver, Canada; [2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada

Introduction: The epigenetic clock leverages DNA methylation (DNAm) to calculate epigenetic age (EA) and may have potential as a biomarker for age-related conditions (e.g. Alzheimer's disease). Currently available clocks demonstrate age and sex-specific biases. For example, the Horvath pan-tissue and skin & blood clocks consistently underestimate EA at ages 65+ while the centenarian clock is less accurate in younger adults and in males.

Objective: Combining clocks and biological sex to develop a more accurate EA predictor

Methods: Using the Illumina EPIC array, DNAm was assayed in blood samples from participants in the Canadian Longitudinal Study on Aging (CLSA, n=1478, ages:45-86 years, 50.54% Female) and the Alzheimer's Disease Neuroimaging Initiative (ADNI, n=1905, ages:55-95.62 years, 44.25% F). In each study, epigenetic age was calculated using the centenarian, pan-tissue and skin & blood clocks. In the CLSA samples (i.e. training set), an elastic net regression with 10-fold cross-validation was carried out: Chronological Age ~ Pan-tissue+Skin&Blood+Centenarian EA+Sex. The resulting model was used to calculated EA in ADNI (i.e. testing set). Accuracy (absolute error (AE= epigenetic-chronological age) and underestimation (ratio=epigenetic age/chronological age) were assessed.

Results: In ADNI, the combined algorithm had a lower mean AE than the three individual clocks overall (2.86± 2.39 years) and in the oldest decade (3.47±2.77). The mean ratio of this approach was 0.998±0.05 and was closer to 1 (i.e. epigenetic=chronological age) than the available clocks.

Conclusions: Combining the results of three epigenetic clocks and sex accounts for the age and sex biases of individual clocks and improves EA accuracy.

123

Multidimensional Analyses of Pedigree, Epidemiologic, Proteomics, and Transcriptomics Data Provide Etiologic Clues for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

Roxana Moslehi*[1], Anil Kumar[1], Amiran Dzutsev[2]
[1]School of Public Health, University at Albany, Albany, New York, United States of America; [2]National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America

Background: Myalgic encephalomyelitis (ME)/chronic fatigue syndrome (CFS) is a complex disabling disorder with no known etiology or approved treatment. We conducted a molecular epidemiologic study to identify risk factors and biologic mechanisms for ME/CFS.

Methods: Our clinic-based case-control study involved 60 ME/CFS patients and 61 healthy controls compared with respect to the prevalence of autoimmune disease (AID) and cancer among their first-degree relatives, prevalence of epidemiologic factors, serum levels of 48 cytokines, and gene expression profiles. We used conventional and machine learning approaches to analyze the data and calculate the relevant associative and predictive metrics.

Results: First-degree relatives of ME/CFS cases were more likely than those of the controls to have AID [Relative Risk (RR)=3.52, *P*=0.0014) and early-onset (diagnosed <60 years of age) cancer (RR=2.24, *P*=0.034) including blood cancers (*P*=0.047). Comparison of epidemiologic factors identified several risk factors such as history of allergies requiring medication [Odds Ratio (OR)=6.00, *P*<0.0001) and exposure to contaminants (OR=4.35, *P*=0.0002) among others. We identified a cytokine signature of ME/CFS, which classified patients with AUC>0.75, sensitivity>80%, and specificity>70% at an optimized threshold in all three tested machine learning models including XGBoost. Key cytokine predictors included IL-27, IP-10, RANTES, and Fractalkine among others. Whole blood RNA-seq analysis identified 115 differentially expressed genes with FDR<0.25 belonging to biologic pathways relevant to infectious diseases and neurologic disorders.

Conclusions: Findings from our multidimensional analysis identify previously unreported risk factors for ME/CFS, links with AID and early-onset cancer, and potential biologic mechanisms, thus providing etiologic clues and druggable targets for treatment.

Keywords: ME/CFS, RNA-seq, Cytokines, Pedigrees, Molecular Epidemiology

124

Hvac: A Fast Hierarchical Bayesian Model With Functional Annotation for Accurate Cross-Ancestry Polygenic Prediction

Zhonghe Shao, Xingjie Hao*

Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

Polygenic risk scores (PRS) face critical limitations in cross-ancestry portability due to Eurocentric biases in genome-wide association studies (GWAS) and population-specific genetic architectures. We present HVAC, a hierarchical Bayesian framework that enhances PRS generalizability through integrated functional annotation guidance and cross-population calibration. Our three-tiered architecture comprises: (1) Ancestry-specific inference using Gaussian mixture priors informed by functional annotations to model population-specific genetic effects under local linkage disequilibrium (LD) patterns; (2) Cross-ancestry calibration via dynamic beta-shrinkage priors that bidirectionally propagate uncertainty between populations; and (3) Polygenic synthesis combining population-specific and cross-ancestry signals through genotype-optimized ensemble weighting.

In comprehensive simulations across African (AFR), East Asian (EAS), and South Asian (SAS) populations, HVAC achieved mean R^2 improvements of 34.2% over LDpred-funct (annotation-integrated method) and 30.6% over BridgePRS (Bayesian hierarchical counterpart), with 52.4% gains in AFR populations compared to MUSSEL (Bayesian correlation-structured method). Real-data analyses on seven UK Biobank traits (five quantitative, two binary) demonstrated HVAC's superior generalizability, attaining up to 57.4% higher R^2 against PRS-CSx (MCMC-based method) and substantial improvements over EUR-derived models (e.g., 89.5% higher R^2 for LDL prediction in AFR populations). HVAC maintained robust performance under varying polygenicity, genetic correlations, and discovery sample sizes. Computationally, variational inference implementation reduced runtime by 86.9% versus PRS-CSx (peak RAM: 11.5 GB) through hybrid code optimization and precomputed LD matrices.

By harmonizing population-specific LD patterns, functional annotations, and cross-ancestry uncertainty propagation, HVAC advances equitable precision medicine, particularly for underrepresented populations. Its computational scalability position it as a transformative tool for global genomic research and clinical translation.

126

Assessing Clustering Performance and Impact on Allele Frequency Estimation Across Sampling Schemes in Fine-Scale Genetic Simulations

Mael Guivarch[1], POPGEN Study Group[2], Emmanuelle Génin[1,3], Anthony F. Herzig*[1], Aude Saint Pierre*[1]

[1] University of Brest, Inserm, Brest, France; [2]Inserm, Brest, France; [3]Centre Hospitalier Universitaire Brest , Brest, France

** Equal Contribution*

Introduction: With the fast expansion of genomic datasets across large geographic areas, understanding fine-scale population structure is crucial to estimate allele frequencies and address stratification. Clustering approaches are widely used to identify such structures by grouping individuals based on genetic similarities. However, choosing the right method from a range of algorithms can be especially challenging.

Methods: We present a comparative study of clustering algorithms, examining how spatial stratification and different subsampling schemes affect clustering results. Specifically, we consider how sampling schemes influence rare-allele sharing across clusters. All methods are applied to simulated datasets designed to closely mimic the fine-scale genetic structure observed in POPGEN, a reference panel of whole genomes representative of French regions.

Results: Building on previous studies, we simulate genetic data under controlled demographic scenarios using a stepping-stone model. This controlled simulation framework enables us to assess clustering accuracy by comparing clustering results to a gold-standard partition. A particular emphasis was placed on evaluating how clustering differences affect allele frequency estimation, both in simulations and in real data from POPGEN.

Conclusion: Our study provides practical guidance for interpreting fine-scale population structure and for selecting appropriate methods. It also addresses the issue of the granularity of allele frequency estimation and the interpretation of allele frequency differences.

Funding: This work was supported by The French Ministry of Research and Innovation in the framework of the French initiative for genomic medicine (Plan France Médecine Génomique 2025; https://www.aviesan.fr/mediatheque/fichiers/version-anglaise/actualites-en/genomic-medicine-france-2025-web). The CONSTANCES cohort benefits from grant ANR-11-INBS-0002 from the French National Research Agency.

127

Impact of Study Sample Composition on Supervised Admixture Modelling

Anthony F. Herzig[1], Maël Guivarch[1], Marc Gros La Faige[1], Gaëlle Marenne1, POPGEN Study Group1, Emmanuelle Genin[1,2]

[1]Inserm, University of Brest, Brest, France; [2]Centre Hospitalier Universitaire Brest, Brest, France

Unsupervised admixture component estimation is a prevalent analysis in exploratory population genetics; but is known to be sensitive to sample size and composition and hence results should be incorporated with care. The aim is to model a given sample of individual genomes as issuing from K hypothetical populations (with K to be chosen by the analyser); with each individual being assigned a set of K admixture components representing the contributions of the K hypothetical population to their genome. The distribution across the whole study-sample of these components describes the broad population structure in the dataset.

When genomes from well-established reference populations are in-hand, supervised admixture can be performed and this is argued to be a more robust approach that is more straightforward to interpret. In this circumstance, the composition of the reference population data is known to greatly impact the final results. Here however, we examine the role of the composition and size of the study sample during supervised admixture by analysing 9,598 individuals from the general population in France either all at once, in groups, or one-by-one against reference populations from

the 1000G and HGDP. We show that by changing the study sample composition, quite different results may be obtained. Therefore, we show that internal population structure within the study sample plays an important role even when the admixture analysis is supervised using known reference samples. We back up our observations through simulation studies with both genetic and spatial data under a stepping-stone model on a five-by-five deme grid.

128

Identifying Multi-Allelic Quantitative Trait Loci Using Empirical Haplotypes

Katelyn A. McInerney*[1], Karen L. Mohlke[1], Michael I. Love[1,2], William Valdar[1]

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Understanding the genetic basis of complex traits remains a critical challenge in biology and genetics across humans, animals, plants, and agriculture. Despite advances in genome-wide association studies and quantitative trait locus mapping, the pathways from genotype to phenotype remain largely unknown for most complex traits. Current quantitative genetic approaches predominantly rely on SNPs as the input variable for modeling these relationships. Although such SNP-based methods have enabled significant discoveries, they also have limitations. A promising alternative is to use haplotypes, allowing more concise modeling of local multi-SNP combinations without requiring a search for complex, higher-order SNP-SNP interactions. In order to identify useful haplotypes for association mapping, we have developed *IDHaplos*, which uses a hierarchical clustering algorithm on the genotyping matrix to group individuals into allele groups. We use the Hamming distance to construct the distance matrix, and we cut the tree based on a pre-specified dissimilarity threshold. We then model these haplotype alleles in a linear mixed model and compare performance with other haplotype clustering approaches (*BigLD, HaploBlocker, tskit*) and SNP-based approaches in simulated, model organism, and human data.

Keywords: haplotypes, statistical genetics, association mapping

129

Mendelian Randomization Study of the On-Target Effects of Long-Term Aromatase Inhibitor Use

Devleena Ray*[1], Philip Haycock[1], Richard Martin[1], James Yarmolinsky[2], Karl Smith-Byrne[3]

[1]University of Bristol, Bristol, United Kingdom; [2]Imperial College London, London, United Kingdom; [3]University of Oxford, Oxford, United Kingdom

Anastrozole, an aromatase inhibitor, was recently approved for breast cancer prevention in high-risk women post IBIS-II trial, that demonstrated reduced breast cancer risk alongside decreased bone density and increased hypertension risk. This study extends on IBIS-II findings by investigating long-term effects of aromatase inhibitor use, focusing on its safety profile and repurposing opportunities. The rs727479 variant, that mimics aromatase inhibition, was used as a proxy to estimate effects on selected outcomes. Instrument validation compared genetically proxied aromatase inhibition effects with IBIS-II established effects, like reduction in oestrogen-receptor positive (ER+) breast cancer risk and bone mineral density (BMD). Secondary IBIS-II outcomes, such as hypertension, were also investigated alongside outcomes lacking definitive associations (e.g., other cancer types) to explore repurposing opportunities. A hypothesis-free phenome-wide association study (PheWAS) was performed to identify novel associations, with top findings examined in clinically relevant subgroups. Genetically proxied aromatase inhibition was associated with reduced ER+ breast cancer risk (OR = 0.78, [95% CI 0.67, 0.92]), decreased heel BMD (β=-0.32, [95% CI -0.36, -0.28]) and increased hypertension risk (β =0.008, [95% CI -0.005,0.023]). In secondary analysis, a reduction in endometrial cancer risk (OR = 0.30 [95% CI 0.21, 0.42]) was observed. The PheWAS revealed reduced risk of endometrial cancer, postmenopausal bleeding and uterine polyps, among other findings. Our findings corroborate associations observed in IBIS-II, demonstrating Mendelian randomization's utility in recapitulating known effects of preventative cancer therapies. The observed reduction in endometrial cancer risk suggests anastrozole's repurposing potential for endometrial cancer prevention, warranting further validation.

132

Direct and Indirect Genetic Effects in the Associations Between Maternal Health and Autism: A Novel, Family-Based Method

Elias Speleman Arildskov[1], Vahe Khachadourian[2], Diana Schendel[3,4,5], Jakob Grove[1], Stefan Hansen[6], Magdalena Janecka[2,7]

[1]Department of Biomedicine, Aarhus University, Aarhus, Denmark; [2]Department of Child and Adolescent Psychiatry, NYU Grossman School of Medicine, New York, New York, United States of America; [3]A.J. Drexel Autism Institute, Drexel University, Philadelphia, Pennsylvania, United States of America; [4]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Aarhus, Denmark; [5]National Centre for Register-Based Research, Aarhus BSS, Aarhus University, Aarhus, Denmark; [6]Department of Public Health, Aarhus University, Aarhus, Denmark; [7]Department of Population Health, NYU Grossman School of Medicine, New York, New York, United States of America

We previously demonstrated that the associations between maternal health in pregnancy and autism are attributable to familial confounding. However, the contribution of direct (DGE) and indirect genetic effects (IGE) to these associations remains unknown. Current methods to estimate IGEs rely on genotype data from family trios. We propose a novel approach to delineate IGEs and DGEs effects using pedigree data and apply it in Danish national registry data.

The source population included 1.13 million children born in Denmark 1998-2015 and their parents. We selected

maternal parallel (related through mother and her sister) and cross cousin (related through mother and her brother) pairs. The outcome was a diagnosis of autism in cousins of the index child; the exposures were pregnancy diagnoses in the mother of the index child (i.e., aunt of children in whom we measure the outcome)- enabling us to index potential genetic liability, rather than the direct impacts of the pregnancy diagnosis. We assumed both cross and parallel cousins share DGEs, but only maternal parallel cousins share prenatal IGEs, as their mothers are sisters (cross cousins' mothers are unrelated).

Diagnosis in the index mother was associated with increased risk of autism in both cross and parallel cousins–indicating DGEs–for e.g., personality disorders and epilepsy. Some diagnoses were associated with autism with significantly different effects in cross vs. parallel cousins–indicating IGEs–for e.g., multiple gestation, depression, and disorders of connective tissue. Results were robust in sensitivity analyses including changes in exposure period and adjustment for additional non-genetic effects.

Keywords: indirect genetic effects; family designs; autism

133

Genetic Regulation of Protein Expression in Prediabetes and Type 2 Diabetes

Archit Singh*[1,2,3], Marlene Ganslmeier*[4], Ozvan Bocher[1], Mauro Tutino[1], Young-Chan Park[1], Norbert Stefan[4,5,6], Andreas Fritsche[4,5,6], Reiner Jumpertz von Schwartzenberg[4,5,6], Eleftheria Zeggini[1,7#], Andreas Birkenfeld[4,5,6#]

[1]Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [2]Munich School for Data Science (MUDS), Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [3]Doctoral Program of Experimental Medicine and Health Sciences, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany; [4]Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the University of Tuebingen, Tuebingen, Germany; [5]Department of Diabetology, Endocrinology, Nephrology, University of Tuebingen, Tuebingen, Germany; [6]German Center for Diabetes Research (DZD E.V.), Neuherberg, Germany; [7]TUM School of Medicine and Health, Technical University of Munich and Klinikum Rechts der Isar, Munich, Germany

*These authors are joint first authors.

#These authors are joint last authors.

Prediabetes and type 2 diabetes (T2D) are characterized by insufficient insulin secretion and poor sensitivity, leading to complications and increased mortality. Genetic predisposition is known to play a key role in these metabolic disorders. Not all individuals with prediabetes develop T2D, highlighting the importance of understanding the molecular mechanisms involved in this progression.

To investigate these differences, we utilized genomics and plasma proteomics data from 450 individuals enrolled in the prediabetes lifestyle intervention study. Through differential protein expression analysis, we identified 185 out of 2523 proteins to be differentially expressed in 321 individuals with prediabetes compared to 88 individuals with T2D.

Using a mixture model framework, we identified 703 shared protein quantitative trait loci (pQTL) effects suggesting overlapping genetic regulation of proteins. Further, we identified three differential pQTLs i.e., genetic variants with significant effect in only one condition, near genes coding for proteins SELE, SCLY, and ACP5, showing significant positive effects on protein levels in prediabetes but not in T2D. These proteins were also found to be differentially expressed in prediabetes compared to T2D. Notably, knockout studies of *SCLY* in mice link it to glucose and lipid homeostasis, implicating pathways associated with fatty liver disease and glucose intolerance. Additionally, we identified 20 shared pQTLs with opposite effects between prediabetes and T2D, underscoring differences in genetic regulation between the two states.

In conclusion, our study provides new insights into the molecular mechanisms underlying progression from prediabetes to T2D and a better understanding of the role of their genetic regulation.

134

An Alternative Analysis Method for Transcriptome-Wide Association Studies

Lei Fang, Wei Pan

Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota, United States of America

Two-stage least squares (2SLS) is by default applied to infer a putative causal association between an exposure, such as a gene or a protein, with an outcome such as a complex disease or trait, in transcriptome- or proteome-wide association studies (TWAS/PWAS). In a typical two-sample setting for TWAS/PWAS, the stage 1 sample size is much smaller than that of stage 2. To reduce the resulting attenuation bias and estimation uncertainty in stage 1 and boost statistical power of the conventional TWAS, we propose a new method, called reverse two-stage least squares (r2SLS): instead of imputing a gene's expression (using genetic variants as instrumental variables, IVs) in stage 1 and then testing the association between the imputed expression and the observed outcome in stage 2 in the conventional 2SLS approach, we propose predicting the outcome (using IVs) and testing the association between the predicted outcome and the observed gene expression. Theoretically, we establish that the r2SLS estimator is asymptotically unbiased with a normal distribution. We also show theoretically when 2SLS and r2SLS are asymptotically equivalent and when r2SLS is asymptotically more efficient than 2SLS. We use simulations and three real data examples based on the GTEx gene expression data, UKB-PPP proteomic data and several GWAS summary datasets to demonstrate some advantages of r2SLS over 2SLS, including possibly better type I error control, higher statistical power and robustness to weak IVs.

135

Deep Learning Based Multivariable Instrumental Variable Regression for Nonlinear Causal Inference With Application to TWAS

Ruoyu He[1], Chunlin Li[2], Zhaotong Lin[3], Xiaotong Shen[4], Wei Pan[4]

[1]Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Department of Statistics, Iowa State University, Ames, Iowa, United States of America; [3]Department of Statistics, Florida State University, Tallahassee, Florida, United States of America; [4]Department of Statistics, University of Minnesota, Minneapolis, Minnesota, United States of America

Instrumental Variable (IV) regression is a foundational approach to causal inference, particularly in contexts where randomized experiments are not feasible. An important and novel application of IV regression is to Transcriptome-Wide Association Studies (TWAS) to identify causal genes (as exposures) for a trait (as an outcome) using genetic variants as IVs. However, current TWAS applications are exclusively based on either linear models or univariable methods, overlooking potential nonlinear gene-trait relationships or pleiotropic/confounding effects of genetic variants. Here we introduce MV-DeLIVR, a robust multivariable IV regression method based on a neural network, accounting for invalid IVs (e.g., due to horizontal pleiotropy of genetic variants) while estimating possibly more complex exposure-outcome (e.g., gene-trait) associations. We offer some theoretical support for the proposed method. We apply the new method with the GTEx gene expression data and UK Biobank GWAS data to conduct simulations and perform real data analyses, focusing on high-density Lipoprotein cholesterol (HDL) as a target trait. Our results demonstrate that MV-DeLIVR controls the Type I error rate at a nominal level and possesses high power in most scenarios, offering significant improvements over existing univariable methods..

136

Class-Specific Lipid Dysregulation and Cardiometabolic Risk in Severe Obesity

Mohammad Yaser Anwar[1], Thy Duong[2], Zhaotong Lin[3], Alexandra B. Palmer[1], Jessica Sprinkles[4], Wanying Zhu[5,6], Rashedeh Roshani[5,6], Elizabeth G. Frankel[5,6], Joshua Landman[5,6], Mariaelisa Graff[1]

[1]Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [2]Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia; [3]Department of Statistics, Florida State University, Tallahassee, Florida, Unites States of America; [4]Department of Nutrition, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [5]Department of Medicine, Division of Genetic Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [6]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Severe obesity (SevO; BMI ≥40 kg/m²) is rapidly increasing worldwide, disproportionately affecting minority populations, yet remains understudied in mechanistic and omics research. Lipid metabolism is central to obesity-related cardiometabolic disease (CMD), but the relationship between molecular lipid species and SevO is poorly characterized. We analyzed 573 participants from the Cameron County Hispanic Cohort (CCHC) with fasting plasma lipidomic and genetic data, comparing SevO cases to normal-weight controls. A total of 830 circulating lipid species across 49 classes were quantified. Associations were assessed using logistic regression, orthogonal projections to latent structures discriminant analysis (OPLS-DA). Mendelian randomization (MR) and Pearson correlations explored relationships between influential lipids (variable influence on projection >1), CMD traits, and body composition. SevO participants had significantly worse cardiometabolic profiles than controls. Lipidomic profiling revealed widespread alterations: elevated shorter, saturated, and monounsaturated triacylglycerols, and reduced lysophospholipids, plasmalogens, cholesteryl esters, and long-chain lipids. OPLS-DA identified over 300 predictive lipid species. MR implicated specific triacylglycerol species as potentially causal. Influential lipids correlated with insulin resistance, liver fibrosis, adiposity, and HDL-C. SevO is marked by extensive, class-specific lipidomic dysregulation with strong cardiometabolic links. Distinct triacylglycerols emerged as key discriminators and potential drivers. These findings highlight a novel lipidomic signature of SevO and emphasize the need to explore its genetic, dietary, and mechanistic determinants.

137

Univariate and Multivariate Genome Wide Association Studies Reveal New Variants Associated With Cytokine Levels in Asthmatic Families

Jessica Martinez*[1], Lucie Troubat[1], Raphaël Vernet[1], Florence Demenais[1] & Emmanuelle Bouzigon[1]

[1]Université Paris Cité, Inserm, HealthFex, Group of Genomic Epidemiology of Multifactorial diseases, Paris, France

Cytokines play a central role in mediating inflammatory and immune responses in asthma.

To characterize the genetic factors influencing key cytokines involved in asthma physiopathology, we conducted univariate and multivariate GWAS (genome-wide association studies) of the serum levels of six cytokines (IL1RA, IL5, IL8, IL10, IL13, and TNFA) measured in 759 subjects from the Epidemiological Study on the Genetics and Environment of Asthma (EGEA) by using linear mixed models implemented in GEMMA software and 1000 Genomes imputed SNPs. Two multivariate GWAS were performed by combining antiinflammatory cytokines (IL-1RA and IL-10) or proinflammatory cytokines (IL5, IL8, IL-13, and TNFA).

Single phenotype analyses identified a genome wide significant locus located in 7p14 region (rs58014320 in *ELMO1* gene, $1.1\times10^{-8} \leq$P value$\leq 4.8\times10^{-8}$) associated with several cytokines (IL5, IL8, IL10, and IL13). This signal was strengthened in multivariate GWAS of antiinflammatory cytokines (P value=3.9×10^{-9}). Multivariate GWAS allowed detection of three additional genome wide significant loci: 1q24 (rs10919178 in *F5*, P value=1.2×10^{-9}) associated with

the combination of IL5, IL8, IL13 and TNFA; and 8q24 (rs76438982 in *ZHX2*, P value=$3.7x10^{-8}$), and 13q13 (rs7335719 between *RFC3* and *NBEA*, P value=$1.7x10^{-8}$) associated with the combination of IL1RA and IL10.

These findings highlight four novel genomic regions associated with variations in serum levels of six key cytokines, and further analyses are underway to explore their relevance in asthma.

Funding: ANR-20-CE36-0009

Keywords: Cytokines, Asthma, GWAS, Multivariate GWAS

139

Multi-trait Genome-Wide and Gene-Based Analyses Implicate Coagulation and Vascular Smooth-Muscle Pathways in Spontaneous Coronary Artery Dissection

T-E. Berrandou[1], A. Georges[1], D. Speed[2] and N. Bouatia-Naji[1]
[1]Université Paris Cité, Paris Cardiovascular Research Center, Inserm, Paris, France; [2]Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark

Spontaneous coronary artery dissection (SCAD) accounts for up to one-third of acute coronary syndromes in women < 50 years, who typically lack classical cardiovascular risk factors. Its imaging-based diagnosis is difficult, leaving the condition under-diagnosed and under-studied, and single-trait genome-wide association studies (GWAS) insufficiently powered.

We harmonized seven GWAS—SCAD, fibromuscular dysplasia, intracranial aneurysm, cervical artery dissection, migraine, coronary artery disease and thoracic aortic aneurysm/dissection—and used GAUSS to impute only variants missing in individual studies. After quality control, 4.94 million autosomal SNPs were shared across all traits. Multi-Trait Analysis of GWAS (MTAG) boosted the SCAD effective sample size from 6,356 to 9,114 participants (+43 %). Gene-level associations were tested with LDAK-GBAT (UK Biobank European reference). Tissue-specific eQTL, chromatin and pathway enrichments provided functional context.

MTAG identified eight new SCAD loci absent from the original GWAS, implicating *GGCX, COL6A3, EDNRA, SLC39A13, LTBP3, BCAR1, SLC39A8* and *SERPINA1.* Variant enrichments peaked in coronary-artery smooth-muscle and fibromyocyte open chromatin. LDAK-GBAT yielded 45 Bonferroni-significant genes; 19 of these (e.g. *ABO, BMP8A, PMS2, ITGA1*) lay in regions without genome-wide significant SNPs, demonstrating added power beyond SNP-based approach. Gene-set analysis highlighted vitamin-K-dependent γ-carboxylation, extracellular-matrix remodeling and smooth-muscle contraction pathways, coherently linking coagulation and vascular integrity to SCAD susceptibility.

Combining multi-trait GWAS with gene-based testing and functional annotation markedly expands the genetic landscape of SCAD and pinpoints a coagulation–smooth-muscle axis centered on *GGCX* and *ABO*. These discoveries provide new mechanistic insights and nominate promising targets for functional validation and precision risk prediction.

Keywords: Complex traits; Cardiovascular; Multi-traits analysis; Gene-based analysis ; functional annotations

141

Using Kolmogorov-Arnold Networks, an ExplainableAI Method, to Integrate Multiomics and Interpret the Relationships Underpinning Complex Traits

James A. Temple*[1], Gayathry Krishnamurthy[1], Nicholas M. Fountain-Jones[2], Shaun P. Brennecke[3], Peter J. Meikle[4], Rae-Chi Huang[5], Fuling Chen[6], John Blangero[7], Eric K. Moses[1], Phillip E. Melton[1,8]
[1]Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania, Australia; [2]School of Natural Sciences, University of Tasmania, Hobart, Tasmania, Australia; [3]Pregnancy Research Centre, Department of Maternal-Fetal Medicine, The Royal Women's Hospital, Parkville, Victoria, Australia; [4]Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia; [5]Nutrition & Health Innovation Research Institute, Edith Cowan University, Joondalup, Western Australia, Australia; [6]The International Centre for Radio Astronomy Research, University of Western Australia, Crawley, Western Australia, Australia; [7]Department of Human Genetics, South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Edinburg, Texas, United States of America; [8]School of Global and Population Health, University of Western Australia, Crawley, Western Australia, Australia

Multiomics datasets have been shown to be informative for a range of complex traits, such as cardiometabolic conditions. Traditionally, each omics layer's contribution has been studied separately and then correlated. However, if omics were integrated, the complex relationships between these layers may allow for more accurate modelling. Multilayer perceptron (MLP) algorithms are an appropriate approach to improve predictive accuracy but provide limited interpretability. We hypothesise that explainable artificial intelligence (xAI) methods could offer improved interpretability in findings for high-dimensional data versus MLPs. In this study we focused on a recently developed deep learning algorithm, Kolmogorov-Arnold Network (KAN), a method where feature rankings are calculated intrinsically. To assess KANs' utility for multiomics, we used genome-wide array and lipidomic data from an Australian population cohort (n=4190) in a two-step process. First, we used 596 lipid species to predict biological sex as assigned at birth. We found that KANs offered an improved performance over random forest for accuracy (77.3 vs. 73.3), R2 (0.36 vs. 0.28), and area under the curve (AUC, 0.87 vs. 0.84). The features ranked as important (largest contribution from phosphatidylethanolamine lipid class at 13%) had non-linear relationships to outcome, but did include those shown to have high significance in previous work that used regression models. Next, we investigated hypertensive disorders of pregnancy in a subset (n=1580), combining array and lipidomic data. These results included an AUC of 0.68 and confirmed the omics had been integrated. Our work indicates xAI to be an emerging methodology for better multiomic integration and interpretation.

143

1-NN Imputation of Missing DNA-Methylation Values

Christelle Kemda Ngueda, Julia Palm, Flavia Remo, André Scherag, Lutz Leistritz

Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Jena, Germany

Introduction: DNA methylation is an epigenetic mechanism influenced by genetic variation [1, 2]. It represents a health risk factor and is of particular interest in medicine because of its reversible nature [3, 4]. For various purposes, DNA-methylation signals can be rapidly measured using either microarray or sequencing protocols in parallel. However, missing values are still a common disadvantage of these technologies, and can significantly affect downstream analyses [5]. To remedy this problem, several approaches have been proposed from both statistics and computer sciences. They all have in common that they can be applied to both DNA methylation microarray and sequencing data, and that they require information from at least two samples.

We propose a time and cost-effective imputation method for replacing missing DNA-methylation values in a single patient methylome, i.e. a method that relies on the personalized medicine idea. Basically, the method replaces a missing value by an available value of its nearest neighbouring CpG.

The proposed method applied to a single methylome yielded an average root mean square error (RMSE) RMSE = 0.27 in β-value units (95%-CI: [0.26, 0.28]) based on publically available 450K BeadChip data set of 3,402 individuals with β-value [6]. It is possible to consider the affiliation of CpGs to CpG islands when imputing missing methylation values. This improves the imputation accuracy. In addition, the imputation accuracy depends on the density of CpG sites on DNA-methylation microarrays and is higher the denser CpG sites are.

The full article has been accepted for publication by BMC Bioinformatics.

References:

1. Hawe JS, Wilson R, Schmid KT, Zhou L, Lakshmanan LN, Lehne BC, Kuhnel B, Scott WR, Wielscher M, Yew YW *et al*: **Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function**. *Nat Genet* 2022, **54**(1):18-29.

2. Villicana S, Bell JT: **Genetic impacts on DNA methylation: research findings and future perspectives.** *Genome Biol* 2021, **22**(1):127.

3. Gupta MK, Peng H, Li Y, Xu CJ: **The role of DNA methylation in personalized medicine for immune-related diseases**. *Pharmacol Ther* 2023, **250**:108508.

4. Jin Z, Liu Y: **DNA methylation in human diseases**. *Genes Dis* 2018, **5**(1):1-8.

5. Dhingra R, Kwee LC, Diaz-Sanchez D, Devlin RB, Cascio W, Hauser ER, Gregory S, Shah S, Kraus WE, Olden K *et al*: **Evaluating DNA methylation age on the Illumina MethylationEPIC Bead Chip.** PLoS One 2019, **14**(4):e0207834.

6. Xiong Z, Li M, Yang F, Ma Y, Sang J, Li R, Li Z, Zhang Z, Bao Y: **EWAS Data Hub: a resource of DNA methylation array data and metadata.** *Nucleic Acids Res* 2020, **48**(D1):D890-D895.

144

Genome-wide Association Study of Post COVID-19 Syndrome in a Population-based Cohort in Germany

Anne-Kathrin Ruß[1], Lennart Reinke[2], Alin Viebke[2], David Ellinghaus[3], Bärbel U. Foesel[4], Christian Gieger[4], Michael Krawczak[1], Jan Heyckendorf[2,5,6], Thomas Bahmer[2,5], on behalf of the NAPKON Study Group

[1]Institute of Medical Informatics and Statistics, Kiel University, University Medical Center Schleswig-Holstein, Kiel, Germany; [2]Department of Internal Medicine I, University Medical Center Schleswig-Holstein, Kiel, Germany; [3]Institute of Clinical Molecular Biology, Kiel University, University Medical Center Schleswig-Holstein, Kiel, Germany; [4]Institute of Epidemiology, Research Unit of Molecular Epidemiology, Helmholtz Munich - German Research Center for Environmental Health, Neuherberg, Germany; [5]Airway Research Center North (ARCN), German Center for Lung Research (DZL), Großhans-dorf, Germany; [6]Leibniz Lung Clinic, Kiel, Germany

Post-COVID Syndrome (PCS), or Long-COVID is diagnosed when symptoms following a coronavirus 2019 infection (COVID-19) persist for 12 weeks or more. Despite the waning of the pandemic by 2024, PCS continues to pose a substantial global health challenge. To explore its potential genetic underpinnings, we conducted a genome-wide association study (GWAS) in 2,247 SARS-CoV-2-infected individuals from the population-based, multi-centre COVIDOM cohort in Germany. We analyzed three severity scores, each reflecting the cumulative presence of up to 12 symptom clusters assessed at least six months after infection. For the purposes of this study, these scores were converted into binary traits based on their median values.

Among 6,383,167 single nucleotide polymorphisms (SNPs) examined, several showed associations with PCS phenotypes, though none reached the conventional genome-wide significance threshold ($p < 5\times10^{-8}$). The most notable signal, rs9792535 ($p = 6.6\times10^{-8}$), is located near NEK6, PSMB7, and ADGRD2, genes not previously linked to PCS. Another variant of interest, rs10893121 ($p = 2.5\times10^{-6}$), lies in an olfactory receptor gene cluster, offering a plausible link to PCS symptoms such as altered smell and taste.

Additional suggestive associations were observed near genes implicated in host-virus interactions, including CHD6 (viral repression), SLC7A2 (immune activation), and ARHGAP44 (viral release). However, most signals did not map to genes currently assumed to influence PCS-related pathways. The generally weak associations observed reinforce the notion that PCS is genetically complex, most likely due to multiple variants with small effects, which is characteristic of most multifactorial conditions.

Keywords: GWAS, Post-COVID Syndrome

145

Polygenic Risk Scores and Arrhythmic Risk in Dilated Cardiomyopathy: Insights From Common Variant Associations and Prognostic Modeling

Ilaria Gandin[1], Maddalena Rossi[2], Paolo Dalena[3,4], Alessia Paldino[2], A. Pio d'Adamo[1,3], Giulia Barbati[1], Marco Merlo[1,2], Matteo Dal Ferro[2], Gianfranco Sinagra[1,2]

[1]Department of Medical Sciences, University of Trieste, Trieste, Italy; [2]Cardiovascular Department, Azienda Sanitaria Universitaria Giuliano Isontina (ASUGI); [3]Institute for Maternal and Child Health-IRCCS "Burlo Garofolo", Trieste, Italy; [4]Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy

Dilated cardiomyopathy (DCM) is a primary myocardial disorder with a strong genetic component, though pathogenic variants are detected in under 40% of cases and show wide clinical variability. Recent genome-wide association studies (GWAS) have investigated the contribution of common genetic variants to DCM susceptibility and cardiac MRI-derived left ventricular measurements, introducing polygenic risk scores (PRS) as potential predictive tools.

This study analyzed 334 individuals with DCM and 718 healthy controls. PRS were derived from GWAS for: reduced left ventricular ejection fraction (PRS-LVEFneg), end-diastolic volume (PRS-LVEDV), end-systolic volume (PRS-LVESV), stroke volume (PRS-SV), and DCM risk (PRS-DCM). Prognostic outcomes included: all-cause mortality; sudden cardiac death, sustained ventricular tachycardia, or ventricular fibrillation (SCD/VT/VF); heart failure–related death, heart transplantation, or left ventricular assist device implantation (DHF/HTx/VAD). Time-to-event outcomes were assessed using cumulative incidence curves and cause-specific Cox regression. Sensitivity analyses explored potential unmeasured confounding.

PRS-DCM was significantly higher in DCM patients compared to controls (OR=1.45 per SD, p<.001), as were PRS-LVEFneg (OR=1.47, p<.001), PRS-LVEDV (OR=1.22, p=.003), PRS-LVESV (OR=1.50, p<.001), and PRS-SV (OR=1.22, p=.003). Pathogenic variants were found in 39.8% of cases. Over a median follow-up of 109 months, 18% of patients died, 26% had SCD/VT/VF, and 14% experienced DHF/HTx/VAD. Only PRS-DCM was associated with SCD/VT/VF risk (HR=0.69, p=.002), even after adjustment. No associations were found between other PRS and outcomes. Higher PRS-DCM levels appeared to be protective against severe arrhythmic events, raising questions about the causal nature of this observation. Sensitivity analyses suggested that a moderate unmeasured confounder could explain the finding.

Keywords: Polygenic Risk Scores, Dilated Cardiomyopathy, Arrhythmic Risk, Arrhythmic Events

146

Improving Causal Effect Estimation in Multi-Ancestry Multivariable Mendelian Randomization With Transfer Learning

Yihe Yang, Xiaofeng Zhu*

[1]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America

Multivariable Mendelian randomization (MVMR) is a widely used approach to estimate the causal effects of exposures on disease outcomes. However, its applications are predominantly limited to European populations due to the much larger sample sizes available in European GWAS, which provide greater statistical power. Joint modeling of multiple ancestral populations has been suggested to improve causal inference in MR, but this approach has so far been limited to univariable MR. Here, we present MRBEE-TL, a novel MVMR that integrates transfer learning with bias-corrected estimating equations, to enhance causal inference in underpowered populations and to assess causal effect heterogeneity across populations. In simulations, MRBEE-TL consistently outperformed MR methods that relied solely on population-specific GWAS data, demonstrating superior estimation accuracy, statistical power, and type-I error control. In real data analysis, we will present exemplary results of causal effect estimation for multiple risk factors on five stroke subtypes in both Europeans and Asians. MRBEE-TL offers a valuable tool to jointly perform MR in multiple populations simultaneously.

147

Genome-Wide Association Meta-Analysis on Type D Personality

Anna D Argoty-Pantoja[1], Harold Snieder[1], Nina Kupper[2] on behalf of Type D Personality Consortium Investigators

[1]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2]Center of Research on Psychological Disorders & Somatic Diseases (CoRPS), Department of Medical & Clinical Psychology, Tilburg University, Tilburg, The Netherlands.

Background: Type D (distressed) personality and its subcomponents, negative affectivity (NA) and social inhibition (SI), are highly heritable (~50%) and linked to cardiometabolic and mental health. However, its genetic basis remains unexplored. We aim to perform the first large-scale genome-wide association study (GWAS) of Type D personality, assessing both the interaction (NA × SI) and the independent effects of NA and SI.

Methods: The analysis will use data from the Type D Personality Consortium, comprising six cohorts with ~120,000 participants: Lifelines, GHS, NTR, NESDA, KORA, and NEO. GWAS outcomes derived from the DS14 questionnaire include total NA score, total SI score, continuous Type D (NA × SI), and binary Type D status. Continuous traits will also be analyzed after using inverse normal transformation. Linear and logistic models will adjust for age, sex, and population stratification via principal components.

Results: GWAS results from GHS (n=15,000) and NTR (n=15,671), using the continuous Type D personality variable, identified suggestive top hits for the untransformed Type D measure in NTR on chromosomes 7, 10, and 22. These associations did not reach genome-wide significance after inverse normal transformation, nor were they replicated in GHS, suggesting possible inflation in the untransformed data. Analyses for Lifelines (n=74,124) are currently ongoing. Increasing the sample size through a meta-GWAS will improve statistical power and likely yield genomic loci robustly associated with Type D personality.

This study represents the first large-scale GWAS on Type D personality and its subcomponents, with potential to identify its underlying genes and biological pathways.

Keywords: Genome-wide association study, Type D Personality, Negative affectivity, Social inhibition.

149

Genomics Education Reframed: Race, Ancestry, Ethnicity, and Culture in the Classroom

Kathryn Shows[1], Frauke Seemann[2], Latrice Landry[3], Tafadzwa Machipisa[4], Howard Koh[5], Cheryl D. Cropp[6]

[1]Department of Biology, Virginia State University, Petersburg, Virginia, United States of America; [2]Department of Life Sciences, Texas A&M University Corpus Christi, Corpus Christi, Texas, United States of America; [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [4]Departments of Genetics and Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania Unitd States of America; [5]Section of Neurology and Developmental Neuroscience, Department of Pediatrics, Baylor College of Medicine, Houston, Texas, United States of America; [6]Department of Pharmacology and Toxicology, Morehouse School of Medicine, Atlanta, Georgia, Unites States of America

Equitable genomics requires population descriptors that extend beyond race and ancestry. Responding to the National Academies' 2023 framework, *"*Who Is *All of Us?",* supported by Baylor College of Medicine, piloted a two module curriculum embedding the NIH *All of Us* Researcher Workbench into undergraduate genetics courses at two minority serving institutions: Texas A&M University–Corpus Christi (TAMU CC; Hispanic serving, n=23) and Virginia State University (VSU; HBCU, n=18). Module 1 delineated the distinctions among race, ancestry, ethnicity, and culture and highlighted their relevance to health disparities. Module 2 provided hands on analyses of genotype–phenotype datasets, allele frequency plots, and logistic regression notebooks within a secure workspace.

A mixed methods evaluation employed a parallel course comparison and a one group pre/post design. IRB-approved surveys assessed knowledge, attitudes toward precision medicine, and self-efficacy in handling population level genomic data. Pre intervention, all participants had a similarly positive perception of their race, ancestry, ethnicity and culture. The students considered race, ethnicity, culture and ancestry equally important for personal identification, genetic epidemiology and medicine. Post intervention, TAMU CC students rated the importance of race, ancestry, ethnicity and culture higher than prior to the intervention and an increased focus on considering disease prevalence in response to the patients' physiological and physical data over race. Qualitative reflections demonstrated enhanced ethical awareness and enthusiasm for community engaged genomic research.

Keywords: population descriptors, undergraduate genomics education, *All of Us* Workbench, health disparities, minority-serving institutions

151

The Metabolic Subtype of Polycystic Ovary Syndrome: A Gene by Lifestyle Interaction Study in Hispanic/ Latinas

Hridya Gardner[1,2], Anne E Justice[‡2], Lindsay E Fernandez-Rhodes[*‡,1,3] on behalf of Hispanic Community Health Study/ Study of Latinos coauthors

*Presenting author

‡Co-authors

[1]Department of Biobehavioral Health, Pennsylvania State University, University Park, Pennsylvania United States of America; [2]Department of Population Health Sciences, Geisinger Health System, Danville, Pennsylvania United States of America; [3]Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America

Polycystic Ovary Syndrome (PCOS) results in subfertility, insulin resistance and other metabolic disturbances. In the United States it affects 2 in 10 Hispanic/Latina (H/L) women—a population with high cardiometabolic burden. The metabolic subtype of PCOS (mPCOS) is characterized by higher fasting insulin (FI), fasting glucose (FG), and body mass indexes (BMI). Herein, we seek to explore mPCOS as an outcome of interest for study in future gene-lifestyle interactions in H/L women.

Among 5259 premenopausal females (median age 35 years) from the Hispanic Community Health Study/Study of Latinos, 28.6% were PCOS cases (by self-report or menstrual cycles >35 days or "too irregular to report"). As hypothesized, FI and BMI were associated with PCOS status (p=3.71e-6 and 1.83e-7), but FG was not (p>0.05). None of the lifestyle factors examined (diet, physical activity, anxiety, depression) were associated with PCOS in our model accounting for age, site and H/L background. The prevalence of mPCOS was 15.9%, defined as FI, FG or BMI >75th percentile. Using summary statistics from a large publicly-available non-Hispanic European study, we a derived polygenic risk score (PRS) in our H/L female sample. This PRS predicted PCOS or mPCOS with a receiver operating characteristics-area under the curve of 0.49-0.50.

Although the European-ancestry PRS did not predict mPCOS in H/L women, we note high cardiometabolic dysfunction among those reproductive-aged, which further investigations can study using gene-environmental interaction models. Results could potentially partition the heterogeneity of mPCOS in H/L populations that is conferred by genes, lifestyle, or their interactions.