INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY

The 2023 Annual Meeting of the International Genetic Epidemiology Society

1

Multi-omics Integration Identifies Genes Influencing Traits Associated with Cardiovascular Risks

S. Acharya[*1,3], L. Liao[1,2], W. Jung[1,2], E. Kang[1,2], V. A. Moghaddam[1], M. Feitosa[1], M. Wojczynski[1], S. Lin[1], J. A. Anema[1], K.Schwander[1], J.O Connell[1], M. Province[1], M. Brent[1,2]

[1]Department of Statistical Genomics, Washington University School of Medicine, St Louis, Missouri, United States of America; [2]Department of Computer Science and Engineering, Washington University, St Louis, Missouri, United States of America; [3]Department of Computational and Data Sciences, Washington University, St Louis, Missouri, United States of America

**Presenting Author*

The Long Life Family Study (LLFS) enrolled 4,953 participants in 539 pedigrees displaying exceptional longevity. To identify genetic mechanisms that protect LLFS participants against age-related cardiovascular risks, we developed a freely available multi-omics integration pipeline and applied it to 11 traits associated with cardiovascular risks. Using our pipeline, we aggregated gene-level statistics from Rare-Variant Analysis, GWAS, and gene expression-trait association by Correlated Meta-Analysis (CMA). Across all traits, CMA identified 51 significant genes after Bonferroni correction ($P \leq 2.8\times10^{-7}$). *CETP*, *NLRC5*, *SLC45A3*, and *TOMM40* lie within 50 Kb of a known trait-associated variant (*previously associated genes*). Analysis of protein-protein interaction (PPI) networks identified another 63 genes (*passing genes*) that (1) have CMA p-value $\leq 5\times10^{-3}$, (2) lie in a PPI module (highly connected subnetwork) enriched for genes with low P-values, and (3) are annotated with a biological process that is enriched among module genes, ten of which were previously associated with the same traits. Permutation analysis showed that passing genes have a false positive rate of 1 in 14876 and are more likely to be previously known than non-passing genes with similar p-values. CMA improved on the 3 input analyses by producing the largest number of modules enriched for genes with low P-values and highly enriched for genes participating in shared biological processes. Overall, module analysis identified highly plausible candidate causal genes whose P-values after CMA alone were merely suggestive.

2

Joint Analysis of Longitudinal Omics Data and Time-To-Event Outcomes in the Context of Aging

Konstantin G. Arbeev[1*], Olivia Bagley[1], Aravind Lathika Rajendrakumar[1], Sheng Luo[2], Igor Akushevich[1], Anatoliy I. Yashin[1], Svetlana V. Ukraintseva[1]

[1]Social Science Research Institute, Duke University, Durham, North Carolina, United States of America; [2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, United States of America

Contemporary longitudinal studies started collecting information on repeated measurements of various omics (e.g., metabolomics) data. Joint analyses of such data and time-to-event (TTE) outcomes require specific approaches to address inherent analytic complications and provide inference relevant in the context of aging. We present the extensions of joint models (JM) for longitudinal and TTE outcomes and the stochastic process models (SPM) which are suitable for these purposes. We applied the approach to longitudinal measurements of 44 lysophosphatidylcholine (LPC) metabolites that showed significant (p-value<0.05, Bonferroni-corrected) association of larger values with higher risk of onset of Alzheimer's disease (AD) in the Alzheimer's Disease Neuroimaging Initiative (ADNI) participants (from entire sample, n=1,161; 360 AD cases) in traditional JM. SPM applications decomposed the observed associations into several components related to different aging-related characteristics: (a) decline in robustness and resilience to deviations of LPC metabolites from optimal levels (those minimizing AD risk at respective ages); (b) age-related changes in mean allostatic ("equilibrium") trajectories of the metabolites; and (c) gaps between the optimal and equilibrium trajectories. These components varied (p-value<0.05, Bonferroni-corrected) by sex and *APOE* e4 carrier status and across the metabolites. The resulting complex interplay of those components results in increased AD risk at older ages compared to younger ages. These findings call for further analyses of age-related dynamics of LPC metabolites to determine the underlying causes and mechanisms (including genetic underpinnings) of the observed associations. The approach can be applied to other types of omics data as well as to joint analyses of those.

3

The Intergenerational Transmission of Mental Health Disorders: A Systematic Review of Molecular Genetic Studies

Michelle Arellano Spano[*1,2], Johanne Hagen Pettersen[3], Evelyn R. Dilkes[4,] Amanda Hughes[1,2,3] and Neil Davies[1,2,3]

[1]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Barley House, Oakfield Grove, Bristol, United Kingdom; [3]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Norway; [4] Institute for Social and Economic Research, University of Essex

Mental disorders capture a range of diseases that affect cognition, emotion, and behavioral control, most commonly including depression, bipolar disorder, psychosis, dementia, and developmental disorders. Parental mental disorder is highly prevalent, 15 to 23% of children live with a parent with a mental disorder worldwide, and children of parents with a mental disorder are at a high risk of developing a mental disorder themselves (Leijdesdorff et al, 2017). This intergenerational transmission not only perpetuates mental health conditions but also places the offspring into a significantly disadvantaged position; poor mental health of parents limits socioeconomic and educational attainment (Currie, 2009). However, factors like variability in parental disorder severity, the mental health of the other parent, and reporting tendencies can affect the child's outcome. The current literature presents evidence for a patterning of mental health mostly as environmental, but there is no clear consensus on whether the transmission of mental health arises from environmental or through direct genetic inheritance. Particularly, designs that do not account for genetic factors in parent-child associations often suffer from confounding and fail to conclude whether the associations are causal (Jami et al, 2021). Molecular genetics provide a novel way of interrogating the genetic and environmental effect. Therefore, intergenerational studies of mental health that use molecular genetics are able to investigate the underlying parent-child mental health associations and interrogate to what extent these association are causal. Here we systematically review the literature and synthesize the findings of studies using molecular genetics to investigate these hypotheses. Our initial abstract search yielded 1,669 papers, of which 1,375 were excluded, as they did not meet our inclusion criteria. Thus, we had 294 papers at the full paper screening stage. Then 280 papers were excluded due to not having molecular genetic information. Finally, data extraction yielded 14 papers that had information on mental health disorders and utilized molecular genetics in their analyses. We systematically compared those studies on transmission of mental disorders that used molecular genetic data, focusing on their instruments, effect estimates presented, and diagnostic metrics used. We first provide a comprehensive overview of mechanisms that might lead to an association between parental mental disorders and offspring mental health. We then summarized the literature using molecular genetics to investigate the intergenerational transmission of mental health disorders.

References: Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic Status, poor health in childhood, and human capital development. *Journal of Economic Literature*, 47(1), 87-122. doi: 10.1257/jel.47.1.8

Jami ES, Hammerschlag AR, Bartels M, et al. (2021) Parental characteristics and offspring mental health and related outcomes: A systematic review of genetically informative literature. *Translational Psychiatry*, Translational Psychiatry 11(1).

4

Polygenic Risk Score for Partial Lipodystrophy Based on Clustered Phenotypes – Modelling and Validation in UK Biobank and Oxford Biobank

Naeimeh Atabaki-Pasdar[1,2,3*], Daniel E. Coral[3], Matt Neville[1,2], Katherine Pinnick[1,2], Eirini Trichia[4,5], Ana Viñuela[6], Fredrik Karpe[1,2]

[1]The Oxford Centre for Diabetes, Endocrinology and Metabolism (OCDEM), University of Oxford, Oxford, United Kingdom; [2]The NIHR Oxford Biomedical Research Centre (BRC), Oxford University Hospitals (OUH) Trust, Oxford, United Kingdom; [3]Genetic and Molecular Epidemiology Unit, Lund University Diabetes Centre, Department of Clinical Science, Lund University, Skåne University Hospital, Malmö, Sweden; [4]Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health (NDPH), University of Oxford, United Kingdom; [5]MRC Population Health Research Unit, CTSU, NDPH, University of Oxford, United Kingdom; [6]Biosciences Institute, Faculty of Medicine, Newcastle University, Newcastle Upon Tyne, United Kingdom

Human fat distribution is strongly associated with the incidence of coronary heart disease (CHD). The rare monogenic familial partial lipodystrophies (FPLD2-3) show this in the extreme. A common phenocopy, FPLD1, has a supposedly polygenic background, but this is poorly characterised. The characteristic phenotypic triad includes loss of leg fat mass, raised plasma triglycerides and insulin resistance. We developed a phenotype-clustered polygenic characterisation of FPLD1 by utilising GWAS summary data from GIANT, MAGIC and GLGC consortia.

Pre-processing GWAS summary data identified 281 linkage disequilibrium-pruned SNPs, all concordantly associated with waist-to-hip ratio, triglycerides and fasting insulin in an FPLD1 directionally consistent manner. These SNPs were utilised to build a unified PRS on UK Biobank DEXA-derived trunk-to-leg fat mass percentage ratio (FMR) (n=26,932). An FPLD1 binary outcome was defined as the top FMR decile (>1.23 women and >1.74 men). Shrinkage regressions were applied, which resulted in building a new FPLD1-PRS of 80 SNPs with a top decile OR of 3.7 (*P value*=2e-16) and validated in the Oxford Biobank (n=4,547) showing OR of 3.86 (*P value*=3.13e-05).

We then explored the association of the FPLD1-PRS with incident CHD in UK Biobank (n=335,964, 6% cases), showing a significant OR of 1.22 (*P value*=2e-16). Finally, to improve the predictability of FPLD1, we used machine learning methods to combine the PRS with other relevant clinical features.

The phenotype-clustered approach to creating a PRS for FPLD1 resulted in an effective definition of the complex condition. The PRS showed a substantial effect on incident CHD.

5

Leveraging Identity by Descent within Biobanks to Elucidate Genetic Architecture of Dilated Cardiomyopathy

James T. Baker[1*], Hung-Hsin Chen[1], Quinn S. Wells[1,2,3], David C. Samuels[1,4], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]

Department of Pharmacology, Vanderbilt University, Nashville, Tennessee, United States of America; [4]Department of Molecular Physiology & Biophysics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

Dilated cardiomyopathy (DCM), characterized by cardiac dilation and contractile dysfunction in the absence of abnormal cardiac loading or advance coronary artery disease, affects up to 1:250 individuals and is responsible for ~40% of cardiac transplants. Genetic testing can be used for diagnosis and clinical care but often fails to identify a casual variant and ~ 1/3 of test return only variants of uncertain significance. We used genomic segments shared identically-by-descent (IBD) to locate causal DCM variants. We identified pairwise IBD segments ≥ 3 cM using hap-IBD within 69,819 individuals of European ancestry as identified by PCA within Vanderbilt University Medical Centers biobank, BioVU. We clustered these individuals who shared an IBD segment containing the rare DCM associated missense mutation p.Arg636His into networks using a random walk where the segment length was used as a probability weight. This approach identified a network of 33 individuals significantly enriched for carrying the p.Arg636His variant (5 carriers, p=4.13e-12). Whole exome sequencing validated all the genotyping calls for all individuals in this network. We then constructed a dendrogram of the network using the inverse of the IBD segment length to calculate local familial distance between pairs. All five p.Arg636His carriers clustered into one branch of the dendrogram indicating that this missense variant mutation occurred on the branch preceding their most recent common ancestor. We identified other cardiac phenotypes (cardiomyopathies, conduction disorders, arrhythmias, heart failure) within their medical records for all 33 individuals and found that these phenotypes were concentrated in the five carriers.

6

A Cloud-Based Bioinformatic Pipeline Tool to Assess Sex-Specific Genetic Effects on Orofacial Cleft Risk using Genome Sequenced Trios

Seth R. Berke[1*], Kanika Kanchan[1,3], Eric Tobin[5], Cera Fisher[5], Debashree Ray[1,2], Claire L. Simpson[4], Alan F. Scott[6], Terri H. Beaty[1,2], Mary L. Marazita[7,8,] Ingo Ruczinski[1]

[1]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America; [2]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America; [3]Division of Allergy and Clinical Immunology, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America; [4]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America; [5]CAVATICA, Velsera, Charlestown, Massachusetts, United States of America; [6]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America; [7]Department of Oral and Craniofacial Sciences, Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [8]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America.

Data storage and scalability for association studies using whole genome sequencing (WGS) data present significant concerns. Research groups often rely on local computing clusters to overcome these issues but still experience limited computational power and substantial expense. Recently, cloud computing has provided a more efficient and cost-effective strategy. Drawbacks of cloud computing include difficult implementation with specialized software and potentially costly "tinkering." We recently addressed these challenges when investigating sex-specific genetic effects on non-syndromic orofacial clefts (OFCs) in multi-ethnic WGS data sets generated by the Gabriella Miller Kids First initiative and stored in the cloud-based environment CAVATICA.

Our multi-step genome-wide association study (GWAS) pipeline originally operated on local computational environments, involving several programming languages and packages including the Bioconductor R package *trio* to perform rapid genotypic transmission disequilibrium tests (gTDTs) to evaluate SNP-by-sex interaction. To become cloud-based, we first centralized these environments in CAVATICA with Docker, which containerizes environments for delivery. Within these ported environments, we utilized CAVATICA's software to construct our processing steps, which were woven together resulting in a complete Common Workflow Language (CWL) workflow. To reduce time and cost, these workflows utilized parallelization and optimal instance types.

We explain here the key components and implementation of our cloud computing approach. We also discuss its findings, including a SNP in the *RFTN1* gene on chromosome three that shows significant differential OFC risk between the sexes (p value = $6.1x10^{-11}$). Our methodology can extend to other software packages and pipelines by investigators interested in scalable high-throughput genomic analysis.

7

Revealing Genetic Signatures of Lung Cancer Histologic Subtype Using Deep Learning

Michael J. Betti[1*], Melinda C. Aldrich[1], Eric R. Gamazon[1,2]

[1]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Clare Hall, University of Cambridge, Cambridge, United Kingdom

Lung cancer is a highly heterogeneous disease, and traditional computational approaches are limited in their ability to model its complex genetic architecture. Deep learning-based approaches provide an opportunity to elucidate lung cancer disease etiology through modeling non-linear relationships within high-dimensional genomic data.

We leveraged tumor gene expression, DNA methylation, miRNA, and histology data from 962 non-small cell lung cancer cases in The Cancer Genome Atlas (TCGA). We trained an initial series of three fully connected neural networks using TCGA genomics data to distinguish between adenocarcinoma and squamous cell carcinoma. Due to the poor performance of the fully connected miRNA-based model, a more complex convolutional network (CNN) architecture was subsequently trained. Genomics-based model performance was compared against that of a state-of-the-art histology image-based CNN.

Finally, we utilized Shapley Additive Explanations to identify genomic features most highly contributing to each model's predictions.

Gene expression and DNA methylation-based models performed well, with area under the receiver operator curves (AUCs) in the test set at 0.89 and 0.96, respectively. However, the fully connected miRNA model achieved poor performance (0.61 AUC). Using a CNN-based architecture, however, the miRNA model achieved an AUC of 1.0. Each genomics-based model achieved higher performance than the image-based model (AUC 0.80). Using Shapley Additive Explanations, we found that although thousands of individual mRNAs and methylated CpG sites contributed to their respective model's predictions, only ~200 of the 1,534 miRNAs seemed to significantly influence histologic prediction. This suggests that miRNAs may be promising targets for future diagnostic gene panel design.

8

Sex-specific Blood DNA Methylation in Rab-regulatory Genes Underly Sex-biased Risk of Recurrence in Unprovoked Venous Thromboembolism

Ohanna C. Bezerra[1]*, Marc Rodger[2], Michael J. Kovacs[3], Gregoire Le Gal[4], Pierre E. Morange[5], Gaëlle Munsch[6,] David-Alexandre Trégouët[6], Celia M. T. Greenwood[7,8], France Gagnon[1]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Department of Medicine, McGill University, Montreal, Canada; [3]Department of Medicine, University of Western Ontario, London, Canada; [4]Department of Medicine, Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Canada; [5]Cardiovascular and Nutrition Research Center, Aix-Marseille University, Marseille, France; [6]University of Bordeaux, Bordeaux Population Health Research Center, INSERM U1219, Bordeaux, France; [7]Lady Davis Institute, Jewish General Hospital, Montreal, Canada; [8]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

Deciding whether to stop oral anticoagulants beyond initial treatment (3-6 months) after an unprovoked venous thromboembolism (VTE) is challenging and controversial, partially due to an intriguingly higher risk of recurrence in men after therapy discontinuation compared to women. In preliminary work, we observed sex-specific blood DNA methylation (DNAm) marks in vitamin K cycle genes, relevant to the coagulation cascade and vascular integrity. We hypothesized that sex-specific DNAm are associated with the observed sex-biased VTE recurrence (rVTE). Using the EPIC array, we performed a sex-stratified epigenome-wide association study with rVTE in 417 Europeans (REVERSE I study). We identified two male (*TBC1D22B* and *ZHX2*) and one female (*DENND3*) hypomethylated CpG sites associated with rVTE (p value < 7×10^{-8}). The female association displayed the same direction of effect in 139 French VTE women, although not significant (MARTHA study). *TBC1D22B* and *DENND3* are regulators of the Rab family proteins involved in vesicle trafficking, corroborating findings on the participation of Rab in rVTE. A follow-up methylation quantitative trait locus (meQTLs) analysis within 250Kb around the associated CpGs in REVERSE I identified *cis*-meQTL variants modulating *DENND3* methylation, also reported as expression QTL in whole blood by GTEx. The synonymous variant rs1045303 significantly decreased DNAm of probe cg03401656 in rVTE, implying that genetic variants may mediate the effect of DNAm on overall recurrence. Our results showing independent DNAm sites associated with sex-specific rVTE may optimize decision-making on prophylaxis after a first event. Replication and functional analyses can expand insights into the molecular mechanisms driving sex-biased rVTE.

10

Integrating Large-scale Priors for In-silico Functionalization of GWAS Associated Loci Using Machine Learning Models

Krittika Bhattacharyya[1,2*], Samsiddhi Bhattacharjee[2]

[1]*Department of Statistics, University of Calcutta, Kolkata, India;* [2]*National Institute of Biomedical Genomics, Kalyani, India*

Identification of causal genetic variants and understanding their regulatory mechanisms is crucial for comprehending the genetic etiology of complex traits. Although genome-wide association studies (GWAS) have successfully identified variants associated with these complex traits, interpreting their causal role remains ambiguous due to three main challenges: (1) complex linkage disequilibrium (LD) among associated variants, making it difficult to distinguish true causal variants from false positives, (2) a majority of associated variants residing in noncoding regions, and (3) limited availability of large gold standard sets of regulatory variants for training and validating machine learning models.

To address these challenges, our proposed novel statistical method combines machine learning (ML) and Bayesian techniques to integrate functional annotations and tissue/cell-type specific epigenomic landscapes, enabling efficient identification of causal variants, target genes, and regulatory mechanisms. Our approach utilizes Colocalization (e.g., Coloc), fine mapping (e.g., Sum of Single Effects Regression Model or SuSiE), along with the machine learning technique XGBoost. The NHGRI-EBI GWAS catalog, HAPLOREG database, GTEx, and summary GWAS datasets are used to train our ML models. By systematically prioritizing causal variants, identifying regulated cis-genes and predicting potential regulatory mechanism, our approach can help in guiding wet-lab validation experiments and reduce resource burden. We apply our method to integrate GWAS summary data, GTEx v8 data, LD, and functional annotations for quantitative lipid traits in relevant tissues. The application illustrates the potential of this approach to interpret the associated loci of lipid traits and prioritize mediating genes, thus shedding light on their genetic causality in cardiovascular diseases.

11

Predicting Gene-Driven Cortical Changes in Neurologic Disease through the Neuroimaging PheWAS

Xavier Bledsoe, Eric Gamazon

Vanderbilt University, Division of Human Genetics, Nashville, Tennessee, United States of America

Neuropsychiatric diseases are associated with alterations in neuromorphology. These changes possess tremendous potential as a means of understanding the natural progression and pathophysiology of disease. Unfortunately, it is difficult to discriminate between changes that drive vs. result from the trait. Here we describe a novel methodology to (1) identify endogenous neurologic consequences of trait associated transcriptomic variation and (2) partition these neurologic changes according to their dominant transcriptomic drivers.

Using JTI-PrediXcan, we generated a transcriptome wide atlas of associations between genetically regulated gene expression and over 3,000 neuroimaging measures. Leveraging schizophrenia GReX associations from the most recent genome wide association study meta-analysis, we perform a phenome wide association study (PheWAS) of neuroimaging features to highlight schizophrenia genes whose endogenous expression is associated with neuroimaging measures in the UKB ($P<0.005$). We consider only imaging measures in the UKB that were independently associated with schizophrenia in clinical studies. For each gene, we use a binarized multiplicative model to predict neuromorphologic change in the context of schizophrenia. We then cross reference the clinical schizophrenia neuroimaging study with the predictions for each region and annotate gene-region pairs according to predictive accuracy.

We identify associations between 22 schizophrenia genes and 40 schizophrenia neuroimaging measures. 65% of these neuroimaging measures are accurately modeled by expression of at least one schizophrenia gene.

As drivers of subclinical trait-associated brain changes in healthy individuals, these genes represent targets for further analyses. This methodology represents a generalizable approach to matching trait genes with trait associated neuroimaging measures.

12

Unravelling the Interplay between Type 2 Diabetes, Genetics, and Metabolite Levels

Ozvan Bocher[1*], Archit Singh[1,2], Ana Luiza Arruda[1,2], Peter Kreitmaier[1,2], Andrei Barysenka[1,3], William Rayner[1,3], Eleftheria Zeggini[1,3]

[1]Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [2]Technical University of Munich (TUM), TUM School of Medicine, Munich, Germany; [3]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine, Munich, Germany

Type 2 diabetes (T2D) represents a major health burden for which genetics has been successfully investigated in large GWAS. The remaining challenge lies in fully understanding the role of these variants in biology giving rise to disease, something which can be investigated through metabolomics. We sought to investigate the interplay between genetics, metabolomics, and T2D risk in the UK Biobank cohort. We first conducted a bidirectional Mendelian randomization study to assess causal relationships between metabolite levels and T2D risk. We found only a few of the 164 absolute metabolite levels tested to be causal of T2D, but half of them to be caused by T2D (with p value down to 10^{-61}), including an increase in amino acids and glucose levels, and a decrease in cholesterol classes. Some of these metabolites are also seen to be associated with specific T2D complications such as HDL cholesteryl esters showing lower values in T2D individuals with kidney complications compared to those without complications (β=-0.55, p=$3.66x10^{-8}$). Secondly, using a differential metabolite QTL analysis, we describe a different genetic regulation of 22 metabolites between individuals with and without T2D, including glycine (β=0.41, p values down to $5x10^{-25}$) and low-density lipoproteins (β=0.53, p=$1.61x10^{-12}$). Additionally, T2D was found to cause changes in eight of these 22 metabolite levels. This work provides a better understanding of the metabolic changes induced by the occurrence of T2D. Although further work is needed to confirm these results, they provide potential directions to investigate T2D metabolic consequences and related subsequent complications.

13

Metabolic Reprogramming Induced by Periodic Veganism in Humans

Ozvan Bocher[1*], Archit Singh[1,2], Ana Luiza Arruda[1,2], Peter Kreitmaier[1,2], Andrei Barysenka[1,3], William Rayner[1,3], Eleftheria Zeggini[1,3]

[1]Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [2]Technical University of Munich (TUM), TUM School of Medicine, Munich, Germany; [3]Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine, Munich, Germany

The biological mechanisms underlying the effects of dietary restriction (DR) on health remain to be elucidated. To address this, we present the FastBio study comprising 200 periodic vegan (PV) and 211 non-vegan (NV) individuals. PV individuals alternate between veganism and omnivory for religious reasons, totaling approximately 200 days of veganism annually, while NV are continuously omnivorous. Molecular profiles for 1,455 proteins and 249 metabolites were measured for all individuals at two timepoints defined by PV diet: during a period of omnivory and during a period where PV had abstained from animal products for 3-4 weeks. Molecular profiles were compared across timepoints for each dietary group using paired differential expression analyses. We report 410 and 201 differentially expressed proteins at FDR<5% for PV and NV groups respectively, with 264 unique to PV, including proteins with a role in browning of adipose tissue, bone degradation, T-cell function, and cognition. For metabolomics profiles, NV and PV individuals display a significant shift in one and 168 metabolites respectively. Significant alterations of these 168 metabolites in the restriction timepoint include decreased percentages of saturated fatty acids, decreased levels of valine and cholesterol classes but also increased levels of alanine and glutamine, most of which are associated with preventive effects from complex diseases. Overall, our study highlights a rapid metabolic shift in PV individuals driven by 3-4 weeks of abstinence from animal products. While further work is needed to elucidate the biological pathways impacted, our results suggest that periodic veganism has mostly positive effects on human health.

14

A Method to Quantify Bias and Identify Spurious Correlations in Biobanks

Lindsay B. Breidenbach[1*], Lea K. Davis[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America

Biobanks collect medical, genomic, and/or survey data across many participants. Biobanks often link back to participants' medical records. These databases are huge and allow researchers better statistical power. Genome-wide association studies (GWAS) and other association studies capitalize on this increased power to find these new smaller correlations and interactions. However, those researchers lose the ability to gather data as part of their study design. This creates a rift between the data a hypothesis needs, and the data a biobank has. For example, hypotheses that require lab values often deal with higher amounts of missing data, as most labs are not uniformly ordered across all subjects. These rifts create bias. Enough uncontrolled bias creates spurious findings that are statistically significant but aren't biologically driven.

To address this, we created an R package that simulates bias in datasets and compares those to unbiased simulated data. We posit that in this comparison, one can quantify how much bias is behind the final effect estimates. With the program, users can model which variables bias and affect others through directed acyclic graphs. Users can then employ Bayesian networks to quantify the extent to which variables impact each other. This method allows researchers to flexibly model a variety of biases that are unique to their hypotheses. Once the bias is quantified, users can easily see how much of their findings can be explained through bias alone. Researchers can then separate the insightful correlations from those which are just statistically significant.

15

Novel Systematic Method for Identifying Congenital Anomaly Cases in the EHR for the Purpose Identifying New Causes of Congenital Anomalies

Elly Brokamp, MS, CGC; Lisa Bastarache, MS; Rizwan Hamid, PhD, MD; Nancy Cox, PhD; Megan Shuey, PhD

Vanderbilt University Medical Center

Congenital anomalies (CAs) affect approximately 3% of live births and are the leading cause of infant death. The cause of approximately 80% CAs is unknown and for the 20% with an identified cause, variability in penetrance suggests additional drivers of risk exist. Research to understand the causes of CAs is hampered by the lack of a uniform identification method in electronic health records (EHRs). We demonstrate the first large-scale effort to define and characterize CAs and multiple CAs (MCA) in the EHR and provide a quantitative way to evaluate associations between CAs and genetically diagnosed conditions.

Using phecodeX, the number of CA phecodes, aggregates of clinical billing codes, increased from 56 to 365. These new codes provide substantially improved granularity of these diagnoses as well as more accurately mirror body system. Using a large clinical cohort, we demonstrate the clinical phenome associated with these conditions and demonstrate that the definition of MCA based on major vs minor conditions is arbitrary and inadequately reflects the severity of conditions.

Further, we provide a quantitative assessment of potential CA causal genetic diseases. Only 85 (19.7%) of genetic disease codes have a known association with CAs based on literature review. For the remaining we performed phenome-wide association studies to identify association of these conditions with CAs. An additional 16 (3.7%) codes significantly associated with CAs (Bonferroni p< 2.75x10-5).

These results are critical for genetic and epidemiological studies of CAs by improving case determination and differentiating idiopathic CAs from those with known cause.

16

Signal Mapping Methods for Region-level Association Adaptive to Local Linkage Disequilibrium (LD)

Myriam Brossard[1*], Kexin Luo[1], Delnaz Roshandel[2], Fatemeh Yavartanoo[3], Yun Joo Yoo[3], Andrew D. Paterson[2,4], Shelley B. Bull[1,4]

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada; [2]Hospital for Sick Children Research Institute, Toronto, Canada; [3]Seoul National University, Seoul, Korea; [4]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Compared to single variant analysis, region-level testing of multiple variants reduces genome-wide testing burden, is robust to genetic heterogeneity, and can be more powerful when there are multiple causal variants, but region-level signal-mapping methods are lacking. To develop and evaluate methods, we simulated a large case-control dataset of melanoma risk (n=40,000) characterized by multiple causal variants (CVs) in *MC1R* (16q24.3) carried by distinct risk haplotypes. Using 1000G European-ancestry haplotypes in 16q with extended linkage disequilibrium (LD), we generated melanoma status (balanced design) under a log-additive model for joint effects of five CVs. To detect region-level association, we partitioned 16q (107,406 SNPs) into 2,394 quasi-independent LD-blocks and applied regression-based tests in each region, including a reduced-dimension multiple-linear-combination (MLC) test. MLC clusters SNPs within a region into LD bins, and combines multi-SNP regression estimates into bin-level effects which are aggregated into a reduced-*df* test adaptive to LD. At $P_{\text{GW-Bonferroni}}\leq 5.4\text{x}10^{-7}$, we detected 39 regions: in the top region, 1126 SNPs clustered into 60 LD bins (with each CV assigned to a bin) while the other 38 regions included CV proxies. These regions lost significance when region-level tests were repeated, conditional on a MLC-derived region-level summary variable from the top region. Within the top region, backward selection of 60 MLC-LD-bin-level summary variables ($P_{\text{Bonferroni}}=8.3\text{x}10^{-4}$) yielded six LD bins, including all five CVs. In the challenging setting of long-range LD and complex genetic architecture, LD-adaptive region-level signal mapping can efficiently prioritize regions and SNP clusters for fine-mapping studies. Evaluations in other genetic architectures are warranted.

17

Adjusting Collider Bias for Disease Progression Trait Using Bivariate Mendelian Randomization

Siyang Cai[1], Frank Dudbridge[1]

[1]Department of Health Sciences, University of Leicester

Genome-wide association studies (GWAS) have provided a large number of genetic markers that can be used as instrumental variables in a conventional Mendelian randomization (MR) analysis to access the true causal effect of a risk factor on an outcome. An extension of MR analysis, multi-variable Mendelian randomization, has been proposed by Burgess et al. (2014) to handle pleiotropy with multiple risk factors. Meanwhile, adjusting or stratifying outcome on a variable that is associated with the outcome of interest involves collider bias. An outcome that represents progression of the disease conditioning by selecting only the cases will cause a biased estimation of true causal effect of the risk factor of interest on the outcome. Recently, Cai et al. (2022) developed Corrected Weighted Least Squares (CWLS) and instrument effect regression to adjust for weak instrument bias and collider bias. In this paper, we highlight the importance of adjusting weak instrument bias and collider bias in a bivariate Mendelian randomization with a risk factor of interest, a disease trait, and the disease progression as the outcome. A generalized version of the CWLS adjustment and instrument effect regression are then proposed based on a multivariate IVW model, followed by simulations demonstrate the performance of the 2-step CWLS adjustment, with illustrations of type-1 errors of adjusted true causals of interest, providing less biased adjustment under null hypothesis with an estimated standard error compared to other existing multivariate MR methods. A further discussion is given based on cases with non-zero casuals between three traits and also the use of adjustment in two-sample MR.

18

Lymphocyte Count-derived Polygenic Score and Inter-individual Variability in CD4 T-cell Recovery in Response to Antiretroviral Therapy

Kathleen M. Cardone[1], Scott Dudek[1], Karl Keat[2], Yuki Bradford[1], Zinhle Cindi[1], Eric S. Daar[3], Roy Gulick[4], Sharon A. Riddler[5], Jeffrey L. Lennox[6], Phumla Sinxadi[7], David W. Haas[8,9], Marylyn D. Ritchie[1,10]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [3]Lundquist Institute at Harbor-UCLA Medical Center, Torrance, California, United States of America; [4]Weill Cornell Medicine, New York, New York, New York, United States of America; [5]University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [6]Emory University School of Medicine, Atlanta, Georgia, United States of America; [7]Division of Clinical Pharmacology, Department of Medicine, University of Cape Town, Cape Town, South Africa; [8]Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [9]Meharry Medical College, Nashville, Tennessee, United States of America; [10]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Among people living with HIV, CD4 T-cell recovery on antiretroviral therapy (ART) varies considerably. We explored whether a polygenic score, derived from summary statistics for absolute lymphocyte count from the general population ($\text{PGS}_{\text{lymph}}$), explained variability in baseline CD4 T-cell count (CD4) prior to ART, and change on ART. We explored associations with pre-ART CD4 (n=4,959) and change from baseline to week 48 on ART (n=3,274) among participants in prospective, randomized ART studies of the AIDS Clinical Trials Group. We examined ancestry specific $\text{PGS}_{\text{lymph}}$, applied to all participants and to African and European ancestral groups separately. Multivariate models that included $\text{PGS}_{\text{lymph}}$, baseline plasma HIV-1 RNA, age, sex, and 15 principal components (PCs) explained approximately 26-27% of variability in baseline CD4, but $\text{PGS}_{\text{lymph}}$ accounted for <1% of this. Models that also included baseline CD4 explained approximately 7-9% of variability in CD4 increase on ART, but $\text{PGS}_{\text{lymph}}$ accounted for <1% of this. In univariate analyses, $\text{PGS}_{\text{lymph}}$ was not significantly associated with either phenotype. Among individuals of African ancestry, African-specific PGS was significantly associated with CD4 increase in the multivariate model but not the univariate model. When applied to lymphocyte count in the general population, PGS explained approximately 6-10% of variability in multivariate models (including age, sex, and PCs) but only about 1% in univariate models. These results highlight the importance of including covariates in PGS models. In summary, a lymphocyte count PGS derived from the general population was not consistently associated with CD4 T-cell recovery on ART, nonetheless, clinical covariates are critical in building polygenic scores.

19

Factors Influencing the Portability of Gene Expression Imputation in Transcriptome-Wide Association Studies across Ancestry

Yung-Han Chang[1,2], Arjun Bhattacharya[3,4], Yi. Ding[3,5], Bogdan Pasaniuc[2,3,5]

[1]Department of Biostatistics, University of California Los Angeles, Los Angeles, California, United States of America; [2] Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America; [3] Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America; [4] Institute for Quantitative and Computational Biosciences, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America; [5] Bioinformatics Interdepartmental Program, University of California Los Angeles, California, United States of America

Transcriptome-wide association studies (TWAS) relies on imputation of gene expression based on models trained in reference expression quantitative trait locus (eQTL) datasets. Previous studies have showed the challenges of TWAS when attempting to predict gene expression across different ancestry. Here, we examine potential factors that drive this lack of portability of expression models.

Across 1,000 simulations at 22 gene loci (one per chromosome), we simulate a training eQTL dataset of European samples and test sets of both European and African populations based on the linkage disequilibrium (LD) patterns in 1000G and across a variety of genetic architectures. We build a sum of single effects (SuSiE) linear regression model in the training dataset and predict expression in the test datasets.

Our findings indicate that, in general, as the proportion of causal eQTLs increases, percent variance explained (PVE) decreases in the European and African test sets, respectively. We also examine differences in PVE across three factors in genetic architecture. First, we examined the impact of varying allele frequencies of the causal eQTLs and found no significant differences. Second, we explored the effect of changing the ratio of heritability between two populations, and observed a decrease in the PVE difference as the ratio increases. Third, we are investigating how differences in LD across the datasets are associated with performance.

In conclusion, our findings emphasize the importance of considering differences in heritability, causal eQTL proportion, and LD when aiming to improve the portability of gene expression imputation across ancestry.

20

Autism Heterogeneity Related to Early-life Exposures: Multi-ancestry Results from the SPARK Sample

Charikleia Chatzigeorgiou[1,2], Behrang Mahjani[1,2], Marina Natividad Avila[1,2], Paul O'Reilly[1,3], Niamh Mullins[1,3], Magdalena Janecka[1,2,3,4]

[1]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York; [2]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [4]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America

Autism spectrum disorder (ASD) is a highly heritable neurodevelopmental disorder with complex etiology. While previous epidemiological studies have identified prematurity as a risk factor for the disorder, the mechanisms underlying its potential effects on neurodevelopment remain unknown. Given that the duration of pregnancy in the general population is influenced in part by genetic factors, and our earlier results suggested a distinct set of medical comorbidities in pre-term born individuals with ASD, we investigated systematic genetic differences between ASD cases born term and pre-term in order to interrogate potential biological differences between these groups of ASD cases.

Individuals including in Simons Foundation Powering Autism Research for Knowledge (SPARK) dataset who had an ASD diagnosis and were born preterm (cases) were compared with ASD individuals born in term (controls) (N_{total}=31,947). Ancestry was estimated through principal component analysis and random forest plot after merging with 1000G and HGDP. Three ethnicity specific GWAS were conducted [African/African American (n_{cases}=196, $n_{controls}$=1722), Admixed American (n_{cases}=396, $n_{controls}$=3955), Non-Finnish European (n_{cases}=2296, $n_{controls}$=19173)] adjusting for the 10 principal components. One SNP from the African/American (rs78395263, p=3.92x10^{-08}) and one from the Admixed American GWAS (rs77818427, p=1.04x10^{-08}) reached the genome-wide level of significance (p<5.0 x 10^{-08}). Metanalysis of the association results from the different ancestries was done with the use of METAL. No SNP was found to reach the genome-wide level of significance.

To advance biological insights among the set of significant variants and map them to disease-relevant genes and pathways, functional genomic resources that provide data on intermediate molecular phenotypes were used to functionally annotate variants and map them to genes (https://fuma.ctglab.nl/). FUMA analyses revealed 24 possible causal genes pinpointed by positional mapping and expression quantitative trait locus mapping, including *OCA2* (p=1.91 x 10^{-7} previously associated with metabolic disorders), *EXOC6B* (p= 1.37 x 10^{-06}, previously shown to associate with immune disease), and *SRP54* (p=1.4 x 10^{-6} previously linked to fetal and blood disorders). Five candidate SNPs in FUMA were associated with educational attainment, triglyceride levels, and serum alkaline phosphatase levels in the GWAS catalogue.

In conclusion, our findings revealed genes that could highlight the biological pathways distinguishing ASD cases born term and pre-term. The analyses are ongoing to explore phenome-wide associations of genetic liability to being born pre-term and gain new insights into autism heterogeneity related to early-life exposures.

21

Inference of Causal Metabolite Networks in the Presence of Invalid Instrumental Variables with GWAS Summary Data

Siyi Chen[1], Zhaotong Lin[1], Xiaotong Shen[2], Ling Li[3], Wei Pan[1]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]School of Statistics, University of Minnesota, Minneapolis, Minnesota, United States of America; [3]Department of Experimental and Clinical Pharmacology, College of Pharmacy, University of Minnesota, Minneapolis, Minnesota, United States of America

We propose structural equation models (SEMs) as a general framework to infer causal networks for metabolites and other complex traits. Traditionally SEMs are used only for individual-level data under the assumption that all instrument variables (IVs) are valid. To overcome these limitations, we propose both a one- and two-sample approaches for causal network inference based on SEMs that can: 1) perform causal analysis and discover causal relationships among multiple traits; 2) account for the possible presence of some invalid IVs; 3) allow for data analysis using only GWAS summary statistics when individual-level data are not available; 4) consider the possibility of bi-directional relationships between traits. Our method employs a simple stepwise selection to identify invalid IVs, thus avoiding false positives while possibly increasing true discoveries based on two-stage least squares (2SLS). We use both real GWAS data and simulated data to demonstrate the superior performance of our method over the standard 2SLS/SEMs. For real data analysis, our proposed approach is applied to a human blood metabolite GWAS summary dataset to uncover putative causal relationships among the metabolites; we also identify some metabolites (putative) causal to Alzheimer's disease (AD), which, along with the inferred causal metabolite network, suggest some possible pathways of metabolites involved in AD

22

Investigating Pleiotropy as an Explanation for the Inverse Association between Cancer and Dementia

Dorothy Chen[1], John S. Witte[2], Thomas J. Hoffmann[1], Jingxuan Wang[1,3], Peter Buto[1,3], Sarah F. Ackley[1,3], Sara Rashkin[4], M. Maria Glymour[1,3], Rebecca E. Graff[1] (corresponding author)

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, United States of America; [2]Department of Epidemiology and Population Health, Stanford University, Stanford, California, United States of America; [3]Department of Epidemiology, Boston University, Boston, Massachusetts, United States of America; [4]Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America

Numerous epidemiological studies have documented an inverse relationship between cancer and dementia, such that diagnosis with either condition is associated with a reduced risk of diagnosis with the other. Little is still known about genetic factors that may be inversely related to the two diseases, both of which are complex polygenic traits. Subset-based association analyses offer a systematic method for investigating pleiotropy, the phenomenon whereby the same genetic factor affects more than one complex trait. We leveraged summary statistics from genome-wide association studies to evaluate the genetic overlap between 18 common cancer types paired with 3 dementia phenotypes (all-cause dementia, Alzheimer's disease (AD), vascular dementia) in the UK Biobank (437,347 European ancestry individuals; 51,046 cancer cases, 8,143 dementia cases, 382,314 controls). We implemented ASsociation analysis based on subSETs (ASSET) to explore all possible subsets of cancer-dementia phenotypic pairs for the presence of association signals, thereby identifying the best combination of traits to maximize test statistics. After adjustment for multiple testing, ASSET did not reveal any statistically or suggestively significant (10^{-6}) genetic variants associated with any phenotype pairs in either the same or opposite directions. Evaluation of sub-threshold signals (10^{-3}) for an inverse association between prostate cancer and AD identified 6 pleiotropic variants mapping to the *TET2* gene (rs6839705, rs7674220, rs1391438, rs9884984, rs9884296, and rs1391439; P value < 0.001 for fixed-effects meta-analyses). Our findings do not provide evidence of pleiotropic signals underlying the inverse relationship between cancer and dementia. Future research should consider alternative methods for assessing local and genome-wide pleiotropy.

23

Summary Statistics from Large-scale Gene-environment Interaction Studies for Re-analysis and Meta-analysis

Duy T. Pham,[1] Kenneth E. Westerman,[2,3,4] Cong Pan,[1] Alisa K. Manning,[2,3,4] Han Chen[1,5]

[1]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [2] Department of Medicine, Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA, United States of America; [3]Metabolism Program, Broad Institute of MIT and Harvard, Cambridge, MA, United States of America; [4]Department of Medicine, Harvard Medical School, Boston, MA, United States of America; [5]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America

Genomic summary statistics have been widely shared in genetic epidemiological research. With numerous methods and tools developed in recent years, genomic summary statistics, largely derived from genome-wide association studies (GWAS), have greatly advanced the field by enabling many valuable downstream analyses that are accurate while being much more efficient than directly analyzing individual-level data. They have also facilitated international collaborations by reducing privacy concerns associated within individual-level data sharing. However, methods and tools for generating, managing, and analyzing genomic summary statistics from gene-environment interaction (GEI) studies have not been well established. As GEI studies generally require much larger sample sizes than GWAS on marginal genetic effects to identify interactions, which often involve up to millions of samples, there is a pressing need for GEI-specific methods on summary statistics. We introduce two tools to facilitate such applications, with a focus on statistical models containing multiple gene-exposure and/or gene-covariate interaction terms, built upon recent GEI analysis software programs; REGEM uses full summary statistics from a single, multi-exposure genome-wide interaction study (GWIS) to derive analogous sets of summary statistics with arbitrary subsets of exposures and interaction covariate adjustments, without having to rerun each model genome-wide using individual-level data. METAGEM performs fixed-effects meta-analysis on summary statistics from multiple studies, with multiple gene-exposure and/or gene-covariate interactions. We demonstrate the value and efficiency of these tools by exploring alternative methods of accounting for ancestry-related population stratification in GWIS in the UK Biobank and show that proper use of interaction covariates can control the type I error rate of a pooled-ancestry GWIS analysis and recapitulate results from the associated multi-ancestry meta-analysis. These programs help to maximize the value of genomic summary statistics from diverse and complex GEI studies, extending efforts at more inclusive and rigorous ancestry-aware genetic epidemiology to the GEI domain.

24

FlexNet: A Flexible Network-Based Framework for Identifying Drug-Disease Relationships for Repurposing Opportunities and Prediction of Adverse Effects

Rui Chen[1], Quan Wang[1], Qiang Wei[1], Yan Yan[1], Yuting Tan[1], Anshul Tiwari[1], Xue Zhong[2], Bingshan Li[1]

[1] Vanderbilt University; [2] Vanderbilt University Medical Center

Predicting drug-disease relationships, such as new indications and adverse drug reactions (ADR) for a drug, has been a growing interest because it has significant value on reducing the drug development and medical care cost. To achieve effective drug repurposing and ADR prediction, integration of heterogenous data at different levels is crucial. In this study, we develop a computational framework that utilizes the Random Walk with Restart algorithm to score relationships among the three types of nodes: genes, diseases, and drugs, integrating multi-layer networks of the nodes. Our framework is flexible and allows for easy addition of customized networks. We Applied the framework to 220 diseases, with integrative networks of 1) Gene Ontology, KEGG, PPI, co-expression networks; 2) phenotypic and semantic similarity networks for diseases; 3) structure similarity networks for drugs; 4) bipartite networks of drug-gene, disease-gene, and drug-disease connections. We observed that the top scoring drug-disease pairs align significantly in clinical trials and off-label uses, suggesting novel repurposing opportunities in our prediction. In parallel to drug repurposing, we found that 51% of reported Drug-Adverse Effect Pairs (DAEPs) are in the top 100 predictions. To further assess the prediction of ADR for new drugs, we collected reported DAEPs for 77 drugs that were approved after

2015 and observed that 40.4% of reported DAEPs are within the top 100 predictions, demonstrating our framework's ability to predict DAEPs. We created a comprehensive scoring table, encompassing 1,520 drugs and 8,120 diseases, which we hope are useful for the drug repurposing and ADR field.

25

Imputation Efficacy Across Global Human Populations

Jordan L. Cahoon[1,2,3], Xinyue Rui[1,*], Echo Tang[2,*], Christopher Simons[2], Jalen Langie[1], Minhui Chen[1], Ying-Chu Lo[1], Charleston W. K. Chiang[1,2,4,†]

[1]*Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America;* [2]*Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, United States of America;* [3]*Department of Computer Science, University of Southern California, Los Angeles, California, United States of America;* [4]*Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America*

* These authors contributed equally

[†] Correspondence: charleston.chiang@med.usc.edu

Genotype imputation is fundamental for association studies but lacks fairness due to the underrepresentation of non-European ancestries. The state-of-the-art imputation reference released by the TOPMed initiative contains admixed African-ancestry and Hispanic/Latino samples but may still lack representation from ancestries outside of North America. To evaluate the efficacy of the TOPMed reference panel to impute diverse global populations, we curated genome-wide array data from 23 publications and imputed over 43k individuals across 123 populations around the world. We identified several populations where imputation accuracy paled in comparison to that of European-ancestry populations. For instance, the mean imputation r-squared (Rsq) for 1-5% alleles in Saudi Arabians (N=1061), Vietnamese (N=1264), and Papua New Guineans (N=776) were 0.79, 0.78, and 0.62, respectively, compared to 0.90-0.93 for European populations matched in sample size and SNP content. Rsq appeared to be inversely correlated to genetic distances to European reference, and may be more inflated when compared to ground truth from sequencing data for non-European populations. We also assessed meta-imputation for improving imputation accuracy by combining results from TOPMed with a reference of 1496 sequenced individuals from Taiwan Biobank. While we found that meta-imputation did not improve Rsq genome-wide, Southeast Asian populations such as Filipino and Vietnamese experience a 0.11-0.16 increase in mean Rsq for population-specific alleles. Taken together, our analysis suggests that meta-imputation may complement a large reference panel such as that of TOPMed for underrepresented cohorts. Nevertheless, reference panels must ultimately strive to increase diversity and size to promote equity within genetics research.

26

Genetic Determinants of Coronary Heart Disease in Hispanic/Latinos: Electronic Health Records

Geetha Chittoor[1*], Navya Shilpa Josyula[1], Yao Tu[2], Lindsay Fernandez-Rhodes[2], Misa Graff[3], Elizabeth Frankel[4], Lauren E. Petty[4], Jennifer E. Below[4], Kari E. North[3], 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, State College, Pennsylvania, United States of America;* [3]*Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America;* [4]*Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America*

Electronic Health Record (EHR) data/biobanks from health systems are increasingly being utilized for common disease research. EHR enhances power to discover novel genetic variants associated with cardiometabolic traits in understudied populations. Hence, we examined genetic variants influencing coronary heart disease (CHD) in Hispanic/Latino (HL) adult participants using EHR data from MyCode Community Health Initiative (MyCode). CHD cases were defined using presence of ICD9 or ICD10 codes in at least two different encounters, or pertinent CHD procedure code from EHR. MyCode data included 4,261 self-identified HL including 322 cases (50% female) and 3,939 controls (73% female). Genome-wide association was conducted in SAIGE adjusting CHD for age, sex, array, and principal components using a generalized linear mixed model. We identified a suggestively significant locus associated with CHD on chromosome 22 (rs146418783, β±SE=4.46±0.87, P=2.86E-07) in *CELSR1*, cadherin epidermal growth factor laminin G seven-pass G-type receptor. These cadherin domains act as homophilic binding regions and the EGF-like domains involved in cell adhesion and receptor-ligand interactions. *CELSR1* was previously reported as a susceptibility gene for familial bicuspid aortic valve and hypoplastic left heart syndrome. We also identified seven novel associations on chromosomes 10 (rs60327547), 20 (rs58041415), 3 (rs75725366, rs4234563), 6 (rs35778273, rs80305422), and 22 (rs12106531) associated with CHD (P<5E-06). We replicated several CHD loci including on chromosomes 16 (rs12444314 [*FOXL1*]), 2 (rs1550115 [*ADCY3*]), and 8 (rs66778572 [*LPL*]). Our results highlight several loci offering insights into CHD etiology in HL. Further, we demonstrated the utility of EHR in identifying susceptibility loci influencing cardiometabolic traits in HL.

27

Multi-ancestry Genome-wide Association Studies of ACE Inhibitor-induced Cough and Chronic Dry Cough Implicate Neurophysiological Functions

Kayesha Coley[1], Jonas Ghouse[2,3], David J. Shepherd[1], Richard Packer[1,4], Catherine John[1], Robert C. Free[4,5], Edward J. Hollox[6], Louise V. Wain[1,4], Martin D. Tobin[1,4], Chiara Batini[1,4]

[1]*Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom;* [2]*Laboratory for Molecular Cardiology, Department of Cardiology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark;* [3]*Laboratory for Molecular Cardiology, Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark;* [4]*National Institute*

for Health and Care Research Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom; [5]School of Computing and Mathematical Sciences, University of Leicester, Leicester, United Kingdom; [6]Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

ACE inhibitors (ACEis) are commonly prescribed for hypertension, an important risk factor for cardiovascular disease. Their most common adverse reaction (AR) is a dry cough which affects 5-35% of users. As clinical guidelines recommend a switch to an angiotensin-II receptor blocker in response to this AR, we have leveraged such drug switches recorded in electronic health records as a proxy for ACEi-induced cough. We investigated the genetic determinants of ACEi-induced cough using a combination of two methodological approaches: (1) a single-trait, two-stage joint meta-analysis including 20,704 cases and 55,793 controls, and (2) a multi-trait genome-wide association meta-analysis with chronic dry cough including 28,339 cases and 144,446 controls. Both analyses included individuals of diverse genetic ancestries from the UK Biobank, EXCEED Study, eMERGE Network and Copenhagen Hospital Biobank. The undertaking of the multi-trait analysis was supported by a strong genetic correlation between the two traits and a significant association between single-trait sentinel rs6062847-T (*NTSR1*) and chronic dry cough in a phenome-wide association study. Functionally informed fine-mapping of sentinel loci was used to inform variant-to-gene mapping. Across both analyses, we identified 12 genome-wide significant (p value $<5\times10^{-8}$) sentinel variants which mapped to 12 protein-coding genes. Among these, seven sentinels and seven genes are novel. Five of the mapped genes encode proteins with neurological functions, including neuropeptide metabolism (*NTSR1*, *PREP*) and neurotransmission (*BTBD11*, *GRIA3*, *KCNIP4*, *NTSR1*). This supports current hypotheses of bradykinin-related sensitivity mediated by ACEi treatment and indicates a neurophysiological pathology of the cough reflex.

28

APOE4-stratified GWAS of Multiple Cognitive Domains in Non-Hispanic White and Non-Hispanic Black Older Adults

Alex G Contreras[1*], Skylar Walters[1], Jaclyn M. Eissman[1,2], Alexandra N. Smith[1], Shubhabrata Mukherjee[3], Michael L. Lee[3], Seo-Eun Choi[3], Phoebe Scollard[3], Emily H. Trittschuh[4,5], Jesse B. Mez[6], William S. Bush[7], Brian W. Kunkle[8], Adam C. Naj[11,12], Katherine A. Gifford[1], Murat Bilgel[28], Amanda B. Kuzma[12], The Alzheimer's Disease Neuroimaging Initiative (ADNI), Alzheimer's Disease Genetics Consortium (ADGC), The Alzheimer's Disease Sequencing Project (ADSP), Michael L. Cuccaro[8], Carlos Cruchaga[9,10], Margaret A. Pericak-Vance[8], Lindsay A. Farrer[6,13,14], Li-San Wang[12]; Gerard D. Schellenberg[12], Richard P. Mayeux[15,16,17], Jonathan L. Haines[7], Angela L. Jefferson[1], Walter A. Kukull[18], C. Dirk Keene[19], Andrew J. Saykin[20,21], Paul M. Thompson[22], Eden R. Martin[8], Marilyn S. Albert[23], Sterling C. Johnson[24], Corinne D. Engelman[24], Luigi Ferrucci[25], David A. Bennett[26], Lisa L. Barnes[26], Julie A. Schneider[26], Susan M. Resnick[28], Reisa A. Sperling[27], Paul K. Crane[3], Timothy J. Hohman[1,2], Logan Dumitrescu[1,2]

[1]Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Medicine, University of Washington, Seattle, WA, United States of America; [4]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, Washington, United States of America; [5]VA Puget Sound Health Care System, GRECC, Seattle, Washington, United States of America; [6]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, United States of America; [7]Cleveland Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America; [8]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, Florida, United States of America; [9]Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri, United States of America; [10]NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, Missouri, United States of America; [11]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [12]Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [13]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [14]Department of Medicine (Biomedical Genetics), Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States of America; [15]Columbia University, New York, New York, United States of America; [16]The Taub Institute for Research on Alzheimer's Disease and The Aging Brain, Columbia University, New York, New York, United States of America; [17]The Institute for Genomic Medicine, Columbia University Medical Center and The New York Presbyterian Hospital, New York, New York, United States of America; [18]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, United States of America; [19]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington, United States of America; [20]Department of Radiology and Imaging Services, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [21]Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, Indiana, United States of America; [22]Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [23]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America; [24]Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [25]Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, United States of America; [26]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, United States of America; [27]Department of Neurology, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts, United States of America; [28]Laboratory of Behavioral Neuroscience National Institute on Aging National Institutes of Health, Baltimore, Maryland, United States of America

Apolipoprotein-E4 (APOE-ε4) is a major genetic risk factor for late-onset Alzheimer's disease (AD). However, the impact of APOE-ε4 carrier status on the genetic architecture of cognitive decline remains unclear. We conducted stratified genome-wide association analyses (GWAS) to identify genetic associations with cognitive decline in APOE-ε4 carriers and non-carriers, as well as across two racial/ethnic groups (non-Hispanic Whites and non-Hispanic Blacks). We analyzed a harmonized cognitive dataset from ten aging and AD cohorts, encompassing a large multi-ancestry population (Ntotal=36,483; NHW_ε4pos=12,047; NHW_ε4neg=20,253; NHB_ε4pos=1,810; NHB_ ε4neg=2,373). GWAS were performed for memory, executive function, and language at baseline and for longitudinal decline, adjusting for age, sex, and the nine within each ancestry group by APOE-ε4 status, followed by fixed-effect meta-analysis. We identified 14 novel associations among APOE-ε4 carriers, including variants near the GRIN3A and NME7 genes among non-Hispanic Whites and variants near the WDPCP gene among non-Hispanic Blacks. In contrast, we found ten associations among APOE-ε4 non-carriers, including variants near the LOC101927668 gene in non-Hispanic Whites and GALNT7 genes in non-Hispanic Blacks. Notably, eQTL evidence for the association on chromosome n among non-Hispanic White APOE-ε4 carriers/non-carriers implicated the GRIN3A gene. Sensitivity analyses, excluding comorbidities/subsetting individuals aged 60 and above, were performed to confirm our findings. These preliminary results have implications for precision medicine approaches targeting cognitive impairment and AD prevention. Future work will explore cross-ancestry analysis, gene-based tests, and genetic correlations.

29

Comparative Analysis of Whole Exome Sequencing in Bipolar Disorder Case-Control Data

Brandon J. Coombes[1*], Nicholas B. Larson[1], Wenan Chen[1], Anthony Batzler[1], Lindsay Melhuish Beaupre[1], Joanna M. Biernacka[1,2]

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America; [2]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, United States of America

Bipolar disorder (BD) is a complex psychiatric condition with a strong genetic component. To gain insights into the impact of rare variants (RVs) on BD, we performed whole exome sequencing (WES) analysis using data from the Mayo Clinic Bipolar Disorder Biobank (n =1,600 cases) and the Mayo Clinic Biobank (n 50K controls). In this study, we adopted two different strategies to explore the rare variants associated with BD, comparing the outcomes of each approach. With the first strategy, we used a standard RV analysis approach, which involved removing a small subset of unrelated individuals (mostly controls) from our dataset and applying burden and sequence kernel association tests (SKAT) to identify genes with RVs that may confer susceptibility to BD. With the second analysis strategy, we implemented the REGENIE method, which employs a novel approach to improve power by adjusting for background genetic signal. The REGENIE method also incorporates related individuals in the analysis which may also improve power. We will compare the results obtained from both analyses, assessing the performance and robustness of each approach in uncovering genes associated with BD. The motivation behind this comparison is to explore the potential advantages and drawbacks of adopting a newer tool in contrast to a conventional approach. By evaluating the performance and outcomes of both methods, we seek to assess the feasibility and potential advancements offered by REGENIE. This comparative analysis will provide valuable insights into the applicability and potential benefits of incorporating REGENIE in future genetic association studies.

30

Exploring Similarities and Differences Between Local Genetic Correlation Methods

Rupal L. Shah[1], Rebecca Darlay[1], Richard M. Dodds[2], Anand T.N. Nair[3], Ewan R. Pearson[3], Heather J. Cordell[1] on behalf of the ADMISSION Research Collaborative

[1]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [2]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [3]Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, United Kingdom

Genetic correlation analysis provides useful insight into the shared genetic basis between pairs of traits or conditions of interest. However, most genome-wide analyses performed using tools such as LDSC only inform about the degree of overall genetic similarity and do not identify the specific genomic regions that give rise to this similarity. Identification of the key contributing regions could allow them to be prioritized for investigation of potential shared biological mechanisms. In recent years, several statistical tools (e.g. rho-HESS, SUPERGNOVA, and LAVA) have been developed to investigate local genetic correlations. These tools partition the genome into multiple segments and provide estimates of the genetic correlation captured by each individual segment. Using publicly available European ancestry genome-wide association study (GWAS) summary statistics as input, local genetic correlations were estimated using each of these tools for the following commonly occurring pairs of conditions: Hypertension and atrial fibrillation and flutter, hypertension and chronic kidney disease, and hypertension and Type 2 diabetes. Despite each of the three methods aiming to address the same question, results were found not to be entirely consistent across tools, with some identified regions overlapping across tools and others implicated only by a single tool. The work presented here highlights the similarities and differences between the results obtained from these methods and attempts to explore the potential reasons underlying these differences.

31

Epigenome-wide Associations with Age and Sex and Asthma

Denise Daley[1,2], Dentisa Vasileva[1], Deep Patel[1], Ming Wan[1], Andrew Sandford[1,2], Edmond Chan[3], Allan Becker[4], Catherine Laprise[5] and Celia M T Greenwood[6]

[1]Centre for Heart Lung Innovation, University of British Columbia and Saint Paul's Hospital, Vancouver British Columbia, Canada; [2]Department of Medicine, Respiratory Division, University of British Columbia, Vancouver, British Columbia, Canada; [3]BC Children's Hospital Research Institute, Faculty of Medicine, Vancouver, Canada;[4] Department of Pediatrics and Child Health, University of Manitoba, Manitoba, Canada; [5]Centre; intersectoriel

en santé durable (CISD) de l'Université du Québec á Chicoutimi, Saguenay, Canada, Centre intégré universitaire de santé et de services sociaux (CIUSSS) du Saguenay–Lac-Saint-Jean, Saguenay, Canada; [6]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [7]McGill University: Gerald Bronfman Department of Oncology, Department of Epidemiology & Biostatistics, and Department of Human Genetics

Diseases such as asthma, autism and auto-immune disorders demonstrate age and sex specific prevalence patterns. Childhood asthmatics are predominately male while the majority of adult asthmatics are women. Mechanisms for sex-specific prevalence patterns are unknown.

Associations with age, sex, and asthma were identified using a targeted methylation sequencing approach that assessed methylation at ~5.2 million CpG sites in 795 samples from two studies. The Canadian Asthma Primary Prevention Study (CAPPS, n=626) is a longitudinal birth cohort recruited during the 2nd and 3rd trimester of pregnancy and followed to age 15. The Saguenay–Lac-Saint-Jean study (SLSJ, n=169 samples) consists of three-generational triads. Associations (p<1x10^{-8}) with asthma (n=13), age (n=559,870) and sex (cord blood n=562, venous blood n=1,268, present in both n=385) were identified using multistage mixed effects regression. Annotation of CpG's differentially methylated by sex identified genes associated with phenotypes that have sex-specific prevalence patterns such as Alzheimer's (*RAP2CP1, RFPL2*), asthma (*MED15P3, PRRT4, PRH1-PRR4*), autism (*RNF168, DPPA5, GRIN1, GABRA5*), Ehlers-Danlos (*PLCH2, HES5, DSE, OLFM1, OBP2B*), Sjögren's (MED15P5, *FOXI2, CCDC177*), imprinted regions of chromosome 15 and Klinefelter's (XXY, *RFPL2, DPPA5*). These data indicate sex specific methylation may contribute to Mendelian, imprinted and complex disorders with sex specific disease prevalence.

32

Integrative Genetic and Exposomic Analysis of Mental Health Problems in Early Childhood

Hayley A. Sowards, Karanvir Singh, Karmel W. Choi, Jessica K. Dennis*

[1]Department of Medical Genetics, University of British Columbia; [2]Department of Psychiatry, Harvard Medical School/Massachusetts General Hospital

Genes and environment contribute to the risk of developing mental health problems, but their relative contribution is dynamic as individuals age, and few studies have focused on early childhood. Moreover, until recently, few methods existed to quantify the collective contribution of many environmental exposures at once (i.e., the "exposome"), and their interplay with the genome in population-based samples. In this study, we use an integrative genetic and exposomic analysis of internalizing (e.g., depression/anxiety) and externalizing (e.g., aggressive behavior) traits measured in a cohort of 2,400 Canadian children at age 5 and 8, participating in the Canadian Healthy Infant Longitudinal Development (CHILD) study.

Our approach quantifies the proportion of variability in internalizing and externalizing problem scores explained by genetic and exposomic factors using a recently developed linear mixed effect model that extends the principles of genetic relationship matrices to the exposome. A unique strength of the CHILD study is the availability of maternal genotype data, which allows us to additionally quantify the variance explained by the prenatal exposome (e.g., pregnancy complications, maternal medication use in pregnancy) while separating out the direct effect of maternal genotype (i.e., genotype transmitted to offspring) from the indirect effect of maternal genotype (i.e., genotype that influences offspring outcomes indirectly through the prenatal environment, "genetic nurture"). Results will advance our understanding of how the genome, exposome, and their interplay contribute to signs of mental health problems in early childhood, and the linear mixed model approach is broadly applicable to cohorts and biobanks with rich exposomic data.

33

APOE-ε4 and Coronary Artery Disease: Effect Modification by Sex and Gender in a Multi-ancestry Sample from the UK Biobank

Tatiana Dessy*[1,2,3,5], Johanna Sandoval[1,2], Louis-Philippe Lemieux Perreault[1,2], Marie-Christyne Cyr[1,2], Sylvie Provost[1,2], Marie-Pierre Sylvestre[1,3,5], Sarah Gagliano Taliun[1,4,6], Marie-Pierre Dubé[1,2,3,4]

[1]Montreal Heart Institute, Montréal, Quebec, Canada; [2]Université de Montréal Beaulieu-Saucier Pharmacogenomics Center, Montréal, Quebec, Canada; [3]Department of Social and Preventive Medicine, Université de Montréal, Montréal, Quebec, Canada; [4]Department of Medicine, Université de Montréal, Quebec, Canada; [5]Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, Quebec, Canada; [6]Department of Neuroscience, Université de Montréal, Quebec, Canada.

BACKGROUND. Biological and socio-cultural risk factors contribute to coronary artery disease (CAD) in men and women. Evidence of an interplay between APOE-ε4, sex, and psychosocial factors relating to sex and gender remains sparse. We aim to evaluate whether sex and gender independently modify the effect of APOE-ε4 on CAD.

METHODS. We evaluated effect modification in n=364,793 participants from the UK Biobank pan-ancestry dataset (return #2442) that were without prevalent cardiovascular disease or neurocognitive impairment. Phased genotypes were used to derive *APOE* diplotypes and define APOE-ε4 carrier status. Gender was modeled using a previously constructed literature-based femininity score (FS) leveraging six psychosocial factors. We estimated adjusted generalized linear models (GLM) of CAD since random effects for the ancestry group variable were not informative in generalized linear mixed models. We estimated GLM stratified by sex including interaction terms for APOE-ε4 and FS. We estimated GLM in the sex-combined sample with interaction terms for APOE-ε4 and sex and adjusted for FS and covariates.

RESULTS: The incidence of CAD was 8.1%. In males, FS (p=0.045) and APOE-ε4 (p<0.001) were associated with increased CAD risk (p_{int}=0.554), whereas FS was associated with increased CAD risk in females (p<0.001), but not APOE-ε4 (p=0.20; p_{int}=0.068). Evidence for modification of the effect of APOE-ε4 on CAD by sex when adjusted for FS shifted from p_{int}=0.091 to p_{int}=0.050.

DISCUSSION: Sex and gender independently contribute to CAD risk. More analyses are needed to clarify the mechanistic role of gender on genetic CAD risk in both sexes.

34

Proteome-wide Mendelian Randomization of Adverse Outcomes in Human Heart Failure

Marie-Joe Dib[1*], Michael Levin[1,2], Lei Zhao[3], Ching-Pin Chang[3], Christina Ebert[3], Dipender Gill[4], Stephen Burgess[5,6], David A. Gordon[3], Thomas P. Cappola[1,3], Julio A. Chirinos[3]

[1]*Division of Cardiovascular Medicine, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, United States of America;* [2]*University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America;* [3] *Bristol-Myers Squibb Company, Lawrenceville, New Jersey, United States of America;* [4]*Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, United Kingdom;* [5]*MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom;* [6] *Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom*

Identifying novel molecular drivers of disease progression in heart failure (HF) is a high-priority goal that may provide new therapeutic targets to improve patient outcomes. We aimed to assess the relationship between plasma proteins and adverse outcomes in HF and assess the causal role of various proteins using Mendelian randomization (MR). We measured ~5,000 plasma proteins (SomaScan assay) among 1,964 participants with HF with a reduced left ventricular ejection fraction enrolled in the Penn HF study (PHFS). We assessed the observational relationship between plasma proteins and: 1) all-cause death; 2) death or HF-related hospital admission (DHFA). Proteins significantly associated with outcomes were the subject of two-sample Mendelian randomization (MR) and colocalization analyses using blood protein quantitative trait loci (pQTL) from the deCODE and Fenland studies to test the putative causal effects of these proteins on HF outcomes. After correction for multiple comparisons, we found 243 and 126 proteins significantly associated with death and DHFA, respectively. Mendelian randomization and colocalization analyses provided converging evidence of potentially causal effects of six proteins (CCDC126, CD55, CCL14, NEGR1, SVEP1 and ADH7) on DHFA and 11 proteins on death (RSPO4, FCN2, IGLL1, HPGDS, FGF23, EFEMP1, STC1, ATOX1, FCN2, SVEP1, ANG) at nominal significance ($p < 0.05$). Our study implicates multiple novel proteins in HF and provides preliminary evidence of a potentially causal association between plasma levels of 17 circulating plasma proteins the risk for adverse outcomes in human HF. Whether these proteins represent suitable therapeutic targets should be the focus of future studies.

35

Genetic Connection between Coronary Artery Disease and Stroke

Kexin Ding[1,2], Kun Wang[1,2], Xiaoyi Li[1,2], Xueying Qin[1,2]†*, Yiqun Wu[1,2]†*

[1]*Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China;* [2]*Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China*

†Equal contribution

**Corresponding to: Xueying Qin (xueyingqin@bjmu.edu.cn) and Yiqun Wu (qywu118@163.com)*

Coronary artery disease (CAD) and stroke are highly complex traits that impose a significant disease burden globally. Both conditions share a range of metabolic risk factors and are similarly influenced by systemic inflammation, ischemia caused by atherosclerosis, and other pathophysiological processes. Genome-wide association studies (GWAS) have identified numerous risk loci for CAD and stroke, many of which display cross-trait associations with metabolic risk factors. While both CAD and stroke involve abnormal conditions in the vascular system, it remains unclear to what extent the genetic architecture underlying coronary and cerebral arteries is shared. Therefore, this study aimed to examine the genetic intersection between CAD and stroke. Summary statistics of GWAS in individuals of European ancestry with the largest sample size for CAD (n=1,165,690) and all strokes (AS, n=446,696) or all ischemic strokes (AIS, n=440,328) were analyzed. The genetic correlations between traits were estimated by LD score regression, while the overlap polygenic components were quantified using MiXeR. Pleiotropic genetic loci were then identified using the conjunctional false discovery rate method, and the causal variants were prioritized through statistical colocalization analysis by HyPrColoc. Our findings reveal that CAD is positively genetically correlated to AS and AIS, with a genetic correlation of 0.53 and 0.50, respectively. We estimate that there are around 1.5K, 1.0K, and 1.3K variants that causally influence CAD, AS, and AIS, respectively. Among these variants, 0.9K are shared between CAD and AS, while 1.0K are shared between CAD and AIS. We then identify 83 independent loci (54 novels) having pleiotropic effects on CAD and AS, while 82 loci (59 novels) for CAD and AIS. Of these, six loci exhibit a posterior probability greater than 0.8 associated with both CAD and AS, while six loci for CAD and AIS. The identified regions contain genes such as *SH2B3*, *CASZ1*, *CDKN2B-AS1*, and *GOSR2*, suggesting the involvement of the neurotrophic signaling pathway, cell cycle regulation, cellular senescence, and transforming growth factor-beta signaling pathway in the pathogenesis of these diseases. Our study supports the notion that the established epidemiological link between CAD and stroke has a genetic underpinning, which sheds light on the common pathogenesis of cardiovascular and cerebrovascular diseases.

37

Construction, Evaluation, and AOP Framework-based Application of the EpPRS as a Genetic Surrogate for Assessing Environmental Pollutants

Silu Chen[1,2], Mulong Du[2,3,*], Junyi Xin[1,4], Rui Zheng[1,2], Zhengdong Zhang[1,2], Meilin Wang[1,2]

[1]*Department of Environmental Genomics, Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China;* [2]*Department of Genetic Toxicology, The Key Laboratory of Modern Toxicology of Ministry of Education, School of Public Health, Nanjing Medical University, Nanjing, China;* [3]*Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China;* [4]*Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, Jiangsu, China.*

Detection of humans' internal exposure to environmental pollutants is cost-intensive, time-consuming, and energy-consuming. Polygenic risk scores (PRSs) have been widely applied in genetic studies of complex trait or diseases. It is urgent to construct a genetically relevant environmental surrogate for pollutant exposure and to explore its utility for disease prediction and risk assessment. This study enrolled 714 individuals with complete genomic data and exposomic data on 22 plasma-persistent organic pollutants (POPs). We first conducted 22 POP genome-wide association studies (GWAS) and constructed the corresponding environmental pollutant-based PRS (EpPRS) by clumping and p-value thresholding (C+T), lassosum, and PRS-CS methods. The best-fit EpPRS was chosen by its regression R^2. An adverse outcome pathway (AOP) framework was developed to assess the effects of contaminants on candidate disease. The C+T method produced the best-performing EpPRSs for seven PCBs and four PBDEs. EpPRSs well replicated the correlations of environmental exposure measurements based on consistent patterns. The diagnostic performance of Type 2 diabetes mellitus (T2DM) was improved upon by the combined model of T2DM-PRS and EpPRS of PCB126/BDE153. Finally, the AKT1-mediated AOP framework illustrated that PCB126 and BDE153 may increase the risk of T2DM by decreasing AKT1 expression through the cGMP-PKG pathway and promoting abnormal glucose homeostasis. EpPRSs can be an equivalent substitute for assessing pollutant internal exposure. The application of the EpPRS to disease risk assessment can finely reveal the toxic pathway and mode of action linking exposure and disease, which provides a decision basis for the environmental pollutant control strategy.

38

Relating Gene Co-expression qtls to PRSs via Individual-Specific Networks

Edoardo Gervasoni[1*], Behnam Yousefi[2,3,4], Federico Melograna[2], Benno Schwikowski[3], Iain R. Konigsberg[5], Katerina J. Kechris[6], Peter J. Castaldi[8,9], Michael H. Cho[8,10], Kristel Van Steen[1,2]

[1]BIO3 Systems Genetics, GIGA-R, Université de Liège, Liège, Belgium; [2]BIO3 Systems Medicine, Department of Human Genetics, KU Leuven, Leuven, Belgium; [3]Computational Systems Biomedicine, Institut Pasteur, Université Paris Cité, Paris, France; [4]Ecole Doctorale Complexite du Vivant, Sorbonne Université, Paris, France; [5]Department of Biomedical Informatics, Anschutz Medical Campus, University of Colorado, Aurora, Colorado; United States of America; [6] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [8]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [9]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America; [10]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

In this work, we assess the complementary value of FEV1 polygenic risk score (PRS) and individual-specific gene expression networks (ISNs - Kuijjer et al. 2019), obtained by reverse-engineering a gene co-expression network derived from RNAseq data from COPDgene (www.copdgene.org) with 4,524 genes and 3,679 individuals. Based on subnetworks that exhibit significant variation across ISNs, we prioritize gene-pairs. Via the ISNs we constructed, such a gene pair has a realization (weight) for each individual and thus can be considered as an outcome in co-eqtl (co-expression qtl) epistasis analysis with MB-MDR (Van Lishout et al. 2016). The resulting statistical epistasis networks show multiple hub genes. For instance, PTPRF for which the predominant function is the regulation of the actin cytoskeleton; dysfunction of the gene may be involved in the greater cellular plasticity observed in human bronchial epithelia in COPD patients. Upregulation of another hub gene, GMDS-DT/GMDS-AS1, has been shown to stop the proliferation of lung adenocarcinoma cells and promote cell apoptosis. One pathway through which this occurs is through upregulation of IL1β of which levels are statistically significantly higher in COPD patients compared to controls. In addition, we demonstrate the clinical utility of our findings by associating ISN based gene-based co-expressions with PRS: Our strategy revealed several instances where co-expression – PRS trends significantly depended on multilocus genotypes.

39

Sex Differences in the Genetic Architecture of Alzheimer's Disease Cognitive Endophenotypes

Jaclyn M. Eissman*[1,2]; Alexandra N. Smith*[1]; Skylar Walters[1]; Shubhabrata Mukherjee[3]; Michael L. Lee[3]; Seo-Eun Choi[3]; Phoebe Scollard[3]; Emily H. Trittschuh[4,5]; Jesse B. Mez[6]; William S. Bush[7]; Brian W. Kunkle[8]; Adam C. Naj[11,12]; Katherine A. Gifford[1]; Murat Bilgel[28]; Amanda B. Kuzma[12]; The Alzheimer's Disease Neuroimaging Initiative (ADNI); Alzheimer's Disease Genetics Consortium (ADGC); The Alzheimer's Disease Sequencing Project (ADSP); Michael L. Cuccaro[8]; Carlos Cruchaga[9,10]; Margaret A. Pericak-Vance[8]; Lindsay A. Farrer[6,13,14]; Li-San Wang[12]; Gerard D. Schellenberg[12]; Richard P. Mayeux[15,16,17]; Jonathan L. Haines[7]; Angela L. Jefferson[1]; Walter A. Kukull[18]; C. Dirk Keene[19]; Andrew J. Saykin[20,21]; Paul M. Thompson[22]; Eden R. Martin[8]; Marilyn S. Albert[23]; Sterling C. Johnson[24]; Corinne D. Engelman[24]; Luigi Ferrucci[25]; David A. Bennett[26]; Lisa L. Barnes[26]; Julie A. Schneider[26]; Susan M. Resnick[28]; Reisa A. Sperling[27]; Paul K. Crane[3]; Timothy J. Hohman[1,2]; Logan Dumitrescu[1,2]

**Equally contributed to this work*

[1]Vanderbilt Memory & Alzheimer's Center, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [4]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, Washington, United States of America; [5]VA Puget Sound Health Care System, GRECC, Seattle, Washington, United States of America; [6]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, United States of America; [7]Cleveland Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America; [8]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, Florida, United States of America; [9]Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri, United States of

America; [10]NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, Missouri, United States of America; [11]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [12]Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [13]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [14]Department of Medicine (Biomedical Genetics), Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States of America; [15]Columbia University, New York, New York, United States of America; [16]The Taub Institute for Research on Alzheimer's Disease and The Aging Brain, Columbia University, New York, New York, United States of America; [17]The Institute for Genomic Medicine, Columbia University Medical Center and The New York Presbyterian Hospital, New York, New York, United States of America; [18]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, United States of America; [19]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington, United States of America; [20]Department of Radiology and Imaging Services, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [21]Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, Indiana, United States of America; [22]Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [23]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America; [24]Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [25]Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, United States of America; [26]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, United States of America; [27]Department of Neurology, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts, United States of America; [28]Laboratory of Behavioral Neuroscience National Institute on Aging National Institutes of Health, Baltimore, Maryland, United States of America.

Multiple domains of cognition show sex/gender differences in aging and Alzheimer's disease (AD). Genetic studies of cognition in older adults have identified significant loci, but these studies have not considered sex/gender. We investigated if genetic contributors to cognition differed by sex, leveraging harmonized memory, executive functioning, language, and global cognition scores across nine cohorts of cognitive aging. For each domain, we conducted sex-stratified and sex-interaction genome-wide meta-analyses among a cross-ancestral sample of ~35,000 (mean baseline age=73 years; 57% female; 14% AD cases; average number of visits>4; 88% non-Hispanic White; 12% non-Hispanic Black), covarying for age and ancestry principal components. Beyond *APOE*, we identified three sex-specific loci: *HS3ST4* associated with global cognitive decline among males (rs10852291: β_M=-5.29x10^{-3}, P_M=4.73x10^{-8}, β_{int}=6.26x10^{-3}, P_{int}=2.77x10^{-6}) but not among females (β_F=7.70x10^{-4}, P_F=0.35). This locus associated with fluid intelligence in males in the UK Biobank (β_M=-2.30x10^{-2}, P_M=7.49x10^{-3}; β_F=1.44x10^{-3}, P_F=0.86). *CTNNAL1* interacted with sex on baseline executive functioning (rs112083348: β_{int}=-0.10, P_{int}=2.67x10^{-8}; β_M=0.05, P_M=5.18x10^{-4}; β_F=-0.05, P_F=4.92x10^{-5}). This locus associated with an executive functioning-related task in the UK Biobank (β_M=1.63x10^{-2}, P_M=6.07x10^{-3}; β_F=5.88x10^{-3}, P_F=0.28) and was an eQTL for *CTNNAL1* in two AD-vulnerable brain regions. *VRK2* associated with language decline among females (rs11898834: β_F=2.75x10^{-3}, P_F=9.99x10^{-10}, β_{int}=2.94x10^{-3}, P_{int}=6.93x10^{-5}) but not among males (β_M=-7.50x10^{-5}, P_M=0.89). rs11898834 is an eQTL for *VRK2* in blood, and *VRK2* associated sex-specifically with developmental stuttering in 23&Me data. Together, our results support a sex-specific component to the genetic architecture of cognition. We look forward next to incorporating bulk and single-cell transcriptomics into our sex-aware exploration of cognition.

40

Coronary Heart Disease and Type 2 Diabetes Metabolomic Signatures in a Middle Eastern Cohort

Mohamed Elshrif[1*], Keivin Isufaj[1], Ayman El-Menyar[2,3], Khalid Kunji[1], Ehsan Ullah[1], Reem Elsousy[4], Haira R. B. Mokhtar[5], Eiman Ahmad[5], Maryam Al-Nesf[6], Alka Beotra[5], Mohammed Al-Maadheed[5], Vidya Mohamed-Ali[5], Mohamad Saad[1], and Jassim Al Suwaidi[4]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar; [2]Clinical Research, Trauma & Vascular Surgery, Hamad Medical Corporation, Qatar; [3]Department of Clinical Medicine, Weill Cornell Medical College, Qatar; [4]Department of Cardiology, Heart Hospital, Hamad Medical Corporation, Qatar; [5]Anti-doping Lab Qatar, Qatar; [6] Department of Internal Medicine, Allergy and Immunology, Hamad Medical Corporation, Qatar

**Presenting author*

Background: The growing field of metabolomics has opened new venues for identifying biomarkers of Type 2 diabetes (T2D) and predicting its consequences such as coronary heart disease (CHD). Middle Eastern populations, despite their large size, are underrepresented in omics research.

Methods: In this study, we used a total of 641 metabolites from a large cohort of 3,679 Qatari adults from the Qatar Biobank (QBB; 272 T2D and 2438 non-T2D individuals) and Qatar Cardiovascular Biorepository (QCBio; all CHD patients; 488 T2D and 481 non-T2D individuals). Univariate analysis was performed to identify metabolites associated with T2D, in the absence or presence of CHD. Multivariate analysis, machine learning (ML) models, and metabolite risk scores were developed to assess the predictive power of the different combination of T2D and CHD.

Results: Many metabolites were significantly associated with T2D in both QBB and QCBio cohorts. Among these, we observed, for example, 1,5-anhydroglucitol (P=1.33×10^{-68} in QBB vs 9.82×10^{-33} in QCBio). Other metabolites were significantly associated with T2D only in one cohort. ML models performed well to predict all T2D and CHD combinations with accuracy reaching 80%. The MRS developed in QCBio and tested in QBB while adjusting for HbA1C yielded an Odds Ratio of 21.18 for the top quintile vs. the remaining quintiles.

Conclusions: Metabolomics profiling has the potential for early detection of metabolic alterations that precede clinical symptoms of T2D and CHD in the presence of T2D. This early detection potential allows for timely interventions and improved management strategies for both T2D and CHD.

41

X Chromosome Association Study of Asthma

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

[1]*Institut Universitaire de Cardiologie et de Pneumologie de Québec, Université Laval, Quebec, Quebec, Canada;* [2]*Department of Social and Preventive Medicine, Laval University, Quebec, Quebec, Canada;* [3]*Department of Molecular Biology, Medical Biochemistry and Pathology, Laval University, Quebec, Quebec, Canada;* [4]*Department of Molecular Medicine, Laval University, Quebec, Quebec, Canada*

Context: Asthma is a common respiratory disease with both genetic and environmental risk factors. Heritability estimates for asthma range from 0.55 to 0.90. Numerous genome-wide association studies (GWASs) have been completed and have identified several single nucleotide polymorphisms (SNPs) associated with asthma. However, because of the difficulty in analyzing the X chromosome, these published asthma GWASs are mostly focused on autosomal variants (chromosomes 1-22, non-sex chromosomes) and ignore the sex chromosomes. The prevalence and course of asthma is known to be sex-specific. It is thus important to take into account the effect of the X chromosome in genetic studies.

Aim: Our research aims to perform X chromosome association in the Quebec City Case-Control Asthma Cohort (QCCCAC) which consists of 1,618 French-Canadian subjects (1,089 asthmatics [660 women and 429 men] and 519 healthy controls [330 women and 199 men]).

Methodology: Genotyping performed using the Illumina Global Screening Array BeadChip. After filtering and imputation (TOPMed reference panel ,we tested more than one million variants in chromosome X for association analysis using SAIGE (Scalable and Accurate Implementation of Generalized mixed model) adjusting for age, sex, and the first 20 ancestry-based principal components.

Results: Five independent variants were identified at a suggestive association threshold of p value $< 1 \times 10^{-5}$ in the X chromosome analysis. The most significant variant is located in *the NHS* gene. The allele frequencies in asthma cases and controls were 0.017 and 0.055, respectively.

Conclusion: Performing genetic association analysis on the X chromosome is a valuable tool for researchers to gain a better understanding of complex diseases. We will compare several statistical methods to study the association in chromosome X.

42

Genomics Policy Analysis in Precision Medicine with Health Disparity Lens

Kelly R. Estilette*

Department of Social, Behavioral, and Population Sciences, Tulane University School of Public Health and Tropical Medicine, New Orleans, Louisiana, United States of America

American Public Health Association, Genomics Forum

Precision medicine is a burgeoning field that will continue to change the very meaning of disease prediction, prevention, detection, and treatment through the use of genomics data. The benefits of this enhanced state of health should be experienced equitably by all. Analyzing the current state of policy within the field of precision medicine will allow for the development of inclusive policy and the design of systems that promote equity. The eightfold Bardach policy analysis method is used to consider current policy, survey current areas of opportunity, and generate recommendations for exploration meant to increase equity and reduce disparities within the field of precision medicine.

43

Estimation of Prevalence and Carrier Frequency of Wilson's Disease in Thai Population Using Whole-genome Population Data Set

Paravee Own-eium[1,2], Sommon Klumsathian[1], Donniphat Dejsuphong[2], Bhoom Suktitipat[3], Jakris Eu-ahsunthornwattana[4*]

[1]*Center for Medical Genomics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand,* [2]*Program in Translational Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand,* [3]*Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand,* [4]*Department of Community Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand*

Wilson's disease (WD) is a rare autosomal recessive disease caused by mutations in the gene *ATP7B* resulting in copper accumulation in major organs leading to chronic hepatic, neurological, and psychological symptoms. Its prevalence varies among different populations. To estimate the prevalence and carrier frequency of WD in the Thai population, we use the SNP array data (Asian Screening Array: ASA; Illumina Corp, CA) of 6,291 individuals from the Genomics Thailand SNP Array project and whole genome sequencing (WGS) data from the Thai 500 WGS cohort and the 1000 Genomes Project. An imputed data set of SNPs within the gene *ATP7B* was also created using the SNP Array and WGS data. SNP data from these sources were compared with ClinVar and Wilson Disease Mutation Database. Disease variants were identified using either "relaxed" definition (variants classified as pathogenic or likely pathogenic in any databases) or "strict" definition (classified as such in both databases). Carrier frequency and disease prevalence were then estimated assuming Hardy-Weinberg Equilibrium. The WD carrier frequency and prevalence in Thai population was estimated to be between 1/24 to 1/62 and 1/2,125 to 1/14,992 using relaxed definition, or 1/78 to 1/92 and 1/24,128 to 1/33,251 using strict definition. These are in line with the previous epidemiological studies, particularly in East Asians. With more genome-wide data becoming available at lower cost, a similar method could be useful for estimating the prevalence and screening for autosomal recessive diseases in the population.

44

Direct and INdirect Effects Analysis of Genetic lOci (DINGO): Increasing the Power of Locus Discovery in GWAS Meta-Analyses of Perinatal Phenotypes

Liang-Dar Hwang[1], Gabriel Cuellar-Partida[2], Loic Yengo[1], Jian Zeng[1], Robin N. Beaumont[3], Rachel M. Freathy[3,4], Gunn-Helen Moen[1,5,6,7,8], Nicole M. Warrington[1,6,8], David M. Evans[1,4,8]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia; [2]Gilead Sciences, Inc, Forest City, California, United States of America; [3]Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, United Kingdom; [4]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [5]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway; [6]Department of Public Health and Nursing, K.G. Jebsen Center for Genetic Epidemiology, NTNU, Norwegian University of Science and Technology, Trondheim, Norway; [7]Population Health Science, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [8]The Frazer Institute, The University of Queensland, Woolloongabba, QLD, Australia

Perinatal traits are influenced by genetic variants from both fetal and maternal genomes. Genome-wide association studies (GWAS) of these phenotypes have typically involved separate fetal and maternal scans, however, this approach may be inefficient as it does not utilize the information shared across the individual GWAS. Here we investigate the performance of three strategies to detect loci in maternal and fetal GWAS of the same trait: 1)the traditional strategy of analyzing maternal and fetal GWAS separately; 2) a novel two degree of freedom test which combines information from maternal and fetal GWAS; and 3) a one degree of freedom test where signals from maternal and fetal GWAS are meta-analyzed together conditional on the estimated sample overlap. We demonstrate through a combination of analytical formulae and data simulation that the optimal strategy depends on the extent of sample overlap/relatedness between the maternal and fetal GWAS, whether loci jointly exhibit fetal and maternal effects, and if so, whether these effects are directionally concordant. We apply our methods to summary results statistics from a recent GWAS of birth weight from deCODE, UK Biobank and the Early Growth Genetics (EGG) consortium identifying >60 novel loci for the trait. We implement our methods in the online DINGO (**D**irect and **IN**direct effects analysis of **G**enetic l**O**ci) software package, which allows users to perform one and/or two degrees of freedom tests easily and computationally efficiently across the genome.

45

Dimensionality Reduction with Network Regularization in Single-cell Expression Analysis

Huaying Fang[1], Weilai Chi[2], Yuedong Zhou[2]

[1]Academy of Multidisciplinary Studies, Capital Normal University, Beijing, China; [2]School of Mathematical Sciences, Capital Normal University, Beijing, China

Single-cell RNA sequencing (scRNA-seq) technology can measure gene expression abundance in single cells. However, there are two challenges for analyzing scRNA-seq data owing to technical limitations and expense budgets. One challenge is the dropout problem that results in zero read counts for many genes in an individual cell; the other challenge is that the number of genes is often much larger than the sample size. In this talk, we will introduce a novel method for simultaneously imputing zeros and reducing data dimensionality for analyzing scRNA-seq data. The proposed method integrates the prior gene-gene interaction network with the observed scRNA-seq data. Using both simulated data and gene expression profiles from real scRNA-seq experiments, we show that the proposed method outperforms existing methods for assisting clustering cells in most cases.

46

Assessing the Causal Effect of SARS-CoV-2 Infection and COVID-19 Severity on Complement System Activation using Mendelian Randomization

Luisa Foco[1], Eva König[1], Reinhard Würzner[2], Florian Kronenberg[3], Peter P Pramstaller[1], Christian Fuchsberger[1], Cristian Pattaro[1], Fabiola del Greco M[1]

[1]Institute for Biomedicine (affiliated to the University of Lübeck), Eurac Research, Bolzano, Italy; [2]Institute of Hygiene & Medical Microbiology, Department of Hygiene, Microbiology and Public Health, Medical University of Innsbruck, Schöpfstr, Innsbruck, Austria; [3]Institute of Genetic Epidemiology, Medical University of Innsbruck, Schöpfstr, Innsbruck, Austria

The complement system (CS) is part of the innate immune system comprising three independently triggering pathways: the classical (CP), alternative (AP) and lectin (LP) pathways. CS has gained popularity in COVID-19 times, appearing from the beginning of the pandemics a putative key player involved in susceptibility to infection and disease severity. However, the role of CS in COVID-19 has never been studied in the general population using a Mendelian randomization (MR) approach.

We used a two-sample MR to test whether the activation of CP, AP, LP can be caused by 1) SARS-CoV-2 infection or 2) COVID-19 severity. We selected SNPs as instrumental variables (IV) leveraging results of the COVID-19 Host Genetics Initiative genome-wide association study (GWAS) release 7 (2021) in European ancestry individuals. As exposures, we selected the C2 "Reported SARS-CoV-2 infection vs. population controls", representing the susceptibility to infection, and the B1 "Hospitalized covid vs. not hospitalized covid," representing disease severity. As outcomes, we retrieved summary statistics from a GWAS of CP, LP and AP activation conducted on the Cooperative Health Research in South Tyrol (CHRIS) study. CS activation was measured using Wieslab assay on n=4,990. Genetic data were imputed against TOPMed and analyzed with Regenie v3.2.1. We conducted MR applying Inverse-Variance Weighted Fixed Effect (IVW-FE) analysis and standard methods robust against pleiotropy, implemented in the Mendelian randomization R package.

We identified a causal effect of SARS-CoV-2 infection on LP activation level (p value=4.51×10^{-9}), with evidence of heterogeneity (I^2=87.4%, Q test p value=9.77×10^{-25}) driven by an IV in the *ABO* gene, suggesting the presence of pleiotropy. Causal effect estimates obtained with robust methods were direction consistent with the IVW-FE estimate: a Rucker statistic of 192.54 indicated the MR Egger to be more appropriate than the IVW-FE (MR Egger p value=0.044).

COVID-19 severity was causally associated with lower activation of CP (p value=0.029). No evidence of causal effects was observed on AP for both C2 and B1.

Our findings suggest a role of CP activation in disease severity. Evidence of causal effect of infection on LP activation was not fully supported due to pleiotropy. This conclusion is in contrast with recent evidence showing that MBL2, the key effector of LP, specifically interacts with the SARS-CoV-2 spike protein, thus activating LP. This inconsistency raises methodological questions regarding MR assumptions, particularly on the assumption of absence of pleiotropy, highlighting the risk of discarding true causal effects.

47

Differential Gene Expression Analysis Reveals Genes Underlying the Transition to T2D in Hispanic/Latino Individuals

E. Frankel[1], W. Zhu[1], H-H. Chen[1], P. Sharma[2], L.E. Petty[1], R. Roshani[1], H.M. Highland[2], E.R. Gamazon[4], J. McCormick[6], S. Fisher-Hoch[6], K.E. North[5], J.E. Below[1]

[1]Vanderbilt University, Nashville, TN; [2]University of North Carolina Chapel Hill, Chapel Hill, North Carolina, United States of America; [4]VUMC Clare Hall University of Cambridge, Nashville, Tennessee, United States of America; [5]University North Carolina, Wake Forest, North Carolina, United States of America; [6]The University of Texas, Houston, Texas, United States of America

Despite an increased prevalence of Type 2 diabetes (T2D) relative to non-Hispanic Europeans, Hispanic/Latino populations are often underrepresented in research studies of T2D. In this study, we sought to measure significant changes in gene expression during the transition to T2D and in quantitative measures such as longitudinal fasting blood glucose (FBG) and glycated hemoglobin (HbA1c) levels regardless of T2D status.

We utilized longitudinal RNA-sequencing data from 1,150 timepoints measured in 575 participants (142 incident cases) from the Cameron County Hispanic Cohort to profile the transition to incident T2D. Incident T2D was characterized using the ADA 2010 criteria, and exclusions were defined as individuals taking diabetes medication at either timepoint. We utilized joint mixed-effect regression models for repeated measures to compare changes in gene expression over time using DESeq2 normalized RNA expression for each individual at each time point as the outcome variable. We developed three longitudinal models of incident T2D, FBG, and HbA1c levels. Each model was adjusted for sex, age, smoking status, body mass index (BMI), and probabilistic estimation of expression residual factors. We also performed an analysis removing BMI as a covariate for incident T2D to capture genes associated with BMI that may contribute to the development of T2D.

Our preliminary longitudinal analyses of incident T2D with and without BMI, FBG, and HbA1c levels identified 3, 9, 50, and 398 significant genes, respectively. Overall, our analyses characterized transcriptomic changes related to incident T2D and longitudinal FBG and HbA1c levels in an understudied population with a high disease burden.

48

The Sex-specific Genetic Architecture of Childhood Asthma

Amelie Fritz,[1,2] Anders U. Eliasen, PhD,[2] Kasper Rasmussen,[2] Casper Emil Tingskov, PhD,[2] Klaus Bønnelykke, MD, PhD[2], Anders G. Pedersen, PhD[1]

[1]Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, Kongens Lyngby,Denmark; [2]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Copenhagen University Hospital, Herlev-Gentofte, Denmark

Childhood asthma is the most common reason for hospitalization in early childhood. From epidemiological studies, it is evident that the prevalence is higher in boys than girls. After puberty, it is more prominent in women than men. The heritability of childhood asthma is estimated to be between 60 and 90%. This suggests that the genetic components driving the development of childhood asthma have a sex-specific effect. Yet, most association studies do not consider gender in their analysis.

In this project, a Bayesian logistic regression model with a variant-sex interaction term was developed in RStan to identify single nucleotide polymorphisms (SNPs) that have a sex-specific effect on childhood asthma. Discovery studies were conducted in a dataset of 1,189 children with severe asthma (2-6 hospitalizations) from Copenhagen Prospective Studies on Asthma in Childhood (COPSAC) and 5,094 non-asthmatic controls. 77 variants have a posterior probability of interaction higher than 95% after removing SNPs in linkage disequilibrium (LD). A subset of individuals with severe asthma (six or more hospitalizations, 372 individuals) suggests 26 variants with a posterior probability higher than 95% of having a sex interaction. Replication was conducted in two independent data sets.

Variants are found to be part of the genes IL1R1 and CLEC16A, known for being associated with asthma previously, and four of the top nine interacting SNPs are expressed in lung tissue.

Further, a sex-stratified analysis of main effects confirms the sex-specific effects in both data sets.

Clinical associations with rs1009360 on chromosome 2 point to an association with sex-specific testosterone levels and their effect on disease in men and women which could influence the production of immunoglobulin E (IgE) differently in both genders. IgE is known to be involved in the outbreak of asthma. Tissue expression analysis shows expression in testis, ovaries, and uterus.

One variant can be replicated UK Biobank (5,581 cases and 88,094 controls). Further replication is planned in the iPSCYH data set.

Sex-specific eQTL analysis of nasal gene expression of another childhood asthma cohort study will further investigate sex-specific effects.

49

Two-sample Mendelian Randomization Study of Circulating Metabolites and Prostate Cancer Risk in Hispanic Populations

Harriett Fuller

Public Health Sciences Division, Fred Hutchinson Cancer Center, Washington, United States of America

While prostate cancer (PCa) is highly heritable, the mechanisms underlying PCa risk are not well understood, particularly in underserved populations. Here, we conducted two-sample Mendelian randomization (MR) to assess whether serum metabolites are causally associated with PCa risk in Hispanic/Latino men.

MR was performed using GWAS summary statistics for 711 metabolites from 3,166 Hispanic Community Health Study/ Study of Latinos participants and PCa GWAS summary statistics for 3,931 cases and 26,405 controls from Hispanic populations in PRACTICAL. SNPs associated at the genome-wide significance level ($p<5x10^{-8}$) were included and pruned by linkage disequilibrium ($R^2=0.2$).

In total, 22 metabolites were significantly associated with PCa risk, including three amino acids, one carbohydrate, and 18 lipids, which included four polyunsaturated fatty acids (PUFAs). All PUFAs were associated with 14-19% reduced odds of PCa: n3 DPA (22:5n-3) (OR=0.81, 95% CI=0.73–0.90, p value=$1.7x10^{-4}$), n6 DPA (22:5n-6) (OR=0.86, 95% CI=0.78–0.95, p value =$2.2x10^{-3}$), EPA (20:5n-3) (OR=0.81, 95% CI=0.77–0.85, p value =$2.8x10^{-15}$), and arachidonate (20:4n-6) (OR=0.85, 95% CI=0.82–0.88, p value =$3.3x10^{-20}$). The most significant associations (p value ≤$3.9x10^{-40}$) observed were phosphatidylcholines 1-stearoyl-2-arachidonoyl-GPC (18:0/20:4) and 1-palmitoyl-2-arachidonoyl-GPC (16:0/20:4n-6), both of which reduced PCa odds by ~15%. Results were largely robust to sensitivity analyses.

This study provides evidence of causal associations between a range of metabolites and PCa risk in individuals of Hispanic ethnicity. Work is ongoing to replicate these findings in individuals of European and African ancestry to further assess the potential for metabolites to serve as PCa biomarkers.

50

Phenome-wide, Metabolomic and Proteomic Association Scan of *SHROOM3* Haplotypes Based on Imputed Exonic Variants

Dariush Ghasemi-Semeskandeh[1,2*], Eva König[1], Luisa Foco[1], Nikola Dordevic[1], Martin Gögele[1], Johannes Rainer[1], Markus Ralser[3], Dorien J.M. Peters[2], Peter P. Pramstaller[1], Cristian Pattaro[1]

[1]Institute for Biomedicine (affiliated to the University of Lübeck), Eurac Research, Bolzano, Italy; [2]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [3]Institute of Biochemistry, Charité - Universitätsmedizin Berlin, Berlin, Germany

SHROOM3 encodes an actin-binding protein involved in cell shaping. Genome-wide association studies (GWAS) identified common variants at *SHROOM3* associated with chronic kidney disease, creatinine-based estimated glomerular filtration rate (eGFRcrea), and serum magnesium (Mg). To understand underlying physiological mechanisms of *SHROOM3*, we conducted a phenome-wide, metabolomic, and proteomic analysis of *SHROOM3* haplotypes in the Cooperative Health Research in South Tyrol (CHRIS) study.

We performed genotype imputation of the whole cohort of 13,389 participants based on whole-exome sequencing data from 3,840 of the sample. We reconstructed haplotypes of 146 functional variants in *FAM47E*, *STBD1*, *CCDC158,* and *SHROOM3*, bounded by the recombination hotspots. The analysis encompassed 74 serum, urine, and anthropometric traits, 172 serum metabolites, and 148 plasma proteins concentrations on 3,423 individuals. We fitted linear models on the inverse normal transformation of each trait, adjusted for age, sex, the first 10 genetic principal components, and expectation-maximization-based haplotype frequencies.

We identified 11 haplotypes (H1 to H11; frequency from 24.36% to 2.03%). H8 (frequency 2.67%) was associated with eGFRcrea (*P*=2.7e-4), the urinary albumin-to-creatinine ratio (p=3.3e-3) and multiple phosphatidylcholines. H6 (11.61%) was associated with serum creatinine and several carnitines. H4 (2.81%) was associated with Mg (*P*=6.7e-4), eGFRcrea and basophils. H10 (2.32%) was associated with Mg, glutamine, putrescine, and afamin. H3 (2.32%) and H9 (3.99%) were associated with thyroid-related traits, carnitines, and immunoglobulins.

Our multiomic, haplotype association analysis highlighted strong pleiotropy at *SHROOM3*, associated with kidney function and other traits. Identification of haplotypes jointly associated with complex traits, metabolites, and proteins highlights molecular pathways warranting further investigations.

51

Exploring Fine-Mapping Using "SuSiE" on Simulated and Real Data

Aida Gjoka, Heather J. Cordell

Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

The main goal of fine-mapping is the identification of relevant genetic variants that affect some trait of interest, usually the presence of a disease. Fine-mapping methods are often challenging to apply because of the presence of LD (linkage disequilibrium), that is, regions of DNA where the variants interrogated have a high correlation structure. From the statistical point of view, fine-mapping can be seen as a variable selection problem.

Many statistical techniques are available to address this issue. In particular, we use the Sum of Single Effects (SuSiE) model for selecting variables that causally affect some trait of interest. This approach provides us with "credible sets", i.e. groups of variables that appear to be correlated with the response variable. This allows us to appropriately acknowledge more uncertainty when selecting the causal effects for the trait.

We focus on using "SuSiE-RSS", which fits the SuSiE model to summary statistics, such as single-SNP z-scores, with the covariance structure estimated either internally (from the same data set used to generate the summary statistics) or else from a reference panel. We apply the method to both simulated and real data. For data simulation we use the HAPGEN2 software. We study the performance of SuSiE-RSS by considering metrics such as power and false discovery rate. For the application to real data, we use summary statistics from a GWAS of autoimmune liver disease, Primary Biliary Cholangitis (PBC), derived from meta-analysis of five cohorts of European ancestry.

52

Cross-ancestry GWAS Meta-analysis of Keloids Discovers Novel Susceptibility Loci in Diverse Populations

Catherine A. Greene[1,2], Gabrielle Hampton[1], Hannah M. Seagle[1], Yuan Luo[3], Gail P. Jarvik[4], Bahram Namjou-Khales[5], Atlas Khan[6], Todd L. Edwards[7], Digna R. Velez Edwards[1,8,9], Jacklyn N. Hellwege[1,10,11]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Preventive Medicine (Health and Biomedical Informatics), Northwestern University Feinberg School of Medicine, Chicago, Illinois, United States of America; [4]Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington Medical Center, Seattle, WA, United States of America; [5]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center (CCHMC), Cincinnati, Ohio, United States of America; [6]Division of Nephrology, Dept of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, New York, United States of America; [7]Division of Epidemiology, Department of Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [8]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [9]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [10]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, United States of America; [11]Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, United States of America

Keloids are benign fibroproliferative growths that form after injury to the skin and grow beyond the original wound boundaries. They can result in pain and disfigurement. Prevalence estimates vary from less than 0.1% in European-ancestry populations to 16% in some African-ancestry populations. Most genetic research has been limited to European and East Asian populations despite African-ancestry populations having a 20-fold increased risk for keloids. We have performed a large cross-ancestry genome-wide association study (GWAS) meta-analysis of keloids, incorporating data from 1,600,846 individuals (7,837 cases and 1,593,009 controls). We detected 142 novel loci in the cross-ancestry meta-analysis, which included seven replicated variants and 1,233 novel variants. The most significant cross-ancestry result ($p= 1.65 \times 10^{-79}$) was at a variant (rs10863683) located downstream of *LINC01705*. *LINC01705* was previously associated with keloids and implicated as a regulatory factor underlying tumorigenesis. Additionally, analysis of genetically-predicted gene expression with S-PrediXcan identified a significant association between decreased risk of keloids and increased expression of *LINC01705* in fibroblasts ($P = 7.62 \times 10^{-21}$), which play an important role in wound healing. Other results include associations with *NEDD4* and *LSP1* in fibroblasts, as well as *PHLDA3* in sun-exposed skin ($p = 7.94 \times 10^{-11}$, 5.96×10^{-8}, and 3.81×10^{-14}, respectively). Keloid SNP-based heritability estimates are 6%, 21%, and 34% for Europeans, East Asians, and Africans, respectively. These results, along with ancestry-specific results, support a potential adaptive origin for keloid disparities and significantly increase the yield of discoveries from keloid genetic association studies.

53

Accounting for Genetic Regulation Alters Pregnancy Exposure-associated Differential Gene Expression Detected in Umbilical Cord Blood

Luke P. Grosvenor[1], Kelly Benke[1], Janine M. Lasalle[2], Rebecca J. Schmidt[3], Nilanjin Chatterjee[4], Christine Ladd-Acosta[5], Heather E. Volk[1], M. Daniele Fallin[1,6]

[1]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America; [2]Department of Medical Microbiology and Immunology, Perinatal Origins of Disparities Center, MIND Institute, Genome Center, Environmental Health Sciences Center, University of California Davis, Davis, California, United States of America; [3]Department of Public Health Sciences and the MIND Institute, School of Medicine, University of California Davis, Davis, California, United States of America; [4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America; [5]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America; [6]Emory School of Public Health, Atlanta, Georgia, United States of America

Detection of transcriptional regulation by exposures may be aided by including information about strength of genetic control. We calculated genetically regulated gene expression (GReX) to capture genetic contribution to assayed expression and included this in models associating assayed expression with maternal gestational diabetes mellitus (GDM). Data used were from the Markers of Autism Risk in Babies-Learning Early Signs (MARBLES) pregnancy cohort, which enrolls infants with a family history of autism spectrum disorder (ASD), for 133 mothers ($n = 27$ with GDM) and their children ($n= 69$ typically developing, 29 with ASD, and 35 other developmental delay). We measured transcript levels in RNA extracted from cord blood and calculated GReX from child genotypes using FUSION and reference transcriptome data from the Genotype-Tissue Expression project. The final expression set consisted of 6,067 genes. Using linear regression models adjusted for neurodevelopmental diagnosis and surrogate variables, 30 genes were associated with GDM ($\log_2$(fold change) > 0.10, p value < 0.01). After including adjustment for GReX, 19 genes remained statistically significant, and ten were newly detected. Genes with the greatest reduction in $\log_2$(fold change) included *LGALS2* and *CCR3*, each implicated in immune response. On average, 1,446 genes with strong associations between GReX and assayed expression (bGReX > 0.10 and p value < 0.01) showed greater reductions in fold change with GDM after adjustment (mean difference in $\log_2$(fold change) = 20% for strong gene set, 1% for weak). Our findings highlight the potential utility of incorporating GReX to evaluate transcriptional variation associated with exposures.

54

Rare Variant Analysis in Small Samples: Improving Power Through Coupling Rare Variant Burden Analysis with Gene Set Enrichment

Sarah E. Guagliardo[*1], Paul C. Dinh[2], Xindi Zhang[3], Matthew R. Trendowski[3], Swetha Nakshatri[3], Robert D. Frisina[4], Regeneron Genetics Center, Lois B. Travis[2], Nancy J. Cox[1], Megan M. Shuey[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Division of Medical Oncology, Indiana University, Indianapolis, Indiana, United States of America; [3]Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [4]Departments of Medical Engineering and Communication Sciences and Disorders, Global Center for Hearing and Speech Research, University of South Florida, Tampa, Florida, United States of America

With the increasing affordability of whole genome sequencing, the ability to incorporate rare variant methodology in human disease studies is crucial. However, these approaches are often limited by power and sample sizes required to identify meaningful results. Using extensive clinical and audiometric data in The Platinum Study, we demonstrate how coupling rare variant burden analyses to gene set enrichment approaches (GSEA) can improve our understanding of cisplatin-associated HL, tinnitus (TINN), and peripheral neuropathy (PN), all of which can severely impact quality of life.

Testicular cancer survivors of genetically-determined European ancestry (n=1,669) underwent high-throughput whole exome sequencing. After QC filters, 1,663 samples remained. Burden testing used aggregation of various types of rare variants (predicted loss of function, missense, deleterious missense, validated deleterious missense) at different minor allele frequencies (< 0.1, 0.5, 0.01, singleton) for a total of 16 burden masks. Prespecified gene sets for Mendelian forms of HL and PN were curated from various resources, and the GWAS Catalog was used for polygenic TINN, HL, and PN gene sets. GSEA used 100,000 permutations of identically-sized randomly-pulled gene sets. Significant enrichment of TINN and HL polygenic sets existed (p<0.02 and p<0.001, respectively), but not for polygenic PN.

Our results demonstrate that this approach may improve power and interpretability of rare variant genetic analyses for clinical studies. Importantly, rare variant risk of both cisplatin-associated HL and TINN appear to be associated with the polygenic forms of HL and TINN, not Mendelian forms.

55

Discovery of Genetic Factors for Atypical Antipsychotic-Induced BMI Change: Obesity Related to Antipsychotic Liability & Exposure (ORAcLE) Consortium

Shreyash Gupta[1], Geetha Chittoor[1], Navya Shilpa Josyula[1], Simon Lee[2], Michael Preuss[2], Rebecca Birnbaum[3,4], ORAcLE Consortium, Ginger Nicol[5], Adam E. Locke[6], Anne E. Justice[1]

[1]Department of Population Health Sciences, Geisinger, Danville, Pennsylvania, United States of America; [2]The Charles Bronfman Institute for Personalized Medicine, Mount Sinai School of Medicine, New York, New York, United States of America; [3]Psychiatry, Mount Sinai School of Medicine, New York, New York, United States of America; [4]Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, United States of America; [5]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, United States of America; [6]Regeneron Genetics Center, Tarrytown, New York, United States of America

Outlining genomic loci linked to antipsychotic induced weight gain (AIWG) might identify individuals inclined towards weight gain and cardiometabolic dysregulation. Thus, we conducted a genome wide association (GWAS) meta-analysis to identify loci associated with BMI change and BMI increase in individuals on atypical antipsychotics using data from 11,230 adults (non-European 12.2%) from two EHR linked biobanks, MyCode and BioME. Employing a linear mixed model, time on medications was incorporated as both fixed and random effects to obtain BMI change (BMIch) and BMI gain (BMIg), stratified by sex and ancestry. These slopes were adjusted for baseline BMI, baseline age, age^2, duration, and principal components. Residuals were subsequently inverse normal transformed. BMIg was defined as a positive residual BMIch. Ancestry and sex specific GWAS were performed using linear regression allowing for heterogeneity of effect by follow-up duration, then meta-analyzed. We identified three distinct suggestively significant (p<5.0e-07) loci: rs7699506 (β=-0.07, p=3.8e-7) near *FAT4*, and rs319454 (β=0.06, p=2.5e-7) in *ATP8B1* for BMIch; and rs115018170 (β=-0.56, p=2.3e-7) near both *MLXP1* and *SCG2* for BMIg. While our loci have not previously been associated with AIWG or cross-sectional BMI, nearby variants have been associated with mental health and neurocognitive disorders. Interestingly, *SCG2* is highly expressed across the human brain (GTEx), expression is increased in response to lithium exposure in mice and may influence appetite through its role in regulating neuropeptides. These results offer new insights and potential mechanisms through which antipsychotics may alter body mass. Further analyses are underway to validate these findings in independent studies.

56

Unraveling Genetic Associations and Ancestral Influences on Fibroproliferative Diseases

Gabrielle Hampton[1], Jacklyn N. Hellwege[2], Megan Roy-Puckelwartz[3], Yuan Luo[4], Digna R Velez Edwards[5], Todd L Edwards[6]

[1]Vanderbilt University, Nashville, Tennessee, United States of America; [2]Division of Genetic Medicine, Dept. of Medicine, Vanderbilt Genetics Inst., Vanderbilt University Med. Ctr., Nashville, Tennessee, United States of America; [3]Center for Genetic Medicine, Northwestern University Clinical and Translational Science Inst. (NUCATS), Chicago, Illinois, United States of America; [4]Institute for Augmented Intelligence in Medicine, Institute for Innovations in Developmental Sciences, Institute for Public Health and Medicine (IPHAM), Center Health Information Partnerships, Northwestern University Clinical and Translational Science Inst. (NUCATS), Robert H. Lurie Comprehensive Cancer Center; [5]DRVE affiliation: Div. of Quantitative Sci., Dept. of Obstetrics and Gynecology, Dept. of BioMed. Informatics, TLE: Div. of Epidemiology, Dept. of Med., Vanderbilt Genetics Inst., Vanderbilt University Med. Ctr., Nashville, Tennessee, United States of America; [6]Vanderbilt Genetics Inst., Div. of Epidemiology; Dept. of Med., Vanderbilt University Med., Nashville, Tennessee, United States of America

Fibroproliferative diseases (FPD), including asthma and lupus, exhibit a common etiology and co-occurrence, suggesting shared genetic risk factors. Racial disparities indicate a higher prevalence in individuals of African ancestry, potentially due to a Th2-favored genome resulting from generational exposure to helminth parasites in Africa. We analyzed global and local ancestry associations with FPDs to investigate this hypothesis using data from electronic health record (EHR) biobanks: BioVU, and the eMERGE Network. Logistic regressions with ancestry proportions from 1000 Genomes Phase 3 populations [West African (WAFR), East African (EAFR), Southern European (SEUR), and Northern European (NEUR)] were performed, adjusting for sex, age, and BMI. Associations were assessed separately for self-identified non-Hispanic blacks (NHB) and non-Hispanic whites (NHW). Positive associations were observed for each FPD. Top associations were seen between sarcoidosis and African ancestries (sarcoidosis: EAFR OR=1.17, WAFR OR=1.21), while European ancestry demonstrated negative associations (sarcoidosis: SEUR OR=0.87, NEUR OR=0.88). Local ancestry analysis in BioVU NHB identified significant peaks at chr2p23.3 for sarcoidosis, chr19p12, chr4p15.33-32 for lupus, and chr10p13-15.1 for asthma. Targeted associations using BioVU NHB GWAS summary statistics identified potential causal SNPs, including CELF2 (asthma: rs543494131 OR=5.33), TWIST2 (sarcoidosis: rs73104883 OR=6.38), and ZNF724 (lupus: rs188752820 OR=7.35). These genes have been associated with T-cell activation and the inflammatory immune response. Our findings highlight shared immune-mediated susceptibility to FPDs in NHB, potentially contributing to the higher prevalence among individuals of African ancestry. European ancestry appears to have a protective relationship with FPD development, possibly due to variations in T-cell regulation and inflammatory responses.

57

Identifying Rare Non-coding Genetic Aggregate Associations for Height in 331,100 Whole-genome Sequences

G. Hawkes[1], R.N.Beaumont[1], Z. Li[2], R. Mandla[3], X. Li[4], J. Locke[1], N. Owens[1], A. Murray, K. Patel[1], T.M. Frayling[1], C.F.Wright[1], A.R. Wood[1], X. Lin[2], A. Manning[3] & M. N. Weedon[1]

[1]Clinical and Biomedical Sciences, University of Exeter, United Kingdom; [2]Department of Biostatistics and Health Data Science, Indiana University, United States of America; [3]Department of Medicine, Harvard Medical School, Broad Institute, Boston, United States of America; [4]Harvard T.H. Chan School of Public Health (HSPH), Harvard T.H. Chan School Of Public Health, United States of America

Most sequence-based association studies for common human phenotypes have focussed on rare variants that reside in the coding regions of the genome. However, the recent release of whole-genome-sequence (WGS) data in 100,000s individuals from several studies provides an unprecedented opportunity to examine rare, non-coding variants and their contribution towards the genetic architecture of common traits.

We performed the largest WGS-based analysis for height to date using 333,100 individuals from three studies: UK Biobank (n=200,003), TOPMed (n=87,652), and All of Us (n45,445). We developed a generalized analytical pipeline with the aim of finding novel rare (<0.1% minor-allele frequency) non-coding genetic aggregate associations. We subsequently tested 75,311,546 variants which had at least 20 carriers in the UK Biobank and performed 52,749,161 genomic aggregates tests split into gene-centric (e.g. proximal) and non-gene-centric (e.g. regulatory), where variants were grouped by measures of conservation, constraint, and deleteriousness. Finally, we performed a hypothesis-free 2kbp sliding window analysis.

We observed 30 independent novel rare variants associated with height at $p<6.3x10^{-10}$, after conditioning on more than 13,000 previously reported loci. Effect sizes ranged from -7cm to +2cm and replicated three rare single variant associations. We also observed evidence for non-coding associations proximal to *HMGA1*, which alter a transcription start site, causing a 5cm increase in height, *GH1* and an association downstream of *C17orf49* overlapping miRNA which have been previously implicated in growth phenotypes. Our approach found novel non-coding associations for height and provides a template for the analysis of non-coding rare variants for common human phenotypes.

58

mv-DeLIVR: A Deep Learning Based Multivariable TWAS Method for Nonlinear Causal Gene Discovery

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

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

Transcriptome-Wise Association Studies (TWAS) have become a critical tool in identifying gene-trait associations, yet current univariable methods often overlook the confounding effects of linkage disequilibrium (LD) and potential nonlinear gene-trait relationships. In this paper, we introduce mv-DeLIVR, a robust multivariable TWAS method based on neural networks, which accounts for horizontal pleiotropy and permits the estimation of complex gene-trait associations. We extend the existing univariable DeLIVR (uv-DeLIVR) model to a multivariable setting and augment it with a direct effect model to account for horizontal pleiotropy. We use GTEx gene expression data and UK Biobank GWAS data to conduct simulations and perform real data analyses, focusing on High-Density Lipoprotein (HDL) as target traits. 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.

Transcriptome-Wise Association Studies (TWAS) have become a critical tool in identifying gene-trait associations, yet current univariable methods often overlook the confounding effects of linkage disequilibrium (LD) and potential nonlinear gene-trait relationships. In this paper, we introduce mv-DeLIVR, a robust multivariable TWAS method based on neural

networks, which accounts for horizontal pleiotropy and permits the estimation of complex gene-trait associations. We extend the existing univariable DeLIVR (uv-DeLIVR) model to a multivariable setting and augment it with a direct effect model to account for horizontal pleiotropy. We use GTEx gene expression data and UK Biobank GWAS data to conduct simulations and perform real data analyses, focusing on High-Density Lipoprotein (HDL) as target traits. 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.

59

Polygenic Scores for Estimated Glomerular Filtration Rate in a Population of General Adults and Elderly

Janina M. Herold[1]*, Jana Nano[2,3], Mathias Gorski[1], Thomas W. Winkler[1], Martina E. Zimmermann[1], Annette Peters[2,3], Ralph Burkhardt[4], Iris M. Heid[1], Christian Gieger[2,5,6], Klaus J. Stark[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]Institute of Epidemiology, Helmholtz Center Munich-German Research Center for Environmental Health, Neuherberg, Germany; [3]Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Munich, Germany;[4]Institute of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Germany; [5]Institute of Genetic Epidemiology, Helmholtz Center Munich, Neuherberg, Germany; [6]Research Unit Molecular Epidemiology, Institute of Epidemiology, Helmholtz Center Munich-German Research Center for Environmental Health, Neuherberg, Germany

Kidney function is commonly assessed by glomerular filtration rate estimated by serum creatinine ($\text{eGFR}_{\text{crea}}$) or cystatin ($\text{eGFR}_{\text{cys}}$). The polygenic score (PGS) for $\text{eGFR}_{\text{crea}}$ has explained less $\text{eGFR}_{\text{crea}}$ variance in elderly compared to general adults. To identify factors determining age-dependent differences on the variance explained by PGS, we evaluated PGS variance, eGFR variance, and beta estimates of PGS association on eGFR.

We used 634 variants known for $\text{eGFR}_{\text{crea}}$ and 204 variants newly identified for eGFR_{cys} to calculate PGSs in two comparable studies with distinct age ranges: KORA-S4 (n = 2,900; age 24–69 years) and AugUR (n = 2,272, age ≥ 70 years).

The PGS for $\text{eGFR}_{\text{crea}}$ explained almost twice as much ($R^2 = 9.6\%$) of age- and sex-adjusted eGFR variance in general adults compared to elderly (4.6%). This difference was less pronounced for the PGS for eGFR_{cys} (4.7% or 3.6%, respectively).

This difference was not explained by differences in the PGS beta estimates between general adults and elderly, neither by the more frequent comorbidities and medication intake in elderly. The allele frequencies, and thus the PGS variance, between general adults and elderly individuals showed no significant differences except for one variant near *APOE* (rs429358); we found no evidence for survival or selection bias by enrichment of eGFR-protective alleles in the elderly.

We concluded that the difference in explained eGFR variance by PGS was primarily due to the higher eGFR variance in the elderly, after adjusting for age. Our results underscore the need for careful interpretation of PGS based R^2 values.

60

The POPGEN Project: Building a French Reference Panel of Genomes

A. F. Herzig[1], M. Guivarch[1], G. Marenne[1], A. Saint Pierre[1], T. Ludwig[1,2], I. Alves[3], C. Dina[3], R. Redon[3], J. M. Sebaoun[4], L. Gressin[4], J. C. Beaudoin[4], V. Morel[4], B. Fin[5], C. Besse[5], R. Olaso[5], D. Bacq[5], V. Meyer[5], F. Sandron[5], A. Ferrane[6], A. Boland[5], H. Blanché[4], M. Zins[7], J. F. Deleuze[4,5], G. Le Folgoc[1], E. Génin[1,2]

[1]University Brest, Inserm, EFS, UMR 1078, GGB, IBSAM, Brest, France; [2]CHU Brest, Brest, France; [3]Nantes Université, CHU Nantes, CNRS, INSERM, L'institut du Thorax, Nantes, France; [4]Fondation Jean Dausset, CEPH, Paris, France; [5]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), , Evry, France; [6]Institut de Santé Publique, Pôle de Recherche Clinique, INSERM, Paris, France; [7]Université de Paris Cité, Université Paris-Saclay, UVSQ, Inserm, Cohortes Epidémiologiques en Population, UMS 11, Villejuif, France

Background/Objectives: The POPGEN project was launched as part of the French genomic medical initiative to build a catalogue of variants found in the different regions of metropolitan France and provide allele frequencies in order to help filter out neutral variants from patient genomes.

Methods: Individuals from the population cohort Constances were asked to complete a questionnaire on birthplaces and birth years of their parents and grandparents. Based on their answers, 15,000 individuals were selected to cover the different regions of metropolitan France and were posted saliva collection kits. Genotyping was successful for 9,772 individuals and 4,000 individuals were selected for whole-genome sequencing. Different methods were used to study fine-scale population structure and rare variants were imputed using public reference panels enriched by 856 whole genomes from the FranceGenRef project.

Results: We demonstrate the fine-scale population structure of French populations and show how it relates to geography. Using these results, we show how the performance of imputation panels can vary across the territory; driven by patterns in haplotype sharing. We also investigate the important impact for downstream genetic epidemiological study designs.

Conclusions: This study proposes a design to sample individuals from the general population to create reference panels that could help improve imputation accuracy for geographically clustered variants. The POPGEN project will contribute to the "Genome of Europe" project.

Grant: French Ministry of Research PFMG2025 and ANR IA-10-LABX-0013 FranceGenRef

61

A Model for Co-occurrent Assortative Mating and Vertical Cultural Transmission and Its Impact on Measures of Genetic Associations

Anthony F. Herzig[1,*], Aude Saint-Pierre[1], Emmanuelle Génin[1,2], Hervé Perdry[3]

[1]Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest, France; [2]CHRU Brest, Brest, France; [3]CESP Inserm U1018, Université Paris-Saclay, UVSQ, Villejuif, France

*Presenting Author

Background/Objectives: Assortative mating is wide-spread in human populations; shortly put, it occurs when there is positive correlation between mates' phenotypes. It induces a positive correlation between the genetic and the environmental values of parents. If in addition parent-offspring environment is shared, it leads to a correlation between the offspring's genetic value and his/her environment. This latter correlation will build up generation after generation, until an equilibrium point is reached. We aim to explore the consequences of this gene-environment correlation on SNP heritability and polygenic score performances.

Methods: In the framework of the polygenic additive model, we assume that there is a correlation r between mate phenotypes, and a correlation ν between parent-offspring environments. We derive the equations governing the evolution of the gene-environment correlation ρ, and its value at equilibrium. The validity of these results is confirmed by realistic genome-wide simulations. We give estimates for the impact of ρ on estimated SNP effect sizes, SNP-heritability estimates, and polygenic score performances.

Results: The gene-environment correlation ρ increases with r and ν, and can exceed ρ = 0.5 when r and ν reach high values (>0.7). The SNP effect sizes estimates and the SNP-heritability estimates can be severely impacted. The prediction ability of polygenic scores is less affected but essentially because polygenic scores are correlated with the environment.

Conclusion: The combination of assortative mating and shared familial environment can induce sizable gene-environment correlations in a population, which affects genetic epidemiology methods.

62

Genetic Susceptibility to Radiation Therapy Side Effects in Childhood Cancer Survivors in Gene-FCCSS Project

Monia Zidane[1], Brice Fresneau[2], Anthony Herzig[3], Ibrahima Diallo[4], Cristina Veres[4], Nadia Haddy[1], François Doz[5], Carole Rubino[1], Hélène Blanché[6], Anne Boland-Augé[7], Jean-François Deleuze[6,7], Emmanuelle Génin[3,8], Florent de Vathaire[1]

[1]University Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Team "Radiations Epidemiology," Villejuif, 94805, France; [2]Department of Pediatric Oncology, Gustave Roussy, Université Paris-Saclay, Villejuif, France; [3]University Brest, Inserm, EFS, UMR 1078, GGB, IBSAM, Brest, France; [4]Inserm France; [5]University Paris Cité, SIREDO Center (Care, Research, Innovation in Pediatric, Adolescents and Young Adults Oncology), Institut Curie, Paris, France; [6]Fondation Jean Dausset, CEPH, Paris, France; [7]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), Evry, France; [8]CHU Brest, Brest, France

The French Childhood Cancer Survivor Study (FCCSS) is the French cohort focusing on the late effects of childhood cancer treatments with the longest follow-up (mean follow-up of 32 years). All patients included in FCCSS have had radiation dose reconstruction per organ and detailed clinical, therapeutic, and follow-up data.

The Gene-FCCSS project was set up to sequence and analyze the genetic data from FCCSS; one of our main objectives is to analyze the genetic susceptibility to childhood cancer treatment side effects, especially iatrogenic cancerous and non-cancerous radiation-related pathologies.

Whole genome sequencing will be performed on available saliva or blood samples from 2,872 FCCSS patients. Of them, 1,518 have been treated with radiotherapy, and median doses in Gy were: 0.5 at the brain, 1.4 at the thyroid gland, 1.6 at the heart, 1.2 at the breast, and 3.0 at the kidney. A total of 2,263 patients have been treated with chemotherapy.

The analyses will investigate the genetic susceptibility to childhood cancer treatment side-effects such as the risk of second primary neoplasm (n=440), severe cardiac disease (n=240), treated diabetes mellitus (320), severe ototoxicity (n=300), surgically treated cataract (n=60), chronic renal failure (n=70), and stroke (n=110). Interaction between genetic variants and radiation/ chemotherapy doses will be tested for each iatrogenic disease. Novel methods for including the inference of large-scale control panels from the general population of France will be explored.

Gene-FCCSS will present a major European source of information about genetic susceptibility to radiation related pathologies.

Grant no: (INCA R21176LP and PL-SHS N°2022140)

63

Genetic Interactions Between Primary Open-Angle Glaucoma loci Add Another Layer of Complexity in Disease Etiology

Jibril Hirbo[1,2], Valeria Lo Faro[3,4,5], Alexandra Scillaci[1], Harold Snieder[6], Nomdo M. Jansonius[3], Karen Joos[7], Nancy J. Cox[1,2]

[1]Department of Medicine, Division of Genetic Medicine, 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 Ophthalmology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [4]Department of Clinical Genetics, Amsterdam University Medical Center (AMC), Amsterdam, The Netherlands; [5]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; [6]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [7]Vanderbilt Eye Institute, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

Glaucoma is an eye disease that poses a large public health and economic burden that is estimated at nearly 5.8 billion annually. There is disparity in Glaucoma prevalence, clinical presentations, and outcomes across ancestries. Considering genetic interactions contribute to phenotypic complexity, we hypothesized that ancestry-specific genetic interactions in Glaucoma might underlie additional etiology of observed disparity. We explored genetic interaction between GWAS lead SNPs that represent all the loci that have been cumulatively identified to date in four cohorts/Biobanks: BioVU (European/ African), Lifelines biobank, GLGS cohort (European), and *All-of-US* (European/African) in a total of 400K and 100K European and African ancestry individuals, respectively. We further determined the effect of the locus that show interaction across cohorts and ancestries, by checking the differences in the measured gene expressions of genes in target loci in GTEx data. We identified a total of 9 and 56 significant pairwise interactions in European and African American ancestry individuals, respectively. These interactions are mainly due to 7 core loci with at least three pairwise interactions with other loci across the two ancestries, which were either shared (single locus) and unique to an ancestry: European (two loci) and Africans (four loci). The lead variant in the locus that show interaction across ancestries, Chr3_rs62250629_*CADM2*, had significant trans effect on *NEAT1* gene in chr11_ rs12789028 locus and interact with local GTEx eQTL to alter the gene's expression pattern. Our findings point to genetic interactions as additional contributors to disease etiology and observed disparities in glaucoma**.**

64

Fast and Powerful Mixed-model Association Analysis for Genome-wide Association Studies

Jasper P. Hof[1], Doug Speed[2]

[1]*Radboud University Medical Center, Nijmegen, The Netherlands;* [2]*Aarhus University, Aarhus, Denmark*

In recent years, mixed model association analysis (MMAA) has emerged as the preferred method for performing a genome-wide association study. MMAA can control type 1 error by accounting for population structure and familial relatedness and can increase statistical power by conditioning on effects of causal loci distal to the SNP being tested. However, existing MMAA software often requires long run times and substantial memory.

We introduce LDAK-KVIK, a new tool for MMAA of quantitative and binary phenotypes. LDAK-KVIK first constructs a genetic prediction model, then includes this model as an offset when testing genetic variants for association with the phenotype. LDAK-KVIK has three novel features. Firstly, it includes a flexible elastic net algorithm that produces state-of-the-art genetic prediction models (e.g., more accurate than those from BayesR, glmnet, and Bolt-LMM). Secondly, LDAK-KVIK never reads in more than 256 genetic variants at once, and therefore has very low computational requirements. Thirdly, we have implemented an empirical saddlepoint approximation method for robust association analysis of unbalanced binary traits, which is orders of magnitude faster than existing methods.

As a result, LDAK-KVIK is both powerful and efficient. For example, LDAK-KVIK takes approximately 12 CPU hours to analyze GWAS data for 420k individuals and 10M SNP and requires less than 10Gb memory. By contrast, REGENIE and Bolt-LMM, two leading methods for MMAA, require 31 and 150 CPU hours, respectively, to analyze the same data. Applied to 20 traits from UK Biobank, LDAK-KVIK finds 15% and 7%, respectively, more genome-wide significant loci than Bolt-LMM and REGENIE.

65

A Pathway Enrichment Approach to Understand the Contributions of Rare Coding Genetic Variants to Sjögren's Disease

Mary K. Horton[1], Joanne Nititham[1], Kimberly E Taylor[2], NIH Intramural Sequencing Center Comparative Sequencing Program[1], Sjögren's International Collaborative Clinical Alliance, Lisa F. Barcellos[3], Caroline H. Shiboski[2], Lindsey A. Criswell[1]

[1]*National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America;* [2]*University of California San Francisco, San Francisco, California, United States of America;* [3]*Division of Epidemiology, School of Public Health, University of California Berkeley, Berkeley California, United States of America*

Compared to other autoimmune conditions, the known genetic contributions to Sjögren's disease (SD) are limited. We aimed to identify classes of genes containing rare high impact variants associated with severe SD.

Whole exome sequences were generated for 387 participants in the Sjögren's International Collaborative Clinical Alliance. A weighted score of five clinical criteria was used for selection: 120 with score=9 (severe cases) and 267 with score=0. Analyses were restricted to putative loss-of-function (pLOF) and deleterious missense variants with a minor allele frequency ≤1%. Gene *P* values were obtained using SKAT-O, which tested for the association between a gene's variant set and severe SD status. Analyses adjusted for common and rare genetic principal components, age, and sex. We assessed whether genes with *P*<0.05 were significantly enriched in biological pathways, using Fisher's exact test and multiple testing correction implemented in *g:Profiler*.

Severe SD cases were 93% female (83% among non-cases) and had an average age of onset of 48 years (46 years among non-cases). On average, severe SD cases carried 46.8 rare pLOF or deleterious variants while non-cases carried 47.5. SKAT-O analyses identified 137 genes with *P*<0.05. Twenty-one pathways were significantly enriched among these genes ($P_{adj}<0.05$), including ion binding (adjusted $P_{adj}=1.82x10^{-4}$) and cytoplasm ($P_{adj}=1.62x10^{-3}$).

The number of rare, high impact variants present among severe SD cases was similar to non-cases. However, ion binding and cytoplasm pathways were significantly enriched among genes most associated with severe SD status. This may improve identification of novel therapeutic pathways and mechanisms of SD pathogenesis.

66

Bayesian Networks Applied to Type 2 Diabetes Data from the IMI DIRECT Project

Richard Howey[1], Ana Viñuela[1], The IMI DIRECT consortium, Heather J. Cordell[1]

[1]*Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom*

Bayesian networks (BN) can be useful for performing exploratory analysis of large complex datasets to identify possible causal relationships between measured variables based on their conditional dependencies and independencies. We apply BN analysis methods to a large Type 2 diabetes dataset taken from the IMI DIRECT consortium. The initial dataset consisted of over 16,000 variables including gene expression, protein, metabolite and clinical variables together with over 9 million SNP data variables. This was reduced to a dataset with around 300 variables based on various association analyses, and the retained SNPs were collated into 121 allele score variables. Of the 795 individuals, only 29 had complete data for every variable. We used our own BN software, BayesNetty, which has been developed to incorporate a novel imputation algorithm that can handle missing data even when the dataset consists of a mixture of discrete and continuous variables. From networks fitted to the data we find that variables of the same kind tend to be more connected (e.g. gene expression with gene expression) and similar clinical variables (e.g. body mass index and weight) are strongly related. Since the resultant BNs are very large and can be difficult to interpret, we also present sub-networks (given by Markov blankets) focused on variables of interest and related variables. A sex-centered sub-network found a causal role of sex on the abundance of glycine and other metabolites. Our BN software package, BayesNetty, is freely available as open-source software.

67

Conditional Generative Adversarial Network-driven Radiomic Prediction of Gene Mutation Status Using Magnetic Resonance Images of Breast Cancer

Zi Huai Huang[1], Qian Liu[1], Lianghong Chen[2], Yan Sun[1,2], Pingzhao Hu[1,2,3,4,5]*

[1]Department of Biochemistry, Schulich School of Medicine & Dentistry, Western University, London, Canada; [2]Department of Computer Science, Western University, London, Canada; [3]Department of Oncology, Schulich School of Medicine & Dentistry, Western University, London, Canada; [4]Department of Epidemiology and Biostatistics, Western University, London, Canada; [5]The Children's Health Research Institute—Lawson Health Research Institute, London, Canada

Radiogenomics is an emerging field that aims to combine medical images and genomic measurements; however, most radiogenomic studies face the challenge of unpaired data comprising of imaging, genomic or clinical outcome data. In this study, the multi-omic profiles (RNA gene expression, DNA methylation, and copy number variation) of breast cancer patients from The Cancer Genome Atlas was integrated and factorized into 17 latent features using a Bayesian tensor factorization approach. A conditional generative adversarial network (cGAN) was trained using the matched patient magnetic resonance images (MRIs) from The Cancer Imaging Archive and their corresponding latent features. Model performance was evaluated using Frechet's Inception Distance (FID) to compare the real images and the predicted images on the test set, which produced a low FID score of 1.53. Using the trained model, we performed MRI predictions for 690 BC patients with only multi-omic profiles but no MRI data. A convolutional neural network was trained to predict mutation status of *TP53*, *PIK3CA*, and *CDH1* from the produced images, which achieved ROC area under curve (AUC) values of 0.93, 0.75, and 0.78, and precision-recall AUC values of 0.90, 0.64, and 0.46 for *TP53*, *PIK3CA*, and *CDH1*, respectively. Our study establishes cGANs as a viable tool for generating synthetic BC MRIs for cancer driver genes' mutation status prediction. Synthetic images have the potential to significantly augment existing MRI data and circumvent issues surrounding data sharing and patient privacy. These findings also have important implications for future machine learning studies focused on BC research.

68

Interaction of Lifestyle and Polygenic Risk Scores on Colorectal Cancer Risk in the Multiethnic Cohort Study

Brian Z. Huang[1], Fei Chen[1], David Bogumil[1], Peggy Wan[1], Lynne Wilkens[2], Loic Le Marchand[2], David V. Conti[1], Christopher A. Haiman[1]

[1]Department of Population and Public Health Sciences, Keck School of Medicine of USC, Los Angeles, California, United States of America; [2]Population Sciences of the Pacific Program-Epidemiology, University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America

Integrating information from lifestyle and genetic risk factors has the potential to greatly improve the prediction of colorectal cancer (CRC) risk. We evaluated CRC risk associated with genetic and lifestyle factors in 68,374 African-American, Native Hawaiian, Japanese-American, Latino, and White participants from the Multiethnic Cohort Study (MEC). Genetic predisposition was assessed using a polygenic risk score (PRS) of 205 established CRC risk variants. We calculated a healthy lifestyle factor score (HLFS) based on smoking status, alcohol consumption, BMI, and physical activity. Associations with CRC risk were evaluated using Cox regression. There were 1,300 incident CRC cases identified over an average 13.9-year follow-up. Each standard deviation increase in the PRS was associated with an increased CRC risk of 44% among Japanese- Americans (p<0.001), 33% among Latinos (p<0.001), 32% among African Americans (p<0.001), 29% among Whites (p<0.001), and a non-significant 16% among Native Hawaiians (p=0.19). The highest quintile of the HLFS was associated with a 37% reduced CRC risk (p<0.001) compared to the lowest quintile. This association also differed by levels of the PRS (low 0-50% vs. high 50-100%), where the highest HLFS quintile had a 47% reduced CRC risk in the high PRS group (HR 0.53, 95% CI 0.41-1.69), and a non-significant 18% reduced risk in the low PRS group (HR 0.82, 95% CI 0.59-1.14) (p-interaction=0.02). Our findings suggest that incorporating both genetic and lifestyle information can help to enhance the risk stratification of CRC, and that a healthy lifestyle may offer greater protection among those more genetically susceptible.

70

Mitochondrial DNA Sequence Analysis of Epithelial Ovarian Carcinomas

Brooke D. Jorgensen[1], Nicholas B. Larson[1*,] Chen Wang[1], Thomas A. Sellers[2], Stacey J. Winham[1], Sebastian M. Armasu[1], Alvaro Monteiro[3], and Ellen L. Goode[1]

[1]Mayo Clinic, Rochester, Minnesota, United States of America; [2]Knight Cancer Institute, Oregon Health Sciences University, Portland, Oregon, United States of America; [3]H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, United States of America

Mitochondria play a key role in energy metabolism, reactive oxygen species generation, and apoptosis, yet the relevance of mitochondrial DNA (mtDNA) variation in complex diseases is understudied. Seeking to explore the spectrum of mtDNA variation in epithelial ovarian carcinoma, we performed targeted mtDNA sequencing of fresh frozen tumors from Mayo Clinic study participants. After quality control, 389 mitochondrial genomes were analyzed, including those from 320 tubo-ovarian high-grade serous carcinomas (HGSC; 231 deaths), 42 endometrioid carcinomas (EC), and 18 clear cell carcinomas (CCC). We derived regional weighted variant burden scores based on mtDNA functional domains, accounting for underlying haplogroup, heteroplasmic fraction, and predicted functional impact. Accounting for predicted haplogroup, we identified a median of two global private heteroplasmies per tumor (range: 0-9). Elevated mutation rates were identified in the mitochondrial control region, as well as among various mt-tRNAs (>3 per 10 kb per sample); the majority of recurrent variants across tumors were seen in the mtDNA control region. Regional mtDNA burden scores differed by histotype for *MT-CO1* (mitochondrially-encoded cytochrome c oxidase I, CO1) and the oxidative phosphorylation complex IV genes (likelihood ratio test P-values<0.005), with higher burdens in HGSCs and CCCs relative to ECs. Among HGSCs, increased variant burden in *MT-CO1* was associated with longer overall survival (covariate-adjusted HR 0.33; 95% CI 0.15-0.74, P-value=0.007). These results suggest tumor mtDNA alterations in epithelial ovarian carcinoma may relate to clinical aggressiveness or treatment response and merit additional study with continued follow-up in this study population and in additional epithelial ovarian carcinoma collections.

71

Genetic Determinants for Differences in Adult Body Mass Change Using Growth Curve Analysis

Anne E. Justice[1*], Navya Shilpa Josyula[1], Geetha Chittoor[1], Yasser Elmanzalawi[1], Mostafa M. Hamza[1], Craig Wood[2], Christopher Still[2]

[1]Department of Population Health Sciences, Geisinger, Danville, Pennsylvania, United States of America; [2]Center for Obesity and Metabolic Disorders, Geisinger, Danville, Pennsylvania, United States of America

Previous work on links between genomics and patterns of adult body mass index change (BMIchg) is limited and primarily relies on cross-sectional measurements stratified by age groups or recalled weights. Further, investigations attempting to identify genetic effects on BMIchg rely largely on small cohort studies with short duration of follow-up and focus on only a single aspect of BMIchg, such as total change over time. We analyzed longitudinal BMI trajectories over a duration of 25 years using a study cohort including 96,934 MyCode adult participants. We used the SuperImposition by Translation and Rotation (SITAR) method to measure three characteristics of BMIchg: increased life course BMI (size), younger age at maximum BMI (tempo), and higher rate of BMI increase (velocity) Genome-wide association analyses identified 63 significant loci (p value < 5x10^{-8}), including 61 for size, seven for tempo, and four for velocity loci. Of the 63, one locus associated with size has not previously been associated with cross-sectional BMI, rs11210887 in *PTPRF*, but has been associated with related neurocognitive, behavioral, and cardiometabolic traits (e.g. dietary patterns, anorexia nervosa, blood pressure). Two well established BMI loci, *FTO* and *MC4R*, were associated with all three metrics, indicating that they influence size, tempo, and velocity. Of note, two loci were associated with only tempo, rs8118253 in *MACROD2* and rs4144233 in *LMX1B*. Our results highlight loci that increase risk of obesity, rapid weight gain, and earlier onset of obesity, and thus may prove beneficial in prioritizing early monitoring and intervention.

72

Multi-ancestry GWAS of Prostate-specific Antigen Levels Identifies Novel Loci and Improves Cross-population Prediction

Thomas J Hoffmann[1,2], Rebecca E Graff[2], Ravi K Madduri[3], Alex A Rodriguez[3], Sonja I Berndt[4], Mitchell J Machiela[4], Jonathan D Mosley[5], David V Conti[6], Linda Kachuri*[7,8], John S Witte[7,8] on behalf of the PSA Genetics Consortium

[1]Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America; [2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America; [3]Data Science and Learning Division, Argonne National Laboratory, Argonne, Illinois, United States of America; [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America; [5]Department of Internal Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [6]Center for Genetic Epidemiology, Department of Population and Preventive Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [7]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, California, United States of America; [8]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, United States of America

Prostate-specific antigen (PSA) is an enzyme encoded by KLK3 and secreted by the prostate gland. PSA levels tend to increase in the presence of prostate cancer. Despite being a reliable marker of prostate cancer recurrence, PSA testing for population-level screening is not recommended due to potential for overdiagnosis of non-aggressive disease. Previous studies have shown that adjusting measured PSA values based on an individual's genetic predisposition may improve test accuracy by removing variation in PSA unrelated to cancer risk. However, these findings were largely based on men of European ancestry.

We extend previous work by conducting the largest multi-ancestry GWAS of PSA levels in men across nine cohorts, including the Million Veteran Program: 211,342 European ancestry; 58,236 African ancestry, 23,546 Hispanic/Latino, and 3,630 Asian ancestry. We identified 319 independent variants (p *value*<5×10^{-8}), including 185 novel associations, 73% of which replicated in an independent GWAS. Notable findings included rs372203682 in *LMTK2*, a gene implicated in spermatogenesis, and rs184476359, a signal in the androgen receptor (*AR*) that was primarily driven by the African ancestry population.

Next, we constructed a genome-wide polygenic score for PSA levels using an approach that infers posterior effect sizes under continuous shrinkage priors coupled across populations. The variance in PSA levels explained by the resulting score was 16.9% (95% CI=16.1%-17.8%) in European ancestry, 9.5% (95% CI=7.0%-12.2%) in African ancestry, 18.6% (95% CI=15.8%-21.4%) in Hispanic/Latino, and 15.3% (95% CI=12.7%-18.1%) in Asian ancestry populations. These results further our efforts towards more personalized and equitable prostate cancer screening.

74

Leveraging the All of Us Biobank to Build Multi-ancestry Polygenic Scores for NSAID-induced Gastrointestinal Bleeding

Karl E. Keat[1]*, Emanuela Ricciotti[2], Kayla J. Barekat[3], Kathleen M. Cardone[4], Garret A. FitzGerald[3], Dokyoon Kim[5,6], Marylyn D. Ritchie[4,5,6]

[1]Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2]Department of Pharmacology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [3]Institute for Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [4]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [5]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [6]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Non-steroidal anti-inflammatory drugs (NSAIDs) are a class of widely used drugs for the management of pain, fever, and inflammation in a broad spectrum of diseases. NSAID use is associated with life-threatening adverse drug reactions (ADRs), including gastrointestinal bleeding and acute coronary syndrome. Given the severity of these outcomes and the large number of patients affected, there is a significant need to predict individual risk of NSAID ADRs. The Clinical Pharmacogenetics Implementation Consortium (CPIC) has published guidelines for clinicians to modify NSAID treatment in the presence of *CYP2C9* loss-of-function variants; these variants result in reduced clearance of NSAIDs and increased risk of ADRs. However, *CYP2C9* alone explains a relatively small proportion of ADR risk, which is currently better predicted using clinical covariates such as age, sex, concomitant drugs, and comorbidities. To gain a better understanding of the heritable risk of NSAID ADRs, we have performed the first large-scale genome-wide association study for gastrointestinal bleeding following NSAID use in a diverse population using the All of Us (AoU) cohort, identifying several significant associations. We also validate these associations in other biobanks including the Penn Medicine BioBank and the UK Biobank and use them to evaluate the performance of a polygenic risk score (PRS) derived from the AoU GWAS summary statistics. Furthermore, to augment the cross-ancestry performance of our PRS, we are integrating it with a transcriptomic risk score (TRS) based on imputed transcriptomes. Discovery of novel genomic risk factors for ADRs improves both our biological understanding and ability to predict NSAID ADRs.

75

Genome-wide Meta-analysis Identifies Novel Risk loci for Uterine Fibroids within and across Multiple Ancestry Groups

Jeewoo Kim[1,2,3], Ariel Williams[4], Hannah Noh[5], Elizabeth A. Jasper[3], Sarah H. Jones[6], Edward A. Ruiz-Narváez[7], Lauren A. Wise[8], Julie Palmer [8], John Connolly[9], Atlas Khan[10], Mohammad Abbass[11], Laura Rasmussen-Torvik[11], Leah Kottyan[12], Wei-Qi Wei[13], Todd L. Edwards[2,14], Digna R. Velez Edwards[2,3,13], Jacklyn N. Hellwege[2,15]

[1]Vanderbilt University School of Medicine, Vanderbilt; [2]Vanderbilt Genetics Institute, Vanderbilt University; [3]Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center; [4]National Human Genome Research Institute, NIH; [5]Medicine, Health and Society, Vanderbilt University; [6]Institute for Medicine and Public Health, Vanderbilt University Medical Center; [7]Department of Nutritional Sciences, University of Michigan School of Public Health; [8]Department of Epidemiology, Boston University School Public Health; [9]Center for Applied Genomics, Department of Pediatrics, Children's Hospital of Philadelphia; [10]Department of Medicine, Division of Nephrology, Columbia University, College of Physicians & Surgeons, Department of Surgery; [11]Northwestern University Feinberg School of Medicine; [12]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati; [13]Department of Biomedical Informatics, Vanderbilt University Medical Center; [14]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center; [15]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center

Uterine leiomyomata (fibroids) are common benign tumors of the uterus with poorly understood etiology. The prevalence of fibroids ranges between 20 to 80% across reproductive ages. Cost estimates for the US due to fibroids range from $5.9 to $34.4 billion.

Previous genome-wide association studies (GWAS) have reported 72 associated loci but featured limited sample sizes for non-European populations. Our objective of this study is to identify novel genetic variants associated with fibroids across and within ancestry groups.

We conducted a meta-analysis of fibroid GWAS summary statistics from adult female participants with 74,294 cases and 465,810 controls across European (EUR), African (AFR), East Asian (EAS), and Central South Asian (CSA) ancestry groups from eight datasets. Bioinformatic analyses included gene-set enrichment and predicted expression.

We identified 371 sentinel SNPs, including nine novel loci and 15 loci not previously discussed in literature but detected in publicly available biobank summary statistics. Functional analysis identified significant gene-set tissue enrichment in the uterus, cervix, esophagus, fallopian tube, ovary, bladder, and sigmoid colon tissues. These genes were also enriched in DNA damage and cell cycle biological pathways.

The predicted expression of 568 gene-tissue pairs at 180 unique genes were significantly associated with fibroids. Of those, 131 were previously unreported gene associations with fibroids. Within uterine tissue analyses, we observed six significant novel gene associations.

We identified consistent and unique associations at SNPs across populations and tissues in predicted uterine gene expression. These new genetic loci and uterine expression factors may provide translational opportunities for novel fibroid treatments.

76

Meta-analysis of Multinomial Genome-wide Association Study (GWAS) of Radiographic Knee Osteoarthritis Progression

Sumin Kim[1*], Liubov Arbeeva[2], Amanda E. Nelson[2,3,4], S. Amanda Ali[5,6,7+], Osvaldo Espin-Garcia[8,1,9,10+]

[1]Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada; [2]Thurston Arthritis Research Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [3]Division of Rheumatology, Allergy, and Immunology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina, United States of America; [4]Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, North Carolina, United States of America; [5]Bone & Joint Center, Henry Ford Health, Detroit, Michigan, United States of America; [6]Department of Physiology, Michigan State University, East Lansing, Michigan, United States of America; [7]Center for Molecular Medicine & Genetics, Wayne State University, Detroit, Michigan, United States of America; [8]Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Canada; [9]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada; [10]Department of Biostatistics & Schroeder Arthritis Institute, University Health Network, Toronto, Ontario, Canada

[*]Presenter

[+]Co-corresponding authors

Introduction: Knee osteoarthritis (KOA) is a common, chronic, functionally limiting disease that is increasing in prevalence. KOA is known to be a complex disease resulting from multiple combinations of genetic and environmental factors. As KOA can progress at different rates, there may be unique genetic variants associated with progression subgroups that are yet to be identified.

Objective: We propose a two-step approach to meta-analyze multinomial GWAS of KOA progression conducted on two longitudinal prospective cohort studies.

Method: We defined four groups of subjects displaying varying progression patterns using data from the Osteoarthritis Initiative (OAI) and the Johnston County Osteoarthritis Project (JoCoOA). Cohort specific TopMed-imputed GWA studies were performed via multinomial logistic modeling. The proposed meta-analysis approach consists of two steps: 1) Aggregated Cauchy Association Test (ACAT) to combine cohort-specific p values from the (global) likelihood ratio tests at each locus; 2) standard meta-analysis to examine regions identified at the first step for consistency in the direction of the effect.

Results: After identifying genome-wide significant variants from the OAI data, the meta-analysis of the OAI and JoCoOA studies is currently underway and is expected to pinpoint additional variants at suggestive significance level with consistent direction of effects.

Discussion: We anticipate the proposed tiered approach in the context of multinomial regression meta-analysis will improve the power to detect significantly associated variants compared to a single step approach comparing all KOA progression groups at each locus. Our study highlights the novel applicability of ACAT to meta-analysis using a four-level nominal response variable.

77

Clinical Utility of Polygenic Scores: A Critical 2023 Appraisal

Sebastian Koch[1*], Jörg Schmidtke[2,3], Michael Krawczak[1], Amke Caliebe[1]

[1]Institute of Medical Informatics and Statistics, Kiel University, University Medical Center Schleswig-Holstein, Campus Kiel, 24105 Kiel, Germany; [2]Amedes MVZ Wagnerstibbe, Hannover, Germany; [3]Institut für Humangenetik, Medizinische Hochschule Hannover, Hannover, Germany

Polygenic scores (PGSs) were first discovered in 2009 for schizophrenia and bipolar disorder and have since been identified for many common complex diseases. However, their usefulness in assessing disease risk or making treatment decisions is likely limited because they only account for the genetic component of a trait and ignore the impact of environmental and lifestyle factors.

Our survey examined the current state of PGSs for diseases such as breast cancer, diabetes, prostate cancer, coronary artery disease, and Parkinson's disease and how clinical scores could be improved by combining them with PGSs. Furthermore, we collected examples of how PGSs can be and are used for clinical purposes.

We found that the diagnostic and prognostic performance of PGSs alone is low for most diseases, and combining them with clinical scores only leads to moderate improvement at best. A notable example is Type 1 diabetes. Despite many PGSs being reported in scientific literature, there are few external validations and prospective studies conducted on their clinical utility.

We concluded that it is still difficult to judge the benefit of PGSs to individual patients or the healthcare system in general. PGSs may be useful for optimizing screening procedures and medication administration, but there is a lack of prospective studies on this matter.

78

Bayesian Mixture Model for the Identification of Loci of Interest from GWAS Summary Statistics

Rachit Kumar[1,2], Rasika Venkatesh[2], Marylyn D. Ritchie[3]

[1]Medical Scientist Training Program, Perelman School of Medicine at the University of Pennsylvania, United States of America; [2]Genomics and Computational Biology, Perelman School of Medicine at the University of Pennsylvania, United States of America; [3]Department of Genetics, Perelman School of Medicine at the University of Pennsylvania

Genome-wide association studies (GWAS) are a popular method for analyzing the association of genetic mutations or alterations with disease or other phenotypes. However, the biological phenomenon of linkage disequilibrium means that a variant that is linked to a causal variant may be mistakenly identified as causing disease, when in reality it is merely likely to be inherited alongside the true causal variant in a particular region of the genome. As such, people have developed downstream analysis methods that make use of GWAS results to try to identify specific causal variants or identify the functional underpinnings of their effect on disease; however, many of these analyses require users to define the bounds of specific regions of the genome that they would like to assess, and these bounds can have quite significant impacts on the results of these analyses. There is therefore a need for a method that identifies these boundaries in a rigorous and reproducible way.

We present a method that uses Bayesian mixture models to perform statistical inference on GWAS summary statistics to identify whether individual genomic positions represent "breakpoints" between regions containing insignificant variants and regions containing significant variants, allowing one to identify probabilistically which regions represent possible regions where variants are in linkage that may be valuable for downstream studies. We will show the results of this analysis on a synthetic set of GWAS summary statistics and on various existing GWAS summary statistics that are publicly available over a variety of hyperparameters.

79

ColoQuium: A Software Package for Performing Colocalization between QTL and GWAS Analyses and Visualizing Colocalization Results

Nimay R. Kumar[1,3], William P. Bone[2], Brian Y. Chen[3], Theodore G. Drivas[4], Anastasia Lucas[2], Yogasudha Veturi[5], Rachit Kumar[2], Marylyn D. Ritchie[1,7], Benjamin F. Voight[6]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [3]School of Engineering and Applied Science , University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [4]Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [5]Department of Biobehavioral Health, Penn State University, State College, Pennsylvania, United States of America; [6]Department of Pharmacology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [7]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Statistical colocalization between genome-wide association studies (GWAS) signals and either expression quantitative trait loci (eQTL) or splicing quantitative trait loci (sQTL) is a popular method among researchers to connect GWAS signals to candidate causal genes. *ColoQuium* integrates three colocalization R packages into one streamlined pipeline for performing colocalization and facilitating analysis through the detailed visualization of colocalization results. *ColoQuium* consists of the following: *ColoGene* performs colocalization between GWAS significant SNPs and corresponding eQTLs for a given tissue on a gene by gene basis and can accommodate multiple causal variants per locus. This is best suited for analyses investigating a set of genes and tissues of interest (e.g. obtained from running transcriptome-wide association studies) using corresponding gene expression and GWAS summary statistics. *ColocQuiaL* is a framework that performs colocalization between SNP based GWAS signals and eQTL/sQTL at scale and returns a summary of the colocalization results across the genome and locus visualization plots. Given a gene trait pair, *eQTpLot* illustrates the colocalization and correlation between GWAS and eQTL p values, enrichment of eQTLs among trait significant variants, the LD landscape of the given locus, and the relationship between the direction of effect of eQTL signals and the direction of effect of colocalizing GWAS peaks. *ColoQuium* lets researchers easily perform SNP or gene based colocalization in one user-friendly tool, providing a better understanding of the interaction between gene expression and trait associations via tabulation and visualization of the results and GWAS summary statistics.

80

Upregulation of Mitochondrial Dynamics is Associated with Human Colorectal Cancer in North Indian Population

Ashok Kumar[1], Kirti Saini[1], Chandan Chatterji[1], Shashank Mathur[2], Lokendra Kumar Sharma[2]

[1]Department of Surgical Gastroenterology & [2]Molecular Medicine, Sanjay Gandhi Postgraduate Institute of Medical Science, Lucknow-226014, India

Background and Aims: Dysregulation of mitochondrial dynamics (fusion and fission) has been linked to the initiation and progression of different types of cancer including colorectal cancer. The present study is an interim analysis of the ongoing project "to study the role of mitochondrial dynamics proteins and phosphatase calcineurin in north Indian Colorectal cancer patients."

Methods: The expression level of mitochondrial fusion markers in 37 colorectal cancer patients was studied by RT-qPCR after the preparation of cDNA both from the tumor and normal tissues. The expression level of target mitochondrial fusion markers mfn1, mfn2, and opa1 and fission marker drp1 was normalized by endogenous control using β-actin. GraphPad Prism software was used for data compilation. An unpaired t-test was performed for comparison between the two groups; p value < 0.05 was taken as statistically significant.

Results: Out of 37 patients, there were 28 males (70%) and 9 females in the age range of 28-71 years. This included 22 (59%) colon cancer and 15 (41%) rectal cancer patients. In our study, both the fusion (mfn1, mfn2, and opa1) and the fission marker (drp1) were significantly upregulated in tumor tissues. However, mfn1 was differentially expressed while comparing stage II and Stage III patients.

Conclusion: Our interim analysis suggests overexpression of mitochondrial fusion and fission markers. These markers may be playing a role in the initiation and progression of colorectal cancer which needs further study.

81

Polygenic Scores for Major Depressive Disorder Provide Insights into Medication use Prediction in EHR-linked Biobank

Sandra Lapinska, Vidhya Venkateswaran, Kristin Boulier, Yi Ding, Aditya Pimplaskar, Loes M Olde Loohuis, Bogdan Pasaniuc

AFFILIATIONS MISSING

Major depressive disorder (MDD) is the most common psychiatric illness. Studies have shown that MDD has a complex etiology with possible population differences in predisposing genetic factors. Investigating the relationship between ancestry and MDD could provide insight into how individuals may respond to certain treatments. We use a publicly available polygenic risk score (PGS) trained on European individuals from the UK Biobank data for major depressive disorder to investigate the clinical utility of PGS in medication use prediction within the diverse UCLA ATLAS biobank (n > 40,000).

First, we found that the standardized MDD-PGS is associated with the phecode for MDD across genetically inferred ancestry (GIA) groups including European American (EUR) (OR: 1.259, CI: [1.215, 1.306]), Hispanic Latin American (AMR) (OR: 1.117, CI: [1.036, 1.204]), African American (AFR) (OR: 1.126, CI: [0.971, 1.304]), and East Asian American (EAS) (OR: 1.425, CI: [1.264, 1.606]). Next, we identified binary medication-use phenotypes based on treatment options for individuals diagnosed with MDD using de-identified patient information within the EHR-linked ATLAS biobank. Medication-use was defined into three categories: long term use of the most recent medication class, treatment resistant, and efficacious to one medication class. For all analyses, we restricted to individuals diagnosed with MDD who have no other psychiatric comorbidities like bipolar, schizophrenia, substance use, and alcohol use disorders. Using a GIA-stratified logistic regression model adjusted for age, age^{2}, sex, and first five principal components, we found MDD-PGS significantly associated with long term serotonin use in MDD diagnosed individuals with AFR ancestry (OR: 0.590, CI: [0.370, 0.941], p value: 0.027). This suggests that higher genetic risk for MDD correlates with decrease in serotonin usage in individuals of AFR ancestries. In addition, MDD-PGS significantly associated with long term antipsychotic use in MDD diagnosed individuals with EAS ancestry (OR: 1.738, CI: [1.015, 2.978], p value: 0.044) suggesting that these individuals have an increase in antipsychotic usage with higher genetic risk for MDD.

Overall, our study demonstrates the potential use of PGS to determine effective treatments for individuals diagnosed with major depressive disorder across ancestries. Further work to understand the relationship between polygenic risk scores, medication use, and ancestry is needed, but our current findings show a promising clinical utility of PGS.

82

The Current Landscape of Chromosome X Integration in Polygenic Risk Scores

Nicholas B. Larson[1*], Anthony Batzler[1], Brandon Coombes[1], Stacey J. Winham[1]

[1]*Department of Quantitative Health Sciences, Mayo Clinic College of Medicine and Science, Rochester, Minnesota, United States of America, United States of America*

With the growing availability of large-scale GWAS summary statistics across various diseases, there is mounting interest in the development, validation, and clinical deployment of polygenic risk scores (PRSs). In addition to autosomes, chrX includes over 800 protein-coding genes and harbors SNP associations with various autoimmune, cognitive, and behavioral phenotypes. However, there is a well-established paucity of chrX inclusion in GWAS. Furthermore, X chromosome inactivation presents quality control, imputation, and analytical challenges that may be heterogenously addressed across studies, leading to potential mismodeling of chrX SNPs or exclusion of chrX altogether during PRS development. One major concern is the handling of chrX male dosage encoding. This is further compounded by the lack of methodology reporting on chrX SNP analysis and variability in default modeling approaches of popular association analysis software. Consequently, there is high risk for differences in male chrX dosage encoding across development, validation, and target datasets. To assess the current state of chrX utilization in PRS modeling, we downloaded all PRS models hosted on The Polygenic Score Catalog (3503 models on 618 distinct phenotypes as of 04/2023). Only 1052 (30.0%) included chrX SNPs; however, among those with at least 30 total SNPs, chrX accounted for a median 5.9% of total SNP weights (interquartile range = [3.8,7.6]), reflective of proportional genomic content. Despite its high biological relevance, chrX utilization in published PRS models remains low. The long-standing heterogeneity in chrX association analysis throughout the GWAS era may propagate into PRS model misspecification, although correcting this bias will improve PRS performance.

83

The Association between rs6859 in *NECTIN2* Gene and Alzheimer's Disease is Partially Mediated by pTau: New Findings from ADNI

Aravind Lathika Rajendrakumar[1*], Konstantin G. Arbeev[1], Olivia Bagley[1], Anatoliy I. Yashin[1], Svetlana Ukraintseva[1]

[1]*Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, North Carolina, 27708-0408, United States of America*

Emerging evidence suggests a connection between vulnerability to infections and Alzheimer's disease (AD). The *NECTIN2* gene for a membrane component of adherens junctions is involved in response to viral infections, as well as in AD. Its SNP rs6859 had been consistently associated with AD risk in observational studies. It is unclear, however, how exactly rs6859 influences the development of AD pathology. The aggregation of hyperphosphorylated tau protein (pTau) is a key pathological feature of AD, which might be induced by infections, among other factors, and influenced by genes involved in AD and vulnerability to infections, including *NECTIN2*. Here we investigated causal relationships between rs6859 in *NECTIN2*, pTau-181, and AD, in a sample of 708 participants of the Alzheimer's Disease Neuroimaging Initiative (ADNI). We conducted a causal mediation analysis (CMA) with rs6859 as a "treatment", pTau-181 levels measured in the cerebrospinal fluid as a mediator, and AD (yes/no) as an outcome. Carrying the rs6859 risk allele (A) was associated with a 7.3% higher probability of AD and 0.15 change per SD of pTau-181 (95% CI: 0.046, 0.253; p value<0.01). In CMA, the proportion of average mediated effect was 17.6%, and 20.0% for the minor allele homozygotes (AA) (95% CI: 6.8%, 44.0%; p value<0.01). We conclude that the effect of rs6859 in *NECTIN2* on AD is partly mediated by the phosphorylated Tau. The vulnerability to infections may play a role in this causal relationship, which warrants further investigation.

84

Localized Multi-Trait Model: Predicting Disease Risks, Identifying Variant Associations, and Mapping Trait Networks

Cue Hyunkyu Lee[1], Atlas Khan[2], Wang Chen[1,2], Joseph Buxbaum[3], Chunhua Weng[4], Krzysztof Kiryluk[2], Iuliana Ionita-Laza[1]

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

Columbia University, New York, New York, United States of America; [3]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [4]Department of Biomedical Informatics, Columbia University, New York, New York, United States of America

Traditional genomic analyses like Genome-Wide Association Studies (GWAS) and Polygenic Risk Scores (PRS) have been instrumental but are often limited by their reliance on binary case-control statuses. The use of simple case-control status may neglect phenotypic heterogeneity and emerge issues like case-control imbalance. To address these limitations, we introduce a novel methodology—Liability Threshold-Based Phenotypic Integration (LTPI).

LTPI aims to improve genomic analyses by leveraging liability threshold models—which account for the latent variables behind observed traits—and by integrating multi-trait phenotypes extracted from Electronic Health Records (EHR). Using LTPI-based risk score estimates, we conducted a GWAS with data from the UK Biobank and eMERGE network, focusing on Chronic Kidney Disease (CKD) and Coronary Artery Disease (CAD). Our model incorporated 30 binary and 20 continuous non-target traits and integrated them by leveraging genome-wide covariance estimates.

Despite its good prediction accuracy, our GWAS yielded an elevated False Positive Rate (FPR), mainly inflated by associations with non-target traits. We observed three distinct patterns in GWAS analysis: (1) true positives (TPs) influenced by both target and non-target traits; (2) TPs solely influenced by target traits; and (3) false positives driven by non-target traits. Notably, the FPR was well-calibrated for variants unassociated with any traits.

To enhance specificity, we propose calculating locus-level disease risk through local genetic covariance estimates. Simulations confirm that the localized version of LTPI ($\text{LTPI}_{\text{Local}}$) outperforms the global approach ($\text{LTPI}_{\text{Global}}$) in both predictive power and resistance to false positives.

85

Evaluating Machine Learning Instrumental Variable Methods to Estimate Conditional Treatment Effects in Mendelian Randomization

Marc-André Legault[1,2], Jason Hartford[3], Michael Lu[1], Archer Y. Yang[4] and Joëlle Pineau[1,2]*

[1]Department of Computer Science, McGill University, Montreal, Canada; [2]Mila, Montreal, Canada; [3]Recursion, Salt Lake City, United States of America; [4]Department of Mathematics and Statistics, McGill University, Montreal, Canada

Mendelian randomization (MR) is a method to estimate the causal effect of an exposure on an outcome in the presence of unmeasured confounding variables by leveraging the framework of instrumental variable (IV) estimation. MR is widely used to predict the effect of interventions on modifiable disease risk factors and to validate drug targets.

Machine learning IV estimators have been developed to estimate nonlinear causal relationships in the presence of statistical interactions. For example, the DeepIV algorithm uses neural networks to model the instrument-exposure and exposure-outcome relationships in two distinct stages, making no additional assumptions about the functional forms and allowing for effect heterogeneity due to observed variables (Hartford *et al.* 2017). However, few of these recent nonparametric IV estimators have been evaluated in the context of MR. This is due in part to optimization and implementation challenges. To bridge this gap, we have developed *ml-mr*, a bioinformatics package that implements various nonparametric IV estimators to enable their use and evaluation in the context of MR. We also provide a framework for simulation analyses enabling the head-to-head comparison of different methods. To assess the precision of these MR estimators, we have included tools to estimate valid prediction intervals from black box machine learning models using conformal inference. Using simulation models, we evaluated the sensitivity of 4 MR estimators to instrument strength, confounding strength, and sample size. We also report on the possible use of these methods to estimate conditional treatment effects for drug target validation in targeted patient populations.

86

Evolutionary Action Analysis of Ultra-rare Genetic Variants in African-American Men with Prostate Cancer from the *All of Us* Research Program

Deyana Lewis[1], Lesley Chapman[2], Kimiko Kriego[3], Victoria Mgbemena[4], Sabur Badmos[5], Jose Nolazco[6], Panagiotis Katsonis[7]

[1]Morehouse School of Medicine, Department of Community Health and Preventive Medicine, Atlanta Georgia, United States of America; [2]National Cancer Institute, Clinical Genetics Branch, Rockville, Maryland, United States of America; [3]Baylor College of Medicine, Department of Molecular and Cellular Biology, Houston, Texas, United States of America; [4]Prairie View A&M University, Department of Biology, Prairie View, Texas, United States of America; [5] University of Texas at El Paso University, Department of Biochemistry, El Paso Texas, United States of America; [6] Harvard School of Medicine, Department of Urologic Oncology, Boston Massachusetts, United States of America;[7]Baylor College of Medicine, Department of Molecular & Human Genetics, Houston, Texas, United States of America

Prostate cancer (PCa) disproportionately affects African American (AA) men, who have higher incidence and mortality rates as compared to other ethnic groups. The underlying reasons for these disparities are not well understood, although extensive evidence exists implicating an important genetic component. In this study, we examined the ultra-rare (singletons) variants in the exomes of AA PCa cases from the *All of Us* Research Program using Evolutionary Action (EA) analysis. It is hypothesized that singleton variants exist randomly in the human genome that is not associated with PCa. Therefore, this analysis aims to identify variants, genes, and gene pathways with more pathogenic singletons than expected by chance, which are most likely the PCa candidate drivers in these cases.

We studied the recently released *All of Us* data containing the exomes of 216 AA PCa cases less than 65 years old using ultra-rare singleton variants that were not found in any of the 250,000 healthy individuals in the All of Us dataset. We performed EA gene pathway analysis which uses sequence homology data and phylogenetic distances to score each variant between 0 (benign) and 100 (pathogenic). This analysis revealed several candidate driver genes including key DNA repair genes *BRAC2*

(EA>50) and *MSH1* (EA=100). Multiple variants of uncertain significance (VUS) were found in genes and gene pathways that carried high EA scores. These findings highlight the need for expanded research to clarify VUS, identify novel genes and variants contributing to PCa risk, and inform precision therapy among AA men.

89

A Robust cis-Mendelian Randomization Method with Application to Drug Target Discovery

Zhaotong Lin[1]*, Wei Pan[1]

[1]*Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America*

Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to investigate a causal relationship between two traits, an exposure and an outcome. Compared to conventional MR using only independent IVs selected from the whole genome, cis-MR focuses on a single genomic region using only cis-SNPs.For example, using cis-pQTLs for each circulating protein as an exposure for a disease opens an economical path for drug target discovery. Despite the significance of such applications, only few methods are robust to (horizontal) pleiotropy and linkage disequilibrium (LD) of cis-SNPs as IVs. In this work, we propose a cis-MR method based on constrained maximum likelihood, called cisMR-cML, which accounts for LD and (horizontal) pleiotropy in a general likelihood framework. It is robust to the violation of any of the three valid IV assumptions with strong theoretical support. We further clarify the severe but largely neglected consequence of the current practice of modeling marginal effects, instead of conditional effects, of SNPs in cis-MR analysis. Numerical studies demonstrated the advantage of our method over other existing methods. We applied our method in a drug-target analysis for coronary artery disease (CAD), including a proteome-wide application, in which three potential drug targets, *PCSK9*, *COLEC11* and *FGFR1*, for CAD were identified.

90

Identification of Epstein-Barr Virus (EBV) Variants in Different Specimen Types in Nasopharyngeal Carcinoma (NPC)

Jingtong Liang * [1], Cheng-Ping Wang * [2], Wan-Lun Hsu [3], Kelly J. Yu [4], Julia Krushkal [5], Allan Hildesheim [6], Yi-Xin Zeng [1], Xihong Lin [7], Miao Xu# [1], Zhiwei Liu#[4]

[1]*State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Sun Yat-sen University Cancer Center, Guangzhou, China;* [2]*Department of Otolaryngology, National Taiwan University Hospital and National Taiwan University, College of Medicine, Taipei, Taiwan;* [3]*Data Science Center, College of Medicine, Fu-Jen Catholic University, New Taipei City, Taiwan;* [4]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America;* [5]*Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville, Maryland; United States of America;* [6]*Agencia Costarriciense de Investigaciones Biologicas, San Jose, Costa Rica;* [7]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts*

** # Authors contributed equally to this work*

Corresponding author: Miao Xu, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Sun Yat-sen University Cancer Center, Guangzhou, China. Email address: xumiao@sysucc.org.cn

Advancements in next-generation sequencing (NGS) have enabled the complete sequencing of an increasing number of Epstein-Barr virus (EBV) genomes. Recent studies have suggested that genetic variants of EBV may contribute to the unique distribution of EBV-associated cancers, such as nasopharyngeal carcinoma (NPC). However, current published studies investigating EBV genetics have primarily relied on tumor tissue, which limits the ability to evaluate and compare the distribution of EBV variants in NPC cases and healthy controls. Since biopsy tissue is unavailable for EBV sequencing of healthy controls, use of alternative specimens would be desirable to permit evaluation of EBV genetics in both diseased and healthy populations. It is unknown, however, whether alternative specimens can reliably detect EBV variants present in NPC tumors. To address this limitation, we conducted a study among a group of well-characterized, histologically confirmed NPC cases to compare DNA variants of EBV sequences detected in tumor tissue with those detected in samples from alternative specimen types collected from the same individuals. Specifically, we aimed to determine whether the EBV strain present in the tumor can be successfully identified in alternative specimen types, such as saliva (the site of EBV shedding during lytic infection and an easily accessible specimen in epidemiological studies) and nasopharyngeal swab (reflects EBV DNA presented in NPC tumor cells *in situ*) samples.

Our study included 33 newly diagnosed NPC patients, with whole EBV DNA genomes available from 22 paired tissue/saliva samples and 16 paired tissue/swab samples, of which 12 had all three specimen types collected. The whole EBV DNA genomes were also available for one paired saliva/swab, five tissue samples only, and one swab sample only. Our results indicate that samples taken from the same individuals (intra-individual) were more likely to have the same variant detected compared to samples from different individuals (inter-individual). The Kappa statistics for paired samples showed high agreement, with Cohen's kappa coefficients of 0.981 for paired tissue/swab samples and 0.994 for paired tissue/saliva samples. These findings strongly suggest that EBV variants in NPC tumors can be reliably detected in nasopharynx swab and saliva samples, supporting the use of these specimens in studies that compare NPC cases and healthy controls to understand the role of EBV variants in NPC pathogenesis.

91

HORNET: Software and methods to perform whole-genome searches for causal gene networks

Noah Lorincz-Comi, Yihe Yang, Xiaofeng Zhu

Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America

The extent to which genes in disease-associated loci actually cause disease risk is mostly unknown. Many non-experimental approaches to investigate causality combine summary statistics from disease and gene expression (eQTL) GWAS. However, these approaches may produce incorrect inferences because of confounding by other genes and misspecified LD structure. We introduce the HORNET software to perform robust multivariable Mendelian Randomization (MR) with variable selection genome-wide using eQTL and phenotype GWAS summary statistics.

HORNET estimates direct causal effects of gene expression, constructs gene regulatory networks, performs Bayesian LD estimation, imputes missing data, and estimates local disease heritability. We applied HORNET to the 500Kb region surrounding the APOE gene and Alzheimer's disease (AD) using eQTL GWAS from lung tissue (n=515) and AD GWAS data (n=455k).

Simulations demonstrate that HORNET can provide valid causal inference across a range of real-world conditions in which other methods cannot. The software itself is computationally fast, spending approximately 1 second for every gene tested. When applied to AD, 5 genes in a 1Mb window around APOE explained 87.6% of the local AD heritability. APOE, APOC2/4, and DMPK all had direct causal effects on AD (P-values<5E-5), whereas PPM1N only caused AD by regulating APOE (P-value=9.4E-8) and APOC4 (P-value<1E-10).

The HORNET software provides researchers with an accessible and robust tool for identifying genes with causal evidence. The software is publicly available and may identify promising candidate genes for follow-up experimental testing.

92

Integrative Proteogenomic Analyses Identify Circulating Protein Abundances Associated with the Risk of Type 1 Diabetes

Tianyuan Lu[1], Lei Sun[1,2], and Andrew D. Paterson[2,3]

[1] Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, Toronto, Ontario, Canada; [2] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [3] Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada

Type 1 diabetes (T1D) is a complex autoimmune disease with limited preventive measures and treatments. Circulating proteins are potential candidates for identifying novel biomarkers and drug targets because they play essential roles in various biological processes, and their abundances can be measured and modulated.

Leveraging genome-wide association studies (GWASs) of T1D (18,942 cases and 501,638 controls of European ancestry) and circulating protein abundances (10,708 individuals of European ancestry from the Fenland cohort), we performed Mendelian randomization (MR) analyses to assess the associations between 1,565 candidate proteins and the risk of T1D. We conducted multiple sensitivity analyses, colocalization analyses, and replications based on a multi-ancestry GWAS of T1D, as well as proteomics studies in the deCODE cohort and the UK Biobank. We performed gene expression enrichment analyses based on GTEx to investigate the tissue(s)-of-origin of each protein.

After validating MR assumptions and colocalization evidence, we found that a one SD increase in genetically predicted circulating abundances of CTSH (OR=1.17; 95% CI:1.10-1.24), IL27RA (OR=1.13; 95% CI:1.07-1.19), SIRPG (OR=1.37; 95% CI:1.26-1.49), and PGM1 (OR=1.66; 95% CI:1.40-1.96) was associated with increased risk of T1D. These findings were consistently replicated in other cohorts. Furthermore, expression of the genes for these four proteins was strongly enriched in whole blood ($p<1x10^{-5}$). *IL27RA* expression in adipose tissues and *PGM1* expression in skeletal muscle were also pronounced. Importantly, CTSH and PGM1 are known drug targets yet the indications did not include T1D. These findings warrant further explorations of these proteins in the context of T1D.

93

Disentangling the Shared Genetic Aetiology of Type 2 Diabetes and Schizophrenia

Ana Luiza Arruda[1,2,3], Golam M. Khandaker[4], Andrew P Morris[5], George Davey Smith[4], Laura M. Huckins[6], Eleftheria Zeggini[1,7]

[1]Institute of Translational Genomics, Helmholtz Munich, Neuherberg, 85764, Germany; [2]Munich School for Data Science, Helmholtz Munich, Neuherberg, 85764, Germany; [3]Technical University of Munich (TUM), School of Medicine, Graduate School of Experimental Medicine, Munich, 81675, Germany; [4]MRC Integrative Epidemiology Unit at the University of Bristol, United Kingdom; [5]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, M13 9PT, United Kingdom;[6]Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut, United States of America; [7]TUM school of medicine, Technical University Munich and Klinikum Rechts der Isar, Munich, 81675, Germany

Multimorbidity represents an increasingly important public health challenge with far-reaching implications for health management and policy. Mental health and metabolic diseases have a well-established epidemiological association. In this study, we investigate the genetic intersection between type 2 diabetes and schizophrenia. We explore potential causal relationships between the two co-morbid diseases and related endophenotypes and find no compelling evidence to support a causal relationship between type 2 diabetes and schizophrenia. Our findings show that higher body mass index has a protective effect against schizophrenia, in contrast to the well-known risk-increasing effect on type 2 diabetes susceptibility. We identify robust evidence of colocalization of association signals for these two conditions at 11 genomic loci, six of which have opposing directions of effect for type 2 diabetes and schizophrenia. To elucidate these colocalizing signals, we integrate multi-omics data from bulk and single-cell gene expression studies, along with functional information. We identify high-confidence effector genes, and find that they are enriched for homeostasis and lipid-related pathways. The top-ranking effector gene is *NUS1*, which plays a role in lipid trafficking regulation. Mendelian randomization analysis suggests that increased expression of *NUS1* in the brain is causal for schizophrenia and protective against type 2 diabetes risk. Our findings provide insights into the biological mechanisms common to type 2 diabetes and schizophrenia, shedding light on the complex nature of this comorbidity.

94

Insights into the Comorbidity between Type 2 Diabetes and Osteoarthritis

Ana Luiza Arruda,[1,2,4] April Hartley,[5] Georgia Katsoula,[1,4] George Davey Smith,[5] Andrew P. Morris,[1,6] Eleftheria Zeggini[1,3,7]

[1]Institute of Translational Genomics, Helmholtz Munich, Neuherberg, Germany; [2]Munich School for Data Science, Helmholtz Munich, Neuherberg, Germany; [3]TUM school of medicine, Technical University Munich and Klinikum Rechts der Isar, Munich, Germany; [4]Technical University of Munich (TUM), School of Medicine, Graduate School of Experimental Medicine, Munich, Germany; [5]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [6]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom

[7]Corresponding author

Multimorbidity is a rising public health challenge with important implications for health management and policy. The most common multimorbidity pattern is for the combination of cardiometabolic and osteoarticular diseases. Here, we study the genetic underpinning of the comorbidity between type 2 diabetes and osteoarthritis. We find genome-wide genetic correlation between the two diseases, and robust evidence for association signal colocalization at 18 genomic regions. We integrate multi-omics and functional information to resolve the colocalizing signals, and identify high-confidence effector genes, including *FTO* and *IRX3*, which provide proof-of-concept insights into the epidemiologic link between obesity and both diseases. We find enrichment for lipid metabolism and skeletal formation pathways for signals underpinning the knee and hip osteoarthritis comorbidities with type 2 diabetes, respectively. Causal inference analysis identifies complex effects of tissue-specific gene expression on comorbidity outcomes. Our findings provide insights into the biological basis for the type 2 diabetes-osteoarthritis disease co-occurrence.

95

Inferring the Length Distribution of Gene Conversion Tracts

Nobuaki Masaki[1], Sharon R. Browning[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America

Recombination is a process in which genetic material is swapped between homologous chromosomes. Crossovers result in a long segment (typically spanning millions of base pairs) from each chromosome being exchanged, whereas non-crossover gene conversions result in a short segment being copied from one chromosome to the other.

Previous studies have detected these gene conversion tracts in humans. Williams et al. looked at haplotypes in three-generation pedigrees in which alleles were descended from the opposite haplotype of the parent relative to their surrounding markers (2015). By measuring the span of these alleles and flanking markers, they also determined lower and upper bounds for each detected gene conversion tract. However, the length distribution of gene conversion tracts in humans has not been studied in very fine detail.

In our study, we devised parametric models to estimate and provide confidence intervals for the mean tract length with the kind of data used in Williams et al. (2015). This model assumes that the length of the true tract follows a known distribution, but accounts for unobserved copying of alleles during a gene conversion event due to the sparsity of the SNP array or homozygosity of the parent. Simulating gene conversion events, we find that our estimator is unbiased under correct model specification. We also propose an AIC-based method to select the form of the true tract length distribution.

96

Smoking-Associated Changes in Gene Expression in Coronary Artery Disease Patients Using Matched Samples

Mohammed Merzah[1]; Szilárd Póliska[2]; László Balogh[3]; János Sándor[1, 4], István Szász[1,4], Shweaye Natae[1] and Szilvia Fiatal[1]

[1]Department of Public Health and Epidemiology, Faculty of Medicine, University of Debrecen, Debrecen, Hungary; [2]Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, Hungary; [3]Cardiology and Cardiac Surgery Clinic, University of Debrecen, Debrecen, Hungary; [4]ELKH-DE Public Health Research Group, Department of Public Health and Epidemiology, Faculty of Medicine, University of Debrecen, Hungary

Background: Smoking is a well-known risk factor for coronary artery disease (CAD). However, the effects of smoking on gene expression in the blood of CAD patients in Hungary have not been extensively studied.

Aim: To identify differentially expressed genes associated with smoking in coronary artery disease (CAD) patients.

Methods: Eleven matched samples based on age and gender were selected for analysis in this study. All patients were non-obese, non-alcoholic, non-diabetic, and non-hypertensive and had moderate-to- severe stenosis of one or more coronary arteries, confirmed by coronary angiography. Whole blood samples were collected using PAXgene tubes. Next-generation sequencing was employed using the NextSeq 500 system to generate high-throughput sequencing data for transcriptome profiling. The differentially expressed genes were analyzed using the R programming language.

Results: The median age of patients was 67 years (range: 54-75). RNA sequencing was performed on two groups: smokers and non-smokers. After quality control and filtering, gene expression data were obtained for all samples. Using DESeq2, we identified 279 differentially expressed genes with a p value ≤ 0.05 and a log2 fold change ≥ 1. Of these genes, 160 were upregulated in the smokers, and 119 were downregulated compared to non-smokers. Gene ontology analysis revealed that the upregulated genes were enriched for pathways related to immune responses and activities (FDR< 0.03). Specifically, upregulated genes were involved in keratinocyte differentiation, cornification, and epidermis development. The downregulated genes were enriched for cell-cardiac muscle cell adhesion (FDR= 0.004) and epithelium development (FDR= 0.001) pathways.

Conclusions: This research sheds light on the complex biological effects of smoking and provides valuable insights into the mechanisms underlying smoking-related diseases. The findings also have implications for personalized medicine, where patients can be stratified based on their gene expression profiles to predict their risk of developing smoking-related diseases and tailor treatment plans accordingly.

97

Comparing the Performance of a Tic Disorder Phenotype Risk Score and Polygenic Score in a Large Clinical Biobank

Tyne W. Miller-Fleming[1*], David A. Isaacs[2], Dongmei Yu[3,4], Carol A. Mathews[5], Jeremiah M. Scharf[3,4], Lea K. Davis[1]

[1]*Division of Genetic Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America;* [2]*Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America;* [3]*Center for Genomic Medicine, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America;* [4]*Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America;* [5]*Department of Psychiatry, Genetics Institute, University of Florida, Gainesville, Florida, United States of America*

The phenotype risk score (PheRS) is a tool for evaluating comorbidity patterns within a disease of interest based on a patient's medical history. We previously generated a tic disorder (TD) PheRS using 69 phenotypic features enriched in TD patients, including known comorbidities (OCD, ADHD, ASD), and several additional neuropsychiatric and neurological phenotypes.

Here we evaluate the TD PheRS compared to the TD polygenic score (PGS) in the Vanderbilt University Medical Center biobank, BioVU. The TD PGS was calculated using PRS-CS with summary statistics from the most recently published TD GWAS. The TD PheRS and PGS were compared within a subset of clinically validated TD cases (n = 319) and controls (n = 1,585) using linear regression models (covariates: age, genetic principal components, sex, diagnosis code density, and genotype batch).

Both the TD PheRS and PGS are significantly higher in the subset of clinically validated TD cases versus controls (PheRS p=< 2e-16, beta=0.13, PGS p= 0.001, beta=0.03); however, the TD PheRS and PGS were only modestly associated with one another (p=0.03, beta=0.09). When tested independently, the prediction performance for the TD PheRS in the clinically validated cases/controls (Nagelkerke's Pseudo r2=0.32) is higher than the TD PGS performance in this same population (r2=0.04). When we combine the PheRS and PGS using a multivariable linear regression model, we find that the combined model modestly improves the prediction value (r2=0.33), suggesting that clinical history in conjunction with genetic information may provide the strongest method at identifying individuals at risk for a tic disorder diagnosis.

99

Underlying Pleiotropic Connections between Alzheimer's Disease and Its Comorbidities May Contribute to Progression

Anni Moore[1], Marylyn D Ritchie[1,2,3]

[1]*Genomics and Computational Biology Group, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America;* [2]*Institute of Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; Division of Informatics, DBEI, Perelman School of Medicine., University of Pennsylvania, Philadelphia, Pennsylvania, United States of America*

Alzheimer's disease (AD) is the most prevalent neurodegenerative disease worldwide, with one in nine people over the age of 65 living with disease in 2023. Despite its prevalence, no effective treatment or single etiologic driver of AD has been discovered. Over 50% of AD cases are brought on by a combination of small-effect common genetic variants, implicated in numerous pathway disruptions that exacerbate each other to promote disease. One strategy to gain insight into AD progression is to investigate shared pleiotropy between AD and other highly co-occurring diseases to uncover overlapping mechanisms. Here we used the results of a phenome wide association study (PheWAS) to find associations between previously associated AD variants with electronic health record (EHR) diseases from the UK Biobank (UKBB) from 361,194 individuals of European ancestry. With 633 phenotypes and 1,838 AD variants, we found 114 significant associations ($p<5*10^{-8}$), primarily in immune and cardiac related diseases. 24 AD variants were significantly associated with intestinal malabsorption (18 occurring within the major histocompatibility complex (MHC) region). When we removed variants within the MHC region, chronic ischemic heart disease was most significantly associated, with 14 AD associated variants. We will replicate these findings in the Penn Medicine Biobank (PMBB) (n=43,624 patients) and NIH's All of Us Biobank (n=245,000) which has yet to be used for investigation in the context of AD. Both biobanks contain genotype and EHR data for all individuals and have higher diverse patient ancestry than UKBB.

100

Unraveling the Genetic Architecture of Autoimmune Gastritis and Pernicious Anemia

Brooke M. Morris[1*], Austin W. Reynolds[1]

[1]*Department of Anthropology, Baylor University, Waco, Texas, United States of America*

Autoimmune gastritis (AIG) is characterized by stomach inflammation due to autoimmune destruction of parietal cells, affecting 0.5-2.5% of the US population. Parietal cells, specialized epithelial cells in the stomach, aid in digestion and nutrient absorption by secreting hydrochloric acid and intrinsic factor. AIG leads to multiple disease end-stage phenotypes, including iron deficiency anemia and pernicious anemia, and increases the risk of cancers; however, it is unclear why some patients develop different end stages of the disease and what risk factors influence this. Pernicious anemia (PA), one end-stage of AIG, causes malabsorption of vitamin B12, a micronutrient that helps maintain healthy blood cells, nerves, DNA synthesis, structural stability, and many other metabolic processes. Quick diagnosis is crucial to prevent permanent damage or death; however, a positive diagnosis can be difficult because of slow progression and non-descript symptoms that can mask the underlying disease, leaving many patients undiagnosed or misdiagnosed for 5 – 10 years. To understand genetic risk factors, we conducted a genome-wide association study (GWAS) using whole-genome sequence data from 3761 AIG cases, 979 PA cases, and 245,388 controls from the *All of Us* research program. Using a statistical fine-mapping approach, credible sets of putative causal variants associated with AIG

and PA. were identified. These findings contribute to ongoing efforts to characterize polygenic and pathway risk scores for AIG, PA, and other autoimmune diseases, laying the foundation for future improvements in clinical guidelines, and diagnostic and therapeutic strategies.

101

Multidimensional Analysis of Pedigree, Epidemiologic, and Molecular Data to Identify Risk Factors for ME/CFS

R.Moslehi[1]*, A. Kumar[1], A. Bhanushali[1], M. Tracy[1], A. Dzutsev[2]
[1]Schoolof 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 compared 59 ME/CFS patients and 54 healthy controls with respect to the prevalence of autoimmune disease (AID) and cancer among their first-degree relatives, prevalence of epidemiologic factors, and serum levels of 48 cytokines. We used logistic regression and cumulative incidence analysis to calculate OR,RR,95%CI, p values, and machine learning approaches to identify a cytokine profile of ME/CFS.

Results: ME/CFS cases were five times more likely than controls to have a family history of AID (OR=5.30, *P*=0.002). Compared to the relatives of controls, first-degree relatives of cases had significantly higher life-time risks of AID (RR=3.72, p=0.0006) and early-onset (diagnosed <60 years of age) cancer (RR=2.81, p=0.03). Comparison of epidemiologic factors identified history of allergies requiring medication (OR=6.00, p<0.0001), exposure to contaminants (OR=4.35, p=0.0002), history of illness requiring hospitalization (OR=4.33, p=0.0004), ≥4 episodes of significant illness requiring hospitalization (OR=24.36, p<0.0001) and ≥2episodes of significant stress (OR=3.07, p=0.03) as risk factors for ME/CFS. We identified a cytokine profile of ME/CFS, which classified patients with 84% accuracy (kappa=0.68, p=0.025, sensitivity=0.75, specificity=1.00) in random forest models.

Conclusions: Findings from our multi-dimensional analysis of pedigree, epidemiologic, and molecular data suggest certain risk factors for ME/CFS and links with AID and cancer ,providing etiologic clues and druggable targets for treatment of ME/CFS.

102

Using Genetics to Explore the Role of BMI as a Shared Risk Factor in Multimorbidity

Ninon Mounier[1], Bethany Voller[1], Carlos Gallego-Moll[2], Elsie Tata[1], Albert Roso-Llorach[2], Mary Mancini[3], Leon Farmer[3], Kate Boddy[4], Sara Khalid[5], Christopher Fox[1], Sarah Lamb[1], David Melzer[1], Concepción Violán[6], Jane Masoli[1,7], João Delgado[1], Frank Dudbridge[8], Luke Pilling[1], Jack Bowden[1,9], Timothy M. Frayling[1]
[1]Department of Clinical and Biomedical Sciences, University of Exeter Medical School, University of Exeter, Exeter, United Kingdom; [2]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol I Gurina (IDIAPJGol), Barcelona, Spain; [3]Patient and Public Involvement Advisor, University of Exeter Medical School, University of Exeter, Exeter, United Kingdom; [4]Patient and Public Involvement Team, NIHR ARC South West Peninsula (PenARC), University of Exeter, United Kingdom; [5]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, United Kingdom; [6]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol I Gurina (IDIAPJGol), Unitat de Suport a la Recerca Metropolitana Nord, Mataró, Spain; [7]Royal Devon University Healthcare NHS Foundation Trust, Healthcare for Older People, Exeter, United Kingdom; [8]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [9]Novo Nordisk Research Centre Oxford, Department of Genetics, Oxford, United Kingdom

Multimorbidity, the co-occurrence of multiple long-term conditions (LTCs), has become an increasingly important area of research in aging populations. Observational studies have already highlighted the importance of obesity as a risk factor in multimorbidity. In this work, we use genetics to quantify shared genetics between pairs of LTCs and investigate the role of body mass index (BMI) in the co-occurrence of 71 LTCs.

We identified LTCs using an established classification of multimorbidity. We meta-analyzed genetic data from up to three sources (UK Biobank, FinnGen, disease-specific GWAS) and used LD-score regression to estimate pairwise genetic correlations. We used GenomicSEM to quantify the contribution of BMI to the genetic correlation between each pair of LTCs and applied a block-jackknife approach to assess the significance of the attenuation when adjusting for BMI genetics.

We found evidence of widespread shared genetics amongst pairs of LTCs and identified 128 pairs with significant genetic correlation attenuation after adjusting for BMI genetics (q value < 0.05). For example, we observed a two-fold reduction for the genetic correlation between sleep apnoea and hypertension (rG=0.41, rG-adjusted=0.19, attenuation q-value=2e-11), the genetic correlation between osteoarthritis and Type 2 diabetes was entirely explained by BMI genetics (rG=0.22, rG-adjusted=-0.01, attenuation q-value=5e-17) but we did not see any attenuation for the genetic correlation between asthma and depression (rG=0.31, rG-adjusted=0.29, attenuation q-value=0.47).

While these results confirmed the role of BMI as a shared risk factor for several pairs of LTCs, most of the genetic correlations remained significant after adjustment, suggesting shared causal pathways beyond BMI.

103

Impact of Genetic Variations in ACE2 and TMPRSS2 Genes on SARS-CoV-2 Infectivity and COVID-19 Disease Variability Among Bangladeshi Population

Md. Mustafizur Rahman[1*], Faria Akter[1], Shrabani Ghosh[1], Sabrina Momtaz[1], Amir Hossain[2], Md. Abdul Mazid[3]
[1]Khulna University, Khulna, Bangladesh, [2]Dhaka International University, Dhaka, Bangladesh, [3]University of Dhaka, Dhaka, Bangladesh
**Presenting Author: dipti0103@yahoo.com*

Background: SARS-CoV-2 uses host cellular angiotensin-converting enzyme 2 (ACE2) as receptor. Viral S protein binds to ACE2 receptor while host cellular transmembrane serine protease 2 (TMPRSS2) cleaves the bound S protein facilitating SARS-CoV-2 entry into host cells. Epidemiological findings highlight that COVID-19 affects different people in different ways and the disease outcomes range from mild malaise to death by sepsis/acute respiratory distress syndrome. This study aims to investigate the association of ACE2 gene (rs2106809 C/T) polymorphism and TMPRSS2 V160M (rs12329760) polymorphism with SARS-CoV-2 infectivity and COVID-19 disease severity among Bangladeshi population.

Methods: The study included 119 SARS-CoV-2 infected (cases) and 104 age- and sex- matched uninfected individuals (controls). Genomic DNAs were extracted from collected peripheral blood using standard protocol and genotyped for the candidate SNPs using PCR–RFLP method.

Results: For TMPRSS2 V160M heterozygous C/T, variant homozygous T/T and combined heterozygous plus variant homozygous (C/T+T/T) genotypes showed 0.20-fold (p<0.0001), 0.43-fold (p= 0.0194), and 0.27-fold (p <0.0001) decreased risk of SARS-CoV-2 infection, respectively, when compared to normal homozygous C/C genotypes. The variant T allele was associated with 0.561-fold decrease (p=0.0028). Similarly, for ACE2 C/T rs2106809 SNP located on X-chromosome, variant C allele exhibited association with 11.03-fold (p<0.0001) increased risk of SARS-CoV-2 infection in male population when compared with normal C allele carriers. However, for females we did not find statistically significant association with heterozygous (C/C) or variant homozygous (T/T) genotypes.

Conclusion: ACE2 C/T rs2106809 and TMPRSS2 V160M polymorphisms may be associated with SARS-CoV-2 infectivity in our studied populations.

Key words: SARS-CoV-2; ACE2; TMPRSS2; rs2106809; rs12329760.

104

Genetic and Medical Correlates of Long-term Buprenorphine Treatment: A Study in the Electronic Health Records

Maria Niarchou, PhD[1], Sandra Sanchez-Roige, PhD[2,3], India A. Reddy, MD, PhD[5], Thomas J. Reese, PhD[4], David Marcovitz, MD[5], Lea K. Davis, PhD[1,3,4,6]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, United States of America; [2]Department of Psychiatry, University of California San Diego, California, United States of America; [3]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center; United States of America; [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, United States of America; [5]Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, United States of America; [6]Department of Molecular Physiology and Biophysics, Vanderbilt University

Despite the benefits associated with longer buprenorphine treatment duration (i.e., >180 days) (BTD) for opioid use disorder (OUD), retention remains poor. Research on the impact of co-occurring psychiatric issues on BTD has yielded mixed results. It is unknown whether genetic risk in the form of polygenic scores (PGS) for OUD and other comorbid conditions, including problematic alcohol use (PAU) are associated with BTD. We tested the association between somatic and psychiatric comorbidities and long BTD and determined whether PGS for OUD-related conditions were associated with BTD. The study included 6,686 individuals with buprenorphine prescription that lasted for less than six months and 1,282 individuals with buprenorphine prescription that lasted for at least six months. Recorded diagnosis of substance addiction and disorders (Odds Ratio (95% CI) = 22.14 (21.88 to 22.41), p = 2.8 x 10^{-116}), tobacco use disorder (OR(95% CI)=23.4 (23.13 to 23.68), p = 4.5 x 10^{-111}), and bipolar disorder (OR(95% CI)=9.70 (9.48 to 9.92), p = 1.3 x 10^{-91}), were associated with longer BTD. The PGS of OUD and several OUD co-morbid conditions, were associated with any buprenorphine prescription. A higher PGS for OUD (OR per SD increase in PGS (95%CI)=1.43(1.16 to 1.77), p =0.0009), loneliness (OR(95%CI)=1.39(1.13 to 1.72), p =0.002), PAU (OR(95%CI)=1.47(1.19 to 1.83), p =0.0004), and externalizing disorders (OR(95%CI) =1.52(1.23 to 1.89), p =0.0001) was significantly associated with longer BTD. Longer BTD is associated with diagnoses of psychiatric and somatic conditions in the EHR, as is the genetic score for OUD, loneliness, PAU and externalizing disorders.

105

A Novel Computational Paradigm for Cost-efficient and Massively Parallel Analysis of Thousands of Genomic Models

Jeffrey R. O'Connell[1]

[1]University of Maryland School of Medicine, Baltimore, Maryland, United States of America

Biobanks such as TOPMed and All of Us (AoU) are generating phenotype, whole-genome sequence (WGS) and omics data on ever larger sample sizes. Analysis of big data is becoming increasingly cloud-dependent, as downloading petabytes of individual level data for local compute is not possible due to cost or permission. The price-tag of cloud compute is becoming a major barrier for researchers seeking to exploit the thousands of potential models in the data to understand biological mechanism.

Although these thousands of models will have some variables (outcome, predictor) in common, our current paradigm of one-model-one-GWAS is forced to recompute the same regression quantities shared between models, thus, incurring the excessive cost of redundant compute.

We have developed a novel compute paradigm that solves the redundant compute problem by combining all model variables into a universal model, then constructing an indexed binary format pre-compute regression core (PCRC) that contains the regression cross-products between variables in the universal model. The PCRC can then be used to build **any** model **on-demand** with **zero redundant** compute. The PCRC can also be extended to handle mixed models with genetic correlation. The cost to build the PCRC is approximately a GWAS with the universal model, but then the cost of any on-demand model scales only with the number of covariates as the sample size cost has been eliminated. Moreover, we have developed a new mixed model approximation and genotype storage format to significantly increase GWAS speed. Extensive benchmarking of this transformative paradigm will be presented.

106

Statistical Approach Leveraging Founder Population Genealogy and Identical by Descent Segments to Identify Rare Variants in Complex Diseases

Samir Oubninte[1,2*], Simon Girard[2,3], Claudia Moreau[3], Alexandre Bureau[1,2]

[1]*Départementde Médecine Socialeet Préventive, Université Laval, Québec City, Québec, Canada;* [2]*Centre derecherche CERVO, Québec City, Québec, Canada;* [3]*Départementdes sciences fondamentales, Université du Québec à Chicoutimi, Canada*

The missing heritability caused by rare variants (RVs) poses a big challenge to pre-established statistical methods. Our study aims at detecting RVs using identical-by-descent segments (IBDS) as a proxy for recent variants in family data from a founder population whose genealogy is available, the distinguishing features of our approach. Inferring IBDS from genotype array data is more accessible than whole-genome sequences and enables application to large sample sizes.

Our approach involves dividing the genome into fixed-length windows, treating each window as a synthetic gene (SG),then identifying groups of affected individuals sharing a specific IBD segment over a SG by analyzing genotype array data to infer pairwise IBDS (PSIBD). PSIBD data is then used to identify densely connected haplotypes as IBD clusters via Dash. Lastly, we adapt, implement and evaluate statistics to test for SG sharing enrichment among affected individuals. The null distribution of the genome-wide maximal value of statistics is obtained by simulation of whole-genome transmission in a genealogy using msprime.

For the sake of application, Eastern Quebec has been studied as an example of population with a founder effect. Using BALSAC database to reconstruct the genealogy of 1,200 subjects across 48 schizophrenia and bipolar disorder multigenerational families led to an 18-generation pedigree with 84% completeness at the 10thgeneration.

Simulations studies realized by msprime in the genealogy support our approach. $SMGS$ for the "most represented SG" has greater power to detect causal SGs compared to alternative methods like GMMAT (generalized linear mixed model association test) applied to IBDS.

107

A Genome-wide Association Study of the Three Complement System Activation Pathways Implicates Causal Biological and Pathological Mechanisms

Pattaro Cristian[1], Noce Damia[1,2], Foco Luisa[1], Orth-Höller Dorothea[2,3], König Eva[1], Barbieri Giulia[1,4], Pietzner Maik[5,6], Coassin Stefan[7], Fuchsberger Christian[1], Gögele Martin[1], Del Greco M Fabiola[1], De Grandi Alessandro[1], Summerer Monika[7], Wheeler Eleanor[6], Langenberg Claudia[5], Lass-Flörl Cornelia[2], Pramstaller Peter P[1], Kronenberg Florian[7], Würzner Reinhard[2]

[1]*Institute for Biomedicine (affiliated to the University of Lübeck), Eurac Research, Via Volta 21, 39100 Bolzano, Italy;* [2]*Institute of Hygiene & Medical Microbiology, Department of Hygiene, Microbiology and Public Health, Medical University of Innsbruck, Schöpfstr. 41, A-6020 Innsbruck, Austria;* [3]*MB-LAB – Clinical Microbiology Laboratory, Franz-Fischer-Str. 7b, A-6020 Innsbruck, Austria;* [4]*Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy;* [5]*Computational Medicine, Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, Germany;* [6]*MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom;* [7]*Institute of Genetic Epidemiology, Medical University of Innsbruck, Schöpfstr. 41, A-6020 Innsbruck, Austria*

Alterations of the complement system (CS), which is a fundamental part of the innate immune response, are associated with both rare and common human diseases. The CS is activated via three distinct pathways: classical (CP), mannose-binding lectin (LP), and alternative (AP) pathways. No hypothesis-free genome-wide screen of the three CS pathways has been conducted so far. We conducted genome-wide association studies of the functional activity of CP, LP, and AP in the Cooperative Health Research in South Tyrol (CHRIS) study (n=4,990). We identified seven loci, including 13 independent and pathway-specific variants ($p<5\times10^{-8}$) located in or near *CFHR4*, *C7*, *C2*, and *MBL2* (known CS genes) and *PDE3A*, *TNXB*, and *ABO* (novel genes). Variants were associated with inflammatory, autoimmune, and coagulation disorders and >400 proteins. We conducted transcriptome- and proteome-wide colocalization analyses based on state-of-the-art datasets, in combination with two-sample Mendelian randomization analysis. We identified three types of results: (1) confirmation of known causal pathways (e.g.: causal role of MBL2 on LP); (2) identification of within-CS feedback loops (e.g.: between AP and complement 7); and (3) identification of novel causal pathways, including: the causal role of ABO protein levels on LP (p value=1.1×10^{-10}; MR-Egger intercept not significant); a causal effect of LP on collectin-11 (p value=6.3×10^{-44}; heterogeneity p value, P_{het}=0.46) and KAAG1 (p value=9.0×10^{-25}; P_{het}=0.27) levels; a causal effect of LP on mouth ulcers' risk (p value=9.5×10^{-6}; P_{het}=0.71). These results depict a first, comprehensive and unbiased map of the role of CS on human health.

109

Identifying Pleiotropy underlying Comorbid Phenotypes Using Binomial Regression

Prasun Panja[1,2], Samsiddhi Bhattacharjee[2]

[1]*Regional Centre for Biotechnology, Faridabad, India;* [2]*National Institute of Biomedical Genomics, Kalyani*

The models suggested in Multiphen and BAMP provide an alternative to study population based genetic association with multivariate phenotypes by exploring the dependence of enotype on phenotype instead of the naturally arising dependence of phenotype on genotype. However, these approaches test the null hypothesis of no association with any of the constituent traits versus the alternative hypothesis of association with at least one of the constituent traits of the multivariate phenotype vector. Thus, such tests do not provide evidence of pleiotropy or common genetic factors underlying all the traits constituting the multivariate phenotype, which might be correlated. With respect to a pair of comorbid phenotypes (both binary, a combination of binary and quantitative or both quantitative), we aim to modify the proposed BAMP (Binomial regression based Association of Multivariate Phenotypes) approach to test the null hypothesis of no association with at least one of the phenotypes versus the alternative hypothesis with both the phenotypes. Identifying the pleiotropy of two diseases is important to decipher the underlying molecular events of two correlated phenotypes. Likelihood ratio test (LRT) will be used for this purpose. Proposed method can also differentiate between mediated pleiotropy and horizontal pleiotropy.

Extensive simulations would be done for different values of correlation coefficients between the two comorbid traits to identify the power of the test and to compare the power of the same model which has used multiple testing correction instead of considering a multivariate phenotype vector.

110

Using Different Analytic Approaches to Identify Genetic Overlap of Idiopathic Pulmonary Fibrosis and Hypertension

Gina Parcesepe[1,2*], Richard J Allen[1,2], Beatriz Guillen-Guio[1,2], Samuel Moss[3], R Gisli Jenkins[3], Louise V Wain[1,2] on behalf of the DEMISTIFI consortium, Ruby M Woodward[1]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR Leicester Biomedical Research Centre, Leicester, United Kingdom; [3]Margaret Turner Warwick Centre for Fibrosing Lung Disease, National Heart and Lung Institute, Imperial College London, London, United Kingdom

Up to 50% of individuals with idiopathic pulmonary fibrosis (IPF) have co-morbid hypertension. High blood pressure can lead to organ fibrosis or can be a consequence of artery stiffening due to fibrosis. Studies have implicated common processes, such as TGF-β signaling, in both traits' regulation. Our goal is to identify shared genetic risk factors for IPF and hypertension using genome-wide and local approaches.

The largest available genome-wide association studies (GWAS) of IPF, and systolic and diastolic blood pressure (SBP and DBP, respectively), were utilized to examine both genome-wide and locus-specific overlap. LD Score Regression was used to conduct genome-wide correlation. Two genetic correlation approaches, LOcal Genetic cOrrelation Dectector (LOGODetect) and Local Analysis of [co]Variant Association (LAVA), were used to identify specific regions of genetic correlation. Colocalisation analysis, using coloc, was applied for any signals present at $p<1x10^{-5}$ in both IPF, and SBP or DBP.

There was no genome-wide correlation between IPF and SBP -0.077 (p=0.022) or DBP -0.027 (p=0.427). Neither of the local genetic correlation approaches identified regions of association. The colocalisation analysis identified seven regions of shared overlap (posterior probability >80%), two increased risk of both IPF and hypertension, whilst five had opposite directions of effect.

These findings support that there may be shared fibrotic mechanisms between IPF and hypertension. The opposite effects of variants at specific loci highlight the need for caution when considering therapeutic targeting of these shared pathways for either disease. The statistical power was affected, due to a significant difference in sample size.

111

Testing for Differences in Hardy-Weinberg Disequilibrium between Groups

Andrew D. Paterson[1,2*], Elika Garg[1], Jaffa Romain[2], Lei. Sun[2]

[1]Hospital for Sick Children, Toronto, Ontario, Canada, [2]University of Toronto, Toronto, Ontario, Canada

Testing for deviation from HWE (Hardy-Weinberg Equilibrium) is a standard quality control approach before conducting genetic analyses. However, deviation from HWE can be caused either by technical biases or violation of numerous assumptions related to population structure. Therefore, assessing HWE differences between groups is important. In our pilot work, we assessed HWE across four super-populations and two sex groups using the Illumina 2.5M array data for 773k SNPs with MAF>5% in 1,604 individuals from the 1000 Genomes Project phase 3 data (doi.org/10.1101/078600). Further, we combined group results by using inverse-variance and second-order meta-analysis. Additionally, we compared group results using Cochran's Q to test for variability among the groups. We identified 26 autosomal SNPs in all super-populations with significant HWE (p value<E-8) in one or more super-population or sex groups, and further with heterogeneity in effect sizes (delta) between ancestry- or sex-based groups. The most common basis for heterogeneity was evidence for deviation in males, but not in females. BLAST of the sequence around such SNPs typically identified similarity to the recently released Telomere-to-Telomere (T2T) Y chromosome sequence, which contains ~30Mb of sequence missing from previous references of the human genome, implying that these variants are paralogous sequence variants (PSV). We will implement a similar analysis using the T2T (telomere-to-telomere v2) aligned high-coverage 1000 Genomes Project whole genome sequence data. We hypothesize that the T2T data will have less technical biases, and thus allow us to better investigate heterogeneity across groups arising due to population structure.

112

Multi-omics Reveal Key Molecular Signatures of Severe Obesity

Lauren E. Petty[1,2], Hung-Hsin Chen[1,2], Priya Sharma[3], Hannah G. Polikowsky[1,2], Jungkyun Seo[3], Mohammad Yaser Anwar[3], Daeeun Kim[3], Mariaelisa Graff[3], Kristin L. Young[3], Wanying Zhu[1,2], Kalypso Karastergiou[4,5], Douglas M. Shaw[1,2], Anne E. Justice[6], Lindsay Fernández-Rhodes[7], Mohanraj Krishnan[3], Absalon Gutierrez[8], Peter McCormick[9], Penny Gordon-Larsen[10], Miryoung Lee[11], Heather M. Highland[3], Eric R Gamazon[1,2,12], Nancy J. Cox[1,2], Susan K. Fried[4,5], Susan P. Fisher-Hoch[11], Joseph B. McCormick[11], Kari E. North[3], Jennifer E. Below[1,2]

[1]Department of Medicine, Division of Genetic Medicine, 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 at Chapel Hill, Chapel Hill, North Carolina, United States of America; [4]Obesity Research Center, Boston University School of Medicine, Boston, MA, United States of America; [5]Diabetes Obesity and Metabolism Institute, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [6]Department of Population Health Services, Geisinger Health, Danville, Pennsylvania, United States of America; [7]Department of Biobehavioral Health, The Pennsylvania State University, University Park, Pennsylvania, United States of America; [8]Department of Internal Medicine, Division of Endocrinology, Diabetes and Metabolism, The University of Texas Health Science Center, Houston, TX, United States of America; [9]Centre for Endocrinology,

William Harvey Research Institute, Barts and the London School of Medicine, Queen Mary, University of London, Charterhouse Square, London, United Kingdom; [10]Department of Nutrition, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [11]Department of Epidemiology, Human Genetics and Environmental Sciences, The University of Texas Health Science Center at Houston School of Public Health, Brownsville Regional Campus, Brownsville, TX, United States of America; [12]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom

Severe obesity [body mass index (BMI) ≥ 40 kg/m^2] is a driver of many cardiometabolic diseases, and disproportionately impacts marginalized populations, including Hispanic/Latino populations, yet we know little about its underlying mechanistic pathways. We deployed integrative multi-omics approaches to enable discovery of genes involved in severe obesity in Mexican-Americans from the Cameron County Hispanic Cohort (CCHC). First, using RNA-sequencing data for 49 severe obesity cases and 81 controls from CCHC, we used DESeq2 to assess differential expression and performed Mendelian randomization to estimate causality. We then used RNA-sequencing data for an additional 52 cases and 59 controls from CCHC to assess replication. We tested for validation of our transcriptomic results using proteomic data for 49 cases and 42 controls from the CCHC and explored the specificity of detected effects by leveraging independent abdominal subcutaneous adipose tissue gene expression data from 19 community volunteers from New York City, New York.

We identified 124 significantly differentially expressed genes after false discovery rate correction, of which 33% replicated in the independent sample from the same population, and 22% of those measured showed differential protein abundance associated with severe obesity. Twenty-six of the differentially expressed genes showed correlation with BMI in abdominal subcutaneous adipose tissue in a diverse independent sample. We provide compelling evidence for genes whose expression is associated with severe obesity in an underrepresented and disproportionately impacted population, observing highly concordant effects in an independent replication, generalization of effects in abdominal subcutaneous adipose with BMI, and translation effects in the proteome.

113

Using Mendelian Randomization to Establish Directionality of Relationships between Fibroids and Associated Genitourinary and Neoplasm Phecodes

Jacqueline A. Piekos,[1-3] Hannah M. Seagle,[1,4] Jacklyn N. Hellwege,[1,5] Nikhil K. Khankari,[1,6] Todd L. Edwards,[1,4] and Digna R. Velez Edwards[2,3]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [6]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Uterine fibroids (UF) are the most common pelvic tumor in women, yet etiology and downstream health effects of UF are unknown. Here we seek to characterize the causes and consequences of phenotypes associated with UF, focusing on genitourinary and neoplasm category phenotypes. From a UF phenome-wide association study, we identified 24 and 28 unrelated phecodes in genitourinary and neoplasm categories respectively, to be associated with UF. We investigated directionality of relationship between UF and associated phecodes using two sample Mendelian randomization (MR). MR instruments were selected by linkage disequilibrium clumping genome-wide significant variants using an r^2 threshold = 0.01. Instruments for UF were selected from FinnGen and tested phecode genetic instruments were selected from UKBioBank. With UF as exposure, 10 genitourinary and 12 neoplasm phecodes were significant consequences of UF (p value < 0.05). The top phenotypes identified as consequences of genetic UF exposure were endometriosis OR=1.30 (95% CI=1.22 – 1.38, p value=3.05x10^{-17}), ovarian cysts OR=1.20 (95% CI=1.14 – 1.26, p value=1.81x10^{-12}), lipoma OR=1.16 (95% CI=1.10 - 1.22, p value=1.01x10^{-7}), and benign neoplasm of skin OR=1.12 (95% CI=1.07 – 1.16, p value=1.19x10^{-7}). Two genitourinary phenotypes were found to be causal towards UF with MR: hemangioma and lymphangioma OR=1.15 (95% CI: 1.10 – 1.21, p value=1.86x10^{-10}) and polyp of female genital organs OR=1.34 (95% CI=1.11 – 1.60, p value=0.002). Understanding the directionality of the relationships between UF and phenotypes may provide insights into possible UF risk factors and downstream consequences of UF which can be leveraged for prioritized UF intervention or prevent downstream consequences.

114

Shared Genetic Architecture and Pleiotropy across Uterine Fibroids and Hypertension

Alexis T. Pigg Akerele [1,2,5,7], Jacklyn N. Hellwege [3,7], Jacqueline A. Piekos, Nikhil K. Khankari [3,7], Todd L. Edwards [4,7], Digna R. Velez Edwards [5,6,7]

[1]School of Graduate Studies, [2]Department of Microbiology, Immunology and Physiology, Meharry Medical College, [3]Division of Genetic Medicine, Department of Medicine, [4]Division of Epidemiology, Department of Medicine; [5]Division of Quantitative Science, Department of Obstetrics and Gynecology; [6]Department of Biomedical Informatics, Data Science Institute; [7]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Uterine fibroids (UF) are among the most common gynecologic diseases in females of reproductive age, having an estimated cumulative prevalence of 70%. There is a need to better understand the genetic liability of risk for UF and risk factors, like hypertension (HTN). In efforts to understand causal relationships between UF and HTN, we conducted a bi-directional two sample Mendelian randomization (MR) analysis and evaluated the genetic correlations across blood pressure

(BP) trait loci and UF. We used data from two cross ancestry genome wide association study (GWAS) meta-analyses, one of UF (44,205 cases and 356,552 controls), and another of BP phenotypes (including diastolic BP [DBP], systolic BP [SBP], and pulse pressure [PP], n=447,758). Linkage disequilibrium score regression (LDSC) was used to evaluate genetic heritability and correlation of BP phenotypes and UF. Genetic instruments for the MR analysis were selected from summary level data of BP traits and UF by linkage disequilibrium clumping of genome wide significant SNP's (p<5e-8) with an r^2 threshold of 0.1. LDSC results indicated a positive genetic correlation between DBP and UF (0.140, p=0.0004), and SBP and UF (0.076, p=0.016), and PP and UF (0.008, p>0.05). MR using BP traits as exposures and UF as the outcome showed that DBP and pulse pressure both increase risk for UF (b=0.024, p=0.002 and b=-0.010, p=0.0008, respectively). Having UF as the exposure and BP traits as the outcomes indicated a relationship between UF and DBP and UF and PP (b=0.46, p=0.005 and b=-0.51, p=0.015, respectively).

115

Characterizing Substructure via Mixture Modeling in Large-scale Summary Statistics

Adelle Price[1,4], Hayley R Stoneman[1,2], Nikole Scribner Trout[1,4], Kristy Crooks[3], Nicholas Rafaels[3], Nikita Pozdeyev[3], Souha Tifour[1,4], Katie M Marker[1,2,3], Christopher R Gignoux[1,2,3], Audrey E Hendricks[1,2,3,4]

[1]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [2]Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [3]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [4]Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, United States of America

Genetic summary data are broadly accessible and highly useful. Nevertheless, collapsing individual-level data into groups masks intra- and inter-sample substructure (e.g., population structure), biasing results. To address this, we developed *Summix v2*, a Bioconductor software suite to estimate and adjust for substructure in summary level data via a computationally efficient mixture model. We benchmarked *Summix v2* with extensive simulations and application to real data. For finer-scale ancestry detection, we find median accuracy over reference groups is high (>99%) and minimum accuracy decreases as reference group similarity increases (i.e., decreases): 97% for =.009; 89% for =.005. We achieve precise and accurate local-ancestry proportion estimates (≥ 99%) using ≥250 variants. We develop a statistical test for regions of ancestry deviation, which maintains 5% type I error and achieves 92% power to detect 400kb regions with 5% difference in local vs global ancestry. In gnomAD v3.1.2, we replicated selection signatures in HLA (African-American, FDR=$6.5x10^{-11}$; Latinx, FDR=$6.68x10^{-9}$) and 1p33/*CYP4A11*, a candidate gene in pharmacogenomics (Latinx, FDR=$9.63x10^{-6}$). At genome-wide significance, we identified dozens of candidate loci containing genes relevant to innate immune response and multiple cancers. Finally, to highlight ability to identify non-ancestry substructure, we used 146 previously associated prostate cancer variants to estimate genetic risk in CCPM biobank summary data reproducing observed case proportions for ≥60 years and identifying yet to be realized risk for 40-60 and ≤40 years. *Summix v2* enables a breadth of substructure estimation improving data harmonization and ultimately increasing the robust use of publicly available summary data.

116

Analysis Approaches for Understanding Cross Trait Effects of *VRK2* in Speech, Language, and Rhythm Phenotypes

[1]Dillon Pruett, [2]Alyssa Scartozzi, [2]Jaclyn Eissman, [1]Yasmina Mekki, [2]Hannah Polikowsky, [2]Ting-Chen Wang, [2]Xavier Bledsoe, [3]Shelly Jo Kraft, [1]Eric Gamazon, [1]Logan Dumitrescu, [1]Timothy Hohman, [1]Reyna Gordon, & [1]Jennifer Below

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Human Genetics Program, Vanderbilt University, Nashville, Tennessee, United States of America; [3]Dept. of Speech Language Pathology & Audiology, Wayne State University, Detroit, MI, United States of America

Stuttering is a neurological speech condition with unknown etiology. Familial and population-based studies show strong genetic influences on stuttering risk. Ancestries- and sex-specific GWAS of stuttering (n = 99,776 cases) identified significant effects of *VRK2* in the European-male cohort, a gene previously implicated in schizophrenia. *VRK2* was also identified in a GWAS investigating musical beat synchronization and language decline in Alzheimer's Disease (AD). Rhythm perception and language are known factors associated with stuttering. Here, we investigate the association of genetic variation in *VRK2* with speech, language, and musicality.

Sentinel SNPs mapping to *VRK2* varied across studies: rs11898834 ($p = 9.990 \times 10^{-10}$) in female AD-related language decline GWAS; rs35609938 ($p = 5.84 \times 10^{-12}$) and rs1040225 ($p = 1.82 \times 10^{-11}$) in the male stuttering GWAS; and rs848293 ($p = 9.23 \times 10^{-18}$) in beat synchronization GWAS. We examined effects at significant SNPs across studies and found alleles associated with increased risk of stuttering were associated with poorer beat synchronization and decreased risk of language decline (p < .003 Bonferroni correcting for four SNPs across four studies).

We utilized NeuroimaGENE, a catalog of neuroendophenotypes derived from transcriptome-wide association studies in 33,000 individuals from UK Biobank, and found the thalamus and somatosensory cortex, areas involved in processing sensory information, are associated with *VRK2* expression (p < Benjamini-Hochberg correction threshold). Phenome-wide association study of genetically regulated *VRK2* expression revealed suggestive associations with neurological and hormonal clinical outcomes, including anxiety, and major depression, and a significant association with congenital anomalies of female genital organs p = .0000235996).

117

Annotation Query (AnnoQ): An Integrated and Interactive Platform for Comprehensive Genetic Variant Annotation on a Large Scale

Bryan Queme[1,2]*, Tremayne Mushayahama[1], Dustin Ebert[1], Anushya Muruganujan[1], Paul D. Thomas[1], Huaiyu Mi[1]

[1]Division of Bioinformatics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California, United States of America; [2]Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California, United States of America.

The Annotation Query (AnnoQ) system (http://annoq.org/) is a comprehensive platform providing up-to-date functional annotations for about 39 million human genetic variants. Distinctively user-friendly, AnnoQ pre-annotates all variants using over 600 annotation types, serving as an intuitive search tool for large datasets. The interactive interface allows for quick queries and reviews before conducting extensive analyses.

Collaborating with renowned resources like the Gene Ontology Consortium, Reactome, and PANTHER, AnnoQ ensures access to the latest functional annotations. Additionally, it utilizes PEREGRINE to extend annotations to non-coding regions, associating variants with respective target genes, functions, and pathways.

Equipped with an optimized Elasticsearch framework, AnnoQ supports real-time complex searches and various search types, catering to diverse research needs. An API for programmatic access to annotated data is also provided, supplementing the web interface. This empowers users to embed annotation queries within scripts using popular programming languages like R.

In essence, AnnoQ serves as an all-encompassing annotation platform, enabling effective exploration and analysis of genomic data.

118

Integrating GWAS Summary Statistics, Individual-level Genotypic and Omic Data to Enhance the Performance for Large-scale Trait Imputation

Jingchen Ren[1,2,*], Zhaotong Lin[3] and Wei Pan[2]

[1]School of Statistics, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [3]Department of Statistics, Florida State University, Tallahassee, Florida, United States of America

Recently a nonparametric method has been proposed to impute the genetic component of a trait for a large set of genotyped individuals based on a separate GWAS summary dataset of the same trait (from the same population). The imputed trait may contain linear, non-linear, and epistatic effects of genetic variants, thus can be used for downstream linear or non-linear association analyses and machine learning tasks. Here we propose an extension of the method to impute both genetic and environmental components of a trait using both SNP-trait and omics-trait association summary data. We illustrate an application to a UK Biobank subset of individuals (n=80K) with both body mass index (BMI) GWAS data and metabolomic data. We divided the whole dataset into two equally-sized and non-overlapping training and test datasets; we used the training data to build SNP- and metabolite-BMI association summary data and impute BMI on the test data. We compared the performance of the original and new imputation methods. As by the original method, the imputed BMI values by the new method largely retained SNP-BMI association information; however, the latter retained more information about BMI-environment associations and were more highly correlated with the original observed BMI values.

119

Longitudinal Gene Expression Changes Associated with Liver Measures in Hispanic/Latino Population at Risk for Metabolic-associated Fatty Liver Disease

Rashedeh Roshani *[1], Hung-Hsin Chen[1], Lizzie Frankel[1], Eric R. Gamazon[1], Wanying Zhu[1], Lauren E. Petty[1], Kari E. North[2], Joseph B. McCormick[3], Heather M. Highland[2], Susan P. Fisher-Hoch[3], Mariaelisa Graff[2], Jennifer E. Below[1]

[1]Vanderbilt University, Nashville, Tennessee, United States of America; [2]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [3]University of Texas at Houston, Houston, TX, United States of America

Metabolic-associated fatty liver disease (MAFLD) is a heterogeneous condition with variable severity and progression and the highest prevalence in Hispanic/Latino populations. To interpret genome-wide association studies (GWAS) signals and identify biomarkers, it is crucial to investigate transcriptomics associated with liver disease progression. This project aims to examine longitudinal gene expression changes related to liver measures in the Cameron County Hispanic Cohort (CCHC), a cohort of Mexican Americans residing at the US-Mexico border with high rates of MAFLD. This study focused on a group of 927 CCHC participants with longitudinal transient elastography phenotyping (FibroScan) to measure liver stiffness and controlled attenuation parameter to measure hepatic fat. Clinical chemistries were used to derive biomarkers, including fibrosis-4 index (FIB-4) and aspartate aminotransferase/platelet index. We employed linear mixed-effect regression models to estimate the effects of liver measures on gene expression using up to 1,294 measures, adjusting for covariates such as age, sex, and ten Probabilistic Estimation of Expression Residuals (PEER) factors. We used a false discovery rate corrected p value (FDR) to account for multiple testing. We identified 800 differentially expressed genes, including several genes within loci that have previously been implicated in GWAS of MAFLD and related traits. These genes include *DDX60L* (FDR_{FIB-4} = 0.038), which has been implicated for alanine aminotransferase levels, and *PARVB* (FDR_{FIB-4}= 0.043), which has been implicated for measures of hepatic fat, alanine aminotransferase, and both adult and pediatric MAFLD. Our results provide evidence of the functional effect driving those GWAS signals, including mapping the effector gene.

120

Ensemble Polygenic Risk Score Development for Coronary Heart Disease in Middle Eastern Populations

Mohamad Saad[1*], Nahin Khan[1], Khalid Kunji[1], Ehsan Ullah[1], Ayman El Menyar[2], Iftikhar J. Kullo[3], Jassim Al Suwaidi[2]

[1]*Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar;* [2]*Hamad Medical Corporation, Doha, Qatar;* [3]*Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota, United States of America;*

[*]Presenting author

Background: We previously validated four existing polygenic risk scores (PRSs) for coronary heart disease (CHD) in Middle Eastern (ME) populations using whole genome sequencing (WGS). Here, we aim at developing ME PRSs and integrating them with existing PRSs to obtain an ensemble PRSs that outperforms individual PRSs.

Methods: Our cohort comprised of 1,014 CHD patients and 6,009 controls with 30x WGS in a Middle Eastern cohort. We split the data into training and testing datasets (70% and 30%). We developed PRSs using pruning and thresholding (P+T), LDpred2, and machine learning models (e.g., XGBoost). We downloaded 35 PRSs from the PGS catalog and evaluated their performance in the testing dataset. We combined developed and existing PRSs into an ensemble PRS using summing and machine learning techniques.

Results: P+T model performed better than LDpred2 (OR=1.8, AUC=0.664 vs OR=1.7, AUC=0.656). The three existing PRSs that performed the best in our data were PGS000337 (OR=1.8, AUC=0.657), PGS003356 (OR=1.6, AUC=0.64), and PGS003355 (OR=1.6, AUC=0.637). Summing these three PRSs improved the performance (OR=1.9, AUC=0.667). Summing these three PRSs, and our P+T and LDpred2 improved the performance further (OR=2.2, AUC=0.698). The best XGBoost model outperformed P+T and LDpred2 (OR=1.8, AUC=0.67).

Conclusions: Our ME PRSs performed better than existing PRSs to predict CHD. Machine learning models showed good performance but not strikingly better than P+T and LDpred2. Combining PRSs developed with different datasets and methods improved prediction performance, which suggests a greater transferability across ancestries.

121

Structural Equation Modeling of Polygenic Risk and Environment in Late-Life Depression Using

Samar Elsheikh, PhD[1], Victoria Marshe, PhD[2], James L. Kennedy, MD[1], Guillaume Pare[3], Corinne E.Fischer, MD[4], Daniel Felsky, PhD[5,6] and Daniel Mueller, MD, PhD[7]

[1]*Centre for Addiction and Mental Health,Toronto, Ontario, Canada;* [2]*Colombia University, New York, New York, United States of America;* [3]*Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada;* [4]*St. Michael's Hospital, Toronto, Ontario, Canada;* [5]*Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario, Canada;* [6]*University of Toronto, Toronto, Ontario, Canada;* [7]*Center for Addiction and Mental Health, Toronto, Ontario, Canada*

Our understanding of the interplay between genetic and environmental factors (Gene x Environment Interaction, or GxE) determining mental health disorders has improved through the proliferation of genome-wide interaction association studies (GWIAS) and targeted GxE analyses. Moreover, multivariate modeling approaches, such as structural equation modeling (SEM) and polygenic risk scores (PRS), offer opportunities for the integration of clinical and genome-wide genotype data in building improved biopsychosocial models of mental illness aetiology and their response to treatment.

We propose to construct a SEM framework to uncover the inter-correlation and directed structure of mental health phenotypes by leveraging the joint predictive capacity of PRS for comorbid traits that share underlying biological and environmental risk pathways. The proposed model will be capable of linking latent constructs to their observed measurements; these will include disease severity, comorbidities and clinical histories, and behaviors and lifestyle factors such as physical and social activity.

Our gene-by-environment SEM (GESEM) will be initially developed and tested using four well-characterized clinical cohorts for older adults diagnosed with late-life depression and treated with antidepressants (CAN-BIND, IRL-GREY, STOP-PD II and IMPACT; n =1,238). The primary outcome will be antidepressant remission. Multiple PRS will be calculated to capture underlying genetic risk across vulnerable pathways which contribute to comorbidities. This selection will be made based on new, largely unpublished work from our group on the impact of PRS and targeted GxE studies on psychiatric outcomes across the lifespan. Each PRS will be calculated using both clumping and thresholding (PRSice-2) and continuous shrinkage (PRS-CS-auto) methods across selected cohorts using well-powered publicly available GWAS summary statistics. The multilevel GESEM model will include interactions between symptoms and comorbidities (i.e., observed measurements), which are caused by unobserved factors (i.e.,latent constructs), and are subject to modification by background PRS. We will compare our GESEM model against existing SEM-based approaches to GxE, including local SEM (LOSEM).

An open-source R package of the analytical code will be created and shared with the research community. This work has the potential to improve upon existing PRS-based predictive models in a clinical setting.

122

Categorization of Alzheimer's Disease Risk Variants Identifies Tissue-Shared and Tissue-Specific Genetic Regulatory Effects

Rebecca L. Sale[1], Michael Betti[1], Garrett Kaas[1], Eric R. Gamazon[1,2]

[1]*Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN;* [2]*Clare Hall, Cambridge University, Cambridge, United Kingdom*

Genome Wide Association Studies (GWAS) of Alzheimer's Disease (AD) have identified over 70 associated loci. Many of these associations lie in intronic and intergenic regions and are assumed to regulate gene transcription. Identifying a causal link between genetic variation and gene expression may identify novel therapeutic targets for AD. Furthermore, AD presentation is heterogeneous and by partitioning AD risk variants into functional and tissue specific categories, one can identify context-specific risk factors for personalized medicine approaches. To understand the tissue and cell type specific genetic regulation of gene expression, we utilized a functional prediction framework developed in the Gamazon lab, CoRE-BED. CoRE-BED leverages cell types and various functional assays from the EpiMap compendium to train a decision tree that categorizes individual variants into cell type specific functional categories, such as promoter or enhancer. These functional annotations are then utilized to estimate cell type contributions to AD heritability and risk by generating partitioned heritability and cell type specific Polygenic Risk Scores (PRS) in each cell type. Finally, to determine causal relationships between cell-type dependent gene expression and AD, we utilized Mendelian randomization with Joint Tissue Imputation (MR-JTI). MR-JTI leverages instrumental variables (Z) to estimate the causal effect of a gene (G) on the trait (Y). This causal framework will provide mechanistic insights into AD genetic predisposition.

123

Colocalization Analysis of Sex-dependent Traits

Eric J. Sanders*[1], Lisa J. Strug[2,3,4], Deb K. Pal[5,6,7]

[1]Dalla Lana School of Public Health, The University of Toronto, M5T 3M7 Toronto, Canada; [2]Genetics and Genome Biology Program, The Hospital for Sick Children, M5G 0A4 Toronto, Canada; [3]Departments of Statistical Sciences and Computer Science and Division of Biostatistics, The University of Toronto, M5G 1Z5 Toronto, Canada; [4]The Centre for Applied Genomics, The Hospital for Sick Children, M5G 0A4 Toronto, Canada; [5]Department of Basic & Clinical Neurosciences, Institute of Psychiatry, Psychology & Neuroscience, King's College London, SE5 9RX London, United Kingdom; [6]MRC Centre for Neurodevelopmental Disorders, King's College London, SE1 1UL London, United Kingdom; [7]King's College Hospital, SE5 9RS London, United Kingdom

Colocalization analysis investigates whether two independent association signals in a genetic region have a shared causal variant. It is commonly conducted using statistics from a genome-wide association study (GWAS) and an expression quantitative train locus (eQTL) analysis, as this can guide follow-up functional studies by informing the responsible gene and tissue of origin for the GWAS signal. However, sex-specific genetic associations pose a challenge for colocalization analysis. For example, in our previous GWAS of absence seizure history in individuals with juvenile myoclonic epilepsy (JME), we identified a protective SNP on chromosome 17 specific to females. Follow-up investigation through colocalization analysis requires choosing to either disregard sex variation or to perform analysis marginally by sex, both of which can reduce power. To address this, we propose two novel approaches extending the colocalization methodology implemented in our software package LocusFocus. The first method derives a novel statistic that considers if genetic association exists in either sex. The second concatenates marginal test statistics from male and female groups for analysis. Simulations were conducted to compare the two novel approaches with an approach that ignored sex variation and an approach analyzing each sex marginally. It was observed that our novel approach based on concatenated statistics consistently met or exceeded the power exhibited in the other three approaches. Thus, the concatenated statistic approach may be the more versatile option when studying colocalization in suspected sex-varying traits. This approach is applied in ongoing investigations of sex-varying associations in absence seizure incidence in those with JME.

124

Large-scale Multilayer Proteomics and Multi-omics Integration Reveal Molecular Networks Related to Alzheimer's Disease in Diabetic Brains

Vishal Sarsani[1,2], Ana W. Capuano[3], Shinya Tasaki[3], Zoe Arvanitakis[3] and Liming Liang[1,2,4]

[1]Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Program in Genetic Epidemiology and Statistical Genetics, Harvard T H Chan School of Public Health, Boston, Massachusetts, United States of America; [3]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, United States of America; [4]Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, Massachusetts, United States of America

Type 2 diabetes mellitus (T2D) and Alzheimer's disease (AD) are two complex diseases that are prevalent in the aging population. Recent research suggests that insulin resistance, a key feature of T2D, is associated with AD pathology, including amyloid-β deposition and cognitive impairment. To better understand the molecular mechanisms underlying insulin resistance in the brain and periphery and relate these to AD neuropathology and cognitive function, we performed a deep multilayer proteomics profiling and multi-omics integration study. We examined 192 post-mortem brain and muscle samples from individuals with and without T2D and with varying levels of AD pathology. We quantified 11,726 phosphopeptides and 29,955 non-phosphopeptides from 3,537 protein groups in frozen brain tissue (dorsolateral prefrontal cortex). Our analysis revealed that subjects with higher levels of AD pathology exhibited hyperphosphorylation in proteins like MAPT, regardless of T2D status. Interestingly, we also identified 105 phosphorylation changes in proteins, including MAP2, SLC43A2, and GIT1, that were common to both AD and T2D. We integrated GWAS, methylation, metabolomics, and transcriptomic profiles of matched samples to build predictive omics signatures using a novel semi-supervised machine learning (ML) framework. To examine the generalizability and utility of our results, we tested our framework-generated signatures on primarily peripheral blood omics data from AD and T2D subjects in UK Biobank. In conclusion, with our multilayer proteomics data and novel ML omics integration framework, we built predictive signatures and molecular network maps of AD and T2D phosphorylation signaling events, thus filling knowledge gaps in metabolism, insulin signaling, and AD research.

125

Multivariable Mendelian Randomization to Disentangle the Alcohol Harm Paradox

Gemma Sawyer[1*], Jasmine Khouja[1,2], Hannah Sallis[1,3], Marcus Munafò[1,2], Liam Mahedy[1]

[1]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [2]School of Psychological Sciences, University of Bristol, Bristol, United Kingdom; [3]Centre for Academic Mental Health, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

The alcohol harm paradox, whereby low socioeconomic position (SEP) groups experience greater alcohol-related harms despite reporting lower alcohol consumption, is yet to be fully understood through observational studies because key drivers are correlated and share similar confounding structures. We conducted multivariable Mendelian randomization (MVMR) to estimate the direct causal effect of drinks per week (DPW) and years of schooling (YOS), whilst accounting for the other, on multiple health outcomes. Previously published GWAS summary statistics for DPW and YOS were utilised, and summary statistics were generated from individual-level data from UK Biobank (N = 462,818) for all outcomes. In inverse variance weighted analyses, we found evidence for direct effects of DPW and YOS on liver diseases, mental and behavioural disorders due to alcohol, and stroke, indicating that alcohol consumption increased the likelihood of outcomes whereas years of education decreased their likelihood. There was also evidence for a direct effect of DPW on depression, anxiety, influenza/pneumonia, and heart disease. In contrast, there was evidence of a total, but not direct, effect of DPW on depression, influenza/pneumonia, epilepsy, and injuries when accounting for YOS, suggesting that education drives these total effects. Although caution is required when interpreting these results due to weak instruments for alcohol, our results may provide indicative evidence that the alcohol harm paradox is partially due to the protective effect of additional years of education, resulting in a reduced likelihood of higher SEP groups developing many alcohol-related outcomes. Replication with strong instruments would be necessary to draw causal inferences.

126

Biobank Scale Analysis of Mendelian Disease

Alexandra Scalici[1,2,*], Tyne Miller-Fleming, PhD[1,2], Dharmendra Choudhary, PhD[1,2], Ela W. Knapik, MD[1,2], Nancy J. Cox, PhD[1,2]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America.

Genome-wide association studies (GWAS) have identified the contribution of common variation to numerous phenotypes. However, population level studies like GWAS are not suitable to study rare, monogenic variation that often causes Mendelian disease due to limitations in statistical power and small sample sizes. We implemented an approach that leverages electronic health records (EHR), genetically predicted gene expression (GPGE) data from an EHR-linked biobank, BioVU, and known Mendelian disease phenotypes cataloged in the Online Mendelian Inheritance of Man (OMIM) to study Mendelian disease at the population level. We applied this framework to LH3 Deficiency, a rare connective tissue disorder linked to mutations in *PLOD3* that is characterized by bone fragility with contractures, arterial rupture, and deafness. We conducted a gene-based phenome-wide association study (PheWAS) to identify phenotypes associated with reduced GPGE of *PLOD3* (< -2 SD units) in BioVU. Our PheWAS identified previously reported LH3 Deficiency phenotypes catalogued in OMIM and novel phenotypes associated with reduced GPGE of *PLOD3*. Using the effect estimates from the associated phenome ($p<0.05$) as weights, we constructed a phenotypic risk score (PheRS) and used it in a transcriptome-wide association study (TWAS) to test whether the *PLOD3* PheRS is associated with changes in GPGE. Conducting pathway and overrepresentation analyses (ORA), we identified pathways associated with reduced GPGE of *PLOD3*. This novel application of PheWAS and PheRS has the potential to allow us to study Mendelian disease at the population level.

127

Mendelian Randomization Approaches Lead to Improved Understanding of Causal Traits and Genes Associated with Stuttering

Alyssa C. Scartozzi[1], Dillon G. Pruett[2], Hannah G. Polikowsky[1], Douglas M. Shaw[1], Hung-Hsin Chen[1], Lauren E. Petty[1], Alexander S. Petty[1], Emily Lowther[4], Yao Yu[5], 23andMe Research Team[6], Heather M. Highland[7], Christy L. Avery[7,8], Kathleen M. Harris[8,9], Reyna L. Gordon[10], Janet M. Beilby[11], Kathy Z. Viljoen[11], Robin M. Jones[2,12], Shelly Jo Kraft[4], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Hearing and Speech Sciences, Vanderbilt University, Nashville, Tennessee, United States of America; [3]Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [4]Communication Sciences and Disorders, Wayne State University, Detroit, MI, United States of America; [5]Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America; [6]23andMe, Inc, Sunnyvale, California, United States of America; [7]Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [8]Carolina Population Center, the University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [9]Department of Sociology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [10]Department of Otolaryngology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [11]Curtin School of Allied Health, Curtin University, Perth, WA, Australia; [12]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Developmental stuttering is a common speech disorder associated with various comorbid traits. Despite population and familial-population efforts, the genetic factors affecting stuttering risk and comorbidity are unknown. Using 23andMe, Inc. summary statistics from GWAS of self-reported stuttering, we explored causal pathways between stuttering and reported comorbidities, and identified candidate causal genes.

To determine if comorbid traits harbor a causal or horizontal pleiotropic relationship, we performed sex-specific Mendelian randomization (MR) of 18 traits with stuttering. We found causal effects of genetic liability associated with BMI, chronotype, walking pace, suicide ideation, and testosterone on stuttering risk. Genetic risk of stuttering showed causal effects on depression (*MR Egger* $p < .05$). These results align with several studies suggesting individuals who stutter experience elevated prevalence of depression.

To identify potential causal genes associated with stuttering, we performed sex-specific MR-joint-tissue expression imputation. For females, we identified 92 unique causally-implicated genes after multiple test correction. Only three of these genes, *NMUR2, MMAB, DCC*, were the reported genes at genome-wide significant loci in the original 23andMe GWAS of stuttering (Polikowsky et al. 2023) and are involved in GPCR signaling, metabolic processes, and axon guidance. For males, we identified 24 causal genes. Only *MMAB* was identified in Poliwkosky et al., 2023, and *ZMAT4* was a top hit in an independent analysis, Shaw et al. 2021, and is involved in nucleic acid binding.

Future studies will investigate neuroendophenotypes associated with these potentially causal stuttering genes. Understanding these genetic relationships may inform clinical manifestations of stuttering that differ between sexes.

128

Examination of Nephrotic Syndrome Genetic Architecture in Large-Scale Biobanks Replicates Known and Identifies 20 Novel Loci

Hannah M. Seagle[1], Joseph H. Breeyear[1,2], Eric S. Torstenson[3], Yanfei Zhang[4], Gail P. Jarvik[5], Ozan Dikiltas[6], Ming Ta Michael Lee[7], Iftikhar J. Kullo[6], Robb Rowley[8], Krzysztof Kiryluk[9], Todd L. Edwards[1,2,10], Katalin Susztak[11], Jacklyn N. Hellwege[1,2,12]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America; [2]VA Tennessee Valley Healthcare System (626), Nashville, Tennessee, United States of America; [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4]Genomic Medicine Institute, Geisinger Health Systems, Danville, Pennsylvania, United States of America; [5]Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington Medical Center, Seattle, WA, United States of America; [6]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota, United States of America; [7]Galatea Bio, Inc. Hialeah, Florida, United States of America; [8]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [9]Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, New York, United States of America; [10]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [11]University of Pennsylvania and Children's Hospital of Philadelphia Kidney Innovation Center, Philadelphia, Pennsylvania, United States of America; [12]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Nephrotic syndrome (NS) is a collection of disorders characterized by excess protein in urine (>3g/day). We integrated genotype data from Vanderbilt University's BioVU and the Electronic Medical Records and Genetics (eMERGE) network with summary statistics from FinnGen, Million Veteran Program (MVP), Biobank Japan (BBJ) and UK Biobank (UKB), for a cross-ancestry total of 5,214 cases (3,060|790|1,364; European (EUR)|African (AFR)|East and Central Asian (ASN)) and 1,601,060 controls (1,279,630|135,664|185,766; EUR|AFR|ASN). Cases were identified with diagnostic codes (FinnGen) or NS phecodes. Controls lacked renal disease diagnosis codes. Associations with NS were modeled as a function of additive genotype, sex, and the top ten principal components, followed by inverse-variance weighted meta-analysis within and across ancestral groups. We identified 31 significant ($r^2<0.1$, >1Mb, $p<5x10^{-8}$) loci in the multi-ancestry analysis. The strongest association in the multi-ancestry analysis was rs1265889 (class II MHC region, $p=1.4x10^{-20}$, OR=1.50 (1.43–1.57)). In the AFR analysis, we identified 16 novel non-HLA loci and replicated *APOL1* G1 variants ($p=4.5x10^{-15}$, OR= 2.56). Associations between NS and genetically-predicted gene expression were evaluated using GTEx v8 tissue models. We identified 28 unique genes among 167 significant ($p<1.55x10^{-6}$) gene-tissue pairs in the multi-ancestry analysis. Among significant SNPs in these genes, 1,080 SNPs in *AGPAT1, C4A, HLA-DQA1, HLA-DQB1, HLA-DQB2, NOTCH4,* and *RNF5* genes were also kidney eQTLs (FDR<0.05) in the Human Kidney eQTL Atlas (susztaklab.org). This study provides insight into NS through identification of novel genetic loci and translation to predicted renal gene expression in the largest multi-ancestry meta-analyses to date.

129

Genome-wide Association Study of Neuropathology Endophenotypes Provides Insights into Dementia Risk Factors

Lincoln M.P. Shade[1], Yuriko Katsumata[1,2], Erin L. Abner[2,3], Steven A. Claas[1], Timothy J. Hohman[4], Shubhabrata Mukherjee[5], Richard P. Mayeux[6], Lindsay A. Farrer[7], Gerard D. Schellenberg[8], Jonathan L. Haines[9], Walter A. Kukull[10], Kwagsik Nho[11], Andrew J. Saykin[11], David A. Bennett[12], Julie A. Schneider[13], Mark T. W. Ebbert[2,14,15], Peter T. Nelson[15,16], David W. Fardo[1,2], the Alzheimer's Disease Neuroimaging Initiative, the National Alzheimer's Coordinating Center, and the Alzheimer's Disease Genetics Consortium

[1]Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, Kentucky, United States of America; [2]Sanders-Brown Center on Aging and Alzheimer's Disease Research Center, University of Kentucky, Lexington, Kentucky, United States of America; [3]Department of Epidemiology and Environmental Health, College of Public Health, University of Kentucky, Lexington, Kentucky, United States of America; [4]Vanderbilt Memory & Alzheimer's Center, Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [6]Department of Neurology, Columbia University, New York, New York, United States of America; [7]Department of Medicine, Boston University, Boston, Massachusetts, United States of America; [8]Penn Neurodegeneration Genomics Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [9]Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio, United States of America; [10]Department of Epidemiology, University of Washington, Seattle, Washington, United States of America; [11]Department of Radiology & Imaging Sciences, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [12]Rush Alzheimer's Disease Center and Department of Neurological Sciences, Rush Medical College, Chicago, Illinois, United States of America; [13]Rush Alzheimer's Disease Center and Departments of Neurology and Pathology, Rush University Medical Center, Chicago, Illinois, United States of America; [14]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, Kentucky, United States of America; [15]Department of Neuroscience, University of Kentucky College of Medicine, Lexington, Kentucky, United States of America; [16]Department of Pathology and Laboratory Medicine, University of Kentucky, Lexington, Kentucky, United States of America

Background: Genome-wide association studies (GWAS) have identified over 70 Alzheimer's disease (AD)-associated loci. However, the clinical- and proxy-based outcomes used in most studies do not reflect the complexity of underlying neuropathologies. We studied the genome-wide association of eleven dementia-related neuropathological endophenotypes using four sources of neuropathological data (pooled N = 7,463).

Methods: GWAS were performed for each endophenotype and followed by downstream functional analyses to identify potential molecular functions of risk loci. To confirm molecular phenotypic association with neuropathologies, we performed targeted analyses with DNA methylation and bulk RNA-Seq data from the dorsolateral prefrontal cortex.

Results: We identified a locus near *APOC2* (rs4803778 OR=1.24, P value=$5.8\text{x}10^{-12}$) associated with cerebral amyloid angiopathy (CAA) after adjusting for *APOE* ε2/ε3/ε4 diplotypes. This locus is also associated with DNA methylation at four nearby CpG sites. Methylation levels at two sites (cg0955818 P value=0.004, cg13119609 P value=0.0007) were associated with CAA. Additionally, we identified two novel neuropathology risk loci (*PIK3R5* and neurofibrillary pathology; *COL4A1* and cerebral atherosclerosis) and confirmed associations of known AD risk loci with multiple neuropathologies.

Conclusions: Our findings highlight the importance of neuropathological endophenotypes as necessary complements to clinical AD studies to understand the genetic risk of clinical AD-related dementias.

130

Network-based Quantitative Trait Linkage Analysis of Microbiome Composition in Inflammatory Bowel Disease Families

Arunabh Sharma[1], Olaf Junge[1], Silke Szymczak[2], Malte Rühlemann[3], Janna Enderle[4], Stefan Schreiber[3,5], Andre Franke[3], Wolfgang Lieb[4], Michael Krawczak[1], Astrid Dempfle[1]*

[1]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [2]Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany; [3]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany; [4]Institute of Epidemiology, Kiel University, Kiel, Germany; [5]Department of Internal Medicine I, University Hospital Schleswig-Holstein, Kiel, Germany

Inflammatory bowel disease (IBD) is characterized by a dysbiosis of the gut microbiome that results from the interaction of the constituting taxa with one another and with the host. At the same time, host genetic variation is associated with both IBD risk and microbiome composition. In the present study, we therefore defined quantitative traits (QTs) from modules identified in microbial co-occurrence networks to measure the consistency of microbial abundance and subjected these QTs to a genome-wide quantitative trait locus (QTL) linkage analysis. Linkage analysis was performed in the Kiel IBD family cohort (IBD-KC), an ongoing study of 256 German families comprising 455 IBD patients and 575 first- and second-degree, non-affected relatives. The analysis revealed five chromosomal regions to be linked to one of three microbial module QTs, namely on chromosomes 3 (spanning 10.79 cM) and 11 (6.69 cM) for the first module (labeled 'blue'), chr9 (0.13 cM) and chr16 (1.20 cM) for module 'yellow,, and chr13 (19.98 cM) for module 'turquoise.' Our study thus illustrates the benefit of combining microbial network and family-based linkage analysis to identify novel genetic drivers of microbiome composition in a specific disease context.

131

Genome-wide Implementation of Multi-population Joint Analysis Marginal Summary Statistics (mJAM) and Its Applications in Polygenic Risk Score Models

Jiayi Shen[1], James Baurley[2], Jingxuan He[1], Gillian King[1], Christopher A. Haiman[3,4], David V. Conti[1, 3, 4]

[1]Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [2] BioRealm Research, Walnut, California, United States of America; [3]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [4]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America

Previously we proposed a multi-population fine-mapping approach called "mJAM". mJAM effectively fits a multi-SNP model within each population with population-specific LD and then performs a fixed-effect meta-analysis of the joint model while incorporating a Bayesian 'g-prior' for robust estimation. The resulting model provides the flexibility to identify noteworthy SNPs using existing Bayesian or frequentist feature selection approaches. We propose a scalable version of the previous region-wise mJAM implementation that can be applied to genome-wide data ("GW-mJAM"). We illustrated GW-mJAM with forward selection as the main approach for selecting index SNPs, applied GW-mJAM on the latest prostate cancer summary statistics with four populations, and constructed PRS using the index SNPs. Resultant PRS built from GW-mJAM can be applied to individuals of any ancestry without estimating ancestry or using self-identifying population information. We evaluated GW-mJAM's PRS–using incident data from the MultiEthnic Cohort Study (five populations, 2,885 cases, 25,904 controls; Kolonel et al., 2000) where the OR of 1-standard-deviation increase in PRS is 1.96 (1.76 - 2.18) among African-ancestry men and 2.25 (2.03 - 2.50) among European-ancestry men. In addition, we incorporate GW-mJAM with other feature selection approaches as well as apply GW-mJAM to three other cancer types (breast, colorectal, and lung) with available multi-population summary statistics. PRS models built from GW-mJAM's results are evaluated and compared with other GW PRS approaches such as PRS-CSx (Ruan et al., 2022) to investigate the impact of trait characteristics, quality of reference panel, and methodology on test performance.

132

Phenome-Wide Association Study of Polygenic Risk for Central Adiposity

Navya Shilpa Josyula[1*]; Geetha Chittoor[1]; Mariaelisa Graff[2]; Emmaleigh Wilson[3], Zhe Wang[4], Meng Lin[5], Hung-Hsin Chin[6], Greg Linchangco[7], Anne E. Justice[1] on behalf of the GIANT Waist Traits Working Group

[1]Population Health Sciences, Geisinger Health, Danville, PA; [2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America;

[3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [4]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, United States of America; [5]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Denver, CO, United States of America; [6]The Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [7]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA

Abdominal obesity has a high comorbidity burden independent of overall obesity, yet abdominal circumference is not typically assessed clinically. With the growing number of EHR-linked biobanks with available genotyping, we assessed the potential of using polygenic risk scores (PRS) for central adiposity as measured by waist-to-hip ratio adjusted for BMI [WHRadjBMI] and performed a phenome-wide association (PheWAS) to identify associations.

Using results from a multi-population GWAS of WHRadjBMI including over 1,187,156 participants (80% European, 11% East Asian, 4% South Asian, 3% Hispanic/Latino, 2% African), we developed a PRS using PRS-CSx with optimization in the pan-ancestry UKBB. The resulting SNPs and weights were used to create a weighted score and evaluate associations between the PRS and diseases/traits, adjusted for age, sex, and principal components, in five independent studies (MyCode, eMERGE, CCPM, BioME and BioVU) including 326,440 individuals. Study-specific PheWAS results were then meta-analyzed using inverse-variance weighted fixed-effects meta-analysis.

Of the 1,792 traits tested; we identified 151 significant associations ($p<2.8e-05$) with most significant results validating expected central adiposity associations with cardiometabolic outcomes. Increasing PRS was positively associated with Type 2 diabetes ($p=1.3e-273$), hyperlipidemia ($P=8.5e-121$), coronary atherosclerosis ($p=3.1e-81$), and hypertension ($p=1.4e-74$). We also identified significant associations with digestive (e.g., chronic liver disease, $P=3.2e-51$), genitourinary (e.g., chronic kidney disease, $p=2.0e-14$), and musculoskeletal (e.g., neurogenic arthropathy, $p=6.6e-12$) illnesses.

Our results demonstrate that central adiposity PRS is associated with an increased risk of several cardiometabolic diseases and suggest a potential value to assess risk in the absence of clinically-assessed body fat distribution.

133

Accurate UV Exposure Measure Reveals Novel Variants in Genome-wide Association and Interaction Analyses of Vitamin D Status

Rasha Shraim[1,2,3], Jos van Geffen[4], Michiel van Weele[4], Ross McManus[2], Lina Zgaga[1]

[1]Department of Public Health and Primary Care, Institute of Population Health, Trinity College Dublin, Dublin, Ireland; [2]Department of Clinical Medicine, Trinity Translational Medicine Institute, Trinity College Dublin, Dublin, Ireland; [3]The SFI Centre for Research Training in Genomics Data Sciences, University of Galway, Galway, Ireland; [4]Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

Sun exposure is the major determinant of vitamin D levels in humans. Recent genome-wide studies have identified up to 143 variants linked to vitamin D status, independently and in interaction with sun exposure. The common approach relies on season of blood sampling, dichotomized as summer/winter, as a proxy for UV exposure. We explored the aetiology of vitamin D status in the UK Biobank, using an exact measure of ambient UV. To capture UV exposure over a period of time, we calculated a weighted UV dose for each participant based on their residence and date of blood sampling. UVB, the relevant wavelength for vitamin D production, was adjusted for cloud cover and the rate of decay of vitamin D in the body. We first carried out a genome-wide association study (GWAS) of 25-hydroxyvitamin-D (25OHD) in 408,820 White British individuals. In addition to age, sex, supplement intake, and population structure, the model was adjusted for the calculated UVB dose. We then carried out a genome-wide gene-environment interaction analysis (GxE) in the same sample where the model additionally included a GxUVB interaction term. The GWAS results identified 54 novel independent variants significantly associated with 25OHD concentration and the GxE analysis identified 37 novel variants. Downstream functional analysis of these variants supports vitamin D involvement in metabolic pathways and hormone and skin phenotypes. Overall, our findings support the importance of accurate environmental exposure measures and accounting for gene-environment interactions in uncovering the genetic architecture of complex traits like vitamin D status.

134

Two-sample Mendelian Randomization of Major Depressive Disorder and Inflammatory Bowel Disease

Claire L. Simpson[1], Steven R. Brant[2]

[1]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America; [2]Robert Wood Johnson Medical School, Rutgers University, New Jersey, United States of America

Inflammatory bowel disease (IBD), including the subtypes Crohn's disease and ulcerative colitis, is a chronic, immune-mediated disease that is typically diagnosed in the late teens or twenties. However, it can begin at any age. Symptoms include persistent diarrhea, abdominal pain, cramping, loss of appetite, weight loss, and fever. A significantly lower quality of life has been reported in many patients. This effect appears to go beyond the simple burden of dealing with the disease as the complex relationship between the gut and the brain may increase the risk to patients with IBD of developing mental illness. Whether the effect is unidirectional, bidirectional, or caused by common external risk factors is unknown. This study aims to explore the causal relationship between IBD and major depressive disorder. Using publicly available genome-wide association study summary datasets, we selected SNPs from GWAS of Major Depressive Disorder (MDD) (N=480,359), Crohn's Disease (CD) (N=51,874), Ulcerative Colitis (UC)(N=47,745) and Inflammatory Bowel Disease (IBD)(N=65,642). We performed inverse-variance weighted (IVW) two-sample Mendelian randomization (TSMR). We observed an inverse association with the outcomes CD (beta=2.23E-02, P=0.04) and UC (beta=0.37, P=0.04) with MDD, but not with IBD (beta=-0.14, P=0.87). Analyses are ongoing and will be presented.

135

A Genetic Association Test Robust to Arbitrary Population Structure

Minsun Song*

Department of Statistics, Sookmyung Women's University, Seoul, Korea

We present a new statistical test of association between a trait and genetic markers, which we theoretically and practically prove to be robust to arbitrarily complex population structure. The statistical test involves a set of parameters that can be directly estimated from large-scale genotyping data, such as those measured in genome-wide associations studies. We also derive a new set of methodologies, called a genotype-conditional association test, shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and non-genetic contributions to the trait. Our proposed framework provides a substantially different approach to the problem from existing methods.

136

Mosaic Loss of Chromosome Y and Telomere Length as Risk factors for Alzheimer's Disease in the Midwestern Amish

Yeunjoo E. Song[1,2,+,*], Yining Liu[1,2,+], Renee A. Laux[1], Weihuan Wang[1], Kristy L. Miskimen[1], Sarada L. Fuzzell[1], Sherri D. Hochstetler[1], Leighanne R. Main[3], Ping Wang[1], Michael B. Prough[4], Daniel A. Dorfsman[4], Susan H. Slifer[4], Larry D. Adams[4], Laura J. Caywood[4], Jason E. Clouse[4], Sharlene D. Herington[4], Audrey Lynn[1,2], Jeffery M. Vance[4,5], Michael L. Cuccaro[4,5], Paula K. Ogrocki[6,7], Alan J. Lerner[6,7], Margaret A. Pericak-Vance[4,5], William K. Scott[4,5], and Jonathan L. Haines[1,2,3]

[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio, United States of America; [2]Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio, United States of America; [3]Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio, United States of America; [4]Hussman Institute for Human Genomics, University of Miami School of Medicine, Miami, Florida, United States of America; [5]Dr. John T Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, Florida, United States of America; [6]Brain Health and Memory Center, University Hospital, Cleveland, Ohio, United States of America; [7]Department of Neurology, Case Western Reserve University School of Medicine, Cleveland, Ohio, United States of America

+ contributed equally

In aging men, mosaic loss of chromosome Y (mLOY) has been suggested as a possible biomarker for increased risk of numerous diseases, including Alzheimer's disease (AD). Telomere length (TL) also is a candidate biomarker for aging and many aging-related diseases. We investigated the familial correlation and heritability of these two biomarkers as risk factors for AD in the Midwestern Amish.

TL was calculated from whole genome sequencing (WGS) data using Telseq. The mean read depth of the male-specific region of chromosome Y was used to measure the degree of mLOY for males only. The multivariate familial correlations were estimated. The association between biomarkers and the Modified Mini-Mental State (3MS) examination and heritability were assessed accounting for relatedness and age at blood sampling.

After QC, 1,047 individuals (mean age=80.76±5.71, 41% male) were available. Heritability of TL was 0.17 (SE=0.07, p=9.4x10^{-3}), 0.38 (SE=0.13, p=2.2x10^{-3}) and 0.04 (SE=0.12, p=0.37) in all, males and females, respectively. Heritability of mLOY was 0.38 (SE=0.16, p=0.01) in males. mLOY and TL was highly correlated in males (r=0.21±0.05, p=2.0x10^{-4}). The TL correlation was higher among brother pairs than in sister pairs (0.306 vs. 0.001). It was higher in father-son pairs than any other parent-offspring pairs. The correlation between TL and 3MS was higher among males than in females (0.109, p=0.08 vs. 0.018, p=0.37). Our results along with the lower prevalence of AD in Amish reinforces mLOY and TL as promising biomarkers for the risk of AD.

137

MagicalRsq-X: A Cross-cohort Transferable Genotype Imputation Quality Metric

Quan Sun[1], Yingxi Yang[2], Jiawen Chen[1], Jia Wen[1], Michael R. Knowles[1], Charles Kooperberg[3], Alex Reiner[3], Laura M. Raffield[1], April Carson[4], Stephen Rich [5], Jerome Rotter[6], Ruth Loos[7], Eimear Kenny[7], Byron C. Jaeger[7], Yuan-I Min[4], Christian Fuchsberger[8], Yun Li[1]

[1]University of North Carolina at Chapel Hill; [2]Yale University; [3]Fred Hutchinson Cancer Center; [4]University of Mississippi Medical Center; [5]University of Virginia; [6]University of California, Los Angeles; [7]Icahn School of Medicine at Mount Sinai; [8]EURAC Research Center.

Since genotype imputation was introduced, researchers have been relying on the estimated imputation quality from imputation software to perform post-imputation quality control (QC). However, this quality estimate (denoted as Rsq) performs less well for lower frequency variants. We recently published MagicalRsq, a machine-learning-based imputation quality calibration metric, which leverages additional typed markers from the same cohort and outperforms Rsq as a QC metric. In this work, we extended the original MagicalRsq to allow cross-cohort model training, named MagicalRsq-X. We removed the cohort-specific estimated minor allele frequency and additionally included LD scores and recombination rates as variant-level features. Leveraging whole genome sequencing data from TOPMed, specifically participants in BioMe, JHS, WHI and MESA studies, we performed comprehensive cross-cohort evaluations for European and African ancestral individuals based on their inferred global ancestry with the 1000 Genomes and HGDP data as reference. Our results suggest MagicalRsq-X outperforms Rsq in almost every setting, with 7.3-14.4% improvement in squared Pearson correlation with true R^2, corresponding to 85-218K variant gains. We further developed a metric to quantify the genetic distances of a target cohort relative to a reference cohort and showed that such metric could largely explain the performance of MagicalRsq-X models. Finally, we found that MagicalRsq-X saved 9-53 GWAS variants in one of the largest blood cell traits GWAS results that would be missed using the original Rsq for QC.

In conclusion, MagicalRsq-X shows clear superiority for post-imputation QC and can greatly benefit genetic studies by rescuing well-imputed low frequency and rare variants.

138

Whole Exome Sequencing Association Study of Familial Bipolar Disorder and Related Conditions in Anabaptist Founder Populations

Heejong Sung[1]*, Layla Kassem[1], Emily Besancon[1], Fabiana Lopes[1], Sevilla Detera-Wadleigh[1], Nirmala Akula[1], Antonio Nardi[2], Thomas G. Schulze[3], Francis J. McMahon[1]

[1]Genetic Basis of Mood and Anxiety Disorder, Human Genetics Branch, National Institute of Mental Health, NIH, Bethesda, Maryland, United States of America; [2]Institute of Psychiatry, Federal University of Rio de Janeiro, Rio de Janeiro, RJ Brazil; [3]Institute of Psychiatric Phenomics and Genomics, University Hospital, LMU Munich, Munich, 80336, Germany; Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York, United States of America; Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

The Amish Mennonite Bipolar Genetics (AMBiGen) study seeks genetic variants that substantially increase the risk for bipolar disorder (BD) and related conditions in founder populations. In this study, the affected participants had diagnoses of BD, schizophrenia, schizoaffective disorder or recurrent major depressive disorder. Whole exome sequencing (WES) was performed at Regeneron Genetics Center. 533,144 variants with a read depth >30x, calling rate >95% in 820 samples with genotyping rate >95%, and no Mendelian errors were retained for analysis of 431 cases and 389 controls. Heterozygous variants, with Anabaptist Variant Server (AVS) minor allele frequency (MAF) <0.01, shared by >5 affected individuals in their first-degree relatives, and belonging to genes with pLI >0.99, were grouped by Variant Effect Predictor (VEP) impact level (High, Moderate, Low, Modifier). The AVS is a database that provides variant annotation information for over 10,000 Amish and Mennonite individuals. Association tests were run in SAIGE-GENE, adjusting for relatedness, sex, and population principal components. Modifier variants were significantly associated with the affection status although other impact level variants were not. Half of the considered modifier variants have an AMBiGen MAF greater than the MAF in AVS or non-Finnish Europeans in gnomAD. The gene list includes some known psychiatric risk genes. The AMBiGen represents the largest WES study of BD in founder populations. The results of AMBiGen WES study suggest increased burdens of rare, modifier variants shared among affected first-degree relatives in genes with a high intolerance to loss-of-function mutations.

139

Drug Repurposing for Alzheimer's Disease: The Use of Genetics-enriched, Neuropathology-associated SPI1 Regulon in Microglia

Tan Yuting[1,2]*, Rui Chen[1,2], Quan Wang[1,2], Anshul Tiwari[1,2] , Yan Yan[3], Bingshan Li[1,2], Xue Zhong[1,4]

[1]Vanderbilt Genetics Institute, Nashville, Tennessee, United States of America; [2]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee, United States of America; [3]Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America; [4]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Alzheimer's disease (AD) is characterized by progressive decline in memory and cognition and involves multiple brain cell types. Among them, microglia play a crucial role and are the major cell type implicated in AD genetics. Evidence accumulates suggesting that drugs with support from genetics are more likely to succeed to the market. Here, we aim to identify genetics-enriched, AD-relevant regulatory programs in microglia to guide the search of approved drugs for repurposing toward AD. We detected 241 regulons from AD single-cell RNAseq data, with 42 of them being significantly enriched in AD heritability based on LDSC on the latest AD GWAS summary statistics. Most of these genetics-enriched regulons are active in microglia, and the SPI1-regulon, named by the master regulatory TF SPI1 (PU.1), ranked with the highest AD heritability. The SPI1-regulon activity score in microglia increased as the AD pathological burden increased, and its regulon genes are enriched for GO terms including mononuclear cell differentiation, wound healing, endocytic vesicle, and GTPase regulator activity etc. Several AD risk genes appear in the SPI1-regulon, including APOE, INPP5D and CD74 etc. By evaluating the proximity between drug target(s) and SPI1-regulon genes on a gene-gene interaction network, we identified several promising candidate drugs for AD, including Salsalate and Baricitinib. In conclusion, we showed that combining single-cell RNAseq, GWAS summary statistics and neuropathological traits of AD enabled the dissection of genetics-enriched, AD-relevant regulatory programs in human microglia, which have value in guiding the drug purposing for AD.

140

Empowering Immunogenetic Analysis with Biofilter 3.0 via Enhanced Annotation and Filtering Capabilities

Van Q. Truong[1-6], Xueqiong Li[2-4], Rasika Venkatesh[1-4], Scott M. Dudek[2-4], E. John Wherry[1,5,6], Marylyn D. Ritchie[1-4]

[1]Graduate Group in Genomics & Computational Biology, Perelman School of Medicine, University of Pennsylvania, United States of America; [2]The Penn Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, United States of America; [3]Biomedical and Translational Informatics Laboratory, Perelman School of Medicine, University of Pennsylvania; United States of America; [4]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, United States of America; [5]The Penn Institute for Immunology and Immune Health, Perelman School of Medicine, University of Pennsylvania, United States of America; [6]Department of Pharmacology & Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, United States of America; Corresponding Author: Marylyn D. Ritchie, marylyn@pennmedicine.upenn.edu

Autoimmune diseases are a growing global health concern, affecting approximately 4.5% of people worldwide. In addition, 25% of affected individuals suffer from multiple autoimmune disorders. While hundreds of genetic variants have been identified for the most common autoimmune diseases, the small effect sizes and limited explanatory power underscore the need for an integrative knowledge-driven approach. The body of knowledge for genetic and immunological information is vast and oftentimes overwhelming. We discuss how incorporating immune-relevant information can empower the discovery and interpretation of combinations of genetic variations based on prior immunological knowledge. In this work, we present Biofilter 3.0 which contains enhanced annotation and filtering capabilities for empowering immunogenetic analyses. Previous versions of Biofilter provided a convenient single interface for accessing multiple publicly available human biological databases stored within the supporting knowledgebase of the Library of Knowledge Integration (LOKI). We integrated additional public databases of autoimmune and immune-mediated information into LOKI to bolster existing information on genomic locations of SNPs, genes, ontological categories, and interaction pairs. Furthermore, we expanded the functionality of Biofilter to make it possible to leverage this prior knowledge in analyses. This approach yields a collection of functions which enable researchers to efficiently annotate, subset, and filter thousands of SNPs, genes, proteins, and genomic locations based on immunological criteria, including relevant immune terms, pathways, and/or diseases.

141

Investigation on the Genetic Landscape of *MT-CO3* in Anemic Patients of Tribal Population in the State of Tamil Nadu, India

Dhivya Venkatesan[1], Nimmisha Eruppakotte[2], Harsha Ganesan[2], Balachandar Vellingiri [2,3,*]

[1]Department of Biotechnology, Karpagam Academy of Higher Education (Deemed to be University), Coimbatore 641021, India; [2]Human Molecular Cytogenetics and Stem Cell Laboratory, Department of Human Genetics and Molecular Biology, Bharathiar University, Coimbatore – 641 046, Tamil Nadu, India; [3]Stem cell and Regenerative Medicine/Translational Research, North block, Department of Zoology, School of Basic Sciences, Central University of Punjab, Bathinda, Punjab 151401, India

Anaemia is a blood disorder where oxygen-carrying capability becomes inadequate to meet the physiologic desires of the body. Low haemoglobin (Hb) concentration or iron deficiency is the most common reason for anaemia. The present study aims to screen *MT-CO3* alterations in anaemic patients of tribal population in the state of Tamil Nadu, India. Institutional ethical clearance was approved to conduct this study. Also, informed consent was obtained from the tribal anaemic patients. The study included n=30 anaemic subjects with an equal number of healthy controls. The study conducted haematological analysis based on that the severity of anaemia was evaluated as mild, moderate, and severe. From the analysis, we found that the parameters concentrations reported with 66.67% with mild anaemic, 26.67% were moderately anaemic, and 6.66% were severely anaemic. The genetic alterations of *MT-CO3* were performed by PCR and sequencing in which the finding T/C transition mutation in nucleotide position 9540 in 19 anaemic patients with significancy at p=0.0016 was observed. The controls resulted with normal haematological reports and no genetic alterations.

In conclusion, the current study delivers the importance of anaemia among Tamil Nadu tribal population, emphasizing the prevalence and mitochondrial DNA alterations in anaemic patients. Though anaemia prevalence is high, the need for genetic alterations is less in India, hence, in the future, the need for advanced molecular techniques is necessary to understand the genetic pattern of anaemia in the Tamil Nadu population.

Keywords: Anaemia; Haematology; Genetic alterations; *MT-CO3*; Tribal population

142

Leveraging Polygenic Scores to Reveal the Interplay of Serum Bilirubin, Smoking, and Cancer Risk in a Diverse Los Angeles Biobank

Vidhya Venkateswaran[1,2*], Ella Petter[1,8], Kristin Boulier[3,7], Yi Ding[1,3], Arjun Bhattacharya[9†], Bogdan Pasaniuc[1,3-6†]

[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America; [2]Department of Oral Biology, School of Dentistry, University of California, Los Angeles, Los Angeles, California, United States of America;[3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, California, United States of America;[4]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America; [5]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America; [6]Institute of Precision Health, University of California, Los Angeles, Los Angeles, California, United States of America; [7]Department of Medicine, Division of Cardiology, University of California, Los Angeles, Los Angeles, California, United States of America; [8]Department of Computer Science, University of California, Los Angeles, Los Angeles, California, United States of America; [9]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

**Corresponding author: vvenkat@g.ucla.edu,*

†Equal contribution

Recent studies identify serum bilirubin (SB) as a metabolic hormone with potent antioxidant effects, linking low SB levels with cancers, and metabolic and cardiovascular diseases. Additionally, tobacco smoking is reported to contribute to low SB levels. Thus, the associations between SB and cancers are theorized to be secondary to interactions with tobacco smoking. Using electronic health records data (EHR) on ~400,000 individuals and a polygenic score for SB on ~60,000 individuals, we examine the interplay of SB with head and neck cancer (HNC) and tobacco smoking within the UCLA ATLAS biobank, a diverse EHR-linked biobank with extensive de-identified phenotypic and demographic information.

We find that SB is inversely correlated with smoking (Linear coefficient: -0.02, CI [-0.0250, -0.0152]) and HNC (Linear coefficient: -0.16, CI [-0.19, -0.13]) after adjusting for age, sex, and self-identified race and ethnicity. Further, in a group of propensity score-matched HNC cases and controls (2040 cases and controls), matched on patient age, sex, smoking history, and self-identified race and ethnicity, we find a similar inverse association with bilirubin (Linear coefficient: -0.16, CI [-0.18, -0.12])

Lastly, we used a polygenic score (PGS) for 'total bilirubin' from the PGS catalog (PGS002160) as a genetic fixed point to identify the direction of the associations between SB, smoking, and HNC. We imputed the PGS in European genetic ancestry individuals in ATLAS and validated the predictive ability against observed total bilirubin from the EHR (Linear coefficient: 0.22, CI [0.21, 0.23]). Next, we tested the association of the validated bilirubin PGS with smoking history and HNC respectively, finding no significant associations (Linear coefficients: -0.004, CI [-0.0116,0.0035] and 0.0002, CI [-0.0014, 0.0018], respectively)

This study is the first reported evaluation of the association between SB and HNC. Our results suggest that low SB is likely secondary to HNC or a common unidentified factor that influences control over both HNC risk and SB and is independent of smoking.

143

Rapid Prediction of COVID-19 Vaccine Effectiveness Against New Genetic Variants of SARS-CoV-2 by Genome Analysis

Lirong Cao[1,2], Jingzhi Lou[3], See Yeung Chan[1,3], Hong Zheng[1,2], Caiqi Liu[1], Shi Zhao[1,2], Qi Li[1,2], Chris Ka Pun Mok[1,4], Renee Wan Yi Chan[5,6], Marc Ka Chun Chong[1,2], William Ka Kei Wu[4,7,8], Zigui Chen[9], Eliza Lai Yi Wong[1,10], Paul Kay Sheung Chan[9,11], Benny Chung Ying Zee[1,2], Eng Kiong Yeoh[1,10], and Maggie Haitian Wang[1,2*]

[1]JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China; [2]CUHK Shenzhen Research Institute, Shenzhen, China; [3]Beth Bioinformatics Co. Ltd., Hong Kong SAR, China; [4]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China; [5]Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong SAR, China; [6]Hong Kong Hub of Paediatric Excellence, The Chinese University of Hong Kong, Hong Kong SAR, China; [7]Department of Anaesthesia and Intensive Care and Peter Hung Pain Research Institute, The Chinese University of Hong Kong, Hong Kong SAR, China; [8]State Key Laboratory of Digestive Disease, The Chinese University of Hong Kong, Hong Kong SAR, China; [9]Department of Microbiology, The Chinese University of Hong Kong, Hong Kong SAR, China; [10]Centre for Health Systems and Policy Research, The Chinese University of Hong Kong, Hong Kong SAR, China; [11]Stanley Ho Centre for Emerging Infectious Diseases, The Chinese University of Hong Kong, Hong Kong SAR, China; Corresponding Author: Maggie Haitian Wang, email: maggiew@cuhk.edu.hk, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China.

Background: Timely evaluation of the protective effects of COVID-19 vaccines against SARS-CoV-2 variants of concern (VOC) is urgently needed to inform pandemic control planning.

Method: Based on 78 vaccine efficacy or effectiveness (VE) data from 49 studies, and 1,984,241 SARS-CoV-2 sequences collected from 31 regions, we analyzed the relationship between genetic distance (GD) of circulating viruses against the vaccine strain and VE against symptomatic infection.

Result: We found that the GD of the receptor binding domain of the SARS-CoV-2 Spike protein is highly predictive of vaccine protection and accounted for 86.3% (*p*-value = 0.038) of the VE change in a vaccine platform-based mixed-effects model and 87.9% (*p*-value = 0.006) in a manufacturer-based model. We applied the VE-GD model to predict protection mediated by existing vaccines against new genetic variants and validated the results by published real world and clinical trial data, finding high concordance of predicted VEs with observed VEs. We estimated the VE against the Delta variant to be 82.8% (95% prediction interval: 68.7 – 96.0) using the mRNA vaccine platform, closely matching the reported VE of 83.0% from an observational study. Among the four sub-lineages of Omicron, the predicted VEs varied between 11.9% to 33.3%, with the highest VE predicted against BA.1, and the lowest against BA.2, using the mRNA vaccine platform.

Conclusion: The VE-GD framework enables predictions of vaccine protection in real time, and offers a rapid evaluation method against novel variants that may inform vaccine deployment and public health responses.

Reference: Cao L, Lou J, Chan SY, Zheng H, Liu C, Zhao S, Li Q, Mok CK, Chan RW, Chong MK, Wu WK. Rapid evaluation of COVID-19 vaccine effectiveness against symptomatic infection with SARS-CoV-2 variants by analysis of genetic distance. *Nature Medicine*. 2022 Aug;28(8):1715-22.

144

Shared Sex-Specific Functional Genetic Risk Factors in Self-Reported Clinical Depression and Alzheimer's Disease

Ting-Chen Wang[1,2], Adam C. Naj[3], William Bush[4], Logan Dumitrescu[1,2], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute and Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Vanderbilt Memory and Alzheimer's Center, Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [4]Department of Population & Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America

In Alzheimer's disease (AD) and self-reported clinical depression, higher disease prevalence in females have been observed. In this study, we investigated the overlapping sex-specific functional genetic architecture of depression and AD to reveal shared biological mechanisms.

We performed tissue-specific transcriptome-wide association analyses (SPrediXcan) on GWAS summary data for AD and sex-stratified depression. We identified genes overlap between AD and depression, investigating if genetically regulated expression (GReX) of these genes shows consistent effects on both phenotypes. The AD GWAS included 111,326 clinically diagnosed/proxy AD cases and 677,664 controls analyzed by Bellenguez et al. The sex-stratified depression GWAS summary data from UK Biobank included 8,166 male cases and 43,675 male controls, also 16,921 female cases and 49,020 female controls. The SPrediXcan analysis included eQTLs from 19 GTEx tissues. The AD SPrediXcan analysis yielded 3,128 significant tissue-specific associations (false discovery rate [FDR] < 0.05). The male depression SPrediXcan model resulted in 24 such associations surpassing the significant threshold (FDR < 0.05), whereas no significant gene associations were observed in females. GReX of three genes, *TMEM106B*, *PPP1R18*, and *ZSCAN9*, was associated with both AD risk and male depression. Interestingly, the predicted expression of *TMEM106B* in whole blood was observed in both datasets but with opposite directions of effect (*TMEM106B* in AD: effect size=-0.057, P_{FDR}=0.013; in male depression: effect size=0.019, P_{FDR}=0.043).

Overall, our results suggest there are genes contributing to depression and AD in a sex-specific manner. We will further investigate causal links between tissue-specific gene expressions and phenotypes and in sex-stratified AD summary statistics.

146

Impact of GWAS Meta-Analysis Heterogeneity on Polygenic Prediction Accuracy

Yuxuan Wang[1*], Ching-Ti Liu[1]

[1]*Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America*

Various PRS methods have leveraged genome-wide association study (GWAS) summary statistics that often derived from GWAS meta-analyses in a large-scale consortium to maximize statistical power. However, the impact of heterogeneity in effect sizes among cohorts on PRS predictive accuracy remains unclear.

We conducted simulations to investigate the implication of heterogeneity on PRS performance. We simulated 20 independent cohorts (n=10,000) for each ancestry (AFR, EUR). Heterogeneity was introduced for true causal effect size by varying genetic correlation levels across cohorts ($r_g = 1, 0.9, 0.6, 0.3$). We performed GWAS for each cohort and meta-analyzed ten distinct cohorts with various proportions of AFR cohorts. We applied clumping and thresholding (C+T) to the meta-GWAS summary statistics and evaluated the predictive performance with the remaining AFR samples. We re-constructed the PRS by excluding heterogeneous ($I^2 > 75$) variants.

The simulation study illustrated that the number of heterogeneous variants increased as genetic correlation decreased across cohorts. The PRS was more predictive with higher proportions of target ancestry-matched cohorts and larger cross-cohort genetic correlations. When r_g=1, excluding heterogeneous variants substantially improved prediction accuracy for PRS built from multi-ancestry meta-GWAS, but not for single-ancestry meta-GWAS. However, when $r_g < 1$, removing the heterogeneous variants reduced prediction accuracy for both single and multi-ancestry meta-GWAS.

In conclusion, our study provides insights into the impact of between-study heterogeneity on polygenic prediction and highlights the potential to enhance prediction accuracy by incorporating heterogeneity measurements.

147

Mediation Analysis with a Categorical Exposure and a Censored Mediator in Genetic Studies

Jian Wang[1]*, Jing Ning[1], and Sanjay Shete[1,2]

[1]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America;* [2]*Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*

Mediation models have been widely used to determine direct and indirect contributions of genetic variants on clinical phenotypes or complex diseases. In genetic studies, the additive genetic model is the most used model because typically the mode of action of susceptibility SNPs is unknown, and the additive model can detect effects from either recessive or dominant models, or any model in between. Moreover, the highly polymorphic loci can also be involved in the mediation model as an exposure. However, the existing approaches for mediation model cannot be directly applied when the genetic model is additive or categorical. Therefore, in this study, we proposed overall measures of indirect, direct, and total effects for a mediation model with a categorical exposure and a censored mediator, accounting for the frequency of different categories of the categorical exposure. The proposed approach provides the overall contribution of the categorical exposure to the outcome variable instead of the relative contribution comparing one category to another. We assessed the empirical performance of the proposed overall measures via simulation studies and applied the measures to evaluate the mediation model for the study of a women's age at menopause on the association between genetic variants and Type 2 diabetes.

148

Longitudinal Data Analyses in UK Biobank Identify Novel Loci Associated with Kidney Function Decline

Simon Wiegrebe[1,2]*, Mathias Gorski[1], Thomas W. Winkler[1], Janina M. Herold[1], Helmut Küchenhoff[2] and 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*

Accelerated kidney function decline can lead to kidney failure requiring renal replacement therapy. Kidney function can be assessed by serum creatinine-based estimated glomerular filtration rate (eGFR). Cross-sectional studies have identified 634 independent variants modulating eGFR levels. However, knowledge on the genetics of eGFR *decline* is limited: nine variants (two in/near *UMOD-PDILT* with genome-wide significance) have been found to be associated with the difference of two eGFR assessments over time by previous meta-analyses (n=350,000).

We investigated the detectability of variants for eGFR decline by testing the 634 variants with various linear mixed models (LMMs) adjusted for age, sex, 20 PCs. For this, we used longitudinal data from UK Biobank augmented with electronic medical records (≥2 measurements for each of n=152,821, European ancestry, unrelated; #measurements=1,356,851).

Via LMM with SNP*time interaction and random intercepts, we found 112 variants associated with eGFR decline with p<0.05/634 (seven of nine known, 105 novel). Interestingly, 30 variants were even genome-wide significant (28 for the first time and the two *UMOD-PDILT* variants). Four were protein-altering (one causal for renal dysfunction: rs3184504, *SH2B3*).

The LMM with SNP*age interaction yielded largely the same variants and highly correlated beta estimates (Spearman r^2=0.99 for 82 variants Bonferroni-corrected significant in both models). When adding random slopes to the SNP*time model, only two variants were Bonferroni-corrected significant (*UMOD-PDILT*), and overall larger SEs suggested lower power.

In summary, we identified new variants and loci for eGFR decline by applying LMMs to UK Biobank eGFR trajectories. Our results underline the power gain from including repeated measurements.

149

Genetic-by-age Interaction Analyses in UK Biobank and Their Potential to Identify Genetic Effects on Longitudinal Biomarker Change

Thomas W. Winkler[1*], Simon Wiegrebe[1,2], Mathias Gorski[1], Helmut Küchenhoff[2], Iris M. Heid[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Franz-Josef-Strauß-Allee 11, 93053 Regensburg, Germany; [2]Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

GWAS have identified thousands of loci for disease-related biomarkers in cross-sectional data. However, identifying genetic effects on longitudinal biomarker change has been hampered by small sample sizes for longitudinal measurements. Such effects, indicating genetic susceptibility to disease progression, can be highly clinically relevant. Under the assumption of no cohort effect, we demonstrate that genetic-by-age interaction observed in cross-sectional data can be indicative of a genetic effect on longitudinal-change that would be observed if the same individuals attended multiple visits. We show that exploiting large cross-sectional sample sizes to pre-screen for genetic-by-age interaction can greatly enhance the power for identifying longitudinal-change effects. Within UK Biobank (UKB), we conducted genome-wide genetic-by-age interaction analyses (Europeans-only, n~370,000, excluding individuals with multiple visits) for nine biomarkers: body-mass-index (BMI), estimated glomerular filtration rate (eGFR), urate, HDL-/LDL-cholesterol, triglycerides, systolic/diastolic blood pressure, and pulse pressure (PP). We identified 76 significant genetic-by-age interaction loci ($P_{GxAge}<5x10^{-8}$; or by a 2-step approach incorporating marginal effects) including 32 for PP, 6 for BMI and 15 for eGFR. We tested these loci for their longitudinal-change effects in independent individuals with longitudinal measurements available (n~52,000 in UKB, n~340,000 for eGFR in CKDGen) and observed significant effects on BMI-change (near *APOE*, *TMEM18*), eGFR-change (near *PDILT*, *TPPP* and *FGF5*) and PP-change (near *FBN1,* implicated in arterial stiffness; all at trait-level Bonferroni-significance; all missed by a screen on longitudinal-change in UKB alone). In conclusion, cross-sectional genetic-by-age interaction can help pinpoint longitudinal-change effects, when the cross-sectional sample size and thus power outnumbers the longitudinal sample size available.

150

The Contribution of the Proteome in Type 2 Diabetes and Osteoarthritis

Ruby Woodward[1], Bethany Voller[2], Gina Parcesepe[1,3] and Frank Dudbridge[1]

[1]University of Leicester, Leicester, United Kingdom [2]Univeristy of Exeter, Exeter, United Kingdom [3]NIHR Leicester Biomedical Research Centre, Leicester, United Kingdom

Multimorbidity (the presence of more than one disease) is on the rise with an increasingly older population. Some clusters of diseases are more common than others, one example supported by observational and genetic data being the co-occurrence of Type 2 diabetes (T2D) and osteoarthritis (OA). Although complex interactions between diseases are often poorly understood, in the case of T2D and OA, BMI is considered the major shared modifiable risk factor; despite this, other shared biological mechanisms (perhaps altered through high BMI) could improve knowledge of disease initiation, progression, and severity.

Circulating proteins play a pivotal role in disease. Proteomic data is crucial for discovering causal mechanisms involved in diseases and shared pathways between diseases. Using publicly available proteomic data from deCODE genetics for ~5000 proteins, we performed LD-Score Regression to assess the genome-wide genetic correlation between all proteins with both T2D and OA. Following this, local genetic correlation was measured for 126 proteins that were genetically correlated (p<0.05) with both diseases using Local Analysis of [co]Variant Association (LAVA).

Our initial results suggest that local genetic correlation is dispersed throughout the entire genome, with no clear differences between cis and trans regions. This provides evidence that some polygenic effects on disease may be explained by their action on protein levels.

151

Pathway-level Rare Variant Burden Scores Aggregate the Effects of Multiple Genes and Associate with Hypertension in the Penn Medicine BioBank

Brenda Xiao[1], Marylyn D. Ritchie[2,3], Dokyoon Kim[3,4]

[1]Graduate Program in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [3]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [4] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Gene burden rare variants analyses have identified numerous genes associated with various phenotypes. However, the low number of rare variant carriers limits power to detect associations for many complex diseases. Grouping similar genes together by analyzing the effects of rare variants at the pathway level could identify candidate pathways that are involved with the development of complex diseases such as hypertension. We performed a gene burden loss-of-function rare variant analysis using SAIGE-GENE+ to identify genes associated with hypertension in European ancestry individuals in the Penn Medicine BioBank (PMBB). We then selected genes that passed a *P* value significance threshold of 0.01. We summed up carrier status for significant genes in each Reactome pathway to represent a pathway-level burden, weighting carrier status per gene by the direction of effect obtained from the gene burden analysis, and tested the association of each pathway's burden on hypertension. *PKD1* was the most significant gene association (P=7.65x10^{-6}), and Reactome pathway *R-HSA-5620916* (VxPx cargo-targeting to cilium) was the most significant pathway association (P=8.45x10^{-9}). Three of the twenty-one genes in this pathway were included in the pathway-level burden, including *PKD1* and *PKD2*. We then used publicly available gene-burden association results from the UK Biobank for hypertension to generate pathway-level burden scores in PMBB to predict hypertension. *R-HSA-5620916* remained the most associated with hypertension, even after addition of a systolic blood pressure polygenic risk score (PRS) (OR=3.87, P=1.32x10^{-6}), and its effect was higher in low PRS groups (<40 PRS percentile group: OR=5.74; >60 PRS percentile group: OR=3.71).

152

A Statistical Learning Method for Simultaneous Copy Number Estimation and Subclone Clustering with Single Cell Sequencing Data

Fei Qin[1], Guoshuai Cai[2], Christopher I Amos[3], Feifei Xiao[4*]

[1]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America; [2]Department of Environmental Health Science, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America; [3]Department of Quantitative Sciences, Baylor College of Medicine, Houston, Texas, United States of America; [4]Department of Biostatistics, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, Florida, United States of America;

** To whom correspondence should be addressed*

The availability of single cell sequencing (SCS) enables us to assess intra-tumor heterogeneity and identify cellular subclones without the confounding effect of mixed cells. Copy number aberrations (CNAs) have been commonly used to identify subclones in SCS data using various clustering methods, since cells comprising a subpopulation are found to share genetic profile. However, currently available methods may generate spurious results (e.g., falsely identified CNAs) in the procedure of CNA detection, hence diminishing the accuracy of subclone identification from a large complex cell population. In this study, we developed a CNA detection method based on a fused lasso model, referred to as FLCNA, which can simultaneously identify subclones in single cell DNA sequencing (scDNA-seq) data. Spike-in simulations were conducted to evaluate the clustering and CNA detection performance of FLCNA benchmarking to existing copy number estimation methods (SCOPE, HMMcopy) in combination with the existing and commonly used clustering methods. Interestingly, application of FLCNA to a real scDNA-seq dataset of breast cancer revealed remarkably different genomic variation patterns in neoadjuvant chemotherapy treated samples and pre-treated samples. We show that FLCNA is a practical and powerful method in subclone identification and CNA detection with scDNA-seq data.

153

Tensor Decomposition of Multi-dimensional Splicing Events across Tissues to Identify Splicing-mediated Risk Genes Associated with Complex Traits

Yan Yan[1*], Rui Chen[2,3], Yuting Tan[2,3], Anshul Tiwari[2,3], Xue Zhong[2,4], Bingshan Li[2,3]

[1]Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Nashville, Tennessee, United States of America; [3]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee, United States of America; [4]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Identifying risk genes associated with complex traits remains challenging. Integration of gene expression data with Genome-Wide Association Study (GWAS) through Transcriptome-Wide Association Study (TWAS) methods has been applied on a variety of complex traits. However, splicing, which explains a comparable heritability of complex traits as gene expression, has not been fully explored due to the multidimensional nature of splicing events. We employed tensor decomposition in conjunction with sCCA (sparse Canonical Correlation Analysis) to extract meaningful information from high-dimensional multiple splicing events across tissues. Leveraging GTEx data, we developed gene-based splicing predictive models and applied them to GWAS summary statistics of Alzheimer's disease (AD). Our analysis identified 174 significant risk genes after applying Bonferroni correction. Gene Ontology analysis revealed a significant enrichment of AD-related functions, such as amyloid-beta-related pathways, endocytosis, and immunity functions. Compared to the models using only the brain frontal cortex tissue, our results demonstrated substantial enrichment of AD related pathways, and identified additional AD risk genes that were not detected in the brain tissue analysis alone, while preserving most of the top genes identified in brain tissue. Given that AD genetics primarily involves microglia, which constitute only a small proportion of brain cells, relying solely on transcriptomics data from brain tissues may not capture the full genetic landscape of AD. Our across-tissue modeling approach allows us to extract splicing information relevant to AD for more comprehensive risk gene discovery. Moreover, these splicing models can be applied to other complex traits to help identify splicing-mediated disease risk genes.

154

AI-Enhanced Integration of Genetic and Medical Imaging Data for Risk Assessment of Type 2 Diabetes

Yi-Jia Huang[1], Chun-houh Chen[2], Hsin-Chou Yang[1,2,*]

[1]Institute of Public Health, National Yang Ming Chiao Tung University, Taipei, Taiwan; [2]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

*Corresponding author

Type 2 diabetes (T2D) is a global public health concern due to its increasing prevalence. Risk assessment and early detection of T2D are vital in improving individuals' health, reducing the burden on national health insurance and enhancing well-being. This study leverages artificial intelligence, specifically eXtreme Gradient Boosting (XGBoost), to develop predictive models for T2D based on genetic and medical imaging data. The models aim to establish a prediction model and identify high-risk subgroups for T2D within a cohort of 68,769 Taiwan Biobank participants.

The approach integrates the Polygenic Risk Score (PRS) and Multi-image Risk Score (MRS) with demographic factors and environmental exposures to assess T2D risk. The model's performance is evaluated using the Area Under the Receiver Operating Curve (AUC). Results demonstrate that genetic information alone is insufficient for accurate T2D prediction (AUC = 0.73), whereas medical imaging data, including abdominal ultrasonography, vertebral artery ultrasonography, bone density scan, and electrocardiography, significantly improves prediction accuracy (AUC = 0.89). The best-performing model integrates genetic, medical imaging, and demographic variables (AUC = 0.94), successfully identifying subgroups at high risk of developing T2D. The study also presents an online risk assessment website for T2D. In summary, this research represents the first integration of whole-genome and medical imaging data for T2D risk assessment. The genetic-only model outperforms previous genetic prediction studies, and integrating genetic and medical imaging information significantly enhances AUC. By utilizing artificial intelligence to analyze genetic, medical imaging, and demographic factors, this study contributes to the early detection and precision health of T2D.

155

Joint Selection of Exposures and Horizontal Pleiotropy in Multivariable Mendelian Randomization with Application to Causal Gene Identification

Yihe Yang, Noah Lorincz-Comi, Xiaofeng Zhu

Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America

Background: Jointly analyzing genome-wide association studies (GWAS) data and expression quantitative trait loci (eQTL) data will potentially identify causal genes of diseases. However, current univariable methods such as transcriptome-wide association studies (TWAS) and univariable Mendelian randomization (UVMR) may not be robust due to high correlations among gene expressions, horizontal pleiotropy, and weak instrument bias in MR analysis.

Methods: We propose a novel multivariable method named MR Joint Outliers-aNd-Exposures Selection (Mr.Jones), which performs multivariable MR (MVMR) analysis in a genome region using multiple gene expressions as exposures. Mr.Jones applies the unbiased estimating function to mitigate weak instrument bias, employs variable selection penalties to select causal genes and identify horizontal pleiotropy simultaneously, and decorrelates the instrument variables using their linkage disequilibrium (LD) matrix.

Results: In simulations, Mr.Jones resulted in unbiased causal effects estimates in the presence of horizontal pleiotropy and many weak instrument variables compared with current MR methods. We applied Mr.Jones to search causal genes for coronary artery disease (CAD), Type 2 diabetes (T2D), and osteoarthritis (OA) in European populations, identifying *TCF7L2* as a protective gene for T2D, *MTAP* as a risk gene for CAD, and *SLC25A13* as a risk gene for OA.

Conclusion: As more GWAS and eQTL data become publicly available, Mr.Jones can serve as a valuable tool in studying causal relationships between gene expressions and diseases, therefore enabling a deeper understanding of disease mechanisms, facilitating precision medicine, and guiding drug development.

156

Use of Genetic Correlations to Examine Selection Bias

Chin Yang Shapland, Apostolos Gkatzionis, Gibran Hemani, Kate Tilling

MRC Integrative Epidemiology Unit at the University of Bristol, United Kingdom

Population Health Sciences at the University of Bristol, United Kingdom

Observational studies are rarely representative of their target population, because there are known and unknown factors that affect an individual's choice to participate (known as the selection mechanism). Selection can cause bias in a given analysis if the outcome is related to selection (conditional on the other variables in the model). However, the selection mechanism usually cannot be detected from the observed data if we have no data on the non-selected sample - for example, when the selected sample is a participant in a research study. Here, we develop methods to examine the selection mechanism by comparing correlations among variables in the selected sample to those expected under no selection. We examine the use of four hypothesis tests to identify induced associations between genetic variants in the selected sample. We evaluate these approaches with Monte Carlo simulations. Finally, these approaches are demonstrated with an applied example, using data from UK Biobank (UKBB), with alcohol intake as exposure to test the presence of selection bias. The proposed tests have identified selection due to alcohol intake into UKBB and the subsample of individuals with weekly alcohol intake. Analyses in UKBB with alcohol consumption as exposure or outcome may be biased by this selection.

157

Learning Portable Polygenic Risk Score Models with Mixtures of Pre-trained Experts to Improve Accuracy across the Continuum of Ancestry

Shadi Zabad[1], Simon Gravel[2*], Yue Li[1*]

[1]School of Computer Science, McGill University, Montreal, Quebec Canada; [2]Department of Human Genetics, McGill University, Montreal, Quebec, Canada

**Correspondence: yueli@cs.mcgill.ca, simon.gravel@mcgill.ca*

In recent years, there has been growing interest in incorporating polygenic scores (PRS) into clinical practice and drug development pipelines. One challenge for realizing the potential of PRS methods, however, is their poor portability to out-of-sample individuals and disparity in prediction accuracy across ancestries. To address this problem, state-of-the-art PRS methods aim to infer ancestry-specific variant effect size estimates. While these methods have been shown to improve prediction accuracy for underrepresented populations, significant challenges remain. To mitigate the PRS portability problem, we propose a Mixture-of-Experts (MoE) modeling framework, which can accommodate heterogeneity of effect sizes and automatically specialize individual experts on various partitions of the data. MoEs consist of an ensemble of K PRS models whose outputs are combined on a per-sample basis using a "gating" model. The gating model takes as input a list of covariates for each sample, such as Principal Components, age, and sex and outputs probabilistic weights for combining the predictions of the PRS models in the ensemble. To showcase the utility of this framework, we illustrate how it can be used to combine pre-trained PRS models from the PGS Catalog. In a 5-fold cross-validation analysis in the UK Biobank and CARTaGENE datasets, we show that MoEs significantly improve prediction accuracy over individual models, with mean improvements of up to 12% in admixed samples. It also produces scores that perform more consistently across a wide range of sociodemographic profiles. These benefits are achieved without imposing any assumptions or arbitrary subdivisions on the data at training time.

158

Identifying the Genetic Etiology of Selected Muscular Dystrophies in Muscular Dystrophy Surveillance, Tracking, and Research Network (MD STAR*net*)

Peter B. Kang, MD[1]; Magali Jorand-Fletcher, MPH[2]; Wanfang Zhang, MS[3]; Suzanne McDermott, PhD[4]; Reba Berry, RN[5]; Chelsea Chambers, MS, LCGC[6]; Kristen Wong, CGC[7]; Kristin M. Conway, PhD[8]; Yara Mohamed, MD[2]; Shiny Thomas, MBBS, MPH[9]; Swamy Venkatesh, MD[10]; Christina Westfield, BSN, MS[9]; Nedra Whitehead, PhD[11]; Nicholas E. Johnson, MD[12]; Muscular Dystrophy Surveillance, Tracking, and Research Network (MD STAR*net*)

[1]Paul & Sheila Wellstone Muscular Dystrophy Center, Department of Neurology, and Institute for Translational Neuroscience, University of Minnesota Medical School, Minneapolis, Minnesota, United States of America; [2]Department of Pediatrics, University of Florida College of Medicine, Gainesville, Florida, United States of America; [3]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, United States of America; [4]Department of Environmental, Occupational, and Geospatial Health Sciences, Graduate School of Public Health & Health Policy, City University of New York, New York, New York, United States of America; [5]Division of Population Health Surveillance, Bureau of Maternal and Child Health, South Carolina Department of Health & Environmental Control, Columbia, South Carolina, United States of America; [6]Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, United States of America; [7]Department of Pediatrics, University of Utah, Salt Lake City, Utah, United States of America; [8]Department of Epidemiology, College of Public Health, University of Iowa, Iowa City, Iowa, United States of America; [9]New York State Department of Health, Albany, New York, United States of America; [10]Department of Neurology, University of South Carolina, Columbia, South Carolina, United States of America; [11]RTI International, Research Triangle Park, North Carolina, United States of America; [12]Department of Neurology, Virginia Commonwealth University, Richmond, VA

Muscular dystrophies are a group of inherited muscle diseases. This study investigated the genetic mutations present in a cohort of 243 individuals diagnosed with four types of muscular dystrophy: Emery-Dreifuss (EDMD, 21 cases), limb-girdle (LGMD, 138 cases), congenital (CMD, 62 cases), and distal (DD, 22 cases), who received medical care between 2008 to 2016 in six regions encompassed by the Muscular Dystrophy Surveillance, Tracking, and Research Network (MD STARnet). Clinical data abstracted from medical records were examined for certainty of diagnosis, demographics, genetic test findings. The findings were reviewed and whenever possible, variants of unknown significance (VUSs) were reclassified utilizing updated information from genetic databases.

We reviewed 144 VUSs from 97 individuals using accepted standard classifications systems and databases. Multiple genes were reported from some individuals. We successfully resolved 60 VUSs to more definitive interpretations, while 84 VUSs remained unchanged. The implications of this research are significant for drug development and the initiation of clinical trials. The attainment of definitive genetic diagnosis not only facilitates timely detection of muscular dystrophy in family members, but also empowers individuals to make informed reproductive decisions.

Additionally, we expected to find a balanced proportion of males and females with these muscular dystrophy subtypes; however, substantially more males (64.6%) were identified, even after excluding diseases caused by the X-linked gene EMD. This study highlights the importance of replicating and revisiting existing data in research, as well as ensuring that outreach efforts emphasize that various muscular dystrophy subtypes can affect both genders.

159

Investigating Ancestry-specific Genetic Variation in Apolipoprotein L Genes Associated with Electronic Health Record Phenotypes in Patient Biobanks

David Y. Zhang[1,2] *, Michael G. Levin[2,3,4], Scott M. Damrauer[4,5], Marylyn D. Ritchie[1], Daniel J. Rader[1,2]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, [2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, United States of America; [3]Division of Cardiovascular Medicine, Perelman School of Medicine, University of Pennsylvania, United States of America; [4]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania, United States of America; [5]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, United States of America

Health care disparities between people of different ancestries and ethnicities are well-documented in every field of medicine. Of the ~6,401 studies currently compiled in the genome-wide association studies (GWAS) catalog as of June 2023, ~95% of all GWAS participants are of European (EUR) ancestry with less than 1% of participants being of African-American (AFR) ancestry. Using the Penn Medicine Biobank (PMBB) and adopting a genome-first approach, we investigated 100 predicted loss-of-function (pLOF) and 737 missense variants in the apolipoprotein L gene family with a specific interest in those more common in non-European populations. We performed phenome-wide association studies (PheWAS) on 62 variants with a MAF > 0.1% in the PMBB AFR population (n = 11,198) against 1,236 binary phenotypes derived from electronic health records data with at least 20 cases. Our results identified a stop-gain variant rs11089781 (p.Gln58*) in the *APOL3* gene found to be significantly associated with increased risk for end-stage renal disease (ESRD) (OR = 1.38, p = 3.64e-08). This variant has a gnomAD minor allele frequency of 0.22 in AFR compared to 3.97e-04 in EUR. It is also in linkage equilibrium ($r^2 < 0.05$) with the *APOL1* G1 and G2 known risk alleles for renal disease. Replication of this association in up to 121,790 AFR individuals from the Million Veterans Program also yielded a significant association with ESRD (OR = 1.16, p = 1.01e-08). Initial hypotheses suggest that *APOL3* may play a protective role against *APOL1* and loss-of-function in *APOL3* increases susceptibility to *APOL1*-induced kidney dysfunction.

160

CellGRN: An Improved Identification Method for Gene Regulatory Networks

Jane Zhao[1], Mingyao Li[2]

[1]School for Advanced Studies - South; Miami Dade College - Kendall, Miami, Florida, United States of America; [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Identifying gene regulatory networks (GRNs) correctly is vital in advancing our understanding of disease mechanisms, drug discovery, and personalized medicine. However, it is a challenging task due to the complexity and dynamic nature of gene regulation. CellNet is a popular and useful tool in identification of GRNs but with a major limitation by using the average gene expression across cells in a tissue and ignoring possible cell-to-cell heterogeneity. That is, CellNet defines the status of a GRN in a query sample as weighted mean of Z scores of genes in the GRN where the absolute gene expression level serves as the weight. Ignoring cell-to-cell heterogeneity can obscure important regulatory mechanisms and functional differences between cells, mask biologically relevant variations, and fail to capture important regulatory relationships. To overcome this limitation, we propose cell-GRN which defines the GRN status score as summation of GRN status scores of all genes in that GRN, while each gene's GRN status score is the summation of cell specific Z scores estimated by cell proportion weighted gene expressions. Cell proportions can be estimated with MuSiC2 software. To assess the performance of Cell-GRN, we applied both Cell-GRN and CellNet to a iPSC fetal liver organoid (FeLO), and designer liver organoids (DesLO) RNA-seq data (GSE159491). Compared with CellNet, Cell-GRN has an increased liver classification score (mean=0.54 of Cell-GRN vs. 0.46 of CellNet). Further analysis with two RNA-seq datasets: 1) 12 autism patients and 12 controls (GSE64018); 2) cerebral organoids differentiated from nine donors with 16p11.2 deletions versus 12 control donors (GSE200851), demonstrated that Cell-GRN can identify more autism associated GRNs than CellNet in both datasets.

161

Using Longitudinal EHRs to Identify Medical Conditions Enriched in Individuals who Later Received Diagnosis of Alzheimer's Disease

Xue Zhong[1,2#], Gengjie Jia[3], Zhijun Yin[4,5], Kerou Chen[5], Andrey Rzhetsky[3,6,7], Nancy J. Cox[1,2]

[1]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Nashville, Tennessee, United States of America [3]Department of Medicine, Institute of Genomics and Systems Biology, Committee on Genomics, Genetics, and Systems Biology, University of Chicago, Chicago, Illinois, United States of America; [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Department of Computer Science, Vanderbilt University, Nashville, Tennessee, United States of America; [6]Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America; [7]Committee on Genomics, Genetics and Systems Biology, University of Chicago, Chicago, Illinois, United States of America

#Corresponding author: xue.zhong@vumc.org

Several mid-life health conditions have been linked to an increased risk of Alzheimer's disease (AD) in later life. Here, we seek to comprehensively identify medical conditions that overrepresent in individuals who later develop AD and assess the genetic basis of these conditions. We use two electronic health records (EHR) datasets (>150 million individuals) for discovery and replication purpose. Using ICD diagnosis codes, we identify AD cases and age- and gender-matched controls at 1:10 ratio, then trace back 10 years in their EHRs and compare between the two groups their medical conditions during the 10-year window. We identify ~43,500 and ~1,300 AD cases, respectively, in MarketScan (MS); in both datasets, 80% of the AD cases have their first diagnosis of AD at age 75 years or older. Comparison of the medical profiles between the AD cases and matched controls, reveals 406 and 102 enriched phenotypes associated with AD status in MS and VUMC, respectively. Mental disorders and neurological symptoms dominate the enriched phenotypes in both datasests. To further identify "causal" phenotypes of AD, we perform PheWAS of AD risk variants in BioVU and UK Biobank. More than 20 of the enriched phenotypes, including *hyperlipidemia*, *hypertension*, *cerebral ischemia*, *memory loss*, and *mild cognitive impairment* etc., are significantly associated with AD risk variants or polygenic risk scores. In conclusion, longitudinal EHRs from millions of individuals enables a comprehensive detection of medical conditions enriched in future AD cases, and a small portion of the phenotypes show evidence for a causal link to AD.

Key words: Alzheimer's disease, PheWAS, EHR, electronic medical records

162

A Multiomics Machine Learning Approach to Characterize Genetic Architecture and Map Traits of Plasma Lipidome in Hispanics/Latinos

W. Zhu[1*], A. Petty[1], L. Petty[1], J. Curran[2], P. Meikle[3,4], J. McCormick[5], S. Fisher-Hoch[5], K. North[6], E. Gamazon[1,7], J. Below[1]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, School of Medicine, Brownsville, Texas, United States of America; [3]Metabolomics Laboratory, Baker Heart and Diabetes Institute, Melbourne, State of Victoria, Australia; [4]Baker Department of Cardiometabolic Health, University of Melbourne, Parkville, State of Victoria, Australia; [5]Department of Epidemiology, Human Genetics and Environmental Sciences, The University of Texas Health Science Center at Houston School of Public Health, Brownsville Regional Campus, Brownsville, Texas, United States of America;[6]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [7]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom

Traditional lipid measures such as HDL and LDL are accessible in many cohorts; however, they are insufficient to characterize the dynamic and subtle changes of lipid species and classes. Alternatively, lipidomics provides a complete view of the underlying lipid metabolism within individuals, thus it has become a valuable resource in recent genetic studies. Although large GWAS studies in Caucasians discovered hundreds of lipid-related variants, these findings may not capture genetic risk factors in minority groups. Similarly, prediction models trained on Caucasian dataset tend to have poor translation across populations.

Our study aims to address existing issues by utilizing lipidomics and genotype data in Cameron County Hispanic Cohort. We performed lipidome wide GWASs in 2289 individual across 49 lipid classes and 830 lipid species, using linear mixed models adjusted for sex, age, age2, BMI, PCs, and relatedness. We constructed a maximum independent set (n=1680) as the training set for prediction models. Prior to model fitting, lipid traits were normalized for sex, age and PCs, and variants were filtered by GWAS P values. Elastic net regressions were run with 10-fold cross validation and grid search for model selection. Remaining samples not involved in training were used to evaluate model performance. Our GWASs identified approximately 9,000 significant variants for 830 lipid species from 49 lipid classes across most chromosomes. We identified novel signals close to genes such as BASP1, PACSIN2, and MCAT, which have been linked to lipid binding and metabolism. The performances of our prediction models aligned with SNP based heritability.

163

Genetic Factors for Differentiated Thyroid Cancer in French Polynesia: New Candidate Loci

Monia Zidane[1], Marc Haber[2,3], Thérèse Truong[4], Frédérique Rachédi[5], Catherine Ory[6], Sylvie Chevillard[6], Hélène Blanché[7], Robert Olaso[8], Anne Boland[8], Éric Conte[9], Mojgan Karimi[4], Yan Ren[1], Constance Xhaard[10], Vincent Souchard[1], Jacques Gardon[11], Marc Taquet[12], André Bouville[13], Jean-François Deleuze[7,8], Vladimir Drozdovitch[14], Florent de Vathaire[1], Jean-Baptiste Cazier[2,3]

[1]University Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Team "Radiations Epidemiology", Villejuif, France; [2]Centre for Computational Biology, University of Birmingham, Birmingham, United Kingdom; [3]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, United Kingdom; [4]University Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Team "Exposome and Heredity", Villejuif, France; [5]Endocrinology Unit, Territorial Hospital Taaone, Papeete, Tahiti, French Polynesia; [6]CEA, Laboratoire de Cancérologie Fondamentale, Institut de Biologie François Jacob, iRCM, SREIT, Laboratoire de Cancérologie Expérimentale (LCE), Université Paris-Saclay, France; [7]Fondation Jean Dausset-Centre d'Etude du Polymorphisme Humain, Paris, France; [8]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, Evry, France; [9]U.S.R. 2003 (CNRS / UPF), Faa'a, Tahiti, France; [10]University of Lorraine, INSERM CIC 1433, Nancy CHRU, INSERM U1116, Nancy, France; [11]Hydrosciences Montpellier, Research Institute for Development, CNRS, University of Montpellier, Montpellier, France; [12]Research Institute for Development, Center IRD on Tahiti, Arue, Tahiti, French Polynesia, [13]National Cancer Institute (retired), Bethesda, Maryland, United States of America; [14]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, Maryland, United States of America

Presenting author: Monia Zidane, INSERM U1018, 114 rue Edouard Vaillant, 94805 Villejuif Cedex, France, monia.zidane@gustaveroussy.fr

Corresponding author: Florent de Vathaire, Cancer and Radiation, INSERM U1018, 114 rue Edouard Vaillant, 94805 Villejuif Cedex, France, florent.devathaire@gustaveroussy.fr

Context: Populations of French Polynesia (FP), where France performed atmospheric tests between 1966 and 1974, experience a high incidence of differentiated thyroid cancer (DTC). However, up to now, no sufficiently large study of DTC genetic factors in this population has been performed to reach a definitive conclusion.

Objective: To analyze the genetic factors of DTC risk among the native FP populations.

Methods: We analyzed more than 300,000 single nucleotide polymorphisms (SNPs) genotyped in 283 DTC cases and 418 matched controls born in FP, most being younger than 15 years old at the time of the first nuclear tests. We analyzed the genetic profile of our cohort to identify population subgroups. We then completed a genome-wide analysis study on the whole population.

Results: We identified a specific genetic structure in the FP population reflecting admixture from Asian and European populations. We identified three regions associated with increased DTC risk at 6q24.3, 10p12.2, and 17q21.32. The lead SNPs at these loci show respective p values of 1.66×10^{-7}, 2.39×10^{-7} and 7.19×10^{-7} and corresponding odds ratios of 2.02, 1.89, and 2.37.

Conclusion: Our study results suggest the role of the loci 6q24.3, 10p12.2 and 17q21.32 in DTC risk. However, a whole genome sequencing approach would be better suited to characterize these factors than genotyping with microarray chip designed for the Caucasian population. Moreover, the functional impact of these three new loci needs to be further explored and validated.

164

Robust Rare Variant Association Testing for Skewed Traits: Application to Model-based Disease Predictions in the UK Biobank

Andrey Ziyatdinov[1], Joseph Herman[1], Joelle Mbatchou[1], Adrian Campos[1], Carlo Sidore[1], Manuel Ferreira[1], Jonathan Marchini[1]

[1]Regeneron Genetics Center, Tarrytown, New York, United States of America

Machine learning derived traits can enhance the performance of genome-wide association studies. However, these traits often exhibit extreme skewness, that leads to the Type I inflation at rare variants and hinders novel discoveries from sequencing data. Continuous disease predictions can look like unbalanced categories and make widely used rank-based trait transformations ineffective. To address this issue, we developed an extension to the Moment-Corrected Correlation (MCC) test and integrated it into REGENIE.

MCC does not rely on the normal approximation for the test statistic r, the Pearson correlation between trait and tested variable. Instead, MCC analytically approximates the distribution of permuted r using the moment-matching approach and a Beta distribution. However, previous implementations of MCC fail when testing rare variants at large sample sizes. Here, we use r^2

to construct our test statistic and a shifted Gamma distribution for approximation. The REGENIE whole-genome regression model accounts for relatedness, population structure, and polygenic effects.

We applied this MCC test to exome-wide association studies in the UK Biobank for several traits, including eye-image-derived class predictions for Age-related Macular Degeneration (AMD), liability-threshold family-history scores for Alzheimer's Disease (AD), the number of hospital visits averaged over 10 years for Chronic Obstructive Pulmonary Disease cases. For each trait, the rare-variant test statistic was severely inflated using the standard score test (even after rank-based transformations), while our MCC-based approach produced well calibrated results. MCC uncovered many rare variant associations which replicated in a larger meta-analysis of related binary traits. Examples include known rare-variant and burden associations in the PSEN1 gene for AD and in the CFI and CFH genes for AMD.