Abstracts

INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

The 2024 Annual Meeting of the International Genetic Epidemiology Society

1

Multi-FISHNET: Finding Significant Hits in Networks

S. Acharya[3], E. Kang[2], L. Liao[2], W. Jung[2], V. A. Moghaddam[1], M. Brent[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

The Bonferroni correction sets a stringent significance threshold in GWAS, necessitating large samples and effect sizes to achieve significance. We propose FISHNET, an algorithm that uses prior biological knowledge embodied in gene interaction networks and gene function annotations to identify genes that do not meet the Bonferroni threshold but replicate nonetheless. Its input is gene-level P-values obtained by omicsWAS or by aggregating SNP P-values. It is based on the idea that genes whose P-values are low due to sampling error are distributed randomly across networks and functions, so suggestive P-values that cluster in densely connected subnetworks and share common functions are unlikely to reflect sampling error and more likely to replicate. FISHNET combines network and gene set enrichment analysis with permutation-based P-value thresholds to identify a small set of exceptional genes that we call FISHNET genes.

We applied FISHNET to 23 GWAS discovery datasets and evaluated the results using 23 replication datasets. In the discovery sets, we identified 583 FISHNET genes. Of these, 268 (46%) were also FISHNET genes in the replication datasets, and 268 (46%) met Bonferroni replication threshold in replication datasets. Notably, 49/268 Bonferroni-replicated genes were not genome-wide significant in the discovery set but were genome-wide significant in replication sets. The replication rate of FISHNET genes matched or exceeded that of all genes at various statistical thresholds in the range ($p \leq 2.9\times10^{-6}$, $p \leq 2.9\times10^{\wedge-2}$). By combining summary statistics with biological knowledge, FISHNET identifies replicable gene-trait associations that fail traditional P-value thresholds.

2

A Methodology for Gene Level Omics-WAS Integration Identifies Genes Influencing Traits Associated with Cardiovascular Risks: The Long Life Family Study

Sandeep Acharya[3], Shu Liao[2], Wooseok J. Jung[2], Yu S. Kang[2], Vaha Akbary Moghaddam[1], Mary Feitosa[1], Mary Wojczynski[1], Shiow Lin[1], Jason A. Anema[1], Karen Schwander[1], Jeff O Connell[4], Michael A. Province[1], Michael R. Brent[2]

[1]Division 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]Division of Computational and Data Sciences, Washington University, St. Louis, Missouri, United States of America, [4]Department of Medicine, University of Maryland, Baltimore, Maryland, United States of America

The Long Life Family Study (LLFS) enrolled 4,953 participants in 539 pedigrees displaying exceptional longevity. To identify genetic mechanisms that affect cardiovascular risks in the LLFS population, we developed a 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 64 significant genes after Bonferroni correction ($p \leq 2.8\times10^{-7}$), 29 of which replicated in the Framingham Heart Study (FHS) cohort. Notably, 20 of the 29 replicated genes do not have a previously known trait-associated variant in the GWAS Catalog within 50 kb. Thirteen modules in Protein-Protein Interaction (PPI) networks are significantly enriched in genes with low meta-analysis p-values for at least one trait, three of which are replicated in the FHS cohort. The functional annotation of genes in these modules showed a significant over-representation of trait-related biological processes including sterol transport, protein-lipid complex remodeling, and immune response regulation. Among major findings, our results suggest a role of triglyceride-associated and mast-cell functional genes *FCER1A*, *MS4A2*, *GATA2*, *HDC*, and *HRH4* in atherosclerosis risks. Our findings also suggest that lower expression of *ATG2A*, a gene we found to be associated with BMI, may be both a cause and consequence of obesity. Finally, our results suggest that *ENPP3* may play an intermediary role in triglyceride-induced inflammation. Our pipeline is freely available and implemented in the Nextflow workflow language, making it easily runnable on any compute platform (https://nf-co.re/omicsgenetraitassociation).

3

Dissecting the Biological Mechanisms Behind Type 2 Diabetes Causal Effects on Non-cardiometabolic Comorbidities

Ana Luiza Arruda[1,2,*], Ozvan Bocher[1,*], Davis Cammann[3,*], Satoshi Yoshiji[4,5], Xianyong Yin[6,7], Chi Zhao[8], Henry J. Taylor[9,10,11], Jingchun Chen[3], Alexis C. Wood[12], Ken Suzuki[13], Josep M. Mercader[4,14,15], Cassandra Spracklen[8], James B. Meigs[4,16,17],

Jerome I. Rotter[18], Marijana Vujkovic[19,20,21], Benjamin F. Voight[19,20,22,23], Andrew P. Morris[24], Eleftheria Zeggini[1,25]
[1]Institute of Translational Genomics, Helmholtz Munich, Neuherberg, Germany; [2]Technical University of Munich (TUM), School of Medicine and Health, Graduate School of Experimental Medicine, Munich, Germany; [3]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, Nevada, United States of America; [4]Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [5]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; [6]Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, China; [7]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America; [8]Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America; [9]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [10]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [11]Heart and Lung Research Institute, University of Cambridge, Cambridge, United Kingdom; [12]USDA/ARS Children's Nutrition Center, Baylor College of Medicine, Houston, Texas, United States of America; [13]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; [14]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [15]Harvard Medical School, Boston, Massachusetts, United States of America; [16]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [17]Division of General Internal Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [18]Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, California, United States of America; [19]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania, United States of America; [20]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [21]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [22]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [23]Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [24]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom; [25]TUM school of medicine and health, Technical University Munich and Klinikum Rechts der Isar, Munich, Germany
*equal contribution

The Type 2 Diabetes Global Genomics Initiative (T2DGGI) has defined eight non-overlapping mechanistic clusters of index variants from multi-ancestry T2D GWAS meta-analysis that represent distinct pathways to disease. To assess putative causal relationships between T2D liability and non-cardiometabolic comorbidities, we performed bidirectional two-sample Mendelian randomization (MR) analyses. In addition, we conducted cluster-specific MR analyses to assess the effect of distinct biological mechanisms driving T2D liability on the different comorbidities. We used matching ancestry data for both diseases, meta-analyzing across ancestries when possible.

We find evidence of a causal effect of overall T2D liability on several comorbidities, including anorexia, cataracts, and osteoporosis. In some cases, the causal effect appears to be driven by specific clusters. For instance, we find evidence that the causal effect of T2D liability on cataracts (OR=1.06 [1.04, 1.08], FDR adjusted $P=4.33x10^{-9}$) is driven by the obesity cluster (OR=1.12 [1.07, 1.17], FDR adjusted $P=3.24x10^{-6}$). We also detect cluster-specific effects in the absence of an overall causal effect. For example, we recapitulate the well-established link between T2D, osteoarthritis, and obesity by showing evidence of a causal effect of the obesity cluster on osteoarthritis (OR=1.18 [1.1.4, 1.22], FDR adjusted $P=2.65x10^{-17}$). Conversely, we find evidence that the T2D liability linked to beta-cell dysfunction positively associated with proinsulin has a protective effect against osteoarthritis (OR=0.96 [0.94,0.99], FDR adjusted $P=1.97x10^{-2}$).T

4

DNA Large Language Models (DNA-LLMs) for Predicting Individual-Level Gene Expression from DNA Sequences

Raghav Awasthi, Xiaofeng Zhu[1]
[1]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University

Gene expression prediction is a challenging task in genetics due to factors like regulatory impacts, epigenetic modifications, and cellular heterogeneity. DNA-LLMs have generated excitement and interest for their capabilities in complex genetics tasks like gene expression prediction. Despite advancements explaining personal transcriptomic variation remains challenging. We fine-tuned DNA-LLMs for individual-level gene expression prediction and benchmarked them against a state-of-the-art Enformer model using data from the Geuvadis consortium. We analyzed 421 samples (224 females, 197 males) from diverse ancestries: British (n=85), Finnish (n=89), Tuscan (n=92), Utah (n=78), and Yoruba (n=77). Selecting 200 genes resulted in 84.2k base pairs of sequences, each 5k long. Two Nucleotide Transformer (NT) models with 2.5 billion and 500 million parameters, pretrained on 3202 human genomes, were utilized. Next generated embeddings from the NT model were fine-tuned using a 1D Convolution neural network. The evaluation included cross-gene and cross-individual correlation metrics and Mean Absolute Error scores. Analysis revealed the Enformer Model exhibited a mean cross-gene correlation of 0.48 (SD = 0.01) and a cross-individual correlation of 0.03 (SD = 0.20), with an MAE of 25.36. The 500 million NT model showed improved performance with a mean cross-gene correlation of 0.96 (SD = 0.01) and a cross-individual correlation of 0.08 (SD = 0.16), with an MAE of 2.91. The 2.5 billion NT model further improved with an MAE of 2.82, achieving a mean cross-gene correlation of 0.97 (SD = 0.01) and a cross-individual correlation of 0.10 (SD = 0.17), approaching the performance of PrediXcan, the model built on eQTLs. Our experiments highlight DNA-

LLMs' potential for enhancing gene-expression prediction, with scope for improvement through larger sequences and fine-tuning on additional genes.

Keywords: Gene expression prediction, DNA-LLMs (DNA large language models), Nucleotide Transformer (NT) models

5

Association Between Alzheimer's Disease Polygenic Protective Score (AD PPS) and Cognitive Function in Centenarians

Harold Bae[1*], Anastasia Gurinovich[2], Anastasia Leshchyk[3], Zeyuan Song[2], Mengze Li[3], Hannah Lords[3], Tanya Karagiannis[2], Stacy L. Andersen[4], Thomas T. Perls[4], Paola Sebastiani[2]

[1]Biostatistics Program, School of Nutrition and Public Health, College of Health, Oregon State University, Corvallis, Oregon, United States of America; [2]Tufts Clinical and Translational Science Institute, Institute for Clinical Research and Health Policy Studies, Tufts University School of Medicine, Boston, Massachusetts, United States of America; [3]Division of Computational Biomedicine, Boston University, Boston, Massachusetts, United States of America; [4]Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, United States of America

Alzheimer's disease (AD) is the most common cause of dementia with a high heritability between 60% and 80%. Centenarians demonstrate exceptional health spans by delaying many age-related diseases. Paradoxically, centenarians also have an increased risk of AD because of their exceptional age, but are often able to delay or avoid AD and related dementias. Therefore, they may serve as an important model to study the mechanisms of AD and identify genetic factors for neuroprotection from AD. In the present study, we used the New England Centenarian Study (NECS) data to construct the AD-specific polygenic protective score (AD PPS) based on the list of 83 published genetic variants that does not include the *APOE* variants. We compared the distributions of the AD PPS between the centenarians, offspring of centenarians, and study controls. Furthermore, the distributions of AD PPS were compared in four age categories--nonagenarians, centenarians (100-104), semi-supercentenarians (105-109), and supercentenarians (110+). We then examined the association between the AD PPS and Blessed Information-Memory-Concentration, Telephone Interview for Cognitive Status, and mortality. The AD PPS offers stronger protection against cognitive declines and mortality in the NECS centenarians. Both the NECS centenarians and offspring had significantly higher AD PPS than the study controls, but there was no significant difference between centenarians and the offspring. The genetic protection from AD significantly increased with more extreme ages with the highest AD PPS among the supercentenarians. This study highlights the value of studying the genetics of participants selected for exceptional longevity in relation to AD.

6

Two Complementary Identity-by-Descent Based Methods for Use Within Biobanks

J.T. Baker[1], Grahame Evans[1], Ryan Bohlender[2], Hung-Hsin[1], Josh Landman[1], Chad Huff[2], David Samuels[1,3], Piper Below[1]

[1] Vanderbilt Genetics Institute, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America;[2] Division of Cancer Prevention and Population Sciences, Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America [3]Department of Molecular Physiology & Biophysics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

Large biobanks offer unique opportunities for Identity-by-Descent (IBD) based tools due to the prevalence of distantly related individuals. Two such tools are IBDMap and DRIVE. IBDMap leverages genome-wide IBD data to identify regions of enriched case-case pair IBD sharing while DRIVE identifies haplotype structures overlapping loci of interest that are enriched for affected individuals. In this work, we present how these two complementary tools, when leveraged together, can elucidate novel genetic variation underlying diseases in large data resources. To accomplish this, we identified pairwise IBD segments using iLASH in 13,452 individuals of African Ancestry as identified by PCA within Vanderbilt University Medical Center's biobank, BioVU. We defined cases, controls, and exclusions within the AFR cohort for 1,436 PheCodes using the PheWAS package requiring that individuals have a code on two or more unique dates. We performed IBD mapping phenome-wide using this cohort and identified 180 PheCodes with one or more genome-wide enriched signals. We then ran DRIVE for these significant PheCodes using the enriched IBD signals as loci to determine the underlying haplotype structures. DRIVE identified 178 haplotypes across 103 of these PheCodes that were enriched for relevant cases. This work highlights how the application of these tools, in tandem, leverages distinct aspects of IBD and, when applied phenome-wide, how these tools have the power to highlight the unique haplotypic variation underlying specific diseases.

7

Biobank-Scale Genetically Regulated Expression Is Predictive of 3D Chromatin Contact Frequency

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

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

Introduction: Research over the past decade has suggested that enhancer RNA (eRNA) transcription plays a key role in mediating enhancer loop formation. We sought to explore whether genetically regulated expression (GReX) of eRNAs and canonical genes could predict chromatin contact frequency between respective transcribed loci.

Methods: We trained *in silico* models of eRNA and canonical gene GReX across 49 tissue types, using them to estimate GReX in a large-scale DNA biobank of European ancestry individuals (N > 70,000). As a baseline, we fit a linear regression model to predict 3D contact frequency between enhancer-enhancer and enhancer-gene pairs, using GReX as the features and high-resolution Hi-C contact frequencies as labels. Using these same data, we next trained a neural network-based regressor to predict pairwise enhancer-enhancer and enhancer-gene contact frequencies in both whole blood and cerebellum. Finally, we tested cross-tissue performance for each model.

Results: The logistic regression model showed virtually no ability to predict contact frequency ($R^2 \approx 0$). However, the neural network-base model showed substantially higher performance (test $R^2 > 0.21$ in whole blood and $R^2 > 0.37$ in cerebellum). The model trained in whole blood showed poor cross-tissue portability ($R^2 = 0.01$ in cerebellum), while the cerebellum-based model achieved markedly higher cross-tissue performance ($R^2 > 0.17$ in whole blood).

Conclusions: Our results suggest that eRNAs play a key role in mediating 3D chromatin contacts. The discrepancy in cross-tissue performance may be a result of whole blood's high heterogeneity and stochastic gene expression relative to other tissues.

8

Evaluation of the impact of selection bias in Canada's nationwide Host Genome Sequencing Initiative (HostSeq) with application to a thromboembolism-informed HD-GWAS of COVID-19 severe outcomes

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

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

Elucidating the true risk factors for a disease is challenging for all observational studies, with exceptional challenges arising in the context of an emerging infectious disease such as COVID-19 where samples are recruited out of convenience. Launching observational studies quickly with patchy lockdowns, cases hotspots and dynamic vaccination schedules amplify the risk of selection bias for COVID-19 determinants of interest. Host Genome Sequencing initiative (HostSeq) performed whole genome sequencing on 10,000 participants from 13 independent studies across Canada. Alongside any heterogeneity due to the diverse study designs, there is likely to be important variability across time and Canadian regions in terms of symptom definitions, disease severity, and SARS-CoV-2 evolution. We are using simulation studies to investigate the effects of rates of comorbidity influencing COVID-19 severity in the studies and different severe disease patterns by SARS-CoV-2 wave, and how these factors influence bias in estimates of association between a SNP and severe disease. Our simulation utilized Canadian data for infection rates by age, sex, hypertension, and virus wave. We sampled from these distributions to match the HostSeq participant characteristics, employing various models for severe disease risk and SNP association, with and without methods to account for study design and anticipated selection bias. So far, logistic regression adjusted by individual study membership outperformed weighted logistic regression. We will use the simulation results to illuminate our thromboembolism-informed hypothesis-driven GWAS of COVID-19 severe outcomes, to test whether COVID-19 morbidity is partially explained by sequenced variants known to be associated with thromboembolic outcomes.

9

Dissecting Ancestry-aware Molecular Causal Effects for Type 2 Diabetes

Ozvan Bocher[1,*], Ana Luiza Arruda[1,2,*], Satoshi Yoshiji[3,4,*], Chi Zhao[5,*], Xianyong Yin[6,7], Davis Cammann[8], Henry J. Taylor[9,10,11], Jingchun Chen[8], Ravi Mandla[3,12], Alicia Huerta-Chagoya[3], Ta-Yu Yang[4], Ken Suzuki[13], Alexis C. Wood[14], Fumihiko Matsuda[4], Jason Flannick[3,15,16], Josep M. Mercader[3,12,17], Cassandra Spracklen[5], James B. Meigs[3,18,19], Jerome I. Rotter[20], Marijana Vujkovic[21,22,23], Benjamin F. Voight[21,22,24,25], Andrew P. Morris[26], Eleftheria Zeggini[1,27]

[1]Institute of Translational Genomics, Helmholtz Munich, Neuherberg, Germany; [2]Technical University of Munich (TUM), School of Medicine and Health, Graduate School of Experimental Medicine, Munich, Germany; [3]Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [4]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; [5]Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America; [6]Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, China; [7]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America; [8]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV, United States of America; [9]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [10]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [11]Heart and Lung Research Institute, University of Cambridge, Cambridge, United Kingdom; [12]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [13]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; [14]USDA/ARS Children's Nutrition Center, Baylor College of Medicine, Houston, Texas, United States of America; [15]Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, United States of America; [16]Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts, United States of America; [17]Harvard Medical School, Boston, Massachusetts, United States of America; [18]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [19]Division of General Internal Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [20]Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, California, United States of America; [21]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania, United States of America; [22]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [23]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [24]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America;

[25]*Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America;* [26]*Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom;* [27]*TUM school of medicine and health, Technical University Munich and Klinikum Rechts der Isar, Munich, 81675, Germany*
*equal contribution

Multiple molecular mechanisms are involved in type 2 diabetes (T2D) pathogenesis, with potentially different effects across ancestries. Recent large-scale efforts by the T2D Global Genomics Initiative have generated novel insights into the genetic architecture of T2D. In this work, we sought to leverage these findings to explore the causal molecular mechanisms leading to T2D in an ancestry-aware manner. We conducted two-sample Mendelian randomization analysis using blood cis-expression and protein quantitative trait loci derived from four major ancestries. We performed cross-ancestry meta-analyses and defined significance at a 5% FDR threshold. To corroborate our findings, we investigated evidence of colocalization using PWCoCo. We detected causal effects of the genetically regulated levels of 78 genes and 2 proteins on T2D risk in the cross-ancestry meta-analysis. Additionally, we found that 249 genes have a significant causal effect on T2D in ancestry-specific analyses only, of which 6 and 12 genes were specific to the African and Hispanic ancestry groups, respectively. Similarly, 11 proteins were only significant in the ancestry-specific analyses, of which 7 proteins were specific to the East-Asian population. Finally, only two signals, SNUPN and PTPN9, were consistently found across the two omics layers in the European population, both with increased levels being protective against T2D. Our findings highlight the power of large-scale meta-analysis and multi-omics MR analyses to identify causal pathways involved in T2D risk. Our results show how meta-analyzing ancestry-specific MR analyses can help to uncover ancestry-specific and shared causal pathways and emphasize the need for expanding investigations into non-European ancestry populations.

10

The Multiethnic Cohort: A Resource for the study of Genetic and non-Genetic Cancer Risk Across Populations

David Bogumil[1], Xin Sheng[1], Peggy Wan[1], Lucy Xia[1], Samantha Streicher[2], Brian Z. Huang[1], Charleston W.K. Chiang[1,3,4], Fei Chen[1], Lynne R. Wilkens[2], Loïc Le Marchand[2], Christopher A. Haiman[1,3,4], David V. Conti[1,3,4]

[1]*Center for Genetic Epidemiology. Department of Population and Public Health Sciences, Keck School of Medicine of University of Southern California, Los Angeles, California, United States of America;* [2]*Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, United States of America;* [3]*Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America;* [4]*Center for Genetic Epidemiology, Keck School of Medicine of University of Southern California, Los Angeles, California, United States of America*

Introduction: The Multiethnic Cohort Study (MEC) was established to identify cancer risk factors across 5 self-reported racial and ethnic groups (African American [AA], Japanese American [JA], Latino [LA], Native Hawaiian [NH] and White [WH]). Here we use the new large-scale MEC genetics data set for genetic risk prediction.

Methods: The MEC includes over 215,000 individuals, aged 45-75 at baseline (1993-1996), from California and Hawaii, with germline data for 73,139 participants (10,962 AA, 24,234 JA, 17,242 LA, 5,488 NH, and 14,649 WH). We estimated the association between weighted genetic risk scores (GRS) and incident cancer, then estimated cumulative risk (CR) with Poisson time-to-event models.

Results: There were 1,880 breast (BC), 1,340 colorectal (CRC), and 2,899 prostate (PCa) cases following biospecimen collection. The largest effect was seen among WH for PCa, where 1-SD increase in GRS was associated with 2.13 (95%CI 1.93-2.34) times the incidence. CR of PCa and BC were considerably higher than CRC. At age 80, an AA male at 95% GRS quantile had 43.1% (36.9%-50.9%) CR of PCa; a WH male with same age and GRS had a 35.9% (29.9%-43.8%) CR. For CRC the CR at 95% GRS ranges from 3.8% (2.9%-5.2%) for WH to 9.0% (7.2%-12.2%) for JA. BC had similar estimates of CR ranging from 19.9% (15.6%-26.7%) for NH to 14.4% (11.9%-18.3%) for JA.

Discussion: These results highlight the unique and valuable characteristics of the MEC genetic database and cohort and the ability to explore genetic and non-genetic risk factors jointly among individuals from diverse populations.

11

GWAS of IgG responses to RV and RSV Genome-wide Association Study of Rhinovirus and Respiratory Syncytial Virus IgG Responses in Children and Adults

Raphaël Vernet[1], Justine Wenta[1], Alicia Guillien[2], Anja Estermann[1], Christophe Linhard[1], Florence Demenais[1], Katarzyna Niespodziana[3], Rudolf Valenta[3,4], Valérie Siroux[2], Emmanuelle Bouzigon[1*]

[1]*Université Paris Cité, Inserm, UMR 1124, Group of Genomic Epidemiology of Multifactorial Diseases, Paris France;* [2]*University Grenoble Alpes, Inserm U 1209, CNRS UMR 5309, Team in Environmental Epidemiology Applied to the Development and Respiratory Health, Institute for Advanced Biosciences, Grenoble, France;* [3]*Division of Immunopathology, Department of Pathophysiology and Allergy Research, Center for Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Vienna, Austria;* [4]*Karl Landsteiner University, Krems, Austria*

Respiratory viral infections, especially from rhinovirus (RV) and respiratory syncytial virus (RSV) are associated with the occurrence and exacerbation of asthma.

We aimed to identify genetic variants associated with RSV- and RV-specific IgG responses in both children and adults while accounting for differential sex effect.

Cumulative RV-specific IgG levels (species A, B and C) and IgG levels to RSV-G protein were measured using micro-array technology applied to blood samples of subjects from the EGEA study. We conducted genome-wide association study of RSV- and RV-specific IgG levels stratified on sex using linear mixed model applied separately to children (292 boys and 215 girls) and adults (555 men and 598 women). Then, we performed: 1) a joint test of genetic main effect and gene-by-sex interaction effect and 2) a test of gene-by-sex interaction alone.

We identified four genome-wide significant loci associated with RSV- or RV-specific IgG levels ($P_{joint}<5x10^{-8}$). Two of them

were detected in children: one was associated with RV-C-specific IgG levels independently of the sex (P_{male}=2.3x10^{-8} and P_{female}=5.3x10^{-3} in *GALNT15*), while the other one was associated with RSV-specific IgG levels in females only (P_{female}=2.4x10^{-8} *vs.* P_{male}=0.67; rs12324881 in *GABRG3*). In adults, one locus was associated with RV-A-specific IgG levels in females only (P_{female}=4.4x10^{-9} *vs.* P_{male}=0.79; rs768382 between *MAP3K7* and *EPHA7*) whereas the other one was associated with RV-B-specific IgG levels in males only (P_{male}=5x10^{-8} *vs.* P_{female}=0.34; rs10542502 between *TRPA1* and *KCNB2*). Multivariate analyses will enable to assess whether immune responses to various RSV and RV share common genetic determinants.

Funding: ANR-19-CE36-0005-NIRVANA

12

FunColoc: A Generalized Functional Regression Model for Genetic Colocalization Analysis of microRNA Counts and Disease-related Outcomes

Myriam Brossard[1,2], Kathleen Zang[2], Thomas G. Wilson[3], Shabana Amanda Ali[3,4*], Osvaldo Espin-Garcia[2,5,6*]

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada; [2]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario, Canada; [3]Henry Ford Health & Michigan State University Health Sciences, Detroit, Michigan, United States of America; [4]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, Michigan, United States of America; [5]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [6]Department of Biostatistics, Krembil Research Institute and Schroeder Arthritis Institute, University Health Network, Toronto, Ontario, Canada

**These authors contributed equally*

MicroRNAs play an important role in regulating gene expression and are increasingly associated with complex diseases such as osteoarthritis. Colocalization analysis of microRNA counts and disease-related outcomes can reveal whether they share the same genetic association signal(s) in a locus, which can point to novel regulatory pathways. Existing colocalization methods do not take full advantage of individual-level data, where both traits are collected in the same set of individuals. In addition, these methods are not applicable to traits with different types of distributions (such as microRNA counts and a binary/quantitative disease-related outcome). To address these challenges, we propose FunColoc, a generalized multivariate Functional regression model for Colocalization analysis between microRNA counts and an outcome with similar or different types of distributions (e.g., binary/quantitative). FunColoc estimates continuous trait-specific functions for the effects of multiple variants in a locus depending on their positions. We construct a colocalization test statistic defined as the product of the variant effect functions between the two traits. We then compute *P*-values for colocalization testing under a composite null hypothesis (i.e., both estimated variant effect functions are different from zero) using empirical approaches. Through simulation studies in a genomic locus randomly chosen on 16q in 5,000 individuals with simulated genotypes, microRNA counts and osteoarthritis severity, FunColoc detected colocalization with good power (60-80%) in various scenarios with colocalization and controlled Type-I errors (<5%) in scenarios without colocalization. Next steps include applying FunColoc to the Osteoarthritis Initiative cohort, a longitudinal observational knee osteoarthritis study.

13

A Novel Multivariable Mendelian Randomization Framework to Disentangle Highly Correlated Exposures with Application to Metabolomics

Lap Sum Chan, Mykhaylo M. Malakhov, Wei Pan

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

Mendelian Randomization (MR) utilizes genome-wide association study (GWAS) summary data to infer causal relationships between exposures and outcomes, offering a valuable tool for identifying disease risk factors. Multivariable MR (MVMR) estimates the direct effects of multiple exposures on an outcome. This study addresses the challenge of highly correlated exposures common in metabolomics data, where existing MVMR methods often become unscalable or suffer reduced statistical power due to multicollinearity. We propose a robust extension of the MVMR framework, named MVMR-cML-SuSiE, which employs constrained maximum likelihood (cML) and incorporates the Sum of Single Effects (SuSiE) model originally developed for fine mapping, to identify independent clusters of exposure signals. Applying MVMR-cML-SuSiE to the UK Biobank's metabolomics data for the largest Alzheimer's disease (AD) cohort through a two-sample MR approach, we identified two independent signal clusters for AD: glutamine and lipids, with posterior inclusion probabilities (PIPs) of 95.0% and 81.5%, respectively. Our findings corroborate the hypothesized roles of glutamine and lipids in AD, providing quantitative support for their potential involvement.

Keywords: Mendelian randomization, Statistical genetics, Metabolomics, Alzheimer's disease

14

StocSum: A Reference-panel-free Summary Statistics Framework for Diverse Populations

Han Chen[1], Nannan Wang[1], Bing Yu[1], Goo Jun[1], Qibin Qi[2], Ramon A. Durazo-Arvizu[3,4], Sara Lindstrom[5,6], Alanna C. Morrison[1], Robert C. Kaplan[2], Eric Boerwinkle[1,7]

[1]Human Genetics Center, Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [2]Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, New York, United States of America; [3]The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, California, United States of America; [4]Department of Pediatrics, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [5]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [6]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, United States of America; [7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United States of America

Genomic summary statistics have been widely used to address various scientific questions in genetic and genomic

research. Applications that involve multiple genetic variants, such as conditional analysis, variant set and gene-based tests, heritability and genetic correlation estimation, also require correlation or linkage disequilibrium (LD) information between genetic variants, often obtained from an external reference panel. While these methods usually have good performance for common variants in populations of only European ancestry, in practice, it is usually difficult to find external reference panels that accurately represent the LD structure for isolated, underrepresented or admixed populations, as well as rare genetic variants from whole genome sequencing (WGS) studies, limiting their applications to European populations. To maximize the applicability of summary statistics-based methods and make them equally beneficial to all human populations, we have developed StocSum, a novel reference-panel-free statistical framework for generating, managing, and analyzing stochastic summary statistics using random matrix algorithms. Using two cohorts from the Trans-Omics for Precision Medicine Program, we demonstrate the accuracy of StocSum-based LD measures as compared to those directly computed from individual-level genotype data, in European-, African-, and Hispanic/Latino-Americans. We also show that for admixed populations such as African- and Hispanic/Latino-Americans, LD measures computed from external reference panels perform much worse, even if all ancestry populations are included in those reference panels. As a reference-panel-free framework, StocSum will facilitate sharing and utilization of genomic summary statistics from WGS studies, especially for isolated, underrepresented and admixed populations.

Keywords: summary statistics, linkage disequilibrium, diversity, whole genome sequencing, random matrix algorithm

15

Exploring Aryl Hydrocarbon Receptor (AHR) Variant Effects in the UK Biobank and Interactions with Diesel Exposure in the Personalized Environment and Genes Study (PEGS)

Uchechukwu S. Chimeh[1,2*], Xiaoran Tong[2], Gregory A. Stamper[3], Farida S. Akhtari[2], John S. House[2], David C. Fargo[4], Charles P. Schmitt[5], Janet E. Hall[6], Alison A. Motsinger-Reif[2], David L. Aylor[1,7,8]

[1]Graduate Program in Genetics, Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, United States of America; [2]Biostatistics & Computational Biology Branch, National Institute of Environmental Health Sciences, Durham, North Carolina, United States of America; [3]Office of Scientific Computing, National Institute of Environmental Health Sciences, Durham, North Carolina, United States of America; [4]Office of the Director, National Institute of Environmental Health Sciences, Durham, North Carolina, United States of America; [5]Office of Data Science, National Institute of Environmental Health Sciences, Durham, North Carolina, United States of America; [6]Clinical Research Branch, National Institute of Environmental Health Sciences, Durham, North Carolina, United States of America; [7]Center for Human Health and the Environment, North Carolina State University, Raleigh, North Carolina, United States of America; [8]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America

The aryl hydrocarbon receptor (AHR) pathway is essential in the metabolism of environmental toxicants like polycyclic aromatic hydrocarbons (PAHs) from cigarette smoke and vehicle exhaust. The effects of exposure to AHR ligands varies among individuals, and we hypothesize that some of this variation is explained by genetic variation in AHR. However, only a fraction of the over 13,000 existing AHR variants have been associated with traits in candidate gene studies (22 variants) and in the GWAS catalog (84 variants). Therefore, to uncover associations between AHR variants and traits in a biobank, we filtered genome-wide association studies of 3273 traits in the UK Biobank for AHR variants below the genome-wide significance threshold ($P < 5x10^{-8}$). We found that AHR variants were primarily associated with traits related to pregnancy, sun exposure, and metabolite levels. Furthermore, we explored gene-environment interactions (GxE) between AHR variants and an index of ambient diesel concentration in the Personalized Environment and Genes Study (PEGS), a North Carolina-based multi-ancestry cohort. In PEGS, we investigated trait-associated AHR variants from the UK Biobank, the GWAS catalog, and candidate gene studies in individuals with fatty liver disease. We found interaction effects between AHR variants and ambient diesel concentration ($P < 0.05$), including one which has previously been associated with idiopathic male infertility and urinary levels of a PAH metabolite. Our results elucidate AHR variant effects on trait variation in the UK Biobank and interaction effects of PAH exposure and AHR on fatty liver disease risk in PEGS.

16

Multi-omics Data Integration for Phenotype Prediction Using Machine Learning Identifies Novel Alzheimer's Disease Risk Factors

Jerome J. Choi[1], Corinne D. Engelman[1]

[1]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

Alzheimer's disease (AD) is a complex, heterogeneous, and multifactorial condition. Multi-omics approaches are instrumental in elucidating such diseases. Machine learning methods provide novel techniques to integrate and analyze diverse omics data, facilitating the discovery of AD risk factors. This study employs machine learning approaches to identify the single and interactive effects of AD risk factors within multi-omics data.

We trained deep learning models to predict AD outcomes, Preclinical Alzheimer's Cognitive Composite 3 (PACC3) and plasma ptau217, using plasma metabolomics and imputed blood transcriptomics from Wisconsin Registry for Alzheimer's Prevention participants. Firstly, we trained single omics deep learning models for transcriptomics and metabolomics separately to select the top 100 genes and top 100 metabolites using gradient-based feature importance for each outcome. Secondly, we integrated these selected genes and metabolites to build and train multi-omics deep learning models. Thirdly, we computed feature importance and interaction scores for these genes and metabolites. Fourthly, we regressed outcomes on the cross-omics pairs in linear regression models, adjusted for age, gender, *APOE* score, and education years. Finally, we identified potential cross-omics AD risk factors that have high interaction scores and significant FDR values.

We identified several cross-omics pairs with high interaction

scores, such as daidzein sulfate (2) and *SPRTN*, and 1-(1-enyl-palmitoyl)-2-arachidonoyl-GPE (P-16:0/20:4) and *TEAD3*, for PACC3 and ptau217, respectively. The interaction effects of these pairs were found to be significant in the linear regression models (FDR<0.05).

Overall, our approaches create an analytic framework for identification of interactions between potential AD risk factors in a multi-omics fashion.

17

Polygenic Risk Scores and Risk of Drug-induced Liver Injury following initiation of antiretroviral therapy in people living with HIV

Zinhle Cindi[1], Katie M. Cardone[1], Yuki Bradford[1], Hannah Kim[2], VA Million Veteran Program[3], Eric S. Daar[4], Roy Gulick[5], Sharon A. Riddler[6], Marijana Vujkovic[7,8], Kyong-Mi Chang[7,8], Philip S. Tsao[9,10], Gary Maartens[11], Phumla Sinxadi[11], David W. Haas[2,12], Marylyn D. Ritchie[1,13]

[1] Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2] Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3] VA Million Veteran Program; [4] Lundquist Institute at Harbor-UCLA Medical Center, Torrance, California, United States of America; [5] Weill Cornell Medicine, New York, New York, New York, United States of America; [6] University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [7] Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania, United States of America; [8] Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [9] VA Palo Alto Health Care System, Palo Alto, California, United States of America; [10] Department of Medicine, Stanford University School of Medicine, Stanford, California, United States of America; [11] Division of Clinical Pharmacology, Department of Medicine, University of Cape Town, Cape Town, South Africa; [12] Meharry Medical College, Nashville, Tennessee, United States of America; [13] Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Background: Drug-induced liver injury affects some people living with HIV (PWH) following initiation of antiretroviral therapy (ART). The contribution of genetic susceptibility to liver injury in this setting is not known. We examined whether polygenic risk score (PRS) models could help predict the likelihood of liver injury within 48 weeks following ART initiation among PWH.

Methods: As a reference population, we derived PRS_{ALT} using summary statistics from the Million Veteran Program. We filtered variants based on ALT associations with $P \leq 0.05$. With the remaining variants, we separately constructed PRS for individuals of European (n=68,725) and African (n=13,387) ancestry using PRS_{csx}. We applied these models to treatment-naïve participants in the AIDS Clinical Trials Group (ACTG, n=5114). ALT elevations grading; grade 1: <2.5X upper limit of normal (ULN), grade 2: >2.5-5X ULN, grade 3: >5X ULN. Participants with baseline ALT grade >2 were excluded from analyses. Cases with liver injury were defined as having ALT grade ≥3 within 48 weeks following ART initiation. Controls had ALT grade ≤1 from baseline to week 48. We assessed for associations in participants of European (31 cases, 1975 controls) and African (23 cases, 1795 controls) ancestry. Multivariable logistic models applied separately within each group included PRS, sex, age, body mass index, and 5 genetic principal components.

Results and Conclusion: In multivariable models, PRS was not associated with grade ≥3 ALT elevation following ART initiation in PWH. This finding suggests that inherent genetic susceptibility may not contribute greatly to drug-induced liver injury in this setting.

18

Genome-wide Association Study of Protein-altering and Regulatory Variants with Resistance to *Mycobacterium Tuberculosis* Infection

Clément Conil[1,2*], Jonathan Bohlen[1,2], Matthieu Chaldebas[3], Monica Dallmann-Sauer[4,5,6], Marlo Möller[7], Marc A. Jean-Juste[8], Jean-Laurent Casanova[1,2,3,9], Laurent Abel[1,2,3], Erwin Schurr[4,5,6], Aurélie Cobat[1,2,3]

[1] Laboratory of Human Genetics of Infectious Diseases, Necker Branch, Inserm U1163, Necker Hospital for Sick Children, Paris, France; [2] Paris Cité University, Imagine Institute, Paris, France; [3] St Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York City, New York, United States of America; [4] Department of Biochemistry, Faculty of Medicine, McGill University, Montreal, Québec, Canada; [5] Program in Infectious Diseases and Global Health., The Research Institute of the McGill University Health Center, Montreal, Québec, Canada; [6] McGill International TB Center, Department of Medicine, Faculty of Medicine, McGill University, Montreal, Québec, Canada; [7] DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa; [8] Haitian Study Group for Kaposi's Sarcoma and Opportunistic Infections (GHESKIO), Port-Au-Prince, Haiti; [9] Howard Hughes Medical Institute, New York City, New York, United States of America

Despite intense exposure to *Mycobacterium tuberculosis* (*Mtb*), some individuals seem resistant to infection, remaining negative to tuberculin skin test (TST) and interferon-gamma release assays (IGRAs). We whole genome sequenced individuals with extreme resistance phenotypes who remained TST and IGRA negative despite high vulnerability due to HIV infection and high exposure to *Mtb*. We enrolled 55 resisters and 100 *Mtb* infected individuals from South Africa and 66 resisters and 57 *Mtb* infected individuals from Haiti. We performed a genome-wide association study of resistance to *Mtb* with candidate common variants (frequency > 5% in gnomAD, African population), defined as protein altering variants or non-coding variants affecting expression (eQTLs) and splicing (sQTLs) in the lungs or whole blood. We identified a locus on chromosome 12 of seven SNPs associated with resistance to infection in South Africa and replicated in Haiti (combined OR[95%CI]=0.22[0.13-0.39], *P value*=1.03×10^{-7}). The *Mtb* resistance allele was significantly depleted from our in-house African TB cohort (9.1% homozygotes in 121 TB vs. 16.4% in 586 controls, *P value*=0.036) and UK Biobank European individuals with TB history (6.4% homozygotes in 581 TB vs. 9.3% in 107,350 controls, *P value*=0.017). The main candidate causal

variant suppresses an upstream open reading frame (uORF) in the 5′ untranslated region of the downstream gene. We show experimentally that suppression of this uORF increases the activity of the downstream gene. This gene is involved in innate lymphoid cells differentiation, important cells in the defense against *Mtb*. These results will help understand the molecular mechanisms involved in *Mtb* infection.

19

Genome-wide Association Study on Metabolic Dysfunction-associated Steatotic Liver Disease (MASLD)

Rebecca Darlay[1]*, Michalina Zatorska[2], Sarah Worthington[2], Quentin M. Anstee[2,3], Ann K. Daly[2], Heather J. Cordell[1] on behalf of the LITMUS consortium

[1]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [2]Translational & Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [3]Newcastle NIHR Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Trust, Newcastle upon Tyne, United Kingdom

Genetic risk factors associated with metabolic dysfunction-associated steatotic liver disease (MASLD) remain incompletely understood. We performed a genome-wide association study (GWAS) involving MASLD cases from the European MASLD Registry with 2,405 biopsy-proven MASLD cases and 17,781 genetically-matched population controls.

The analysis resulted in six genome-wide significant signals, including four reported by us previously in PNPLA3 on chromosome 22, chromosome 19 (TM6SF2), HSD17B13 (chromosome 4) and GCKR (chromosome 2), and two additional signals on chromosome 1 (LEPR and MARC1). The LEPR signal has not been reported previously elsewhere. The MARC1 signal is in line with recent reports using non-histological criteria to confirm cases.

Further subgroup GWAS involved (i) 1312 cases with activity scores >4 as a metabolic-associated steatohepatitis (MASH) cohort and (ii) 732 cases with advanced fibrosis with the 17,781 population controls used for both groups. PNPLA3 and chromosome 19 signals were found to be relevant to both initial risk and development of MASH and advanced fibrosis. A relatively rare variant on chromosome 6 was associated with the MASH cohort. A variant on chromosome 13 was also genome-wide significant and may be an additional risk factor for advanced fibrosis. Replication studies and construction of polygenic risk scores to predict MASLD progression are in progress.

20

DNA Repair Genes XRCC1(Arg399Gln) and XRCC3 (Thr241Met) Polymorphisms in Elevate Cervical Cancer Risk among Bangladeshi Females

Laboni Das[1], Sadia Rahman[1], Amir Hossain[2], Razia Sultana[3], Md. Abdula Mazid[4], Md. Mustafizur Rahman[1*]

[1]Khulna Univ., Khulna, Bangladesh, [2]Dhaka Intl. Univ., Dhaka, Bangladesh, [4]Decent Dental Clinic, Khulna, [4]Univ. of Dhaka, Dhaka, Bangladesh

**Presenting author: dipti0103@yahoo.com; mmrahman0103@pharm.ku.ac.bd*

Background: Cervical cancer is the second most common cancer in women worldwide, and is both a preventable and a curable disease especially if identified at an early stage. Sequence variations in DNA repair genes can cause aberration in cellular functions leading to cancer. Genetic polymorphisms in XRCC1 (Arg399Gln) and XRCC3 (Thr241Met) genes result in individual variation in their DNA repair capacity. The aim of this study was to identify the association between XRCC1 Arg399Gln and XRCC3 Thr241Met single nucleotide polymorphisms (SNPs) and susceptibility to cervical cancer in Bangladeshi populations.

Methods: The case-control study comprised 124 cervical cancer patients and 148 healthy controls. Genomic DNAs were isolated from peripheral blood and genotyped for candidate SNPs using polymerase chain reaction-restriction fragment length polymorphism (PCR–RFLP) method.

Results: For XRCC1, heterozygous Arg/Gln and combined heterozygous plus variant homozygous Gln/Gln genotypes showed 1.78-fold (95% CI 1.0037 to 2.8771, p=0.0484) and 1.8627-fold (95% CI 1.1470 to 3.0250, p = 0.0119) increased risk of cervical cancer, respectively, when compared with normal homozygous Arg/Arg genotype. The variant Gln allele was positively associated with cervical cancer by 1.68-fold increase (95% CI 1.1732 to 2.3980, p = 0.0046). Similarly, for XRCC3, Thr/Met heterozygous and combined Thr/Met + Met/Met genotypes were found to be associated with 1.6993-fold (95% CI 1.0398 to 3.0166, p=0.0354) and 1.8312-fold (95% CI 1.0890 to 3.0791, p = 0.0225) higher risk, respectively, when compared with normal homozygous Thr/Thr genotypes. The variant Met allele showed significant association with 1.71-fold increased risk.

Conclusion: XRCC1 (Arg399Gln) and XRCC3 (Thr241Met) polymorphisms may be associated with increased cervical cancer risk in Bangladeshi females.

Keywords: XRCC1 (Arg399Gln), XRCC3 (Thr241Met), Cervical cancer, Bangladesh

21

Deciphering the Genetic Architecture of Lung Cancer Survival: GWAS Meta-analysis with 10K Patients and Integrative multi-omics study

Mulong Du[1,2,†], Junyi Xin[3], Li Su[1], David C. Christiani[1,4,†]

[1]Departments of Environmental Health, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, Massachusetts, United States of America; [2]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China; [3]Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China; [4]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America.

[†]Correspondence author

Genome-wide association studies (GWASs) have identified over 40 loci involving lung cancer susceptibility, but its polygenic risk score in predicting lung cancer prognosis remains limited, which highlight the importance to decipher the genetic architecture of lung cancer survival. We performed large-scale lung cancer survival GWAS through over 10,000 patients from Boston Lung Cancer Study, International Lung Cancer Consortium, The Cancer Genome Atlas, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial and UK Biobank

cohorts of European ancestry, followed by pathway enrichment analysis, transcriptome-wide association study (TWAS), and single-cell and spatial transcriptomics annotation, as well as causal factors identification by phenome-wide Mendelian randomization analysis (MR-PheWAS). We observed three independent genetic loci associated with lung cancer overall survival, including rs77676649 (5q12.3, HR = 1.26, $P = 4.81\times10^{-8}$), rs12257527 (10p13, HR = 1.31, $P = 1.28\times10^{-8}$) and rs79829102 (22q13.31, HR = 1.91, $P = 4.46\times10^{-9}$), which were mainly enriched in pathways of selenocompound and lipoic acid metabolism, overlapped with enhancer histone modification, and functionally involved in lung tumor progression. Further, among six candidate lung cancer prognostic genes ($P_{TWAS} < 0.001$), *ANXA13* was prioritized in epithelial cells ant its higher expression was associated with poorer prognosis. In addition, MR-PheWAS identified 15 potential causal factors of lung cancer survival, among which patients with higher education was causally associated with a decreased risk of death ($P_{IVW} = 2.06\times10^{-4}$). This large-scale GWAS study provided additional genetic biomarkers involving lung cancer progression, which would advance the understanding of genetic architecture of lung cancer survival.

Keywords: lung cancer; survival; GWAS; omics; genetic biomarkers.

22

Trans-Ancestry Meta-Regression Analysis MR-MEGA Replicates European IBD Loci Using Multiple Ancestries

Ellyn Dunbar[1], Claire Simpson[1], Steven Brant[2]

[1]University of Tennessee Health Science Center, Tennessee, United States of America; [2]Rutgers Robert Wood Johnson Medical School, New Jersey, United States of America

Inflammatory bowel disease (IBD) affects the intestinal tract with two subtypes: Crohn's disease and ulcerative colitis. Most known IBD risk loci have been identified in European ancestry individuals (EUR). Some have been replicated in East Asian (EA) and African American (AA) populations; however, much of the heritability remains unexplained. The purpose of this study was to use the trans-ancestry method MR-MEGA to leverage the power of larger sample sizes by meta-analyzing multiple populations to replicate known loci in IBD.

MR-MEGA, a trans-ancestry meta-regression analysis method was used to combine summary statistics of IBD GWAS. Five studies including four populations were included: AA, EA, Finnish, and Non-Finnish European. MR-MEGA's genomic control options were used to adjust for inflation of the summary statistics. Only SNPs available in all five studies remained in the analysis. Our target genome-wide significance threshold was 5e-08. FUMA annotated the results. Resulting loci were compared to 241 known loci compiled by de Lange, et al. in a 2017 IBD study using EUR to identify replicated loci.

FUMA annotation of the MR-MEGA results identified 122 genomic risk loci. Of these, 95 overlapped/replicated known loci from the previous IBD study in EUR. 146 known loci did not replicate.

We used MR-MEGA to replicate loci previously identified only in EUR. Replication of single ancestry loci across ancestries increases confidence in the validity of those loci and allows for future generalization across ancestries. Loci not replicated may be specific to Europeans or may not have had significant SNPs included in MR-MEGA.

Keywords: Inflammatory bowel disease, MR-MEGA, trans-ancestry

23

gmmcoda: Graphical Model for the Mixture of Compositional Data and Absolute Abundance Data with Applications to Microbiome Studies

Shen Zhang[1], Huaying Fang[2,3,#*], Tao Hu[1,#]

[1]School of Mathematical Sciences, Capital Normal University, Beijing, China; [2]Beijing Advanced Innovation Center for Imaging Theory and Technology, Capital Normal University, Beijing, China; [3]Academy of Multidisciplinary Studies, Capital Normal University, Beijing, China

[#]Corresponding authors.

Probabilistic graphical models provide efficient approaches to exploring the relationship of variables in various applications. Gaussian graphical model (GGM) is popular for constructing the conditional dependence network of interested variables. However, GMM is inappropriate in some applications such as microbiome studies, in which only relative abundances (referred to as compositional data in statistics) can be observed for some variables. Recently, some algorithms have been proposed to deal with the graphical modeling problem for compositional data. Nevertheless, there is a lack of statistical methods for inferring the interaction network for the mixture of compositional data and absolute abundance data. In this study, we propose a probabilistic graphical model for modeling interactions of variables for the mixture of compositional data and absolute abundance data. A novel maximum penalized likelihood estimator, called gmmcoda, is introduced for inferring the network from the mixture data. We develop a majorization-minimization algorithm to solve the optimization problem involved in gmmcoda. The performance of gmmcoda is evaluated and compared with other existing methods by simulation studies. Additionally, we apply gmmcoda to one microbiome data including microbial abundance as compositional data and gene expression data as absolute abundance data. The microbe-gene interactions detected by gmmcoda are further validated using previous studies.

24

Univariate and Mmultivariate Proteome-wide Association Studies to Identify Causal Proteins for Alzheimer's Disease in the Presence of Invalid Instruments with GWAS Summary Data

Lei Fang[1], Haoran Xue[2], Zhaotong Lin[3], Wei Pan[4]

[1]Division of Biostatistics and Health Data Science, University of Minnesota; [2]Department of Biostatistics, City University of Hong Kong; [3]Department of Statistics, Florida State University; [4]Division of Biostatistics and Health Data Science, University of Minnesota

Alzheimer's disease (AD) is a complex and progressive neurodegenerative disorder that accounts for the majority of dementia cases. Here we aim to identify causal plasma proteins for AD, shedding light on physiological and pathological processes of AD. We utilized the latest large-scale plasma proteomic data from UK Biobank Pharma Proteomics Project (UKB-PPP) and AD GWAS summary data from the International

Genomics of Alzheimer's Project (IGAP). Via a new and robust univariate instrumental variable (IV) regression method, we identified 15 causal proteins through both cis- and trans-pQTLs (called trans-pQTLs for simplicity here) and 7 causal proteins through cis-pQTLs. To further reduce potential false positives due to high LD of some pQTLs and high correlations among some proteins, we developed a novel multivariate IV regression method for fine mapping to distinguish direct and mediating (causal) effects of proteins; some key features of the method is its robustness to invalid IVs and applicability to GWAS summary data. Application of our method further reduced the total number of causal proteins to nine. Among those identified proteins, APOE, CR1, CD2AP, TREM2 were also validated based on previous studies. Our work highlights some key differences between trans-pQTL and cis-pQTL analyses, and critical values of multivariate IV analysis for fine mapping causal proteins, providing new insights into plasma protein pathways to AD.

25

Genome-wide Association Study of Multiple Neuropathology Endophenotypes Identifies Novel Risk Loci and Provides Insights into Genetic Risk of Dementia

Lincoln M. P. Shade[1], Yuriko Katsumata[1,2], Erin L. Abner[2,3], Khine Zin Aung[1,2], Steven A. Claas[1], Qi Qiao[1,2], Bernardo Aguzzoli Heberle[2,4], J. Anthony Brandon[2,4], Madeline L. Page 0000-0001-9990-1500[2,4], Timothy J. Hohman[5,6], Shubhabrata Mukherjee[7], Richard P. Mayeux[8], Lindsay A. Farrer[9,10,11,12,13], Gerard D. Schellenberg[14,15], Jonathan L. Haines[16,17], Walter A. Kukull[18], Kwangsik Nho[19,20,21], Andrew J. Saykin[20,22], David A. Bennett[23,24], Julie A. Schneider[23,24,25], the National Alzheimer's Coordinating Center, the Alzheimer's Disease Genetics Consortium, Mark T. W. Ebbert* [2,4,26], Peter T. Nelson*[2,27], David W. Fardo*[1,2]

*co-senior authors

[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]Department of Neuroscience, University of Kentucky College of Medicine, Lexington, Kentucky, United States of America; [5]Vanderbilt Memory and Alzheimer's Center, Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [6]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [7]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [8]Department of Neurology, Columbia University, New York, New York, United States of America; [9]Department of Medicine (Biomedical Genetics), Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, United States of America; [10]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, United States of America; [11]Department of Ophthalmology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, United States of America; [12]Department of Biostatistics, School of Public Health, Boston University, Boston, Massachusetts, United States of America; [13]Department of Epidemiology, School of Public Health, Boston University, Boston, Massachusetts, United States of America; [14]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [15]Penn Neurodegeneration Genomics Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [16]Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio, United States of America; [17]Department of Population & Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America; [18]National Alzheimer's Coordinating Center, Department of Epidemiology, University of Washington, Seattle, Washington, United States of America; [19]Department of Radiology & Imaging Sciences, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [20]Indiana Alzheimer's Disease Research Center, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [21]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [22]Center for Neuroimaging, Department of Radiology & Imaging Sciences, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [23]Department of Neurological Sciences, Rush Medical College, Chicago, Illinois, United States of America; [24]Rush Alzheimer's Disease Center, Rush Medical College, Chicago, Illinois, United States of America; [25]Department of Pathology, Rush Medical College, Chicago, Illinois, United States of America; [26]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky College of Medicine, Lexington, Kentucky, United States of America; [27]Department of Pathology & Laboratory Medicine, University of Kentucky College of Medicine, Lexington, Kentucky, United States of America

Background: Genome-wide association studies (GWAS) have identified >80 Alzheimer's disease and related dementias (ADRD)-associated loci. However, the clinical outcomes used in most prior studies belie the complex nature of underlying neuropathologies. As a complementary approach, we performed GWAS on eleven ADRD-related neuropathology endophenotypes from three sources (n=7,804 total participants with autopsy confirmation).

Method: Endophenotypes were harmonized across data sources and included Alzheimer's-related pathologies (tau neurofibrillary tangles, neuritic plaques, amyloid plaques, cerebral amyloid angiopathy [CAA]), non-Alzheimer's proteinopathies (Lewy body and TDP-43), vascular pathologies (microinfarcts, gross infarcts, cerebral atherosclerosis, and brain arteriolosclerosis), and hippocampal sclerosis. We conducted GWAS meta-analysis and tested association between known ADRD-associated variants and neuropathology endophenotypes. Finally, we performed in silico functional genomic and epigenomic studies to identify potential functional pathways by which genetic variants affect neuropathologies.

Result: Eight independent loci were identified, including a locus near APOE (APOC2) associated with CAA independently of APOE e diplotype. 19 of the remaining known ADRD loci were significant for at least one neuropathology after FDR adjustment, 18 of which were concordant in effect direction with ADRD. Genetic colocalization analyses identified pleiotropic effects and quantitative trait loci (QTL). Multiple

methylation QTL and expression QTL also colocalized to either TMEM106B or APOC2. Methylation at two CpG sites colocalizing with CAA were in turn significantly associated with CAA in ROSMAP.

Conclusion: Our findings highlight genetic studies of neuropathology endophenotypes as a necessary next step to understand the mechanisms underlying genetic risk of ADRD.

Keywords: Alzheimer's disease, cerebral amyloid angiopathy, neuropathology

26

Exploring the Potential Causal Association of Gut Microbiota on Panic and Conduct Disorder: A Two-sample Mendelian Randomization Approach

Abiodun Fatoba [1], and Claire Simpson [1]*

[1]*Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America*

*Correspondence: csimps17@uthsc.edu

The potential causal association of gut microbiota with panic disorder and conduct disorder remains unexplored. This study uses detailed Mendelian randomization (MR) analyses to unravel the causal association of specific bacterial taxa with these disorders. Genome-wide association studies summary-level dataset for gut microbiota (n = 18,340), panic disorder (n = 200,486), and conduct disorder (n = 216,179) were retrieved from the MiBioGen and FinnGen consortium, respectively, and subjected to two-sample Mendelian randomization analysis. The inverse-variance weighted method was used as the primary analysis to estimate the causal effect complemented by other MR methods. Sensitivity analyses were also carried out to assess the validity of our results. Based on IVW MR analysis, thirteen bacterial taxa were detected to be causally associated with panic and conduct disorders. Among these 13 bacterial taxa, the genera *Eubacterium hallii* (OR = 1.39, 95%CI = 1.09-1.78; $P = 0.0077$) and *Alistipes* (OR = 1.74, 95%CI = 1.22-2.47; $P = 0.0018$) increased the risk of panic disorder while genus *Coprococcus* (OR = 2.39, 95%CI = 1.16-4.92; $P = 0.0179$) and class *Coriobacteriia* (OR = 2.20, 95%CI = 1.01-4.77; $P = 0.045$) increased the risk of conduct disorder. The other 9 bacteria taxa function as a protective factor as they reduce the risk of the two psychiatric disorders. There was also the absence of horizontal pleiotropy and heterogeneity. Our study provides genetic evidence for the causal association of gut microbiota with panic and conduct disorder. This provides a bedrock for future prevention and treatment of these disorders.

Keywords: Mendelian randomization; Psychiatric disorders; gut microbiota; GWAS.

27

One Population-based Biobank, Two Cohorts: Selection Effects in Biobank Cohorts from the Estonian Perspective

Krista Fischer[12], Māra Delesa-Vēliņa[1], Estonian Biobank Research Team[2]

[1]*Institute of Mathematics and Statistics; University of Tartu, Estonia,* [2]*Institute of Genomics, University of Tartu, Estonia*

Estonian Biobank is one of the oldest and largest population-based biobanks in Europe, with first participants recruited already in 2002. By the end of 2011, first 50000 participants were recruited. In 2018-2019 a new large recruitment campaign was conducted, resulting in the recruitment of additional 150000 participants. Although the final cohort of more than 20% of adult population in the country, linked to electronic health records and national registries and with different layers of available omics data, provides a rich data source for research, it also poses some challenges. We have noticed that the cohorts from the different recruitment periods have remarkably different risk profiles regarding mortality and major complex diseases. Possible reasons for such differences are related to somewhat different recruitment processes for the two cohorts, but as we will show in the talk, also to the differences in follow-up times. For instance, the risk profile in the first cohort was initially different from the general population, but has become more similar to it during time.

In the talk we are discussing the implications of combining cohorts with different risk profiles on the analyses of the risk of incident diseases and mortality, showing that parameter estimates can be seriously biased, if this issue is ignored. We discuss possible solutions for the problem and show that the existence of two cohorts can also be seen as a strength, if used wisely.

28

Cross-omic Characterization of the Molecular Profile of T2D in Hispanic/Latinos

E. Frankel[1], L. Petty[2], R. Roshani[1], W. Zhu[1], H-H. Chen[3], M. Yaser[4], M. Graff[4], M. Krishnan[4], V. Buchanan[4], M. Lee[5], A. Gutierrez[6], H. Highland[4], K. Young[4], J. McCormick[5], S. Fisher-Hoch[5], K. North[4], J. Below[2]

[1]*Vanderbilt University, Nashville, Tennessee, United States of America;* [2]*Vanderbilt University Medical Center, Nashville, Tennessee, United States of America;* [3]*Academia Sinica, Taipei, Taiwan;* [4]*University of North Carolina Chapel Hill, Chapel Hill, North Carolina, United States of America;* [5]*University of Texas School of Public Health, Brownsville, Texas, United States of America;* [6]*University of Texas, Houston, Texas, United States of America*

Relevant categories: OMICs, Statistical Modeling, Diversity & studies of multiple ancestries, Common disease genetics

Type 2 diabetes (T2D) is a chronic metabolic condition characterized by persistently elevated blood glucose with normal or elevated insulin concentration. T2D is a major driver of global health burdens, with a disproportionate impact on US Hispanic or Latinx (HL) populations. Our research leveraged a comprehensively phenotyped HL cohort, the Cameron County Hispanic Cohort (CCHC), in conjunction with robust cross-sectional transcriptomic, proteomic, and metabolomic data to elucidate the biological mechanisms contributing to T2D risk. We evaluated single mRNA, protein, and metabolite linear regressions to compare the expression and abundance of genes, proteins, and metabolite levels between individuals with T2D and controls without diabetes, excluding individuals with prediabetes. The models were adjusted for age, sex, three principal components (PC) of ancestry, and ten surrogate variables (for RNAseq and proteomics). We observed associated 749 genes, 348 proteins, and 19 metabolites differentially expressed surpassing Bonferroni significance ($p < 0.05$ for all analyses) and abundant in T2D cases and controls. Our most compelling study findings were those that generalized across

the multi-omics considered; thirteen genes were implicated in both the transcriptomic and proteomic analyses, including *CD93*, *AOC3*, and the novel gene *TMED8*, which has not been previously associated with T2D. Our preliminary results demonstrate not only how multi-omic studies may elevate understanding of T2D risk in an HL population, but also implicate functional genes, proteins, and metabolites as well as a mechanism of effect of T2D risk in HL individuals.

29

The Role of Pre-Diagnostic Circulating Metabolites in Prostate Cancer Risk: A Cross-Population Meta-Analysis of Untargeted Metabolomic Studies

Harriett Fuller[1], Orietta P. Agasaro[1,2], Peggy Wan[3], Loreall Pooler[3], Lynne R. Wilkens[4] Loic Le Marchand[4], Demetrius Albanes[5], David V. Conti[3], Christopher A. Haiman[3], Burcu F. Darst[1]

[1]Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, Washington, United States of America; [2]Department of Epidemiology, University of Washington, Seattle, Washington, United States of America; [3]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, United States of America; [4]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America; [5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland, United States of America

Introduction: Prostate cancer (PCa) is the 2nd most common cancer in men and presents a major health disparity. While studies have shown that metabolic dysregulation may contribute to pathogenesis, pre-diagnostic metabolite evidence has not been quantitatively aggregated.

Methods: We performed a metabolome-wide association study (MWAS) of pre-diagnostic circulating metabolites (n=831) and PCa risk in African ancestry individuals (355/360 cases/controls) from the Multiethnic Cohort (MEC). Logistic regression models adjusted for batch, birth year and sample collection year were performed. Following a systematic review of pre-diagnostic untargeted circulating metabolomic studies in any population, we meta-analyzed identified studies with MEC findings.

Results: In African ancestry individuals from MEC, 16, 7, and 7 metabolites were nominally ($P<0.01$) associated with overall, aggressive, and lethal PCa risk. No metabolites were significant following multiple testing correction. Our systematic review identified 14 studies (11,768/12,184 cases/controls) of predominately European ancestry. In total, 498, 297, and 591 metabolites were meta-analyzed for overall, aggressive, and lethal PCa. In European ancestry studies, four metabolites were significantly associated with lethal PCa risk: three lipids (3-hydroxybutyrate, glycerol, and stearoyl ethanolamide) were associated with increased risk and one nucleotide (dihydroorate) was associated with decreased risk. In cross-population meta-analyses, these four metabolites were nominally associated with lethal PCa risk and the fatty acid ethylmalonate was significantly associated with lethal PCa (OR=1.45, CI=1.24-1.69, $P_{adj}=0.002$).

Conclusions: We found evidence of associations between five pre-diagnostic metabolites and lethal PCa risk in Europeans and African ancestry populations. Future work should evaluate the clinical utility of these metabolites as lethal PCa biomarkers.

30

Use of Genome-wide Polygenic Risk Scores for Specific Organ Impairment in Coronavirus Disease 2019 (COVID-19) Non-recovery

Anne F. Goemans[1]*; Olivia C. Leavy[1,2]; Beatriz Guillen-Guio[1,2]; Erola Pairo-Castineira[3]; Konrad Rawlik[3]; Rachael A. Evans[2]; Christopher E. Brightling[2]; Louise V. Wain[1,2]; Tim CD. Lucas[1]; on behalf of PHOSP-COVID Collaborative Group

[1]Department of Population Health Sciences, University of Leicester, United Kingdom; [2]The Institute for Lung Health, National Institute for Health and Care Research Leicester Biomedical Research Centre-Respiratory, University of Leicester, Leicester, United Kingdom; [3]Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, United Kingdom

The ongoing symptoms seen post-COVID are still poorly understood and predicting those at highest risk of ongoing health sequelae, and understanding the underlying mechanisms, remains challenging. Individuals who experience ongoing symptoms are commonly defined as having 'long COVID'. However, this label fails to recognise the considerable heterogeneity seen in this patient population and could obscure efforts to detect informative genetic associations. We hypothesised that polygenic risk scores (PRS) for organ impairment might be associated with specific post-COVID symptoms.

Here we apply genome-wide PRS calculated from publicly available lung-related genome wide association data to respiratory-related post-COVID outcomes. We initially applied this methodology using the Post-HOSPitalisation COVID-19 study. We optimised the PRS for association with overall COVID-19 non-recovery and respiratory-related outcomes at 5 and 12 months, adjusting for pre-existing respiratory comorbidities.

A PRS of forced vital capacity (FVC) was not associated with overall non-recovery at 12 months, but an association with reduced breathlessness was nominally significant (OR= 0.82; 95% CI=0.71 – 0.96; *P*=0.0135; Number of variants=472910). At 5 months, a PRS for pulmonary fibrosis was associated with non-recovery (OR= 1.15; 95% CI=1.01 – 1.31; *P*=0.0359; Number of variants=5653), and increased self-reported cough (OR= 1.16; 95% CI=1.02 – 1.33; *P*=0.0235; Number of variants=2464), but reduced breathlessness (OR= 0.82; 95% CI=0.68 – 0.97; *P*=0.0226; Number of variants=63086).

We show that genome-wide PRS can be used to help identify genome-wide genetic overlap between genetic risk and specific outcomes and variant sets for further pathway specific analyses in a poorly understood complex multi-system disease.

Keywords: polygenic risk scores (PRS), long COVID, genetic overlap

31

Statistics to Prioritize Rare Variants in Family-based Sequencing Studies with Disease Subtypes

Christina Nieuwoudt[1,2], Fabiha Binte Farooq[1], Angela Brooks-Wilson[3,4], Alexandre Bureau[5,6], Jinko Graham[1]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada; [2]EMMES Canada, Burnaby, British Columbia, Canada; [3]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, British Columbia, Canada; [4]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, British Columbia, Canada; [5]Département de Médecine Sociale et Préventive, Université Laval, Québec City, Québec, Canada; [6]Centre de recherche CERVO, Québec City, Québec, Canada

Family-based sequencing studies are increasingly used to find rare genetic variants of high risk for disease traits with familial clustering. In some studies, families with multiple disease subtypes are collected and the exomes of affected relatives are sequenced for shared rare variants. Since different families can harbour different causal variants and each family harbours many rare variants, tests to detect causal variants can have low power in this study design. Our goal is rather to prioritize shared variants for further investigation by, e.g., pathway analyses or functional studies. The transmission-disequilibrium test prioritizes variants based on departures from Mendelian transmission in parent-child trios. Extending this idea to families, we propose methods to prioritize rare variants shared in affected relatives with two disease subtypes, with one subtype more heritable than the other. Global approaches condition on a variant being observed in the study and assume a known probability of carrying a causal variant. In contrast, local approaches condition on a variant being observed in specific families to eliminate the carrier probability. Our simulation results indicate that global approaches are robust to misspecification of the carrier probability and prioritize more effectively than local approaches even when the carrier probability is mis-specified.

32

Examining the Shared Genetic Architecture of Keloid Scars and Uterine Fibroids

Catherine A. Greene[1,2], Toni J. Lewis[3], Gabrielle Hampton[1], Jacklyn N. Hellwege[1,6,7,8], Todd L. Edwards[4], Digna R. Velez Edwards[1,2,5]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, [3]Clinical Pharmacology and Gene Department, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4]Division of Epidemiology, Department of Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Department of Biomedical Informatics, 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; [7]Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, [8]VA Tennessee Valley Healthcare System (626), Nashville, Tennessee, United States of America

Keloid scars and uterine fibroids are benign fibroproliferative disorders hypothesized to share genetic etiology. They are highly heritable, frequently comorbid, and characterized by overgrowth of connective tissue. Given these parallels, we aimed to investigate the shared genetic basis of keloids and uterine fibroids, leveraging summary statistics from recent genome-wide association studies (GWAS). We estimated global genetic correlation and examined commonalities in partitioned heritability using Linkage Disequilibrium Score Regression (LDSC). We conducted local correlation analyses using Local Analysis of CoVariant Association (LAVA), filtering 2495 pre-defined LD blocks to 124 with significant variants for at least one trait. Finally, we performed gene set enrichment analysis using Functional Mapping and Annotation of GWAS (FUMA)'s GENE2FUNC module. We observed strong global correlations between keloids and uterine fibroids. The cross-ancestry result was 0.63 (0.05, $p=3.98\times10^{-39}$), with similar ancestry-stratified values (mean=0.67 [0.11]). Both traits exhibited significant ($p<9.4\times10^{-4}$) functional enrichment in annotations associated with increased enhancer activity (histone marks H3K4me1 and H3K27ac, SuperEnhancers), suggesting gene regulation via enhancers may be important in fibroproliferative disease. Univariate analyses with LAVA found significant ($p<4.03\times10^{-4}$) local heritability for both traits in 81 loci. Two were significantly ($p<6.2\times10^{-4}$) locally correlated, while 15 loci were nominally significant ($p<0.05$). We mapped suggestive ($p<1\times10^{-5}$) GWAS variants within nominally correlated regions and identified overlaps with GWAS Catalog gene sets for obesity and immune measures, in addition to enrichment of genes involved in oncogenic pathways. This preliminary characterization provides functional insight into common origins of keloids and uterine fibroids.

Keywords: genetic correlation, fibroproliferative, functional annotation

33

Populations and Methods Used to Define Prostate Cancer Polygenic Risk Score Categories Impact the Interpretation of Risk

Boya Guo[1], Ali Sahimi[2], Xin Sheng[3], Fei Chen[3], Christopher A. Haiman[3], David V. Conti[3], Burcu F. Darst[1]

[1]Fred Hutchinson Cancer Center, Seattle, Washington, United States of America; [2]New York University Grossman School of Medicine, New York City, New York, United States of America [3]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, United States of America

Polygenic risk scores (PRS) have the potential to identify individuals at higher risk of prostate cancer (PCa) and inform screening. However, PRS interpretation depends on the comparison population. We investigated how different methods to define PRS risk categories affect interpretation. We evaluated a multi-ancestry PCa PRS in 5,288 cases and 12,754 controls from self-identified European (EUR), African (AFR), East Asian (EAS), and Hispanic (HIS) populations in the Multiethnic Cohort (MEC). The PRS was based on 451 variants identified in

a prior PCa GWAS across four ancestries. MEC participants were categorized into PRS deciles within their populations using: 1) MEC controls' PRS distribution; 2) MEC controls' PRS distribution after adjusting the PRS for the first 10 principal components (PCs); 3) External reference populations' PRS distribution matched by continental ancestry; and 4) personalized PRS deciles calculated as a linear combination of ancestral proportions and external reference populations' decile cutoffs. Reference populations for methods 3 and 4 included 1,890 EUR, AFR, EAS, and AMR individuals from 1KGP and the PAGE Global Reference Panel, chosen for their close resemblance to respective populations based on PCA. Logistic regression models adjusted for age and the first 10 PCs were used to evaluate the association between PRS deciles and PCa risk. We found high variability in the PRS predictive ability across methods and populations. For instance, compared to the 40%-60% PRS category, the OR for men in the top 90%-100% PRS decile ranged from 2.72 (95% CI: 1.31-5.64; method 3) to 5.51 (95% CI: 2.07-5.51; method 4) in AFR. PRS interpretation and individual risk categorization depend on the populations and methods used to define PRS risk categories, which have clinical implications and are crucial to consider for implementing PRS clinically.

Topic areas/categories: General polygenic trait genetics, diversity & studies of multiple ancestries

Keywords: polygenic risk scores, prostate cancer, diverse populations, genetic epidemiology, multi-ancestry

34

Epigenome-wide Association Study of Placental DNA Methylation and Cumulative Maternal Glycemic Levels Throughout Pregnancy

Tesfa Dejenie Habtewold[1], Prabhavi Wijesiriwardhana[1], Richard J. Biedrzycki[2], Fasil Tekola- Ayele[1]

[1]*Epidemiology Branch, Division of Population Health Research, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, United States of America;* [2]*Glotech, Inc., contractor for Division of Population Health Research, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, United States of America*

Background: Placental DNA methylation (DNAm) is a potential mechanism underlying the association between maternal prenatal glycemic dysregulation and pregnancy complications. We investigated whether maternal cumulative glycemic trait concentration throughout pregnancy is associated with placental DNAm in multiethnic women.

Methods: Cumulative glycemic exposure across four time point measurements of pregnancy was estimated using area under the curve (non-fasting glucose (AUC*gluc*), insulin (AUC*insl*), and HbA1c (AUC*hba1c*)) in 301 pregnant women from the NICHD Fetal Growth Studies-Singletons cohort. Association of cumulative glycemic trait concentration with epigenome-wide placental DNAm (450K array) was tested using linear regression models adjusted forcovariates. We tested significantCpGs' correlation with nearby gene (±200KB) expression in placenta in a subset of 75 samples, and enrichment for biological pathways was assessed. The associated CpGs were looked-up in the EWAS catalog to check their overlap with previously reported diseases/traits.

Results: Maternal AUC*gluc* and AUC*insl* were associated with placental DNAm at 7 and 178 CpG sites (p*FDR* < 0.05), respectively. Ten AUC*insl*-associated CpGs were correlated with expression of 10 genes, the strongest correlation between cg21233754 and *GPD2* (r = -0.45, p = 5.7x10-5). The 10 genes include transcripts highly expressed in brain, reproductive, and cardiovascular tissues. Significant CpGs were enriched for various pathways, including nitric oxide signaling in cardiovascular system and insulin secretion signaling pathway. We observed overlap with CpGs previously associated with diseases/traits, including cardiometabolic and psychiatric disorders.

Conclusions: Our analyses suggest a potential placenta-mediated epigenetic mechanismin the link between glycemic dysregulation throughout pregnancy and maternal and offspring health.

35

Cell Type Prediction in Spatial Transcriptomics by Projection of Validated Single-Cell RNA Reference Datasets in Seurat

Gideon B. Hallum*[1], Courtney G. Montgomery[1], Nathan Pezant[1], Christopher A. Bottoms[1], David R. Stanford[1]

*Presenting Author

[1]*Center for Biomedical Data Sciences, OMRF, Oklahoma, United States of America*

Spatial Transcriptomics (ST) is a rapidly expanding field that is burgeoning with new methods and technologies. These practices allow investigators to get the resolution needed for spatial inferences and answer new questions within their desired field. The cost of this spatial resolution however, is lack of transcriptomic coverage. That is, researchers must narrow the genes to be probed within their tissue of interest from a few thousand to as few as 180. This becomes a problem when investigators want to answer a question specific to certain cell types within their spatial context. Specifically, probe sets for ST of this size often do not allow for inclusion of enough genes for both cell-type identification and specific to a hypothesis of interest. Single Cell RNA (scRNA) sequencing data can provide a resource to help solve this problem. More and more validated scRNA datasets are becoming available to serve as atlases for the expression data from a variety of cell types found within a given tissue. These atlases, in conjunction with Seurat tools, have been used for the projection of other scRNA datasets, but little has been done to evaluate their use in the context of ST. In this study, we leveraged Seurat to project data from the greatly reduced gene sets from ST onto known atlases to obtain cell-type specific assignment then placing cells back into spatial context. We assessed the performance of this approach and present comparisons across multiple tissue and data types.

36

Methods for Accurately Estimating Hereditary Cancer Prevalence in Diverse U.S. Biobanks in the Presence of Participation Bias

Sarah C Hanks[1*], Ruhollah Shemirani[1], Christa Caggiano[1], Kathleen DM Ferar[1], Emily R Soper[1], Eimear E Kenny[1]

[1]*Institute for Genomic Health, Icahn School of Medicine at Mount*

Sinai, New York, New York, United States of America

Electronic health record-linked biobanks have facilitated cost-effective genetic discovery and translational efforts, including the development of genomic screening tools for medically actionable diseases. However, the nonrandom selection of study participants into biobanks can bias estimates of disease prevalence and genetic associations derived from these resources. This participation bias can lead to misinformed disease risk management and incomplete or inefficient genomic screening guidelines. Here, we estimate the impact of participation bias on the prevalence of CDC-designated Tier 1 medically actionable genetic diseases, including hereditary breast and ovarian cancer (HBOC) and Lynch syndrome (LS), in the Bio*Me* and All of Us biobanks. We will begin by identifying and comparing drivers of participation across the biobanks. We will then compare multiple strategies to derive survey weights for calculating the adjusted prevalence of HBOC and LS in the biobanks. These strategies will include leveraging information about demographics, disease, and social determinants of health from external, nationally representative epidemiological data sources and implementing methods to detect the "genetic footprint" of participation from related individuals within the biobanks. The latter strategy, developed in the UK Biobank, has not yet been tested in ancestrally heterogenous populations present in Bio*Me* and All of Us. To implement this method, we have called IBD (Identity-by-descent) tracts in 859 full sibling pairs and 1,735 parent-offspring pairs in the Bio*Me* biobank. We will evaluate the adjusted HBOC and LS prevalence estimates against the SEER cancer registry. We expect our findings to inform genetic screening guidelines, biobank recruitment efforts, and community risk awareness.

37

Proteome-Wide Mendelian Randomization Identifies Potential Causal Circulating Proteins for Colorectal Cancer Risk

Sihao Han [1*], Jiemin Yao [1], Cory Lumsdaine [1], Lang Wu [2], Hajime Yamazaki [3,4], Samantha A. Streicher [2], Roch A. Nianogo [1], Jian-Yu Rao [1,5], Zuo-Feng Zhang [1], and Brian Z. Huang [6,7]

[1]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, California, United States of America; [2]Cancer Epidemiology Division, Population Sciences in the Pacific Program, University of Hawaii Cancer Center, University of Hawaii at Manoa, Honolulu, Hawaii, United States of America; [3]Section of Clinical Epidemiology, Department of Community Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan; [4]Center for Innovative Research for Communities and Clinical Excellence (CiRC2LE), Fukushima Medical University, Fukushima, Japan; [5]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America; [6]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America; [7]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America

Background: Colorectal cancer (CRC) remains a major global health concern, necessitating investigations into novel biomarkers and therapeutic targets. Plasma proteins are pivotal in cancer dynamics and present a promising avenue for CRC prevention and treatment. This study aims to identify circulating proteins that might be causally associated with CRC risk using Mendelian Randomization (MR).

Methods: We employed two-sample MR to evaluate the associations of >4,000 circulating proteins with CRC risk. Genetic instruments for protein concentrations were obtained from three large-scale genome-wide association studies (GWAS) with proteomics (deCODE Health Study, Fenland Study, and UK Biobank). The association between these instruments and CRC risk was obtained from CRC GWAS summary statistics from FinnGen. We performed *cis*-MR and *cis*+*trans* MR analyses and validated the identified associations using Bayesian colocalization.

Results: Proteome-wide *cis*-MR analyses identified four unique circulating proteins with putative causal associations with CRC risk (false discovery rate $P<0.05$): greminlin-1 (GREM1), lactase/phlorizin hydrolase (LPH), cGMP-specific 3',5'-cyclic phosphodiesterase (PDE5A), and LIM domain and actin-binding protein 1 (LIMA1). GREM1, PDE5A, and LIMA1 were positively associated with CRC risk, whereas LPH showed an inverse association. Fifteen additional proteins were identified in the *cis*+*trans* MR analysis. We also observed suggested evidence of colocalization for LPH and PDE5A, indicating potential shared genetic etiology between these proteins and CRC risk.

Conclusions: This study identified multiple plasma proteins that may play causal roles in CRC pathogenesis and can perhaps serve as promising therapeutic targets. Future research is needed to validate these proteins as potential targets in CRC prevention and treatment.

38

The Problem with Proteins: What Mendelian Randomization Can (and Can't) Reveal

Emma Hazelwood[1]*†, Matthew A. Lee[2]†, Gibran Hemani[1], Emma Vincent[1]

[1]University of Bristol, Bristol, United Kingdom; [2]International Agency for Research on Cancer, Lyon, France

* Presenting author

† These authors contributed equally and are co-first authors

The use of Mendelian randomization (MR), a statistical approach which uses genetic variants to instrument exposures, has increased exponentially in recent years. However, the inherent limitations of using genetic variants to instrument highly complex molecular traits (molecular MR) – which often include feedback loops, protein-protein interactions, and tissue-specific effects – have received little attention. Here, we discuss considerations and applications for molecular MR, using protein expression as an example.

We (1) highlight challenges in distinguishing proteins as causes or consequences of disease; (2) consider how causal estimates may be misinterpreted when obscured by protein function; and (3) question the value of protein expression as an exposure.

Integrating data across MR and genetic colocalization analyses, we show that systemic and tissue-specific expression are not equivalent and may be driven by distinct pathways. We further show, using structural proteins with a role in tissue integrity as an example, that protein function can

mask underlying causal pathways in MR analyses. Finally, we challenge the utility of protein expression as a biological measure and suggest protein activity as an informative alternative for understanding causal effects.

Given these considerations we ask what can (and cannot) be ascertained from current molecular MR analyses. We highlight complimentary methods (e.g., genetic colocalization), approaches (e.g., single-cell MR), and under-utilised tools (e.g., prior-knowledge-networks and variant effect prediction) which can aid in deciphering underlying molecular mechanisms. We advocate for triangulation and validation of molecular MR results across disciplines, incorporating evidence from intervention trials, conventional observational research, and laboratory-based analyses.

39

Detecting Phase Effects Using Long-Read Sequencing Data

Gengming He[1,2], Scott Mastromatteo[2,3], Katherine Keenan[2,3], Lisa Strug[1,2,4,5,6]

[1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto; [2]Program in Genetics and Genome Biology, The Hospital for Sick Children; [3]Program in Translational Medicine, The Hospital for Sick Children; [4]The Centre for Applied Genomics, The Hospital for Sick Children; [5]Department of Statistical Sciences, University of Toronto; [6]Department of Computer Science, University of Toronto

Accurate phasing of genetic variants across extensive regions, i.e., determining whether they are on the same homologous chromosomes (in-*cis*) or different ones (in-*trans*), is now made possible by long-read sequencing. Phase can affect disease outcome but is often neglected by traditional genotype-based interaction analysis. We propose a regression approach to make inference on the *cis*/*trans* effects of two variants with an additive model for their phase. In simulation the method has correct type 1 error when both loci have main effects and are in linkage disequilibrium (LD), and demonstrates greater power than genotype-based tests, especially for detecting *trans* effects with low allele frequencies. We show even greater power when analyzing phase effects across different LD blocks.

We applied our test to phased sequence from n=564 individuals with cystic fibrosis (CF) using the 10X Genomics linked-read technology at two modifier loci, *SLC6A14* and *PRSS1/PRSS2*. We found significant *cis* effects between an enhancer variant and a promoter variant at the *SLC6A14* locus that associates with CF lung disease and meconium ileus (MI), respectively, with elevated *SLC6A14* CF airway expression (n_males=40, p_males=0.05; n_females=39, p_females=0.03) and early onset bacterial infection (n_males=105, p_females=0.24; n_females=112, p_females=0.05). The MI-associated variants at the *PRSS1/PRSS2* locus include a 20 kb deletion and a missense variant rs62473563, both regulating *PRSS2* expression in the pancreas. No significant phase effect between the two variants was detected (n=309, p=0.95), suggesting they may contribute to MI independently. These analyses on cis/trans effects help understand how phase affects the joint effects of genetic variants.

40

100,000 Genomes of Europe: Unlocking Genetic Variability Across Europe for Science and Health

Anthony F. Herzig[1], Astrid Vicente[2,3], Hugo Martiniano[2,3], Emmanuelle Génin[1,4], Helen Ray-Jones[5], Jeroen van Rooij[5], André G. Uitterlinden[5], on behalf of the GoE/1+MG consortium

[1]Inserm, Univ Brest, EFS, UMR 1078, GGB, 29238 Brest, France; [2]Instituto Nacional de Saúde Doutor Ricardo Jorge, Av Padre *Cruz, 649-016 Lisbon Portugal; [3]Biosystems and Integrative Sciences Institute (BioISI), Faculdade de Ciências da Universidade de Lisboa, Lisboa Portugal; [4]CHRU Brest, 29200 Brest, France; [5]Laboratory for Population Genomics, Erasmus MC, Rotterdam, The Netherlands*

Background/Objectives: As part of the 1 million genomes (1+MG) initiative, the Genome of Europe (GoE) project will create a pan-European reference database comprising whole-genome sequencing (WGS; 30x) data from >100,000 European citizens. Spanning 51 contributing partners from 29 countries providing existing and de novo datasets (short-read and long-read WGS), GoE presents enormous potential for understanding the genetic dimension to public health in Europe; including through interactions with the European Rare Disease Research Alliance (ERDERA).

Methods: Challenges for GoE are evident across three axes: technological, statistical, and ethical, in order for GoE to provide wide-ranging utility. Here we present the scope of the "use-cases" of GoE, which will establish the envisaged applications of GoE.

Results: The following pilot studies are outlined: (1) Building a fine-scale map of genetic structure across Europe to inform aggregating individual-level data. (2) Creating a reference database for individual variant look-ups; with particular attention on clinically relevant variant frequencies in actionable-disease and pharmacogenetic genes. (3) Building reference panels for ancestry-informed genetic imputation and the required secure informatics environments. (4) Deriving distributions of polygenic (risk) scores, calibrated for ancestry gradients in Europe; with a focus on cancer phenotypes in collaboration with the Can.Heal project. (5) Understanding the added value of long-read sequencing in the general population and exploring the 'dark regions of the human genome'.

Conclusion: We lay out here the methodological advances and adjustments necessary to best unlock the potential of GoE; particularly in the context of the prospective GoE federated data embedding in 1+MG.

41

Metabolomic Analysis Indicates Adenine, Oleic acid, and Glutamic Acid are Associated with Flare Remission, DNA Methylation Changes, and Clinical Subtypes in Systemic Lupus Erythematosus

Mary K. Horton[1], Dylan Johnson[2], Joanne Nititham[1], Kimberly E Taylor[3], Patricia Katz[3], Chun Jimmie Ye[3], Jinoos Yazdany[3], Maria Dall'Era[3], Lisa Barcellos[4], Jian-Liang Li[2], Kirsten E. Overdahl[5], Alan K. Jarmusch[5], Lindsey A. Criswell[1], Cristina Lanata[1]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [2]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, National Institutes of Health,

Durham, North Carolina, United States of America; [3]Division of Rheumatology, Department of Medicine, University of California, San Francisco, San Francisco, California, United States of America; [4]Division of Epidemiology, School of Public Health, University of California, Berkeley, Berkeley California, United States of America; [5]Metabolomics Core Facility, National Institute of Environmental Health Sciences, National Institutes of Health, Durham, North Carolina, United States of America

Systemic lupus erythematosus (SLE) treatments neither adequately prevent flares nor disease-related organ damage. We previously identified longitudinal changes in DNA methylation associated with flare remission and defined 3 patient clusters using methylation changes. We aimed to integrate metabolomics into this analysis to identify whether changes in metabolites were associated with flare remission and methylation.

We utilized 40 multi-ethnic SLE patients recruited during an active flare with whole blood methylation profiles (EPIC array) and untargeted metabolomics from plasma (UHPLC-HRMS/MS) at a baseline flare and follow-up visit ~3 months later. We observed metabolites (with annotation) that changed between visits differentially by whether a patient remitted at the follow-up visit (disease activity score=0) or not using *limma*. Significant metabolite changes (FDR q<0.05) were tested for their association with methylation changes using correlation.

Sixteen patients remitted by follow-up. Remitters and non-remitters did not differ by race, ethnicity, age, clinical features at flare, or medications. Three metabolite changes were associated with remission status: oleic acid (difference in change comparing remitters to non-remitters: 2.52, 95%CI: 1.39, 3.35), adenine (1.61, 95%CI: 0.88, 2.34), and glutamic acid (-1.63, 95%CI: -2.42, -0.83). Seven methylation changes were highly correlated with metabolite changes (correlation >|0.7| and Bonferroni p<0.05). These included cg17847344 (within *EBF1*, a B-cell transcription factor) and oleic acid (correlation=-0.79, p=1.06x10^{-7}) and cg10685380 (within *IL12B*, encodes a T cell cytokine) and adenine (0.70, p=1.26x10^{-6}).

The integration of DNA methylation and metabolomics might help us better understand underlying biological pathways relevant for SLE flare remission and result in targeted therapies.

42

Genomic Co-localization, Child Proteomics and Brain Imaging Support a Link Between Obesity-associated Genotype and Child Language Development.

Jian Huang,[1,2,3,*] Michelle Z.L Kee,[1] Jonathan Yinhao Huang,[1,4,5] Dennis Wang[1,2,6]

*[1]Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR), Singapore; [2]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore; [3]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London, United Kingdom; [4]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore; [5]Thompson School of Social Work & Public Health, Office of Public Health Studies, University of Hawaii at Mānoa, Honolulu, Hawaii, United States of America; [6]National Heart & Lung Institute, Imperial College London, London, United Kingdom*

Exploring the link between genetic predisposition to obesity and language development, our study analyses longitudinal data from a Singaporean birth cohort. We employed a trans-ancestry polygenic risk score (PRS) method, PRS-CSx, to enhance the estimation of genetic effect sizes across different populations. Specifically, we constructed PRS for obesity for the GUSTO children and parents using an East Asian genome-wide association study (GWAS) of body mass index (BMI) from BioBank Japan (N=163,835) and a European GWAS of BMI from a meta-analysis of GIANT and UK BioBank (N=681,275). Child PRS for obesity was inversely associated with child language scores as assessed by the Wechsler Individual Achievement Test, Third Edition (WIAT-III) composite score assessed at age 9. These associations were stronger in boys (β=-0.56, 95%CI -0.81 to -0.31, P value=2.5×10^{-5}) compared to girls (β=-0.27, 95%CI -0.53 to 9.8×10^{-5}, P value=0.050). Genetic correlations and colocalization suggest a complex interaction between obesity-related traits and language-related skills. Specifically, investigation in the neurology-related proteins and brain imaging suggested potential roles of inflammation mechanisms. However, we did not identify a causal relationship using two-sample Mendelian randomization. Intriguingly, we identified a connection between obesity predisposition and elevated MSR1 protein levels, whereas EFNA4, VWC2, and CNTN5 protein levels were associated with a higher WIAT-III composite score. Our study highlighted neurology-related proteins influenced by obesity-associated genotypes and those implicated in language development. Expression levels of MSR1, EFNA4 and VWC2 were associated with fractional anisotropy in white matter tracts, indicating potential mechanisms involving myelination.

Keywords: obesity, language development, polygenic risk score, colocalization, Mendelian randomisation

43

Bridging Histology to Spatial Transcriptomics: A Pathology Foundation Model-driven Contrast Learning for Predicting Spatial Transcriptomic Profiles from Histology Images

Zi Huai Huang[1], Ziyang Xu[1], 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

*Contact email: phu49@uwo.ca

Spatial transcriptomics (ST) has revolutionized cancer research by providing detailed insights into cellular arrangements and gene expression patterns within tumor microenvironments. This spatial information holds immense potential for refining patient stratification and tailoring treatment strategies. However, the prohibitive costs and expertise required for ST data acquisition hinder its widespread clinical adoption. In contrast, conventional hematoxylin and eosin (H&E) stained histology images, entrenched in routine clinical practice, exhibit predictive capacity for ST data.

We hypothesize that a deep learning-based framework will be able to predict ST data from unlabeled breast cancer (BC) patients in The Cancer Genome Atlas (TCGA). Furthermore, the predicted ST data can be used to predict the immune checkpoint inhibitor (ICI) treatment response of BC patients. We developed a digital pathology foundation model-driven contrastive learning to predict spatial transcriptomic profiles from histology images, which is named FMCL2ST. Comprehensive testing on multiple datasets shows that the FMCL2ST significantly outperforms existing methods in terms of gene expression prediction and spatial region identification, and better retains biological information. By employing the BC subtype-specific cell types identified from predicted ST data, we generated a list of BC subtype-specific and cell type-specific immune-related gene signatures. These gene signatures were used to build classification models that can accurately predict the response status (responder vs no responder) of BC patients treated with immune checkpoint inhibitors (ICI) in I-SPY2 clinical trials. Our study bridges histology and spatial transcriptomics, offering a promising avenue for precision oncology in BC.

44

Genome-wide Association Studies in a Large Korean Cohort Identify Novel Loci for 36 Quantitative Traits and Illuminate Their Genetic Architectures

Yon Ho Jee[1*], Ying Wang[2,3], Keum Ji Jung[4], Ji-Young Lee[4], Heejin Kimm[4], Rui Duan[5], Alkes L. Price[1,5,6], Alicia R. Martin[2,3,7], Peter Kraft[1,8]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [3]Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [4]Institute for Health Promotion, Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea; [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America; [6]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [7]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [8]Transdivisional Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Maryland, United States of America

Genome-wide association studies (GWAS) have been predominantly conducted in populations of European ancestry, limiting opportunities for biological discovery and generalizability. We report GWAS findings from 153,950 individuals across 36 quantitative traits in the Korean Cancer Prevention Study-II (KCPS2) Biobank. We discovered 616 novel genetic loci in KCPS2. Thyroid-stimulating hormone, which is not typically measured at baseline in biobanks, shows a particularly high novelty rate in KCPS2 (55/100 loci); one such novel locus includes a missense variant in *CD36*. Meta-analysis of 21 traits across KCPS2, Korean Genome and Epidemiology Study, Biobank Japan, Taiwan Biobank, and UK Biobank (N_{total}=928,679) identified 11,861 loci, 3,524 of which were not significant in any of the contributing GWAS. We assessed the genetic architecture of these traits and demonstrated different heritability estimates across East Asian cohorts as well as across East Asian and European ancestry populations, reflecting differences in study design, sample size, and linkage disequilibrium. We also highlight associations with alleles that are common in East Asian but rare in European populations, including a known pleiotropic missense variant in *ALDH2*, which fine-mapping identified as a likely causal variant for liver enzymes, blood pressure, and lipid values. Our findings provide insights into the genetic architecture of complex traits in East Asian populations and highlight how broadening the population diversity of GWAS participants can aid discovery. By increasing the sample size and ancestral diversity of GWAS samples, our analysis may help identify novel targets for prevention and treatment and offer equitable access to precision medicine to diverse populations.

Keywords: Genome-wide association, Quantitative traits, Genetic diversity, Complex traits Genetic epidemiology

45

Rare Variant Intensity Estimation for Genetic Mapping of Complex Traits

Jing-Rong Jhuang[1, #], Wan-Yu Wei[1, 2, #], Yin-Chun Lin[1], Chao-Yu Guo[2], Hsin-Chou Yang[1, 2, *]

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

[#]Contributed equally

[*]Corresponding author

Rare causal variants often underlie complex disorders within the general population. With the advent of next-generation sequencing technologies, there's been a significant stride in assessing the impact of these rare variants on complex traits at a large scale. However, analyzing such copious sequencing data poses both opportunities and challenges, primarily revolving around the limited statistical power of rare-variant association analyses. This study introduces a novel approach for genetic mapping in complex traits through rare variant intensity estimation. The methodology involves two key steps. Firstly, employing the nearest neighbor method to generate sliding windows along each chromosome. Secondly, utilizing local polynomial regression within each window to estimate the rare variant intensity. This regression incorporates local weights, reflecting the proximity of genetic markers to an anchor point, and locus weights, accounting for the pronounced effect of a rare variant on the complex trait. Subsequently, the generalized linear model is employed to ascertain associations between the rare variant intensity estimates and the complex trait. Through simulation studies, the performance of the proposed method is compared against conventional approaches such as the burden test, the variance component tests, and the combination of burden and variance component tests. Results indicate that the proposed method exhibits superior statistical power and better type I error control. Additionally, analysis of a dataset from Genetic Analysis Workshop 19 using the proposed method identifies several genes associated with hypertension and systolic and diastolic blood pressure.

Keywords: next-generation sequencing, complex traits, association testing, rare variant intensity estimation, local polynomial regression.

46

Genetic Analysis of Fibrotic Multi-morbidity

Joof E[1,2], Massen GM[8], Leavy OC[1,2], Parcesepe, G[1,2], Stewart I[3], Aithal GP[4,5], Scotton CJ[6], Auer DP[4,5,7], Francis S[5,7], Jenkins RG[3], Quint JK[8], Wain LV[1,2], Allen RJ[1,2], Longhurst H[9], DEMISTIFI Consortium

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, United Kingdom; [3]National Heart & Lung Institute, Imperial College London, London, United Kingdom; [4]Mental Health & Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom; [5]NIHR Nottingham Biomedical Research Centre, Nottingham University Hospitals NHS Trust and the University of Nottingham, Nottingham, United Kingdom; [6]Respiratory Medicine, University of Exeter, Exeter, United Kingdom; [7]Sir Peter Mansfield Imaging Centre, University of Nottingham, Nottingham, United Kingdom; [8]School of Public Health, Imperial College London, London, United Kingdom; [9]Dyskeratosis Congenita (DC) Action, United Kingdom

Fibrosis (scarring or hardening of tissue) is a pathological feature that affects many organs, and individuals are often affected by fibrotic disease in more than one organ; but there is no good treatment. One way to understand and address this phenomenon is to target genetic mechanisms associated with fibrotic multi-morbidity (FMM). We sought to identify genetic variants associated with FMM.

We included unrelated European ancestry individuals from UK Biobank. Diseases were assigned as "always fibrotic" or "broad fibrotic" (i.e. can develop fibrosis) based on a published Delphi survey and were assigned to one of 11 organ groups (pulmonary, liver, biliary, skin, cardiovascular, intestinal-pancreatic, skeletal, systemic, reproductive, renal and diabetes). Individuals with an "always fibrotic" disease in at least one organ and any fibrotic disease (always or broad) in at least one other organ were defined as FMM cases. All other individuals were eligible for selection as controls, and matching by age and sex was used to select 20 controls per case. Genome-wide association study (GWAS) was performed to test the association between each genetic variant across the genome with FMM.

There were 4,681 FMM cases and 93,620 controls included in the GWAS. We identified four FMM associated signals meeting genome-wide significance ($p<5x10^{-8}$) and two additional signals meeting suggestive significance ($p<5x10^{-7}$). Two of the signals, one near *CACNA2D1* and the other near *FLRT2*, have never been reported for association with fibrosis.

Functional follow-up of these signals may reveal potential causal mechanisms of fibrotic multi-morbidity. We are seeking replication using independent cohorts.

47

KidneyGenAfrica: Genome-wide Association Studies for eGFR on 110,000 Africans and Assessment of Polygenic Prediction

Abram B Kamiza, June Fabian, Jean-Tristan Brandenburg, Segun Fatumo and members of KidneyGenAfrica

[1]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences,University of the Witwatersrand, Johannesburg, South Africa; [2]Medical Research Council/Wits University Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [3]Department of Non-commincable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

Background: Most genome-wide association studies (GWAS) on estimated glomerular filtration rate (eGFR) have been performed in Europeans. To increase discovery of additional loci we performed a meta-analysis in 110,000 individuals of African ancestry.

Methods: We performed a three-stage meta-analysis: (1) regional meta-analyses within East, West, and South Africa; (2) Continental African meta-analysis of East, West, and South Africa; and (3) Pan-African meta-analysis using data from continental Africa, MVP, UKBB, and CKDGEN. Meta-analyses were performed using GWAMA, METAL and Metasoft, respectively. We also performed fine-mapping, colocalization, and PheWAS. Polygenic risk scores (PRSs) were developed from GWAS summary statistics using data from continental Africans, Africans of Diaspora, and Pan-African ancestry, and tested and validated in a Malawi cohort.

Results: We identified six novel loci in regional meta-analysis. Pan-African meta-analysis detected an additional 20 independent loci, including six novel loci. These loci were mapped to genes involved in kidney function. Our fine mapping reduced the credible set size and identified eight loci with a posterior probability of causality > 0.99. Colocalization recapitulated known eGFR-related genes, and PheWAS identified 26 loci, in addition to loci associated with cardiometabolic and immunological traits. The pan-African ancestry-derived PRSs performed and predicted better than continental and African diaspora PRSs.

Conclusion: We identified novel region-specific loci in continental Africa. By incorporating data from continental Africa and the diaspora into PRSs, we enhanced their accuracy, underscoring the critical role of genetic diversity in ensuring the equitable application of PRS across different populations.

Keywords: eGFR, Kidney Function, meta-analysis, Africa, GWAS, PRSvycdef5 rgfdcx

48

Mendelian Randomization Borrowing Strength via Shared Pleiotropic Pathways Across Populations

Bowei Kang[1], Yihao Lu[1], Ke Xu[1], Lin S. Chen[1*]

[1]Department of Public Health Sciences, The University of Chicago, Chicago, Illinois, United States of America

[*] Correspondence author

Two-sample Mendelian randomization (MR) assesses the causal effect of exposure on outcome by utilizing genetic variants as instrumental variables (IVs) and integrating summary statistics from genome-wide association studies (GWASs). Classic MR assumptions are violated when IVs are associated with unmeasured confounders, creating correlated horizontal pleiotropy (CHP). Identifying invalid IVs with CHP is challenging when sample sizes are limited. In this work, we propose a multi-population MR method for jointly estimating the causal effect

across two or more populations and identifying IVs with CHP, leveraging at least partially shared pleiotropic pathways across populations. Our approach improves causal estimation and inference in populations with limited sample sizes by borrowing information from populations with larger samples. Using the IVs identified with CHP, we map the cis-genes, which in turn inform pathways and etiology shared across populations. We applied our method to identify risk factors for complex diseases across multiple ethnic groups. Additionally, we used our method to identify risk factors for AD in the oldest-old, defined here as age above 95, by borrowing information from samples of all ages. Our method can also be applied to analyze multiple correlated outcomes.

Keywords: Mendelian randomization, correlated horizontal pleiotropy, multi-population

49

A New Multi-trait Fine-mapping Method Using a Non-local Prior, with Applications in Circulating Metabolic Biomarker Level Analysis

Ville Karhunen[1], Johannes Kettunen[2], Stephen Burgess[1,3,], Marina Evangelou[4,*], Mikko J. Sillanpää[5,*]*

*[1]MRC Biostatistics Unit, University of Cambridge; [2]Research Unit of Population Health, University of Oulu; [3]Cardiovascular Epidemiology Unit, University of Cambridge; [4]Department of Mathematics, Imperial College London; [5]Research Unit of Mathematical Sciences, University of Oulu. *Equal contribution.*

Fine-mapping aims to identify independent causal variants within a particular genomic locus, adjusting for linkage disequilibrium (LD) patterns. Despite a number of existing fine-mapping methods for summary-level data, there is still room for improvement in the joint analysis of multiple traits, dealing with a misspecified LD reference, and incorporating functional annotation of the variants.

We introduce Multi-FiniMOM (multi-trait fine-mapping using inverse-moment priors), a new Bayesian method for simultaneous fine-mapping of multiple traits based on summary-level data. The method builds on the previously published fine-mapping approach FiniMOM, which uses a non-local inverse-moment prior for the effect sizes. In Multi-FiniMOM, the model formulation includes a hyperparameter which allows controlling for LD misspecifications, and includes the possibility to include variant-specific prior information. Rapid Markov Chain Monte Carlo sampling from the posterior distribution is achieved by using a recently proposed version of Laplace's method for integral approximation, with running times comparable to variational inference approaches.

We evaluate the performance of the method in simulations, and compare the results to another multi-trait fine-mapping method mv-SuSiE. We demonstrate that a particular strength of our method is its robustness to misspecifications in the LD reference panel. Finally, through an applied example of genetic associations for circulating metabolic biomarker levels within the *APOB* gene locus, we show that the method can distinguish both shared and distinct causal variants for several phenotypes. The method is implemented in a freely available R package: https://vkarhune.github.io/finimom.

Keywords: Fine-mapping, Bayesian inference, linkage disequilibrium

50

Maternal Health in Pregnancy and Autism Risk – Genetic and Non-genetic Mechanisms

Vahe Khachadourian[1], Elias Speleman Arildskov[2], Jakob Grove[2], Paul F O'Reilly[3], Joseph D Buxbaum[1,3], Abraham Reichenberg[1,4], Sven Sandin[1,5], Lisa A. Croen[6], Diana Schendel[7,8,9], Stefan Nygaard Hansen[10], Magdalena Janecka[1,3,11,12]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [2] Department of Biomedicine, Aarhus University, Aarhus, DK; [3] Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [4] Department of Environmental Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [5]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [6]Division of Research, Kaiser Permanente Northern California, Oakland, California, United States of America; [7]A.J. Drexel Autism Institute, Drexel University, Philadelphia, Pennsylvania, United States of America; [8]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark; [9]National Centre for Register-Based Research, Aarhus BSS, Aarhus University, Aarhus, Denmark; [10]Department of Public Health, Aarhus University, Aarhus, Denmark; [11]Department of Child and Adolescent Psychiatry, NYU Grossman School of Medicine, New York, New York, United States of America; [12]Department of Population Health, NYU Grossman School of Medicine, New York, New York, United States of America

Efforts to identify modifiable risk factors for autism have revealed multiple associations between maternal health in pregnancy and autism risk offspring. However, the mechanisms underlying these associations remain unknown, including the role of transmitted and non-transmitted genetic variation.

We used the Danish registry data (N=1.13M), with 3 generation family linkage, and genotype data (N~130k, iPSYCH) to test the associations between all maternal diagnoses and autism. To triangulate evidence for genetic confounding, we implemented (1) family-based (paternal negative control; sibling comparison; risk in maternal cross and parallel cousins), and (2) polygenic risk score (PRS) analyses.

We identified 30 maternal diagnoses associated with autism risk, after accounting for their comorbidity and chronicity, familial correlations and potential confounders, and controlling for multiple testing. We observed evidence for genetic confounding for most of those associations, including through both transmitted (psychiatric and neurological disorders) and non-transmitted (obstetric complications) variation. For example, for epilepsy: (i) autism risk was the same irrespective of which parent was affected; (ii) children born to affected mothers had higher autism PRS, compared to children of unaffected women; and (iii) autism risk was the same in offspring of sisters and brothers of women with epilepsy, and for both higher than in the general population. We found no evidence for genetic confounding for very few associations, including predominantly medical codes indicating severe maternal injury in pregnancy.

In summary, associations between maternal health in pregnancy and autism risk are largely attributable to genetic factors, both transmitted and non-transmitted.

51

Rapid Long-Range Linkage Disequilibrium Calculations at Biobank Scale using GPU Acceleration

Rachit Kumar[1,*]; Pankhuri Singhal[1]; Chris Carson[1]; Mitchell Conery[1]; Alex Rodriguez[2]; Tarak Nath Nandi[2]; Marylyn D. Ritchie[1]; Mathialakan Thavappiragasm[2]; Ravi K. Madduri[2]; Benjamin F. Voight[1]; Anurag Verma[1]

[1]*Perelman School of Medicine of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America;* [2]*Argonne National Laboratory, Lemont, Illinois, United States of America*

Linkage disequilibrium (LD) information from one's own dataset (in-sample LD) is considered optimal for downstream analyses such as statistical fine-mapping. However, the computational complexity ([N^2]/2 computations for N variants) lead most studies to use external reference panels, such as from 1000 Genomes. To capture LD across all variant pairs in biobank-scale whole-genome sequencing (WGS) datasets with hundreds of millions of variants, new computational strategies are essential.

We present a novel approach that uses multi-GPU distributed computing to compute R^2 for every variant pair in a dataset. On chromosome 22 (1.8 million variants) of the 30x WGS 1000 Genomes dataset, our method using a single consumer-level 12GB GPU (NVIDIA RTX 2080TI) takes <30 minutes, while the same calculation with PLINK using a high-end 64-threaded CPU (Intel Xeon Gold 6338) takes >3 hours, corresponding to a ~6x speedup. When using 4 such GPUs, we observe a ~24x speedup, indicating linear scaling.

With this method, we successfully computed on the aforementioned 30x WGS 1000 Genomes dataset (~120 million variants and ~2500 samples) the entire LD matrix (>1e16, or 10 quadrillion elements) in under 6 hours using 512 NVIDIA 40GB A100 GPUs on the Department of Energy Argonne Leadership Computing Facility Polaris Supercomputer. We make this tool, coded in Python using CuPy, publicly available. Using this tool, researchers can leverage the full extent of their genomic data without relying on external LD reference panels and acquire more accurate, population-specific findings, particularly for groups underrepresented in existing databases.

52

Statistical Considerations for Cross-fitting Two-sample Mendelian Randomization Across Biobank Datasets Using Summary Statistics

Nicholas B. Larson

Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic College of Medicine and Science

Large-scale biobanks with comprehensive phenotyping and genome-wide profiling are becoming increasingly available for performing a variety of genetic analyses, including Mendelian Randomization (MR). A common MR analysis strategy is to conduct two-sample MR using GWAS summary statistics, whereby two independent datasets serve as respective sources for exposure and outcome trait SNP associations. If GWAS results for both outcome and exposure traits are available in both datasets, as may be the case with two independent biobanks, an immediate study design question is in regard to the source selection for the respective SNP associations. In general, it makes sense to choose the source for exposure SNP associations that corresponds to the highest power to ensure sufficient effect estimate precision and avoid weak instrument bias. However, for sufficiently large biobank datasets where weak instrument bias is not a major concern, a "converse" MR analysis could be conducted whereby the respective sources for exposure and outcome SNP associations are switched, producing a different MR causal effect estimate. A natural question is then if and how these two related MR results could be aggregated into a singular causal estimate. Herein, we work through some of the statistical considerations of such aggregation, including deriving the correlation of these converse MR estimates and drawing connections to cross-fitted one-sample MR. We additionally examine illustrative examples through two-sample MR analyses conducted using UK Biobank and FinnGen GWAS results. Finally, we review some considerations for scenarios where we have more than two datasets available for analysis.

Keywords: mendelian randomization, biobank, genetic correlation

53

Quantile IV Relaxes Parametric Assumptions and Enables Conditional Average Treatment Effect Estimation in Mendelian Randomization

Marc-André Legault[1,2,*], Jason Hartford[3], Benoît J. Arsenault[4,5], Archer Y. Yang[2,6], Joelle Pineau[1,2]

[1]*Department of Computer Science, McGill University, Montreal, Canada;* [2]*Mila, Montreal, Canada;* [3]*Valence Labs, Montreal, Canada;* [4]*Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec, Quebec, Canada;* [5]*Department of Medicine, Faculty of Medicine, Université Laval, Quebec, Canada;* [6]*Department of Mathematics and Statistics, McGill University, Montreal, Canada*

Mendelian Randomization (MR) enables estimation of causal effects while controlling for unmeasured confounding factors. However, traditional MR relies on strong parametric assumptions and is susceptible to bias if these are violated. We recently introduced a new machine learning MR estimator named Quantile Instrumental Variable (IV). Quantile IV is a two-stage nonparametric estimator that uses neural networks to estimate the IV—exposure and the exposure—outcome relationships.

Our estimator is distinctive in its ability to estimate nonlinear and heterogeneous causal effects offering a flexible approach for subgroup analysis. For instance, Quantile IV can be used to estimate conditional treatment effects without pre-specifying an interaction model. Confidence intervals for treatment effects are estimated using bootstrap aggregation or frequentist inference.

We demonstrate that Quantile IV performs competitively in realistic MR simulations and apply it in the drug target MR context. We investigated the impact of circulating sclerostin levels on heel bone mineral density, osteoporosis, and cardiovascular outcomes in the UK Biobank. Sclerostin is the drug target of romosozumab, a therapeutic antibody developed to prevent fractures in osteoporosis. The cardiovascular safety of sclerostin inhibition is debated. Employing various MR

estimators and colocalization analysis, we observe that a genetically predicted reduction in sclerostin levels significantly increases heel bone mineral density and reduces the risk of osteoporosis, with no increased risk of ischemic cardiovascular diseases. Quantile IV contributes to the advancement of MR methodology, and the case study on the impact of circulating sclerostin modulation contributes to our understanding of the on-target effects of sclerostin inhibition.

54

Effect Modification by Sex of Genetic Associations with Vitamin C Related Metabolites in the Canadian Longitudinal Study on Aging

Rebecca Lelievre,[1*] Mohan Rakesh,[1] Pirro G. Hysi,[2] Julian Little,[1] Ellen E. Freeman,[1] Marie-Hélène Roy-Gagnon[1]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada; [2]Section of Ophthalmology, School of Life Course Sciences, King's College London, London, United Kingdom

Vitamin C is an essential nutrient. Sex differences in serum vitamin C concentrations have been observed but are not fully known. Investigation of levels of metabolites may help shed light on how dietary and other environmental exposures interact with molecular processes. O-methylascorbate and ascorbic acid 2-sulfate are two metabolites in the vitamin C metabolic pathway. Past research has found genetic factors that influence the levels of these two metabolites. Therefore, we investigated possible effect modification by sex of genetic variant-metabolite associations and characterized the biological function of these interactions. We included individuals of European descent from the Canadian Longitudinal Study on Aging with available genetic and metabolic data (n= 9004). We used linear mixed models to tests for genome-wide associations with O-methylascorbate and ascorbic acid 2-sulfate, with and without a sex interaction. We also investigated the biological function of the important genetic variant-sex interactions found for each metabolite. Two genome-wide statistically significant (P value < $5x10^{-8}$) interaction effects and several suggestive (P value < 10^{-5}) interaction effects were found. These suggestive interaction effects were mapped to several genes including *HSD11B2*, associated with sex hormones, and *AGRP*, associated with hunger drive. The genes mapped to O-methylascorbate were differently expressed in the testis tissues, and the genes mapped to ascorbic acid 2-sulfate were differently expressed in stomach tissues. By understanding the genetic factors that impact metabolites associated with vitamin C, we can better understand its function in disease risk and the mechanisms behind sex differences in vitamin C concentrations.

Keywords: vitamin C, metabolites, GWAS, gene-environment interaction, CLSA

55

A Community of Networks Approach for Multi-Omics Integration

Anastasia Leshchyk[1], Paola Sebastiani[2]

[1]Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America; [2]Tufts Medical Center, Boston, Massachusetts, United States of America

Recent studies reveal that long-lived individuals experience a significant delay in age-related diseases such as Alzheimer's, dementia, and heart disease. Investigations into centenarians' genetics indicate that individuals carrying the APOE e2 allele are more likely to achieve longevity. This allele is associated with serum proteins and metabolites that provide insights into its influence on health outcomes. However, modeling the mechanisms linking genetic, molecular, and phenotypic data is challenging. We developed "Community of Networks," a method using Bayesian networks to integrate data from genetics, proteomics, metabolomics, and phenotypes. This method handles missing data through multiple imputation, creating a community of networks for downstream analyses. To reduce the dimension of multiple omics features, we summarize their effects into "omics-scores," providing a comprehensive summary of an individual's omics profiles. The method uses probabilistic reasoning with Bayesian networks to understand paths from genotypes to phenotypes. We applied this method to a multi-omics dataset from the New England Centenarian Study, including 362 centenarians, 672 centenarian offspring, and 435 controls. This approach identified paths connecting APOE alleles with molecular signatures of genetic variants and aging traits. For example, individuals with the APOE e2 allele have lower metabolomics risk scores for dementia compared to those with e4 and e3 alleles. Proteomic risk scores for dementia show a 40% change in the odds for a high score comparing e4 versus e2 carriers. Our research offers a framework for understanding the molecular underpinnings of longevity and age-related diseases, paving the way for targeted interventions to promote extended health span and quality of life in aging populations.

56

Inferring Causal Direction Between Two Traits Using R^2 with Application to Transcriptome-Wide Association Studies

Huiling Liao[1], Haoran Xue[2], Wei Pan[1]

[1]Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Department of Biostatistics, City University of Hong Kong, Kowloon, Hong Kong

In the framework of Mendelian randomization (MR), two single SNP-trait Pearson correlation-based methods have been developed to infer the causal direction between an exposure and an outcome: the popular MR Steiger's method and Causal Direction-Ratio (CD-Ratio). Steiger's method uses a single SNP as an instrumental variable (IV), while CD-Ratio combines the results from each of the multiple SNPs.

We propose a new method using R^2, the coefficient of determination, to combine information from multiple (possibly correlated) SNPs to simultaneously infer the presence and direction of a causal relationship. It generalizes Steiger's method from using a single to multiple SNPs as IVs, making it useful in transcriptome-wide association studies (TWAS) with typically small sample sizes for gene expression data and applicable to GWAS summary-level data with a reference panel. We further introduce a new method incorporating an invalid IV selection step to enhance robustness.

To illustrate advantages of our proposed method, we compared its performance to TWAS, Steiger's method and CD-

Ratio in simulations and in identifying causal genes for high/low-density lipoprotein cholesterol (HDL/LDL) using individual GTEx (V8) gene expression data and UK Biobank GWAS data. Our method confirmed some well-known causal genes, such as LPL, LIPC, and TTC39B for HDL, and identified novel gene-trait relationships, suggesting its power gains from using multiple correlated SNPs as IVs. Additionally, we showed its application with GWAS summary data through inferring causal relationships between HDL/LDL and stroke/coronary artery disease (CAD).

57

Genetic Ancestry-specific eQTL in Healthy Lung Tissue

Jonathan Lifferth[1], Hung Hsin Chen[2], Alex Petty[1], Jennifer Below[1], Bríd Ryan[3], Melinda C. Aldrich[1]

[1]Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Academia Sinica, Taipei, Taiwan; [3]National Cancer Institute, United States of America

Genetic variation between populations can result in distinct gene expression patterns. Prior studies have identified lung eQTL, but the role of ancestry-related differences in lung eQTL remains unexplored. To study the relationship between genetic ancestry and eQTL variation, we apply a novel method to identify local ancestry-specific eQTL in tumor-adjacent lung tissue collected from self-identified African American individuals. We describe ancestry patterns and RNA seq data for this unique resource.

Blood and lung tissue were collected from self-identified African American participants in a lung cancer case-control study. Samples were genotyped with the Illumina MEGA array and bulk RNA sequencing was performed on lung tissue (n=80). Global ancestry was inferred using ADMIXTURE. Local ancestry was estimated using RFMix2 with 1,757 reference samples (gnomAD v3.1.2). Genotype data were phased and imputed with the TOPMed Server and allele-specific gene expression was determined using phASER Gene AE.

Among self-identified Black/African American individuals, median global African ancestry proportion was 0.74 (sd = 0.34). Sixteen individuals had low African global ancestry (less than 3%). Using RFMix2, we inferred African and European local ancestry at 9,910 chromosomal segments. Using local ancestry, median African ancestry was nearly identical to global African ancestry. After performing quality control using RNA-seQC, we detected 29,242 genes with at least 5 unambiguous reads. Mean mapping rate was 94.08% across all samples. 87.45% of mapped reads were of high quality.

These data constitute a valuable resource for assessing the impact of genetic variation on gene expression in lung tissue of ancestrally diverse groups. These data will inform future work to identify ancestry-specific eQTL that may influence racial disparities in lung cancer risk and outcomes.

58

The Effect of Alzheimer's Disease risk Genes on Limbic White Matter Microstructure

Anna Lorenz[1,3,*], Aditi Sathe[1], Dimitrios Zaras[1], Yisu Yang[1], Alaina Durant[1], Niranjana Shashikumar[1], Bennett A. Landman[5,6,7,8], Logan Dumitrescu[1,2,3,4], Timothy J. Hohman[1,2,3,4], Derek B. Archer[1,2,3,4]

*Presenting Author: Anna Lorenz (anna.s.lorenz@vanderbilt.edu)

[1]Vanderbilt Memory and Alzheimer's Center, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [2]Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4]Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, Tennessee, United States of America; [6]Department of Radiology & Radiological Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [7]Vanderbilt University Institute of Imaging Science, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [8]Department of Biomedical Engineering, Vanderbilt University, Nashville, Tennessee, United States of America

White matter (WM) integrity is essential for brain function, but it declines with age and is compromised in neurodegenerative diseases that have strong genetic components such as sporadic Alzheimer's Disease (AD). Diffusion MRI quantifies WM by measuring water diffusion in the brain, providing metrics including fractional anisotropy (FA), axial diffusivity (AxD), and radial diffusivity (RD). Using bi-tensor models, we refine these measurements by accounting for free water (FW) content, enhancing the distinction between extracellular and intracellular spaces (Pasternak, 2009). By integrating genetic data with FW-corrected metrics, we aim to understand how AD genes affect age-related changes in limbic WM tracts which are crucial for memory function. For this purpose, we selected 94 SNPs linked to AD from various GWAS and analyzed their relationship with WM tracts. Participants (N=2,614) included in the analysis were non-Hispanic White individuals aged 50 to 100 years (mean=73.66, SD=9.76).

Our analysis identified 205 associations (P value<0.05) with AD loci. The genes *RBCK1*, *TNIP1*, *TREM106B*, *NTN5*, *CD33*, and *ADAM10* displayed more than five significant associations with WM microstructure. These genes were positively correlated with FA and AxD measures and negatively with RD and FW, suggesting a protective effect on WM integrity. In contrast, the genes *APOE*, *PTK2B*, and *MME* demonstrated positive correlations with RD and FW and negative correlations with FA and AxD, indicating harmful impacts on WM integrity.

These findings suggest that genetic drivers of AD contribute to alterations in WM microstructure, observed in a cohort where only 3% are clinically diagnosed with the disease.

59

A Multi-phenotype colocalization framework in *LocusFocus*

Hua Lu[1,2,4*], Rae S.M. Yeung[2,3,7,8], Lisa J. Strug[1,4,5,6]

[1]Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; [2]Cell Biology Program, The Hospital for Sick Children Research Institute, Toronto, Canada; [3]Division of Rheumatology, Department of Paediatrics, The Hospital for Sick Children, Toronto, Canada; [4]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [5]Department of Statistical Sciences, University

of Toronto, Toronto, Ontario, Canada; [6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada; [7]Department of Immunology, University of Toronto, Canada; [8]Institute of Medical Science, University of Toronto, Canada

Genome-wide association studies (GWAS) have identified genetic loci associated with complex diseases. A next step is to elucidate their mechanism of action. Colocalization analysis, which integrates GWAS data with other GWAS/omics data, can suggest biological mechanisms and evaluate pleiotropic effects. We present a multi-trait extension of our Simple Sum 2 (SS2) frequentist-based colocalization methodology, implemented in our *LocusFocus* software. This extension includes an omnibus test to determine if at least one trait colocalizes with the GWAS summary statistics and a separate test assessing trait-set colocalization. In simulations, our omnibus test demonstrated superior power compared to the minimum p-value approach, matching the performance of the Cauchy combination test (CCT) while maintaining a type one error rate at 5%. This advantage holds as the number of traits increases. For instance, when six out of 10 secondary traits colocalize with the primary trait, both the omnibus test and CCT achieve 80% power, surpassing the 70% power of the minimum p-value approach. From a computational standpoint, our new method outperforms both the minimum p-value and the CCT. The method can test 10 traits within a 500 SNP region in under three seconds, which is 10 times faster than methods. This efficiency increases linearly with the number of traits and cubically with the length of the genomic position. We plan to refine our method to address summary statistics coming from overlapping samples and apply it to our Multisystem Inflammatory Syndrome in Children (MIS-C) cohort to explore colocalization in related phenotypes and omic measurements.

60

Enhanced Mapping of Gene-Environment Interactions for Vitamin D through Variability Quantitative Trait Loci

Tianyuan Lu[1,2,3], Wenmin Zhang[4], Lei Sun[3], and Andrew D. Paterson[5]

[1]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [3]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada; [4]Montreal Heart Institute, Montreal, Quebec, Canada; [5]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada

Understanding gene-environment (GxE) interactions is crucial for personalized nutrition and public health strategies for vitamin D-deficient individuals. Existing GxE studies for vitamin D only identified interactions between several genetic variants and season of measurement.

Since GxE interactions are enriched in variability quantitative trait loci (vQTLs), we conducted vQTL discovery for serum vitamin D using a newly developed, powerful quantile integral linear model in UK Biobank European, African, East Asian, and South Asian ancestry populations, respectively. We identified 24 independent vQTLs in the European ancestry population (N=313,514), which were verified by multiple sensitivity analyses and replicated in All of Us (N=7,258 European ancestry individuals). No vQTL was identified in non-European ancestry populations due to limited sample sizes.

We tested interactions between the 24 vQTL lead variants and 20 environmental factors. We detected 55 GxE interactions involving 17 loci (FDR<0.05). For the first time, apart from season of measurement, we identified GxE interactions in several loci with modifiable risk factors, including body mass index, time spent outdoors, and fish intake. These interactions did not attenuate upon adjusting for other covariates. We also identified gene-sex interaction affecting *DHCR7*, which regulates vitamin D biosynthesis. Integrating gene expression profiles, we illustrated the sex-differentiated genetic effects on vitamin D likely act through sex-biased expression of a protein isoform of DHCR7 in the skin due to alternative splicing.

In summary, this work has greatly expanded known GxE interactions affecting serum vitamin D, and may help to tailor interventions for vitamin D deficiency and promote equity in healthcare.

61

A Statistical Cautionary Tale: Estimating Correlations When Inferring the Causal Direction Between Two Traits in Genetic Association Studies

Sharon M. Lutz[1,2], Christoph Lange[2]

[1]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, United States of America; [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, United States of America

In genetic association studies, Mendelian Randomization (MR) has gained in popularity as an approach to assess the causal relationship between an exposure and outcome using single nucleotide polymorphisms (SNPs) as instrument variables for the exposure. Recently, MR approaches have been extended to infer the effect direction between two phenotypes. The causal direction (CD) methods aim to establish both the existence and the direction of a causal relationship. These methods include the causal direction-ratio (CD-Ratio), causal direction-GLS (CD-GLS), causal direction-Egger (CD-Egger), and constrained maximum likelihood approaches (CD-cML, and MR-cML). In order to infer the effect direction between two phenotypes (denoted phenotype 1 and phenotype 2), these approaches first estimate the correlation of the SNP and phenotype 1 and the SNP and phenotype 2 using summary statistics from regression analyses. However, the formula used to estimate the correlations may not be valid depending on the type of model used to obtain the summary statistics. Through simulation studies, we assess/quantify the impact of the correlation estimates on the results of the CD methods in the presence of pleiotropy and unmeasured confounding. For simulations when unmeasured confounding was generated and there is a no causal effect of phenotype 1 on phenotype 2, the results of the CDcML and MRcML approaches differed depending on the regression used to estimate the correlations. Estimation of the correlation of the SNP and phenotype 1 and the SNP and phenotype 2 using summary statistics from different regression analyses may impact the results of the CD methods in the presence of pleiotropy or unmeasured confounding. Therefore, how the correlation is estimated may result in the incorrect effect direction being concluded.

62

Analysis of Follow-Up Data in Large Biobank Cohorts: A Review of Methodology

Merli Mändul[1,2], Anastassia Kolde[1,2], Krista Fischer[1,2]
[1]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [2]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

In recent years the sample sizes in population based biobanks have increased rapidly. Linking the genomics data with follow-up data from electronic health records (EHR) provides databases that can be used for Genome-Wide Association Studies (GWAS) for the risk of incident diseases. The standard approach is then to use the Cox Proportional Hazards (CPH) model, which accounts for censored time to event data. As biobanks are mostly volunteer-based, the underlying proportionality assumption in CPH model is not the only thing to bear in mind. Firstly, in biobanks there is no natural time origin for the start of follow-up, therefore one needs to adjust for left-truncation when choosing the timescale. Secondly, the participants of the biobanks are often (closely) related, which violates the independence assumption. Thirdly, the choice of additional covariates can be complicated - the number of available variables can be large, but the data is often incomplete. However, a CPH model is known to give biased estimates, when important covariates are omitted. We have conducted a simulation study mirroring the Estonian Biobank cohort. The main aim was to assess the magnitude of bias and power in realistic GWAS settings, where the "naive" CPH model is used, while ignoring left-truncation, relatedness and/or omitted covariates. Our results show that accounting for left-truncation is crucial for unbiasedness , but not for variant discovery. In addition, relatedness is not a concern in discovery studies. However, the omitted covariates can lead to significant bias as well as decreased power.

63

Blood Metabolome and Transcriptome Integration in Bullous Pemphigoid: A Case-Control and Case-Only Cohort Study

Louis R. Macias[1]*, Silke Szymczak[1], Christian Sadik[2]
[1]Institute of Medical Biometrics and Statistics, University of Lübeck; [2]Department of Dermatology, University of Lübeck

Introduction: Bullous pemphigoid (BP) is an autoimmune blistering skin disease with clearly defined autoantigens. Its pathogenesis is, however, still poorly understood and effective treatment strategies are needed. Determining BP characteristic metabolite abundance and gene expression in case-control and pre-post treatment comparisons is a promising strategy to highlight previously unknown pathogenic pathways.

Methods: Mass spectrometry and mRNA sequencing was performed on serum and whole blood samples, respectively, of BP patients and their age- and sex-matched controls. Metabolite abundances and transcript counts were analyzed with linear models and negative binomial regression, respectively. Designs for pre-post analyses were paired. Pathway information was obtained from the Small Molecule Pathway Database[1] and enriched pathways were identified by the simultaneous feature-set competitive test.[2]

Results: Thirty-six metabolites were differentially abundant between 65 cases and 43 controls. Eight pathways were enriched, 5 among them included both pyruvic acid and glutamine. The number of differentially expressed genes was 11,452 (50.25% of measured genes). CD40L signaling pathway and starch and sucrose metabolism pathway were enriched. Additionally, the Toll-like receptor pathway was enriched when metabolome and transcriptome results were combined. Regarding pre-post analysis, only 1/3 of patients had a follow-up metabolome sample. In these, 24 metabolites were differentially abundant and 18 genes differentially expressed. No pathway enrichment was found.

Conclusion: Metabolites and genes involved in immune activation, energy metabolism, and amino acid metabolism are differentially abundant or differentially expressed in the blood of BP patients.

1. Jewison T, et al. Nucleic Acids Res. 2014;42:D478-84.
Ebraimpoor M, et al. Brief Bioinform. 2020;21:1302-1312.

64

The Impact of Genetic Ancestry on Survival Outcomes in Pediatric Rhabdomyosarcoma: A Report from the Children's Oncology Group

Christina L. Magyar*****, BSN[1,2,3,4]; Ekene A. Onwuka*****, MD, MS[5]; Bailey A. Martin-Giacalone, PhD[4,6]; Michael E. Scheurer, PhD, MPH[4,7]; Deborah A. Marquez-Do, BS[4]; Mark Zobeck, MD, MPH[4]; Erin R. Rudzinski, MD[8]; Michael A. Arnold, MD, PhD[9]; Donald A. Barkauskas; PhD[10,11]; David Hall, MS[11]; Javed Khan, MD[12]; Jack F. Shern, MD[13]; Brian Crompton, MD[14,15]; Corinne Linardic MD, PhD[16]; Douglas S. Hawkins, MD[8]; Rajkumar Venkatramani, MD, MS, MBA[4]; Lisa Mirabello, PhD[17]; Chad Huff, PhD[18], Melissa A. Richard, PhD[4,7]; Philip J. Lupo, PhD, MPH[4,7,19]
[1]Graduate Program in Genetics and Genomics, Baylor College of Medicine, Houston, Texas, United States of America; [2]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX [3]McNair Medical Institute, The Robert and Janice McNair Foundation, Houston, Texas, United States of America; [4]Department of Pediatrics, Section of Hematology-Oncology, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas, United States of America; [5]Department of Pediatric Surgery, Surgical Oncology, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas, United States of America; [6]Graduate Program in Translational Biology and Molecular Medicine, Baylor College of Medicine, Houston, Texas, United States of America; [7]Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America; [8]Division of Hematology-Oncology, Seattle Children's Hospital, University of Washington, Seattle, Washington, United States of America; [9]Department of Pathology and Laboratory Medicine, Children's Hospital Colorado, University of Colorado School of Medicine, Aurora, Colorado, United States of America; [10]Department of Population and Public Health Sciences, Keck School of Medicine of the University of Southern California, Los Angeles, California, United States of America; [11]QuadW Childhood Sarcoma Biostatistics and Annotation Office, Children's Oncology Group, Monrovia, California, United States of America; [12]Oncogenomics Section, Genetics Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [13]Pediatric Oncology Branch, Center for Cancer Research, National Cancer

Institute, Bethesda, Maryland, United States of America; [14]Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Harvard Medical School, Boston, Massachusetts, United States of America; [15]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America; [16]Departments of Pediatrics and Pharmacology & Cancer Biology, Duke University School of Medicine, Durham, North Carolina, United States of America; [17]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Massachusetts, United States of America; [18]Department of Epidemiology, Division of Cancer Prevention and Population Sciences, The University of Texas, MD Anderson Cancer Center, Houston, Texas, United States of America; [19]Epidemiology and Population Sciences Program, Texas Children's Cancer and Hematology Center, Houston, Texas, United States of America
*co-first author

Childhood rhabdomyosarcoma (RMS) is an aggressive form of pediatric cancer with two main histological subtypes: alveolar and embryonal. Both alveolar RMS and *PAX3/7::FOXO1* gene fusion products are associated with poor prognosis. Recently differences in treatment outcomes were noted across genetic ancestry groups within some pediatric cancers, but its effect on RMS outcomes remains unknown. Therefore, we sought to determine the role of genetic ancestry on event-free (EFS) and overall survival (OS) in children with RMS. 920 individuals with RMS underwent sample collection and genome-wide genotyping as part of Children's Oncology Group protocols. The cohort was subdivided into African American, European, Latin American, South Asian/Asian Pacific Islander, and other ancestry groups using Grafpop software. We estimated hazard ratios (HRs) and 95% confidence intervals (CIs) across ancestry groups using multivariable Cox regression models, adjusting for relevant clinical covariates. Stratified analyses by histological subtype and *PAX::FOXO1* fusion were also conducted. Among our cohort, no significant differences in clinical characteristics were noted across ancestry groups yet embryonal RMS patients with South Asian/Asian Pacific Islander (SA/API) ancestry exhibited both inferior EFS (HR: 2.06; 95% CI: 1.07-3.97; p=0.031) and OS (HR:2.33; 95% CI: 1.09-4.84; p=0.028). The SA/API subgroup also exhibited similar trends in cases of fusion-negative RMS for both EFS (HR: 2.01; 95% CI: 1.07-3.76; p=0.029) and OS (HR: 2.33; 95% CI: 1.15-4.70; p=0.019). This work highlights the importance of considering genetic ancestry as a potential risk factor in select pediatric cancer outcomes.

65

Co-expression-Wide Association Studies Implicate Protein-Protein Interactions in Complex Disease Risk

Mykhaylo M. Malakhov[1]*, Wei Pan[1]
[1]Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America

Transcriptome-wide association studies (TWAS) have proven highly successful in prioritizing genes and proteins whose genetically regulated expression modulates disease risk. Standard TWAS approaches consider each gene or protein independently of the rest, and although recent work has extended TWAS to a multi-exposure setting, no currently available methods explicitly model the genetic regulation of co-expression. Here we introduce the co-expression-wide association study (COWAS) method to identify pairs of co-expressed genes or proteins that are associated with complex traits. COWAS trains models to predict the expression of each functional unit in the pair as well as their genetically regulated co-expression, which we estimate as conditional correlation. The model weights are subsequently used to impute expression and co-expression levels into genome-wide association study (GWAS) summary-level data for any trait of interest, enabling COWAS to test for association between co-expression and the trait while also accounting for direct expression-trait effects. We applied our method to plasma proteomic concentrations from the UK Biobank (N = 36k), focusing on 26,433 pairs with known protein-protein interactions comprised of 2,603 proteins coded by autosomal genes. We tested for association with low-density lipoprotein cholesterol, Alzheimer's disease, and Parkinson's disease, demonstrating that COWAS can successfully identify protein pairs whose co-expression impacts complex traits. Notably, our results demonstrate that co-expression between proteins may affect disease risk even if neither protein influences the disease when considered on its own. Our contribution provides a novel framework for studying genetically regulated co-expression, facilitating interrogation of the phenotypic consequences of gene-gene and protein-protein interactions.

Keywords: co-expression, genetic prediction, plasma proteome, quantitative trait loci, statistical genetics

66

A Genome-Wide Association Study of Preferences for 18 Bitter-tasting Foods and Beverages in the UK Biobank

Tongzhu Meng[1]*, Daiva E. Nielsen[1]
[1]School of Human Nutrition, McGill University, Sainte-Anne-de-Bellevue, Quebec, Canada

Genetic variation is implicated in individual preferences for bitter-tasting foods/beverages. However, previous studies have focused on candidate genes or limited varieties of bitter-tasting foods/beverages. The present aim was to conduct a genome-wide association study (GWAS) to identify loci associated with preferences for a comprehensive set of bitter-tasting foods/beverages.

UK Biobank data on self-reported preferences for 18 bitter-tasting foods/beverages were analyzed (n=123,682 unrelated individuals). Alcoholic beverages, coffee, cruciferous vegetables, and foods with bitter and other taste modalities were evaluated. Preference scores were provided on a 9-point scale. Item scores above the median value ("liking") were compared to scores at/below the median. GWAS was performed in PLINK using logistic regression models adjusted for age, sex, array type and batch, and the first 10 principal components. Study-wide significance was *P* value below 4.23×10^{-9}.

A total of 10 SNPs were identified with three having been reported in previous GWAS for the same trait. Single SNPs in *ADH1B*, *TAS2R1*, and *TAS2R38* were associated with lower liking of ale, red wine, and white wine, respectively. One SNP in *TET2* was associated with liking of whisky. Two SNPs in *FTO* were associated with liking coffee without sugar. One novel SNP in *CAMKMT* was associated with liking of green olives and single SNPs in *CRHR1* and *NSF* were associated with lower liking of

black olives. One SNP in *TAS2R19* was associated with liking of grapefruit.

This GWAS confirmed previously known SNPs and identified new loci associated with preferences for a variety of bitter-tasting foods and beverages.

Keywords: GWAS, Bitter Foods, Bitter Beverages, United Kingdom Biobank

67

Identifying Frequently Pleiotropic SNPs in Mendelian Randomization using MRBEE

Mengxuan Li, Yihe Yang, Xiaofeng Zhu*
*Corresponding author (xxz10@case.edu)
Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University

Introduction: Pleiotropy refers to the genetic variants influencing multiple phenotypic traits. Horizontal pleiotropy reflects a separate path of a genetic affecting a trait through a mediation. Many methods for identifying pleiotropic effects, like multi-trait analysis of GWAS, cannot distinguish horizontal pleiotropy from putative pleiotropy.

Method: We performed univariable Mendelian randomization (MR) and horizontal pleiotropy test using MR bias-correction estimating equation (MRBEE) on 38 traits in the UK Biobank across blood and urine biomarkers, and blood pressure, to identify horizontally pleiotropic variants. Each trait was alternately considered as an exposure and an outcome. Independent significant SNPs were identified using the clump algorithm, excluding SNPs within 500 KB that were highly correlated with the lead SNP ($r2 > 0.01$).

Result: From 1,406 pairwise univariable MR analyses across 38 traits, we calculated pleiotropic frequency (number of times an SNP was identified as pleiotropic / number of times it was selected as an IV) for 7,880 genome wide significant SNPs and identified 1,441 pleiotropic SNPs. The average frequency was 0.056 and top SNPs with the highest frequencies were rs1260326 in *GCKR* (0.55), rs1260333 near *GCKR* (0.51), and rs9273363 in *HLA-DQB1-AS1* (0.46). Notably, rs1260326 was selected as an instrumental variable 851 times out of 1,046 MR iterations.

Discussion: GCKR regulates lipid metabolism. The pleiotropic rs1260326 variant aligns with previous findings, exhibiting horizontal pleiotropic effects on blood pressure, lipid metabolism, and metabolic disorders. Our work suggests some genes are pleiotropic hot genes, independently contributing to studied traits.

68

Polygenic scores of blood cell profiles predict gastrointestinal adverse events in non-small cell lung cancer patients treated with immune checkpoint inhibitors

Pooja Middha[1], Rohit Thummalapalli[2], Zoe Quandt[3,4], Princess Margaret Lung Group[5], Matthew A. Gubens[6,7], Christine M. Lovly[8], Geoffrey Liu[9], Melinda C. Aldrich[10], Adam J. Schoenfeld[11], Linda Kachuri[12,13], Elad Ziv[1,14]

[1]Department of Medicine, University of California San Francisco, San Francisco, California, United States of America; [2]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America; [3]Division of Endocrinology and Metabolism, Department of Medicine, University of California San Francisco, San Francisco, California, United States of America; [4]Diabetes Center, University of California San Francisco, San Francisco, California, United States of America; [5]Princess Margaret Cancer Centre, Toronto, Ontario, Canada; [6] Medical Oncology, University of California San Francisco, San Francisco, California, United States of America; [7]Department of Medicine, Weill Cornell Medical Center, New York, New York, United States of America; [8]Department of Medicine, Division of Hematology and Oncology, Vanderbilt University Medical Center and Vanderbilt Ingram Cancer Center, Nashville, Tennessee, United States of America; [9]Princess Margaret Cancer Centre, Temerty School of Medicine, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [10]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [11]Thoracic Oncology Service, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America; [12]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, California, United States of America; [13]Stanford Cancer Institute, Stanford University of Medicine, Stanford, California, United States of America [14]Helen Diller Family Comprehensive Cancer Center, Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America

Adverse events from immune checkpoint inhibitors (ICI), such as gastrointestinal immune-related adverse events (GI-irAEs), including colitis (ICI-C) and hepatitis (ICI-H), pose a significant burden to cancer patients. Routinely measured hematologic indices reflect systemic immune responses and may be dysregulated in cancer patients, even prior to treatment initiation. In this study, we examined the potential of polygenic scores (PGS) for blood cell profiles to predict GI-irAEs in 1,349 non-small cell lung cancer (NSCLC) patients treated with ICIs.

PGS for blood cell counts and ratios were developed in disease-free individuals of European ancestry in UK Biobank (n=335,332). Associations between PGS for blood cell counts/ratios and GI-irAEs (determined by manual chart review) were evaluated using Cox regression models, adjusted for age, sex, ICI type, recruitment site, histology, and 5 genetic ancestry principal components.

Among 1,349 patients, 4% had ICI-C, and 5.3% had ICI-H. PGS for higher lymphocyte counts was associated with an increased risk of ICI-C (HR per SD=1.46, p=0.01) and ICI-H (HR per SD=1.71, p=0.0006) However, PGS for higher neutrophil count was associated with higher risk for ICI-H (HR per SD=1.45, p=0.01), but not ICI-C (p=0.56). PGS for neutrophil-to-lymphocyte ratio was inversely associated with ICI-C (HR per SD=0.70, p=0.01) and ICI-H (HR per SD=0.74, p=0.04).

Overall, our findings underscore the complex role of different immune cells and their relative proportions in the pathogenesis of GI-irAEs. Therefore, genetic predictors of peripheral blood cell profiles may be informative for predicting adverse events and guiding risk-benefit decisions in settings with high clinical uncertainty.

69

Human Protein-Small Molecule Interaction Networks Reveal the Cross-Talks Between Genomic Components and Metabolome Across Complex Metabolic Traits

Vaha Akbary Moghaddam[1*], Sandeep Acharya[2], Shu Liao[3], Woo-Seok Jung[3], Bharat Thyagarajan[4], Kaare Christensen[5], Michael R. Brent[1,3], Gary J. Patti[6], Ping An[1], Michael A. Province[1]

[1]Department of Genetics, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, United States of America; [2]Division of Computational & Data Sciences, McKelvey School of Engineering, Washington University in St. Louis, St. Louis, Missouri, United States of America; [3]Department of Computer Science & Engineering, McKelvey School of Engineering, Washington University in St. Louis, St. Louis, Missouri, United States of America; [4]Department of Laboratory Medicine & Pathology, University of Minnesota Medical School, University of Minnesota, Twin Cities, Minnesota, United States of America; [5]Department of Public Health, Southern Denmark University, Odense, Denmark; [6]Department of Chemistry, School of Arts & Sciences, Washington University in St. Louis, St. Louis, Missouri, United States of America

Small molecules (SMs) modulate biological processes through influencing macromolecules, gene expression, and epigenetic changes. However, their functions remain largely unknown due to their diverse origins and technological limitations. Here, we introduce a framework to construct statistical protein-SM interaction networks (PSI-Nets), enrich them with biological information, and use them to study complex metabolic traits. We used transcriptomics, lipidomics, and metabolomics data from 3839 healthy participants of the Long Life Family Study (LLFS) to construct statistical PSI-Nets with a combined linear mixed model / correlated meta-analysis (CMA) approach. Next, we enriched gene-SM connections by known gene-SM information and common regulatory and structural signatures of SM-associated genes to make biologically informed PSI-Nets. We constructed 529 PSI-Nets containing 5080 gene-SM interactions with 90% enrichment for biological features among 3347 genes and 401 SMs. Network association analysis for insulin sensitivity (HOMA2-S) in non-diabetic LLFS participants revealed nine significant PSI-Nets (*P* value $<9.45 \times 10^{-5}$). Our findings suggest a protective effect of mast cell genes (*FCER1A, MS4A2, HDC, CPA3, GATA2*) on HOMA2-S, which can be downregulated by elevated dimethylguanidino valeric acid in higher BMI and triglyceride (TG) conditions, decreasing HOMA2-S and increasing glucose and L-alanine levels. This axis was also replicated in the Framingham Heart Study. Lastly, combining whole genome sequencing association scans of HOMA2-S, genes, and SMs by CMA identified 26 genome-wide significant SNPs across five gene-SM pairs for HOMA2-S. Our framework provides a promising paradigm to uncover the biological effects of SMs and their role in complex traits.

Keywords: *Omics Integration, Network Inference, Multi Omics Networks, Insulin Sensitivity, Metabolic Health*

70

Decoding the Genetic Basis of Autoimmune Gastritis and Pernicious Anemia

Brooke M. Morris[1*], Austin W. Reynolds[2]

[1]Department of Anthropology, Baylor University, Waco, Texas, United States of America; [2]School of Biomedical Sciences, University of North Texas Health Science Center, Fort Worth, 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.

71

Integrative Proteogenomic Analyses Reveal Insights into Subtype-Specific Glioma Risk

Taishi Nakase[1*], Karl Smith-Byrne[2], Quinn T. Ostrom[3], Geno A. Guerra[4], Beatrice S. Melin[5], Margaret Wrensch[4], Robert B. Jenkins[6], Melissa L. Bondy[1,7], Stephen S. Francis[4,8,9], Linda Kachuri[1,7]

[1]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, California, United States of America; *[2]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom*; *[3]Department of Neurosurgery, Duke University School of Medicine, Durham, North Carolina, United States of America*; *[4]Department of Neurological Surgery, University of California San Francisco, San Francisco, California, United States of America*; *[5]Department of Diagnostics and Intervention, Oncology, Umeå University, Umeå, Sweden*; *[6]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America*; *[7]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, United States of America*; *[8]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America*; *[9]Weill Institute for Neurosciences, University of California San Francisco, San Francisco, California, United States of America*

Gliomas are a heterogeneous group of brain tumors with few known risk factors. Tumors with *IDH* mutations and 1p19q co-deletion have a favorable prognosis, while IDH-

wildtype glioblastomas have a median survival of 14 months. We undertook an integrative analysis of the plasma and brain proteome using protein quantitative trait loci (pQTL) and GWAS data (11,304 glioma cases, 304,523 controls).

Plasma proteome-wide association study (PWAS) using models for 1,362 proteins trained in 3 cohorts identified 29 candidate proteins (FDR<0.1), including 9 with high posterior probability (PP) of colocalization (>70%). We found three novel risk proteins for IDH-mutant 1p19q-intact glioma: immunoglobulin glycoprotein BCAM (P=9.6×10^{-8}), neurotrophin receptor SorCS2 (P=3.6×10^{-5}) and immune checkpoint PD-1 (P=1.1×10^{-5}). Interestingly, PD-1 inhibitors have not demonstrated clear benefits in glioblastoma patients, which aligns with our observation that PD-1 has a larger effect on IDH-mutant 1p19q-intact tumors (OR=2.75, 95% CI=1.61-4.69) than IDH-wildtype glioblastomas (OR=0.73, 95% CI=0.51-1.05). For all colocalized candidates, plasma protein abundance was cis-regulated by transcriptional activity in brain tissue (e.g. SorCS2: z=16.1, P=1.1×10^{-68}).

Analyses of the brain proteome using pQTL data for 1,776 proteins identified 9 associations (PP>0.70), including known and new risk proteins for IDH-wildtype glioblastoma such as EGFR (OR=0.63 P=1.9×10^{-35}) and ENPP6 (OR=0.62, P=7.6×10^{-5}), respectively. Novel findings for glioma overall included galectin-3 (OR=1.37, P=1.3×10^{-5}), a validated drug target and tumor fitness gene confirmed by CRISPR screens of glioma proliferation.

Our analysis identified candidate proteins involved in immune response and neuronal proliferation, which may enable prioritization of therapeutic targets and provide insight into disease mechanisms for glioma.

72

Accounting for Heterogeneity Due to Environmental Sources in Meta-analysis of Genome-wide Association Studies

Oyesola O. Ojewunmi[1*], Siru Wang[2], Abram Kamiza[3,4,5], Michele Ramsay[3], Andrew P Morris[6], Tinashe Chikowore[7,8,9], Segun Fatumo[1,4,10], Jennifer L Asimit[2]

[1]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom; [2]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [4]The African Computational Genomic (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda; [5]Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi; [6]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom; [7]MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [8]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [9]Harvard Medical School, Boston, Massachusetts, United States of America; [10]Precision Healthcare University Research Institute Queen Mary University of London

Meta-analysis of genome-wide association studies (GWAS) across diverse populations offers power gains to identify loci associated with complex traits and diseases. Often, heterogeneity in effect sizes across populations will be correlated with genetic ancestry and environmental exposures (e.g. lifestyle factors). MR-MEGA (Meta-Regression of Multi-Ethnic Genetic Association) has increased power to detect associations with heterogeneity in allelic effects between populations due to genetic ancestry compared to fixed and random-effects meta-analysis. However, to allow for the impact of environmental exposures that differ across GWAS, we have developed a novel environment-adjusted meta-regression model (env-MR-MEGA) to detect genetic associations by adjusting for and quantifying environment- and ancestry-correlated heterogeneity between populations.

Our extensive simulations assessed the power to detect association and heterogeneity of allelic effects due to ancestry and/or environment using env-MR-MEGA. The results showed that env-MR-MEGA gained greater or comparable association power than MR-MEGA across various heterogeneity scenarios. Notably, when environmental factors were strongly correlated with the trait than ancestry, the power gains of env-MR-MEGA over MR-MEGA became more apparent. We then applied env-MR-MEGA to summary statistics of twelve sex-stratified African GWAS for LDL-cholesterol in 19,000 individuals using sex, mean body mass index, and the proportion of study participants with urban status as environmental variables. We identified additional heterogeneity beyond ancestry-correlated effects for nine variants. Env-MR-MEGA provides a powerful approach to account for environmental effects using summary-level data, making it an important tool for GWAS meta-analyses.

73

Transferability of a Single- and Cross-Tissue Transcriptome Imputation Models across Ancestry Groups

Inti A. Pagnuco[1*], Stephen Eyre[1], Magnus Rattray[2], Andrew P. Morris[1]

[1]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Manchester, United Kingdom; [2]Division of Informatics, Imaging and Data Sciences, The University of Manchester, Manchester, United Kingdom

Transcriptome wide association studies (TWAS) explore the associations between genetically regulated gene expression and complex traits and diseases. TWAS methods start by imputing gene expression using expression quantitative trait loci (eQTL) as predictors, followed by testing the association of the imputed expression with the disease or trait. The power of TWAS to detect association depends on the predictive power of the gene expression imputation models. Training these models requires genotype and gene expression data from the same samples. However, publicly accessible transcriptomics resources, such as the Genotype Tissue Expression (GTEx) Project, are biased towards individuals of European ancestry, potentially leading to less accurate gene expression prediction models for individuals from other ancestry groups. This study examined eQTL transferability across ancestries by comparing the performance of two gene expression imputation models:

PrediXcan (tissue specific approach) and UTMOST (cross tissue approach).

Both models were trained using a dataset comprising 49 tissues from GTEx, exclusively composed of European ancestry individuals but were subsequently tested on two distinct datasets representing European ancestry and African American individuals. The findings indicate that, for most tissues, both approaches perform better when the training and testing datasets share the same ancestry, with the cross-tissue approach generally outperforming the single-tissue approach. The study highlights that eQTL detection by gene expression imputation models is influenced by ancestry and tissue context. Developing population specific reference panels across tissues can improve gene expression prediction accuracy, enhancing TWAS analysis and understanding of the biological processes underlying complex traits and diseases.

74

Enhancing Non-linear TWAS Performance via Trait Imputation with Applications to Alzheimer's Disease

Ruoyu He[1,2], Jingchen Ren[1,2], Mykhaylo M. Malakhov[2], Wei Pan[2,*]
[1]School of Staistics; [2]Division of Biostatistics and Health Data Science, University of Minnesota

Genome-wide association studies (GWAS) performed on large biobank datasets have identified numerous genetic loci associated with Alzheimer's disease. However, the younger demographic of biobank participants relative to the typical AD age of onset has resulted in an insufficient number of AD cases, limiting the statistical power of GWAS and any downstream analyses. To mitigate this limitation, several trait imputation methods have been proposed to impute the expected future AD status of individuals who may not have yet developed the disease. This paper explores the use of imputed AD status in nonlinear transcriptome-wide association studies (TWAS) to identify genes and proteins whose genetically regulated expression is associated with AD risk. In particular, we considered the TWAS method DeLIVR, which utilizes deep learning to model the nonlinear effects of expression on disease. We trained transcriptome and proteome imputation models for DeLIVR on data from the Genotype-Tissue Expression (GTEx) Project and the UK Biobank (UKB), respectively, with imputed AD status in UKB participants as the outcome. Next, we performed hypothesis testing for the DeLIVR models on diagnosed AD outcomes using the Alzheimer's Disease Sequencing Project (ADSP) dataset. Our results demonstrate that nonlinear TWAS trained with imputed AD outcomes in UKB successfully identified known and putative AD risk genes, and that training with imputed outcomes yielded more discoveries than training with diagnosed cases alone. Notably, we found that different AD imputation methods yield complementary results in association analyses, suggesting that investigations relying on a single imputation method may miss important pathways involved in neurodegeneration.

75

Subtle stories in GWAS data: multiomics implicate GPX3 at the *TNIP1* locus in Alzheimer's disease

Daniel J. Panyard[1,2], Lianne M. Reus[3,4,5], Muhammad Ali[6,7,8], Jihua Liu[9,10], Yuetiva K. Deming[2,11,12], Qiongshi Lu[9,10], Gwendlyn Kollmorgen[12], Margherita Carboni[13], Norbert Wild[12], Pieter J. Visser[3,4,15,16], Lars Bertram[17,18], Henrik Zetterberg[19,20,21,22,23], Kaj Blennow[19,20], Johan Gobom[19,20], Dan Western[6,7,8], Yun Ju Sung[6,7,8], Cynthia M. Carlsson[10,1124,25], Sterling C. Johnson[6,724,25], Sanjay Asthana[11,12,25], Carlos Cruchaga[6,7,8], Betty M. Tijms[3,4], Corinne D. Engelman[2], Michael P. Snyder[1]
[1]Department of Genetics, Stanford University School of Medicine, Stanford University, Stanford, California, United States of America; [2]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [3]Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, The Netherlands; [4]Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands; [5]Center for Neurobehavioral Genetics, University of California Los Angeles, Los Angeles, California, United States of America; [6]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, United States of America; [7]NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, Missouri, United States of America; [8]Hope Center for Neurological Disorders, Washington University School of Medicine, St. Louis, Missouri, United States of America; [9]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [10]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [11]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [12]Department of Medicine, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [13]Roche Diagnostics GmbH, Penzberg, Germany; [14]Roche Diagnostics International Ltd, Rotkreuz, Switzerland; [15]Department of Psychiatry, Maastricht University, Maastricht, the Netherlands; [16]Department of Neurobiology, Care Sciences and Society, Division of Neurogeriatrics, Karolinska Institutet, Stockholm, Sweden; [17]Lübeck Interdisciplinary Platform for Genome Analytics, Institutes of Neurogenetics and Cardiogenetics, University of Lübeck, Lübeck, Germany; [18]Department of Psychology, University of Oslo, Oslo, Norway; [19]Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden; [20]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden; [21]Department of Neurodegenerative Disease, UCL Institute of Neurology, London, United Kingdom; [22]UK Dementia Research Institute at UCL, London, United Kingdom; [23]Hong Kong Center for Neurodegenerative Diseases, Hong Kong, China; [24]Wisconsin Alzheimer's Institute, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [26]William S. Middleton Memorial Veterans Hospital, Madison, Wisconsin, United States of America

Introduction: A common issue in GWAS is that loci associated with disease are identified without any clear mechanism of effect. Such is the case In Alzheimer's disease (AD), where a recent GWAS reported a genetic association with Alzheimer's disease (AD) at the *TNIP1* locus, but without a clear mechanism.

Methods: We used multiple, independent cerebrospinal fluid (CSF) proteomics data to test (n = 137) and replicate (n = 446) the association of GPX3 protein levels with CSF biomarkers of Alzheimer's disease (i.e., amyloid and tau levels). We then

used available genomic and transcriptomic data to investigate the mechanism linking variant to disease.

Results: CSF GPX3 levels decreased with amyloid and tau positivity (ANOVA P = 1.5 x 10^{-5}; replication P = 2.56 x 10^{6}) and higher CSF phosphorylated tau (ptau) levels (P = 9.28 x 10^{7}; replication P = 4.38 x 10^{-9}). Incorporating the genotype for a GWAS-implicated variant (rs34294852) into these models identified a significant effect of the genotype and amyloid/tau status on GPX3 levels, suggesting a potential gene-by-environment interaction. *GPX3* is expressed in multiple cell types in the brain, and *GPX3* transcript levels are significantly associated with AD diagnosis in brain tissue (P = 7 x 10^{-6}). The genetic variant is a known eQTL for *GPX3* in microglia (P = 0.038), and it is near a predicted enhancer region connected with *GPX3* expression.

Discussion: These results suggest variants in the *TNIP1* locus may affect the oxidative stress response in AD via altered glutathione peroxidase 3 (GPX3) levels.

76

New Genetic Discovery for Diseases with Shared Pathology but Low Genome-Wide Genetic Correlation

Gina Parcesepe[1,2*], Ebrima Joof[1,2], Richard J. Allen[1,2], Jennifer K Quint[3], Iain D Stewart[4], Guruprasad Aithal[5,6], Chris Scotton[7], Hilary Longhurst[8], R Gisli Jenkins[4], Louise V. Wain[1,2] on behalf of the DEMISTIFI consortium

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR Leicester Biomedical Research Centre, Leicester, United Kingdom; [3]School of Public Health, Imperial College London, London, United Kingdom; [4]Margaret Turner Warwick Centre for Fibrosing Lung Disease, National Heart and Lung Institute, Imperial College London, London, United Kingdom; [5]Faculty of Medicine & Health Sciences, University of Nottingham; [6]NIHR Nottingham Biomedical Research Centre, Nottingham, United Kingdom; [7]Medical School, University of Exeter, Exeter, United Kingdom; [8]Dyskeratosis Congenita (DC) Action, United Kingdom

Fibrosis is a pathological feature of many diseases and can occur in almost all organs. However, genome-wide genetic correlation (r_g) between fibrotic diseases is often low, likely due to disease-specific processes. Multi-Trait Analysis of GWAS (MTAG) jointly analyses GWAS summary statistics from different traits to discover new genetic associations but typically assumes a minimum r_g>0.7. GWAS of lung, liver, and renal-system fibrosis show r_g~0.2. We hypothesised that, despite low r_g, MTAG would still detect signals related to shared fibrotic pathology.

Lung, liver, and renal-system fibrosis GWAS summary statistics were used. MTAG was utilised to jointly analyse the three traits, and conditional analysis and fine-mapping of 95% credible sets were performed. Signals reaching genome-wide significance (P<5x10^{-8}) were identified, and novel signals were defined as those with an association significance of P<5x10^{-8} in the MTAG, but not significant at this threshold in the contributing organ GWAS. Credible sets were considered improved if either the top sentinel's posterior probability had increased or the number of SNPs in the 95% credible set had reduced.

Two novel liver and three novel renal-system fibrosis association signals were identified, including some previously reported for specific fibrotic diseases, but not reaching genome-wide significance in the organ-level GWAS. For 10/21 lung, 2/5 liver, and 11/22 renal-system signals reaching genome-wide significance in both the GWAS and MTAG, the fine-mapping resolution had improved.

By successfully identifying novel fibrosis signals and refining previous association signals, we showed that MTAG is able to detect genetic signals even in the presence of low r_g.

77

The Polygenic Architecture of Hidradenitis Suppurativa and Its Clinical Implications

L. Petukhova[1], A. Khan[2], Y. Luo[2], L. Tsoi[3], L. Wheless[4], A. Hung[4], B. Kirby[5], N. Dand[6], J. Barker[7], M. Simpson[6], J. Saklatvala[6], L. F. Thomas[8], M. Løset[8], B. Brumpton[8], K. Hveem[8], A. Braun[9], S. Ripke[9,10], A. Saeidian[11], H. Hákonarson[12], M. March[12], Y. Bradford[13], T. Drivas[13], M. Ritchie[13], Penn Medicine BioBank, Regeneron Genetics Center, C. Weng[2], M. Teder-Laving[14], K. Kingo[14], M. Hayes[15], P. Sleiman[12], F. Mentch[12], J. Connolly[12], G. Hripcsak[2], S. Gaddam[16], A.E. Oro[16], E. Prens[17], K. van Straalen[17], J. Gudjonsson[3], C. Weng[2], K. Kiryluk[2], Hidradenitis Suppurativa Genetics Consortium

[1]New York University, New York, New York, United States of America (corresponding author); [2]Columbia University, New York, New York, United States of America; [3]University of Michigan-Ann Arbor, Ann Arbor, Michigan, United States of America; [4]Vanderbilt University Medical Center, Nashville, Tennessee, United Statse of America; [5]University College Dublin, Dublin, Ireland; [6]Kings College London, London, United Kingdom; [7]St John's Institute of Dermatology, London, United Kingdom; [8]Norwegian University of Science and Technology, Trondheim, Norway; [9]Charite Medical University, Berlin, Germany; [10]Massachusetts General Hospital, Boston, Massachusetts United States of America; [11]Baylor College of Medicine, Houston, Texas, United States of America; [12]Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America; [13]University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [14]University of Tartu, Tartu, Estonia; [15]Northwestern University, Chicago, Illinois, United States of America; [16]Stanford University, Palo Alto, California, United States of America; [17]Erasmus University, Erasmus, Netherlands

Hidradenitis suppurativa (HS) is a prevalent and debilitating inflammatory skin disease that has been understudied and has many unmet needs. HS arises from aberrant interactions between the immune system and terminal hair follicles in intertriginous skin. The HS Genetics Consortium provides a platform to facilitate global collaborations for conducting and translating HS GWAS using standardized protocols. Here, we conducted a GWAS meta-analysis across 10 cohorts encompassing three major ancestry groups and including 6,500 cases and 1.2M controls. We identified associations at 11 loci, which includes independent replication of two previously reported HS risk loci at 17q24.3 and 13q22.1. Additionally, we discovered a second independent association at chromosome 13q22.1 and new risk loci at 1q32.1, 2p22.2, 2q22.1, 6p21.32, 9q31.3, 14q24,3, 18q11.2, 18q12.2, and 22q12. Statistical fine-mapping in disease relevant tissues allowed us to prioritize a set of 50 genes that map onto a small set of co-regulated gene modules. Integration with transcriptomic data implicates cell types and pathways, revealing new disease mechanisms that we experimentally validate. We developed and evaluated a PRS using an independent dataset and then used it to estimate

risk in our meta-analysis cohort, obtaining reliable estimates across ancestries. We conducted PheWAS and genetic correlation analyses that identified disease associations with mechanistic and clinical relevance, including cardiovascular outcomes, infection susceptibility, and an inverse relationship with IBD. Our results identify disease mechanisms involving both intrinsic and immune mediated hair follicle biology as drivers of HS pathogenesis, while also suggesting new priorities in drug repurposing efforts. There are also immediate clinical implications in terms of comorbidity screening among individuals with HS.

78

Leveraging Biobank-linked Electronic Health Records for Drug-induced Stuttering Discovery

Dillon G. Pruett[1], Christine Hunter[2], Alyssa Scartozzi[1], Douglas M. Shaw[1], Shelly Jo Kraft[3], Robin M. Jones[4], Megan M. Shuey[1], and Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center; [2]Lipscomb University College of Pharmacy and Health Sciences; [3]Department of Communication Sciences and Disorders, Wayne State University; [4]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center

Drug-induced stuttering (DIS) is a condition where stuttered speech is caused by exposure to medications. Its observable features are very similar to developmental stuttering and causal mechanisms for both DIS and developmental stuttering are unknown. Although pharmaceuticals implicated in DIS are widely prescribed, DIS is rare. Precision medicine approaches may elucidate possible gene x exposure interactions and help explain stuttering at large. This investigation involves an electronic health record (EHR) review to characterize DIS cases.

Suspected cases of DIS were identified within approximately 3 million de-identified EHRs at Vanderbilt University Medical Center. This study examines the 40 resultant suspected DIS cases to describe: 1) the level of evidence for the implicated drug as a causal agent, 2) name, class, and mechanism of action of suspected drug, 3) other drugs present, 4) therapeutic measures taken, and 5) progression or remission of stuttering.

Eighteen different drugs were linked to possible DIS in 22 individuals. Antiseizure agents (8 cases), CNS stimulants (4 cases), and antidepressants (4 cases) were the most common drug classes implicated. Effects on GABA, glutamate, norepinephrine, serotonin and dopamine neurotransmitters and receptors were observed across multiple drugs and drug classes.

Disfunction in connections between the cerebral cortex, basal ganglia, and thalamus, termed the cortico-basal ganglia-thalamocortical (CBTC) loop, has been studied in the pathogenesis of both developmental stuttering and neurogenic stuttering and may provide a model for DIS. Further studies using pharmacogenetic approaches are needed to both characterize the etiology of DIS and to provide insight into stuttering in general.

79

Cell Type-Specific Transcriptome-Wide Association Studies Identify Susceptibility Genes in Various Cancers

Fei Qin, Kevin Wang, Xing Hua, Jianxin Shi, Kai Yu

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America

Background: Transcriptome-wide association studies (TWAS) were proposed to discover novel susceptibility genes associated with complex traits by integrating gene expression data with genome-wide association studies (GWAS). The availability of single cell RNA sequencing (scRNA-seq) data now enables the exploration of associations between different cell types and susceptibility genes in different cancers.

Methods: To identify cell-type specific susceptibility genes, we conducted TWAS using a scRNA-seq dataset of peripheral blood mononuclear cell (PBMC) samples from 982 individuals (14 cell types and 1.27 million cells) in the OneK1K cohort and GWAS summary data for six cancers: breast (118,474 cases, 96,201 controls), prostate (79,194 cases, 61,112 controls), lung (29,266 cases, 56,450 controls), ovarian (25,509 cases, 40,941 controls), endometrial (12,906 cases, 108,979 controls), and non-Hodgkin's lymphoma (3,857 cases, 7,666 controls). In prediction models, we leveraged shared information across cell types to enhance the gene expression prediction accuracy for each cell type. Additionally, we replicate our TWAS findings for breast and prostate cancers in the UK Biobank.

Results: We identified 339 novel genes for breast cancer, 89.4% of which were cell type-specific. For prostate cancer, we identified 92 novel genes, with 85 being cell type-specific. Replication in the UK Biobank data revealed enriched SNP-association signals among SNPs driving the TWAS findings. We also identified 20, 12, and 2 novel genes for lung, ovarian, and non-Hodgkin's lymphoma cancers, respectively.

Conclusion: Cell type-specific TWAS identified novel cancer loci, demonstrating distinct cell type-specific patterns for these novel loci across different cancers.

Keywords: Cell type, Transcriptome-wide association studies, Cancer

80

Impact of Bias in Variance Analysis: From SNPs to Pathways

Bryan Queme[1,2]*, Tremayne Mushayahama[1], Anushya Muruganujan[1], Mingzhi Ye[1], Huaiyu Mi[1]

[1]Division of Bioinformatics, Department of Population and Public Health Sciences; [2]Division of Cancer Biology and Genomics, Keck School of Medicine of the University of Southern California, Los Angeles, California, United States of America

Single Nucleotide Polymorphisms (SNPs), the most prevalent genetic variations, are pivotal in understanding human genetics and disease susceptibility. Accurate SNP annotation is essential for elucidating their functional implications, impacts on gene expression, and associations with phenotypic traits, crucial for advancing personalized medicine and genetic disorder research. Previous studies, using subsets of SNPs, have shown that different annotation tools lead to varying gene mappings. However, a comprehensive, genome-wide study comparing these tools has not been conducted until now.

In this study, we perform a genome-wide analysis to evaluate the consistency of SNP-to-gene annotations using SnpEff, ANNOVAR, and VEP. Statistically significant differences in the gene mappings produced by these tools (p-value < 0.001) were identified. These discrepancies highlight a potential source of bias that can significantly impact downstream analyses. Specifically, pathway overrepresentation analyses yield varying results depending on the annotation tool used, leading to inconsistent and potentially misleading conclusions Our findings underscore the importance of considering tool-specific biases in SNP-to-gene annotation processes. We illustrate the gaps between these tools and their downstream consequences in pathways overrepresentation analysis. Addressing these biases is critical for ensuring genomic research's accuracy, reliability, and reproducibility, enhancing the interpretation of genetic data and its applications in personalized medicine and genetic disorder studies.

To mitigate these biases, we have developed a pipeline that allows users to choose a comprehensive approach integrating multiple tools or select a specific tool tailored to their needs; providing researchers with a more holistic view of their results, promoting more robust and reliable genomic analyses.

Keywords: SNPs, annotation tools, tool-specific bias, genomic research reliability

81

Hybrid Design of Case-Parent Trio and Control Parents Under Mating Asymmetry

Mohan Rakesh[1*], Kelly M. Burkett[2], Marie-Hélène Roy-Gagnon[1]
[1]*School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada;* [2]*Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada*

Maternal genetic effects acting directly on the intra-uterine environment have been observed in early onset disorders. Analyses of case-trio data that condition on parental genotypes can disentangle the offspring from the maternal genetic effects and allow for estimation of maternal genotype relative risks. One source of bias in studying maternal effects is mating asymmetry (MA): the non-random assortment of a set of alleles between mothers and fathers in a population. MA can lead to over/under representation of risk alleles among case mothers without affecting Hardy-Weinberg equilibrium, which can either cause spurious associations or mask a true maternal effect. The hybrid design (HD) (Weinberg & Umbach, 2005) can remedy this by including control parents in the model to estimate additional mating parameters. However, differences in allele frequencies or MA levels between case and control samples would invalidate this design. Weinberg & Umbach proposed a population stratification test to screen for this. However, this test would not detect differences in MA. Here we used simulations to investigate type 1 error and power of the HD design in estimating a maternal effect in the presence of varying levels of population stratification and MA based on observed levels in 1000 genomes trios. Our results describe the tolerable levels of differences in allele frequencies and MA between cases and controls, the bias caused by differences in MA in combination with small differences in allele frequencies, and the performance of the population stratification test as a screening tool for the similarity of case and control samples.

82

Comprehensive Analysis of the Genetic Variation in the *LPA* Gene from Short-Read Sequencing

Raphael O. Betschart [1], Georgios Koliopanos [1], Paras Garg [2], Linlin Guo [3], Massimiliano Rossi [4], Sebastian Schönherr [5], Stefan Blankenberg [3,6,7], Raphael Twerenbold [3,6,7], Tanja Zeller [3,6,7] and Andreas Ziegler [1,3,6,8,*]
[1]*Cardio-CARE, Medizincampus Davos, Herman-Burchard-Str. 1, 7265 Davos, Switzerland;* [2]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Hess Center for Science and Medicine, New York, New York, United States of America;* [3]*Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Ham-burg-Eppendorf, Hamburg, Germany;* [4]*Illumina Inc., San Diego, California, United States of America;* [5]*Institute of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria;* [6]*Centre for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany;* [7]*German Center for Cardiovascular Science (DZHK), Partner Site Hamburg/Kiel/Lübeck, Hamburg, Germany;* [8]*School Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa*

Lipoprotein (a) (Lp(a)) is a risk factor for cardiovascular diseases and mainly regulated by the complex LPA gene. We investigated the types of variation in the LPA gene and their predictive performance on Lp(a) concentration. We determined the Kringle IV-type 2 (KIV-2) copy number (CN) using the DRAGEN LPA Caller (DLC) and a read depth-based CN estimator in 8351 short-read whole genome sequencing samples from the GENESIS-HD study. The pentanucleotide repeat in the promoter region was genotyped with GangSTR and ExpansionHunter. Lp(a) con-centration was available in 4861 population-based subjects. Predictive performance on Lp(a) con-centration was investigated using random forests. The agreement of the KIV-2 CN between the two specialized callers was high (r = 0.9966; 95% CI 0.9965–0.9968). Allele-specific KIV-2 CN could be determined in 47.0% of the subjects using the DLC. Lp(a) concentration can be better predicted from allele-specific KIV-2 CN than total KIV-2 CN. Two single nucleotide variants, 4925G>A and rs41272114C>T, further improved prediction. The genetically complex LPA gene can be analyzed with excellent agreement between different callers. The allele-specific KIV-2 CN is more important for predicting Lp(a) concentration than the total KIV-2 CN.

Keywords: GWAS, Kringle IV-2 repeat, WGS, Lipoprotein (a)

83

A Novel Statistical Test of Pleiotropy Between Traits Using GWAS Summary Statistics

Jiwon Park[1], Debashree Ray[1,2*]
[1]*Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America;* [2]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America*

Pleiotropy, the phenomenon where a genetic region confers risk to multiple traits, is widely observed, even among seemingly unrelated traits. Knowledge of pleiotropy can

improve understanding of biological mechanisms of diseases/traits, and can potentially guide identification of molecular targets or help predict side-effects in drug development. However, statistical approaches for identifying pleiotropy genome-wide are lacking, particularly for two correlated traits measured on the same or different individuals or between case-control traits with unknown sample overlap. We proposed a new method based on GWAS summary statistics from two traits by considering a composite null hypothesis that a variant is associated with none or only one of the traits. We assumed an inflated variance model, and our test statistic involves mixtures of product of correlated normal variables that allow for fractions of variants to be associated with none or only one trait. Our method is genome-wide scalable, where analytical *P* value is computed as a weighted sum of extreme tail probabilities of variance gamma distribution. Our method may be used to prioritize variants underlying a GWAS hit and also regulating an omics trait. Simulations demonstrate well-calibrated type I errors at stringent levels and substantially improved power over conventional approaches. Application to triglyceride and high-density lipoprotein levels reveal pleiotropic regions that conventional approaches missed and all of which were replicated in a larger GWAS of these lipid traits. This demonstrates our method may be used to identify novel genetic associations in traits with modest sample sizes by leveraging information from another correlated trait.

84

Exploring the genetic link between Diabetes Mellitus and Sarcoidosis

Jacob Makadsi[1], Olga D. Chuquimia[1], Susanna Kullberg[1], Leonid Padyukov[2], Natalia V. Rivera[1,2]

[1]Respiratory Division, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden; [2]Rheumatology Division, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden

Introduction: Sarcoidosis and diabetes mellitus are two polygenic, complex, and multifactorial diseases. The prevalence of type 2 diabetes mellitus (T2D) has reportedly increased in sarcoidosis patients. However, the causal relationship between sarcoidosis and T2D remains largely unstudied.

Aims: To elucidate the causal effects of the genetics of T2D and its glycemic traits on the risk of sarcoidosis clinical phenotypes Löfgren's syndrome (LS) and non-Löfgren syndrome (non-LS).

Material and Methods: This retrospective, cross-sectional epidemiological study utilized summary statistics from GWAS (genome-wide association studies) on sarcoidosis consisting of 7,000 Swedish individuals (outcome) and T2D and related-glycemic traits, including ~500,000 individuals (exposures). Glycemic traits studied were fasting glucose (FG), randomized glucose (RG), fasting insulin (FI), glycated hemoglobin (HbA1c), and 2-hour postprandial glucose (2hGlu). Mendelian randomization (MR) approaches were applied using the TwoSampleMR R package.

Results: T2D was not found to have a causal relationship with the sarcoidosis risk based on 39 SNPs (single nucleotide polymorphisms) as independent variables (IVs). Glycemic traits, including FG (20 SNPs, OR = 6.33, 95% CI: 1.41-28.5) and 2hGlu (3 SNPs, OR = 4.08, 95% CI: 1.01-16.4) were found to have a causal association with non-LS sarcoidosis risk.

Conclusions: Glucose levels may be implicated in non-LS sarcoidosis's genetic architecture and pathogenesis. Validating these results in independent cohorts can strengthen funding and be helpful for public health interventions.

85

Epigenome-Wide Mediation Analysis of the Relationship between Psychosocial Stress and Cardiometabolic Risk Factors in the Health and Retirement Study

Lauren O. Rogers[1]*, Wei Zhao[1,2], Scott M. Ratliff[1], Jessica D. Faul[2], Lauren L. Schmitz[3], Xiang Zhou[4], Jiacong Du[4], Belinda L. Needham[1], Jennifer A. Smith[1,2]

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America; [2]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, United States of America; [3]Robert M. La Follette School of Public Affairs, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America

Exposure to psychosocial stress is linked to a variety of negative health outcomes, including cardiovascular disease and its cardiometabolic risk factors. DNA methylation has been associated with both psychosocial stress and cardiometabolic disease; however, little is known about the mediating role of DNA methylation on the association between stress and cardiometabolic risk. Thus, using the high-dimensional mediation testing (HDMT) method, we conducted an epigenome-wide mediation analysis of the relationship between psychosocial stress and 10 cardiometabolic risk factors in a multi-racial/ethnic population of older adults (n=2,668) from the Health and Retirement Study (mean age=70.4 years).
A total of 50, 46, 7, and 12 CpG sites across the epigenome mediated the total effects of stress on body mass index, waist circumference, high-density lipoprotein cholesterol, and C-reactive protein, respectively. When reducing the dimensionality of the CpG mediators to their top 10 uncorrelated principal components (PC), the cumulative effect of the PCs explained between 35.8% and 46.3% of these associations.
A subset of the mediating CpG sites were associated with the expression of genes enriched in pathways related to cytokine binding and receptor activity, as well as neuron development. Findings from this study help to elucidate the underlying mechanisms through which DNA methylation partially mediates the relationship between psychosocial stress and cardiometabolic risk factors.

Keywords: social epigenomics, psychosocial stress, cardiometabolic risk factors, epigenome-wide mediation analysis, DNA methylation

86

Evaluating Kidney Size as a Marker for Renal and Hepatic Health

Rashedeh Roshani[1*], Lauren Petty[3], Alex Petty[3], Cassianne Robinson-Cohen[4], Jennifer E. Below[2]

[1]Vanderbilt Genetics Institute of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of

America; [2]Vanderbilt Genetics Institute and Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4]Division of Nephrology and Hypertension within the Department of Medicine at Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Kidney disease impacts 37 million Americans, yet traditional markers for kidney function assessment, such as estimates of glomerular filtration rate (eGFR) and glomerular integrity (albuminuria), are often inadequate due to the complexity of renal pathology. Assessing kidney size offers a complementary perspective, as changes in size can precede alterations in function. This study aims to understand the role of kidney size in health and disease by analyzing phenotypic and genetic data from individuals with radiology report or magnetic resonance imaging (MRI)-derived kidney size measurements. Previous Genome-Wide Association Studies (GWAS) on kidney volume have identified 10 risk loci affecting eGFR, chronic kidney disease (CKD), body mass index (BMI), and type 2 diabetes, including loci near *UMOD* and *FTO*. We performed linkage disequilibrium score regression analysis (LDSC) on traits from European ancestry and kidney volume GWAS summary statistics of about 38,000 individuals from the UK Biobank (UKBB). Our analysis revealed substantial genetic correlations between body size, kidney function, and total kidney volume (*P* value < 10^{-8}). Additionally, we conducted stratified Phenome-Wide Association Studies (PheWAS) on approximately 43,000 diabetic and non-diabetic individuals from Vanderbilt University medical center's biobank (BioVU). PheWAS adjusts for age and BMI to explore associations between kidney size and various clinical phenotypes. The results demonstrated significant associations between kidney length and both subclinical and clinical cardiovascular outcomes, acute kidney injury, and liver diseases. These findings highlight the importance of kidney size as a relevant trait for kidney, liver, and cardiovascular health.

87

Mendelian Randomization of Sex-Dependent Traits

Eric Sanders[1,2]*, Lisa Strug[1,2,3,4]

[1]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [2]Program in Genetics and Genome Biology, The Hospital for Sick Children Toronto, Ontario, Canada; [3]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada; [4]Departments of Statistical Sciences and Computer Science, University of Toronto, Toronto, Ontario, Canada

Mendelian Randomization (MR) is a method that leverages genetic instruments to identify causal exposure-outcome relationships while overcoming common issues of confounding. This can guide investigations of biological mechanisms driving a disease. Two-sample MR has been implemented frequently in the recent literature, conducted using statistics from two independent studies such as genome-wide association studies (GWASs), one for the exposure and one for the outcome. However, sex-specificity in exposure-outcome relationships can pose challenges when applying MR, as state-of-the-art methods produce a single effect estimate per genetic locus. This leaves researchers with the option of either performing MR marginally by sex or disregarding the possibility of sex-specific associations, both of which can reduce power. To address this, we propose novel extensions to both single-locus and multi-locus MR techniques such as MR-Egger and the median- and mode-based estimates. Our methods are based on constructing underlying models incorporating sex-specific genotype-outcome/genotype-exposure relationships, and jointly estimating more than one effect parameter of interest. We investigate via simulation how these novel MR implementations perform relative to typical approaches, investigating situations with and without sexual dimorphism and situations with typical assumption violations such as horizontal pleiotropy or incorrectly chosen genetic instruments. We observe that the novel approach introduces some unmeasured confounding and heteroskedasticity, but that no bias is introduced to the main effects of interest and that power is comparable to sex-stratified analyses. Our novel methods offer researchers a convenient implementation of MR that can identify sex-specificity, with no need to stratify analyses by sex and without sacrificing power.

88

Pleiotropic Effects of Pathway-partitioned Genetic Risk Scores for Asthma in UK Biobank

Matthew J. Saward[1*], Robert J. Hall[2], Richard J. Packer[1,3], Liam G. Heaney[4], Ian Sayers[2], Katherine A. Fawcett[1]

[1]Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Centre for Respiratory Research, National Institute for Health Research Nottingham Biomedical Research Centre, School of Medicine, Biodiscovery Institute, University of Nottingham, Nottingham, United Kingdom; [3]Leicester NIHR Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom; [4]Wellcome-Wolfson Centre for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

Asthma is a genetically and clinically heterogeneous respiratory disease, but it is unclear to what extent distinct underlying pathobiological mechanisms contribute to clinical differences. To investigate this, we partitioned the genetic component of asthma risk based on known biological pathways and investigated their pleiotropic effects.

Through literature searches, we identified a comprehensive list of independent asthma-associated variants and mapped them to genes based on genomic position, gene expression and chromatin interaction. We ascertained enriched biological pathways using QIAGEN IPA and constructed pathway-partitioned genetic risk scores, with variants weighted by their effect size on asthma risk in non-UK Biobank cohorts. Scores were then tested for association with up to 1,909 traits and 2,923 plasma proteins in UK Biobank.

In total, we found 255 independent signals mapping to 1,438 candidate causal genes and 16 enriched canonical pathways (predominantly cytokine and inflammatory response). Pathway-partitioned genetic risk scores were generally associated with an eosinophilic profile but two (C-type lectin receptors and Keratinization) showed strong association with neutrophils. One pathway (*IL15* Signalling) was more significantly associated with adult-onset than childhood-onset

asthma; the opposite was true for all others. *IL15* Signalling also showed significant association with hypothyroidism.

We identified key pathways underlying common genetic risk of asthma, including novel pathways such as C-type lectin receptors. Pathway-partitioned genetic risk scores for asthma show distinct associations with respiratory traits and other comorbidities, suggesting a genetic basis for different clinical presentations of asthma. These data provide opportunities for a more tailored approach to asthma management and treatment.

89

Candidate Gene-based Rare Variant Analysis of Developmental Stuttering

Alyssa C. Scartozzi[1], Yao Yu[2], Hannah G. Polikowsky[1], Douglas M. Shaw[1], Lauren E. Petty[1], Dillon G. Pruett[3], Janet M. Beilby[4], Kathy Z. Viljoen[4], Shelly Jo Kraft[5], Chad D. Huff[2], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America; [3]Hearing and Speech Sciences, Vanderbilt University, Nashville, Tennessee, United States of America; [4]Curtin School of Allied Health, Curtin University, Perth, Australia; [5]Communication Sciences and Disorders, Wayne State University, Detroit, Michigan, United States of America

Developmental stuttering is a common speech disorder characterized by blocks, repetitions, and speech sound prolongations. Stuttering is strongly enriched in families and influenced by common and rare genetic variation. Although six familial candidate causal stuttering genes have been identified, this variation does not explain stuttering risk in other families or general populations.

Here, we leverage whole-exome sequenced data from The International Stuttering Project to categorize rare functional variation in the six candidate genes. The current study uses Variant Annotation Analysis and Search Tool (VAAST) to identify stuttering risk genes in European genetic ancestry ("EA", ncase= 862, ncontrol= 4856) and African genetic ancestry datasets ("AA", ncase= 82, ncontrol= 1012).

In preliminary data, we joint called VCFs from unrelated participants using XPAT and then performed VAAST analyses to identify enrichments of rare functional variation in the six candidate causal stuttering genes. We found that *GNPTAB* reached nominal significance in the AA sample ($p = 0.015$) and *NAGPA* approached nominal significance ($p = 0.087$). No genes surpassed significance in the EA dataset.

Next, we performed analogous analysis testing for rare variant effects in the known 60 stuttering genes identified using population-based methods. In the EA dataset, three genes (*SGCD*, $p = 0.0097$, *ADCY5*, $p = 0.029$, and *SLC39A8*, $p = 0.045$) reached nominal significance. In the AA dataset, one gene (*SORCS1*) reached nominal significance ($p = 0.038$).

Future work will leverage family-based rare variant analysis tools (pVAAST) and complete pedigree information to increase power and identify novel stuttering risk genes enriched in families.

90

Prioritization of Icosapent Ethyl for Potential Reversal of Metabolic Dysfunction Associated Fatty Liver Disease

Hannah M. Seagle[1-4]*, Nikhil K. Khankari[1,3], Alexis P. Akerele[1,5,6], Jacklyn N. Hellwege[1,3,7], Megan M. Shuey[1,3], Michael Levin[8], Kyung Lee[9], Jennifer S. Lee[10], Kent Heberer[11], Donald R. Miller[12,13], Peter Reaven[14], Kyong-Mi Chang[15,16], Julie A. Lynch[9], Todd L. Edwards[4,7], Marijana Vujkovic[8]

[1]Vanderbilt University Genetics Institute, Department of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America; [2]Joseph Maxwell Cleland Atlanta VA Medical Center, Atlanta, Georgia, United States of America; [3]Division of Genetic Medicine, Department of Medicine, 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]School of Graduate Studies and Department of Microbiology, Immunology and Physiology, Meharry Medical College, Nashville, Tennessee, United States of America; [6]Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [7]VA Tennessee Valley Healthcare System (626), Nashville, Tennessee, United States of America; [8]University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [9]Salt Lake City VA Medical Center, Salt Lake City, Utah, United States of America; [10]Stanford University, Stanford, California, United States of America, [11]Palo Alto VA Medical Center, Palo Alto, California, United States of America; [12]VA Center for Medication Safety, Department of Veterans Affairs, Chicago, Illinois, United States of America; [13]Center for Population Health, Department of Biomedical and Nutritional Sciences, University of Massachusetts, Lowell, MA, United States of America; [14]Phoenix VA Health Care System; University of Arizona, Phoenix, Arizona, United States of America; [15]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania, United States of America; [16]University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America

Metabolic dysfunction-associated steatotic liver disease (MASLD) affects over 100 million American adults, yet there are currently no approved pharmacologic treatments. We present a genetically-informed pipeline to prioritize existing drug-target pairs for potential MASLD reversal. Our dataset included 90,408 MASLD cases and 128,187 controls from a large multi-ancestry Million Veteran Program GWAS. Genetically predicted gene expression (GPGE) profiles were generated from 47 GTEx v7 tissues using S-PrediXcan. Drug-gene pairs were identified using directional mapping in publicly available databases. Mendelian randomization (MR) estimated genetic effects of candidate genes on alanine aminotransferase (ALT) levels to proxy therapeutic effects. We identified 212 significant genes, 81 in the MASLD GWAS and 131 using GPGE. Of these genes, 13 encoded protein targets for 81 drugs and showed beneficial pharmacological modulation. For these 81 drugs, we obtained GWAS summary statistics for 19 primary indications. Fourteen were excluded due to non-significance, yielding seven final drug targets with five primary indications. The effect of icosapent ethyl (IPE) on ALT level was striking, a 1.2-1.4 units/L reduction. Significant effects were observed when IPE was proxied via increased *FADS1* expression and decreased *FADS2* expression

($p = 6.8x10^{-160}$, IVW beta = -1.26 and $p = 3.4x10^{-127}$, IVW beta = -1.39, respectively). Using genetic information from the largest MASLD GWAS study to date, we identified IPE as a possible drug-repurposing target. Our analysis suggests that reducing triglycerides via IPE, as proxied by genetically-predicted *FADS1* and *FADS2* expression, may lower ALT with reduced liver injury. Further exploration of IPE may be warranted.

91

Shared Genetic Associations in Chronic Viral Infection and Vaccine Failure

Mark Seielstad[1*]

[1]Institute for Human Genetics, University of California, San Francisco, California, United States of America

HepatitisB (HBV) infects >250 million people. Some 5%-10% of exposed adults develop a chronic infection, causing liver damage, cirrhosis, or cancer in 15%-25% of cases. Vaccination has reduced infections, but there is no cure. We identify genetic variation impacting vaccine response, and the ability to spontaneously clear HBV in a cohort of Chinese individuals living in Taiwan.

The TaiwanBiobank comprises 120,552 adults tested serologically for HBsAg, anti-HBs, anti-HBc, and anti-HBe. Previously vaccinated, chronically, and formerly infected individuals were identified serologically. Genotyping was performed with a custom array containing 591,048 SNPs.

We performed an unadjusted GWAS analysis on an early release of 2,000 chronically and 2,000 previously infected individuals. Highly associated SNPs ($P < 5x10^{-16}$) fall within a 36 kb region, encompassing HLA-DPB1 and HLA-DPA1. There was no inflation of P values due to population stratification. We have observed similar Class II associations with vaccine response.

Shared HLA associations suggest those least protected by vaccination may be most at risk of chronic infection, a significant clinical concern. The inability to respond to vaccine and the risk of chronic infection are both characterized by absence of HBs antibodies. This in turn suggests that HLA Class II protein sequences vary in their ability to bind and present s-antigen to B-cells, resulting in variable antibody responses. Future studies intend to directly examine binding affinities for s-antigen with HLA alleles identified here.

92

Developing a Polygenic Score for Idiopathic Parkinson's Disease: Insights into Mutation Penetrance

Sebastian Sendel[1*], Zied Landoulsi[2], Katja Lohmann[3], Björn-Hergen Laabs[4], Inke R. König[4], Patrick May[2], Christine Klein[3], Amke Caliebe[1]

[1]Institute of Medical Informatics and Statistics, Kiel University, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany; [2]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg; [3]Institute of Neurogenetics, University of Luebeck, University Medical Center Schleswig-Holstein, Campus Luebeck, Luebeck, Germany; [4]Institute of Medical Biometry and Statistics, University of Luebeck, University Medical Center; Schleswig-Holstein, Campus Luebeck, Luebeck, Germany

Background: The cause of Parkinson's disease (PD) is largely unknown, with a contribution of genetic and environmental factors. Less than 15% of PD cases are due to monogenic causes, but these often have reduced penetrance with largely unknown drivers. Nalls et al. (2019) developed a polygenic score (PGS) for idiopathic PD using clumping and thresholding, but newer methods like penalized regression and Bayesian inference might better capture SNP contributions.

Aim: To develop and validate an advanced PGS for idiopathic PD and assess its contribution to penetrance of monogenic PD.

Methods: Using genotype data from 1,762 PD patients and 4,227 controls (European ancestry, age at onset ≥ 40) from the ProtectMove cohort (www.protectmove.de), we developed a PGS with a custom pipeline using advanced methods. The best PGS was validated in the COURAGE-PD (10,854 patients, 8,626 controls) and LuxPARK (822 patients, 814 controls) cohorts and applied to 771 carriers of pathogenic variants (335 patients, 436 controls).

Results: The best-performing PGS (928,814 SNPs, LDpred2) achieved an area-under-curve (AUC) of 0.680 [0.665, 0.695] in our dataset, 0.718 in COURAGE-PD, and 0.667 in LuxPARK, outperforming existing PGS. In carriers of pathogenic variants, the PGS had an AUC of 0.646, indicating that polygenic factors influence monogenic PD penetrance, with AUCs of 0.605 for *PRKN* and 0.647 for *GBA1* mutation carriers.

Conclusion: Our PD-PGS shows high performance and potential for individual risk analysis in idiopathic PD and suggests that genetic risk factors for idiopathic PD also influence the penetrance of monogenic forms of PD.

Keywords: polygenic score, Parkinson's disease, monogenic disease, method comparison

93

Population Descriptors Variation in Plasma Proteomic and Metabolomic Profiles and their Association with Type 2 Diabetes Risk

M. Sevilla-Gonzalez[1], N. Wang[2], K. Westerman[1], S. Cromer[3], Y. Zhang[2], S. Hsu[4], C. Patel[5], K. Taylor[6], J. Rotter[6], C. Kooperberg[7], J. Meigs[8], A. Manning[1]

[1]Clinical and Translational Epidemiology Unit. Massachusetts General Hospital, Boston, Massachusetts, United States of America; [2] School of Public Health. Boston University, Boston, Massachusetts, United States of America [3]Diabetes Unit. Massachusetts General Hospital, Boston, Massachusetts, United States of America [4]Metabolism Program. Broad Institute of MIT and Harvard. Cambridge Massachusetts, United States of America [5]Harvard Medical. School., Boston, Massachusetts, United States of America [6]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, California, United States of America; [7]Fred Hutchinson Cancer Ctr., Seattle, Washington, United States of America; [8]Department of Medicine. Massachusetts General Hospital, Boston, Massachusetts, United States of America [9]Clinical and Translational Epidemiology Unit. Massachusetts General Hospital, Boston, Massachusetts, United States of America

Background: It is critical to understand the roles of both genetic and non-genetic factors in molecular biomarkers that contribute to disease risk. We aimed to investigate variations in 'omics profiles (proteomic and metabolomic) across self-

reported "race and ethnicity", as well as genetic ancestries and test their impact on type 2 diabetes risk (T2D).

Methods: We studied proteomic (N=2,339) and metabolomic data (N=4,900) from individuals within the Women's Health Initiative and Multi-Ethnic Study of Atherosclerosis cohorts. We used linear models adjusting by confounders; site of recruitment, age, continental genetic ancestry, income, education, diet, smoking, BMI and estimated glomerular filtration rate to identify race- and ethnicity-specific associations. We employed structural equation modeling to examine the mediation of these identified 'omics in the relationship between self-reported race/ethnicity and prevalent type 2 diabetes (T2D). Models were stratified by sex; results were meta-analyzed and corrected for multiple comparisons.

Results: A total of 61% (N=200) proteins and 74% (N=1415) metabolites were significantly associated with race and ethnicity population groups after accounting for covariates including genetic ancestry. Additionally, sets of proteins and metabolites displayed unique associations across genetic ancestries even after accounting for self-reported race and ethnicity population groups ($P<10^{-5}$). Furthermore, seventeen proteins and three metabolites, mediated the association between self-reported race and ethnicity and T2D risk ($P<10^{-6}$) when comparing self-reported Black vs. White individuals.

Conclusion: Our study reveals that both race and ethnicity, as well as genetic ancestry, significantly impact the proteomic and metabolomic profiles; particularly variations by the construct of race and ethnicity affect T2D risk.

94

Circular RNA: The Evolving Potential of Circular RNA in the Disease World

Aarti Sharma[1], Cherry Bansal[2], Kiran Lata Sharma[3], Ashok Kumar[4]
[1]Post-Doctoral Research Fellow, Mayo clinic Arizona, United States of America; [2]Professor, Dr SS Tantia medical college hospital & research center, Sriganganagar Rajasthan India; [3]Staff Scientist, Baylor College of Medicine Houston, Texas United States of America; [4]Professor, Department of Surgical Gastroenterology, SGPGIMS Lucknow India
Correspondence to: Ashok Kumar Doc.ashokgupta@gmail.com

Circular RNAs (circRNAs), a new star of noncoding RNAs belong to a group of endogenous RNAs that form a covalently closed circle and occur widely in the mammalian genome. Most of the circRNAs are conserved throughout species and frequently show stage-specific expression during various stages of tissue development. CircRNAs were mystery discoveries as they were initially believed to be because of splicing error; though, subsequent research displays that circRNAs can perform various functions and help in the regulation of splicing and transcription including their role as microRNA (miRNA) sponges. With the application of high throughput next-generation technologies, circRNAs hotspots were discovered. There are emerging indications that explains about circRNAs association with human diseases, like cancers, developmental disorders, and Inflammation, and can be a new potential biomarker for diagnosis and treatment outcome of various diseases including cancer. After the discoveries of miRNA and long noncoding RNA (lncRNA), circRNAs are now acting as a novel research entity of interest in the field of RNA biology of diseases. In the present review article, we have focused on major updates on the biogeny and metabolism of circRNAs, along with their possible/established roles in major human diseases.

95

Predictive Performance of a Multi-Ancestry Polygenic Risk Score for LDL with Gene-Environment Interactions of Smoking and Air Pollution in Diverse Groups: Results from the PAGE Study

Jayati Sharma[1], Yanwei Cai[2], Zhe Wang[3], Meng Lin[4], Christopher Haiman[5], Kari North[6], Ulrike Peters[2], Eimear Kenny[3], Miranda R. Jones[1], Genevieve L. Wojcik[1]
[1]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America; [2]Public Health Sciences Division, Fred Hutch Cancer Center, Seattle, Washington, United States of America; [3]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [4]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States; [5]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States; [6]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States; [7]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, New York, United States

Introduction: The degree to which environmental contexts such as smoking and air pollution may influence polygenic risk scores (PRSs) and biomarkers, such as low-density lipoprotein cholesterol (LDL) is understudied. We examined the performance of a multi-ancestry PRS on LDL in the Population Architecture Using Genomics and Epidemiology (PAGE) Study (n=31,729) and assessed smoking and particulate matter exposure ($PM_{2.5}$) as modifiers in 11,398 BioMe participants.

Methods: We modeled the PRS_{LDL}-LDL relationship via a linear regression including age, sex, BMI, and HDL cholesterol. Smoking status was assessed as a modifier to the PRS_{LDL}-LDL relationship in the full sample and across racial and ethnic groups in PAGE. Stratification was aligned with a broader goal of investigating cardiovascular health disparities. $PM_{2.5}$ was assessed using 2016 national data, split into tertiles, and evaluated as an additional modifier.

Results: Smoking and $PM_{2.5}$ exposure improved model performance differentially by race and ethnicity [incremental R^2 range: 0.004 to 0.053]. The interaction of PRS_{LDL}-smoking was statistically significant for PAGE former vs. never smokers (β_{former}=-1.24, P=0.025): a smaller mg/dL LDL change per unit increase in PRS. The PRS_{LDL}-$PM_{2.5}$ interaction in BioMe participants was statistically significant for those in the highest tertile of $PM_{2.5}$ exposure ($\beta_{tertile3}$=13.20, P=0.047).

Conclusion: The LDL-PRS_{LDL} relationship and PRS_{LDL}-smoking interaction varied across racial and ethnic groups, with stronger PRS_{LDL} influence in non-smokers. The PRS_{LDL}-$PM_{2.5}$ interaction showed higher effect at higher $PM_{2.5}$ exposures. Incorporating intersectional environmental variables in genetic risk models enhances our understanding of health disparities across settings and populations.

96

Accounting for the Effects of Recent Population Structure on Heritability Estimation in Biobank Studies

Ruhollah Shemirani[*1], Christa Caggiano[1], Eimear E. Kenny[1]
[1]*Institute for genomic Health, Icahn school of Medicine at Mount Sinai, New York City, New York, United States of America*

Estimated narrow-sense heritability (h^2) informs statistical genetics analyses that require accurate estimation of both genetic and environmental effects. This directly affects important genetic estimators, such as Polygenic Risk Scores (PRS). Confounding by recent population structure remains a major source of inaccuracies in these estimates, even in more homogeneous populations, and phenotypes with strong environmental or rare variant effects. We explored the efficacy of the Spectral Components of Identity-by-Descent relatedness (SPCs) in addressing the effects of population structure.

We used simulations to demonstrate the inflation of h^2 in GREML analysis. For example, we simulated a non-heritable phenotype with a h^2 of 1.0. SPC were able to reduce the estimate to 0.06 (std=0.01). We tested SPCs in multiple biobanks, including 350,000 individuals with White British ancestry in the UK Biobank, where the h^2 of birth-place was lowered by 53.2% [52.3%-54.1%]. SPCs reduced h^2 for socio-economic factors used as covariates in association studies. SPCs decreased the correlation of 15 socio-economic and anthropometric outcomes and their PRS with geographical locations (p-value=0.001). SPC-adjusted h^2 was closer to LDSC h^2 in low heritability phenotypes. There, SPC adjustment also reduced intercept of LD score regression analysis compared to PCs.

We empirically show the benefits of SPC adjustment for the accuracy of heritability estimation in improving the accuracy of estimates. It ultimately contributes to improved understanding of the genetic causes of medical and behavioral genetic phenotypes and accuracy of downstream analyses, predictive models, and transferability of results.

97

Applications and Evaluations of Polygenic Risk Scores in Cancer Outcomes: Context Matters

Jiayi Shen[1], James Baurley[2], Gillian King[1], David Bogumil[3,4], 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*

Performance of multi-population polygenic risk score (PRS) models are often validated using a summary measure such as R^2 or AUC among out-of-sample individuals. Here, we demonstrated that PRS performance should be carefully interpreted as multiple factors could affect PRS validation performance. We constructed various multi-population PRS models for four cancer outcomes (prostate, breast, lung, colorectal) using their latest multi-population GWAS summary statistics, and evaluated their out-of-sample performance in the Multiethnic Cohort Study (MEC, including individuals of European, African, East Asian, Hawaiian, and Hispanic ancestry). We compared several types of PRS methodology: selected independent SNPs that are significantly associated with the outcome, and an approach that incorporates all measured variants across the genome. We found that PRS performance (in terms of AUC and OR of 1SD-PRS) varies not only by construction methods and ancestral population, but also by the genetic distance of an individual from the overall diversity of the discovery GWAS. Additionally, PRS effects are generally better among younger individuals. For example, the OR of 1SD increase in prostate cancer PRS among the Hispanic is 1.82 (95% CI: 1.60 to 2.07) marginally in the MEC, but ranges among individuals aged 60 from 2.13 to 1.72 as the genetic distance increases. For individuals aged 70, the 1SD OR ranges from 2.03 to 1.64 for the same increase in genetic distance. These general trends hold for all four cancer types investigated suggesting interpretation of PRS performance should be done carefully by taking into the account the specific contexts for evaluation.

Keywords: polygenic risk score, GWAS, diverse population

98

Alternative Genetic Models for Discovery and Characterization of Genetic Associations with Lung Function in UK Biobank

Nick Shrine[1], Jing Chen[1], Abril G. Izquierdo[1], Alex Williams[1], Anna L. Guyatt[1], Catherine John[1], Richard Packer[1], Louise V. Wain[1,3], Ian P. Hall[2] and Martin D. Tobin[1,3]
[1]*Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom;* [2]*Division of Respiratory Medicine and NIHR Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom;* [3]*National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, University of Leicester, Leicester, United Kingdom*

In genome-wide association studies (GWAS) an additive model of association is most commonly used as it is well-powered to detect effects arising from a range of genetic models. GWAS of quantitative lung function to date have discovered 1020 associated signals under an additive model. We aimed to determine whether alternative genetic models reveal novel loci.

We included 320,656 European UK Biobank samples with genotype data and four quantitative lung function phenotypes: forced expiratory volume in 1 second (FEV_1), forced vital capacity (FVC), FEV_1/FVC and peak expiratory flow (PEF). Phenotypes were adjusted for age, age^2, sex, height and dichotomous ever/never smoking status before rank inverse-normal transformation. Genome-wide association of 63.5 million imputed variants with minor allele count >=5 was done with regenie under dominant and recessive models including 10 ancestry principal component covariates.

We selected 1320 sentinel variants passing $P < 5x10^{-9}$ in ±1 MB regions across 4 lung function traits. 775 variants were most significant for the dominant model and 545 for the recessive. 25 sentinels were outside ±1 MB of a previously reported lung function sentinel and a further 172 were in low linkage disequilibrium ($r^2 < 0.2$) with a previous sentinel, giving 197 association signals detected using alternative models that are independent from the 1020 discovered under an additive

model. Of the 1020 previously reported signals, 103 and 22 were more significantly associated under dominant and recessive models respectively which could influence experimental design of functional follow up studies.

99

A Genome-wide Association Study Identifies Genetic Regulators of Metabolism at Birth

Brittney M. Snyder[1], Christopher G. McKennan[2], Nathan Schoettler[3], Tebeb Gebretsadik[4], William D. Dupont[4], Carole Ober[5], Tina V. Hartert[1,6] for the ECHO-CREW investigators

[1]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [3]Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [4]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America; [6]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Background: Integration of multiple levels of molecular data can provide a more comprehensive assessment of disease mechanisms through identification of upstream/downstream targets, intermediate processes, and mediating effects. We aimed to determine the genetic contribution of variation in metabolite concentrations at birth, a period critical to understanding disease origins.

Methods: Our study population included infants enrolled in the INSPIRE population-based birth cohort (n=1,414) with linked targeted newborn screening metabolic data and genotypes. Separate GWASs for each metabolite (n=39) and related metabolite ratios (n=14) were conducted using multivariable linear or ordinal regression, depending on metabolite distributions. We performed analyses for each ancestral group (European, African, and Hispanic Americans) and reported the meta-analyzed results. Lead variants in each genomic region were tested for replication in two independent birth cohorts of European ancestry (MAAP: n=29, WISC: n=139; meta-analyzed results reported).

Results: Of the 284,247,533 variant-metabolite associations tested (5,363,161 variants*53 metabolites/ratios), 1,934 associations reached genome- and metabolome-wide significance ($p<5x10^{-8}/53=9x10^{-10}$). The significantly associated variants mapped to six genomic regions (1p31.1, 5q31.1, 6q21, 9q34.11, 11q13.2, 12q24.31). The most significant variants, at 12q24.31, were associated with C4 acylcarnitine concentrations (lead variant: rs12829722, $\beta=0.30$, $p=1.7x10^{-57}$). The lead variants in the 1p31.1, 9q34.11, 11q13.2, and 12q24.31 regions replicated ($p<0.05/6=0.008$).

Conclusions: We characterized the genetic contribution to inter-individual differences in metabolite concentrations at birth. Our findings suggest that metabolite concentrations are regulated by genotype and highlight the promise of metabolites as functional intermediate phenotypes in disease pathways originating at birth.

Keywords: newborn, metabolite, GWAS

100

Mosaic Chromosomal Alterations in Peripheral Blood Cells from Whole-Genome Sequencing and Alzheimer's Disease in the Amish

Yeunjoo E. Song[1,2,*], Renee A. Laux[1], Sarada L. Fuzzell[1], Sherri D. Hochstetler[1], Kristy L. Miskimen[1], Audrey Lynn[1,2], Weihuan Wang[1], Leighanne R. Main[2,3], Ping Wang[1], Yining Liu[1], Noelle Moore[1], Michael B. Prough[4], Daniel A. Dorfsman[4], Laura J. Caywood[4], Jason E. Clouse[4], Sharlene D. Herington[4], Alex V. Gulyayev[4], Susan H. Slifer[4], Larry D. Adams[4], Patrice Whitehead[4], Jeffery M. Vance[4,5], Michael L. Cuccaro[4,5], Paula K. Ogrocki[6], Alan J. Lerner[6], Margaret A. Pericak-Vance[4,5], William K. Scott[4,5], William S. Bush[1,2] 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, FL, United States of America; [5]Dr. John T Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, United States of America; [6]Department of Neurology, Case Western Reserve University School of Medicine, Cleveland, Ohio, United States of America

Mosaic chromosomal alterations (mCAs) are structural somatic alterations (Gain, Loss and copy-neutral loss of heterozygosity (CN-LOH)) and have been suggested as prognostic markers for a host of human diseases. We investigated the autosomal mCAs estimated from whole genome sequencing (WGS) to gain a better understanding of rates and the association with Alzheimer's disease (AD) in the Midwestern Amish, a founder population with homogeneous lifestyle.

The calling of mCAs was performed using the Mosaic Chromosomal Alterations (MoChA) pipeline, utilizing long-range phase information to search for imbalances between maternal and paternal allelic fraction in a cell population. The cognitive status of each individual was assigned via consensus review of medical history and neuropsychological testing. The cognitively-impaired (CI) group combined AD, MCI and Unclear individuals and compared to the cognitively-unimpaired (CU) group.

After extensive QC, a total of 920 individuals were included (mean age=82.19±5.69, 60% female), with a total of 402 (43.7%) having at least one detectable autosomal mCA event and a total of 1,273 mCAs detected on the autosomes. Of the detected mCA events, 493 (38.7%) were CN-LOH, 73 (5.7%) were losses, 704 (55.3%) were gains, and 3 (0.2%) had undetermined state. The rate of mCAs was higher for males than that for females (p<2e-16). No mCA events were associated with developing CI in females and combined data. In males, Gain and CN-LOH decreased the probability of CI

We found that male carriers of Gain and CN-LOH had a decreased risk of subsequent CI.

101

Survival-Associated Tumor Expression Quantitative Trait Loci (eQTLs) in Pediatric Hepatoblastoma

Josey Sorenson[1*], Andrew Raduski[2], Lauren Mills[2], Lindsay Williams[2], Erin Marcotte[2**], Stephanie R. Huang[3], Logan G. Spector[2] , Tianzhong Yang[1]

[1]Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Department of Pediatrics, Division of Epidemiology and Clinical Research, University of Minnesota, Minneapolis, Minnesota; United States of America; [3]Department of Experimental and Clinical Pharmacology, College of Pharmacy, University of Minnesota, United States of America

Hepatoblastoma, a rare pediatric liver cancer, is known to have low mutation burden and be driven by cis-regulatory regions. In this study, we examined germline expression quantitative trait loci (eQTLs) that regulate liver tumor expression, i.e., tumor eQTLs, and affect patient survival. We used patient genotype and tumor RNA-seq data from a Japanese cohort of hepatoblastoma patients. A total of 78 samples remained after filtering for contaminated samples or with too few read counts. Standard quality control pipeline was applied to the genotype data, except that we filtered out SNPs with minor allele frequency less than 0.1 due to the limited sample size. We used a linear model adjusting for tumor impurity and the interaction between genotype and tumor impurity effects to ensure robustness of tumor eQTL discovery along with additional batch effects, genetic principal components and baseline characteristic variables. The Wald test identified 216 tumor eQTLs, 49 of which were significantly modified by interaction effects, i.e., differentially regulating gene expression in tumor vs normal cells (FDR using EigenMT method < 0.05). Subsequently, we identified 3 of these tumor eQTLs SNPs significantly associated with overall survival and/or 5-year event-free survival outcomes using the Log-rank test. Furthermore, stratification by sex uncovered 6 additional noteworthy eQTLs, underscoring the importance of considering sex in survival analysis for hepatoblastoma. Overall, our findings contribute to the understanding of genetic determinants underlying hepatoblastoma survival via eQTL analysis and shedding light on potential prognostic markers in hepatoblastoma.

Keywords: tumor eQTL, survival, hepatoblastoma, childhood cancer

102

Simulation Study: Aggregating SNPs to Spikes Enables Better Preservation of True Positive Associations in Filtering for Imputation Quality

Katharina Stahl[1]*, Heike Bickeböller[1]

[1]Department of Genetic Epidemiology, University Medical Center Göttingen, Göttingen, Germany

A high threshold for imputation quality measures filters out a large number of false positive associations caused by inflated *P* values, but often sacrifices true associations in the process. This poses a challenge, especially for studies where a majority of SNPs is imputed after using SNP arrays for genotyping. In practice, true associations register as spikes of SNPs in close physical proximity due to LD, but post-imputation filtering is usually conducted on SNPs independently. We simulated 1266 small case-control studies on chromosome 19 with different settings influencing power and imputation quality based on the 1000 Genomes Project data, both to quantify false positive signals caused by imputation and wrongfully discarded associations, and to optimize filtering. We evaluated, among several imputation quality measures, the established IMPUTE info thresholds of 0.3 and 0.8 and compared association results on the likeliest genotype and the genotype dosage, both on SNP level and on spike level, where we aggregated significant SNPs by position. We found the interval between those thresholds is of interest to optimize filtering. Dosage is more likely to induce false positive associations, but also identifies more true signals. For SNPs between thresholds, 49% (likeliest genotype) and 61% (dosage) are true associations on SNP level; grouped in spikes, 50% (likeliest genotype) and 67%(dosage) are true associations. Finally, we present a filtering method based on spikes, which discards false positive and preserves true positive spikes below the higher threshold by leveraging differences between imputation output formats.

Keywords: Genotype Imputation, Quality Control, Simulation Study, Association

103

Multi-ancestry Proteome-wide Mendelian Randomization Offers a Comprehensive Protein-disease Atlas and Potential Therapeutic Targets

Chen-Yang Su[1,2,*], Adriaan van der Graaf[3], Wenmin Zhang[4], Susannah Selber-Hnatiw[2,5], J. Brent Richards[5,6,7,8,9], Vincent Mooser[2,5], Jason Flannick[10,11,12,13], Sirui Zhou[2,5], Tianyuan Lu[14,15,16], Satoshi Yoshiji[2,5,10,11,12,13]

*Presenting author

[1]Quantitative Life Sciences Program, McGill University, Montreal, Canada; [2]Victor Phillip Dahdaleh Institute of Genomic Medicine, McGill University, Montréal, Québec, Canada; [3]Department of Computational Biology, University of Lausanne, Lausanne Switzerland; [4]Montreal Heart Institute, Université de Montréal, Montreal, Canada; [5]Department of Human Genetics, McGill University, Montreal, Canada; [6]Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada; [7]5 Prime Sciences, Montréal, Québec, Canada; [8]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; [9]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom; [10]Programs in Metabolism and Medical & Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [11]Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, United States of America; [12]Harvard Medical School, Boston, Massachusetts, United States of America; [13]Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts, United States of America; [14]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America [15]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [16]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Background: Circulating proteins influence disease risk and are valuable drug targets. To increase discovery of protein-

phenotype associations for diverse populations, we conducted multi-ancestry proteome-phenome-wide Mendelian randomization (MR).

Methods: We performed ancestry-stratified two-sample MR scanning 2,265 unique proteins from SomaLogic and Olink platforms—2,110 proteins from four European cohorts (max. n=35,559), 1,144 from two African cohorts (max. n=1,871), and 581 from a novel East Asian cohort (n=1,823). We curated the largest GWAS for 179 traits in Europeans, 26 in Africans, and 206 in East Asians. To minimize horizontal pleiotropy risk, we used *cis*-pQTL instruments specific to the protein of interest with the highest Open Targets variant-to-gene score. We performed sensitivity analyses including colocalization using PwCoCo and SharePro. We further evaluated shared causal effects of prioritized proteins across ancestries and assessed druggability with the druggable genome, Open Targets, DrugBank, and DGIdb.

Results: We tested 726,035 protein-phenotype pairs in Europeans, 33,078 in Africans, and 115,352 in East Asians. Notably, 119 proteins were instrumentable only in Africans, and 17 only in East Asians, highlighting the value of multi-ancestry inclusion. We identified causal effects for 4,028 protein-phenotype pairs in Europeans, 55 in Africans, and 325 in East Asians (FDR<5%). Of 62 protein-phenotype pairs present in multiple ancestries, 51 had concordant effects, including ANGPTL4-triglycerides, SWAP70-HbA1c, INHBB-HDL-cholesterol, and IL1RL1-eczema. Importantly, 71.4% of these proteins are targeted by licensed drugs or those under development.

Conclusion: We provide a comprehensive atlas of protein-disease associations across three ancestries, offering insights into disease etiology and opportunities for prioritizing therapeutic targets.

Keywords: Proteomics, Mendelian randomization, Colocalization, pQTL, GWAS

104

Copy Number Variants in Familial Bipolar Disorder Ascertained from 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], Alan Shuldiner[4], 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, 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; [4]Regeneron Genetics Center, Tarrytown, New York, United States of America

The Amish Mennonite Bipolar Genetics (AMBiGen) study aims to identify genetic variants that increase the risk for bipolar disorder (BD) and related conditions within Anabaptist founder populations. Participants were ascertained through clinical settings or advertisements and assessed using semi-structured psychiatric interviews. Affected individuals diagnosed with BD, schizoaffective disorder, recurrent major depression, or schizophrenia were compared with their unaffected relatives. Copy number variants (CNVs) were detected based on whole exome sequencing by the Regeneron Genetics Center using CLAMMS algorithm with 508 CNVs meeting stringent quality control criteria. CNV overlaps across samples were called based on shared genes. A total of 300 unique CNVs (88 deletions and 212 duplications) in 913 individuals were analyzed further. 253 participants met narrow criteria for BD, 396 met broad criteria that included BD and related conditions, 367 were psychiatrically healthy and 150 were unphenotyped. The effective number of independent samples was estimated using the genetic relationship matrix. Fisher's exact tests evaluated the association between BD status and carrier status for each CNV, and the carrier status of the most loss-of-function intolerant gene within all CNVs detected. Some known neuropsychiatric CNVs were detected in different families, including on 1q21.1, 15q11.2, 16p11.2, and 22q11.2. A duplication overlapping the ***CES1*** gene at 16q12.2 was significantly associated with BD after Bonferroni correction. Duplications of intolerant genes (pLi>0.9 and LOEUF<0.03) were significantly enriched in narrowly-affected individuals. These results suggest a novel association with ***CES1*** and support a role for CNVs overlapping highly intolerant genes in familial BD.

105

Association Between Eye Disease and Cognitive Function Modified by a *KIBRA (WWC1)* Genetic Variant

Emily F. Tran[1]*, Mohan Rakesh[1], Gisele Li[2], Ellen E. Freeman[1,3,4], Marie-Hélène Roy-Gagnon[1]
[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada; [2]Maisonneuve-Rosemont Hospital, Montreal, Canada; [3]Ottawa Hospital Research Institute, Ottawa, Canada; [4] Department of Ophthalmology, University of Ottawa, Ottawa, Canada

Age-related eye diseases are inconsistently associated with cognitive decline, which could be due to effect modification. This study aimed to investigate whether a genetic variant (rs17070145) in the *KIBRA (WWC1)* gene, previously found to be associated with cognitive decline, modifies the association between eye disease and cognitive function. We used data from a Montreal hospital-based cross-sectional study (n=302) for the primary analysis. Adjusted linear regression models were used to test for interactions between rs17070145 and eye diseases affecting the relationship between eye disease (age-related macular degeneration (AMD) or glaucoma) and cognitive function measured by six oral cognitive tests. Replication analysis was done in the Canadian Longitudinal Study on Aging (CLSA; n=22,622) using adjusted linear mixed models to test for interactions between eye diseases and genetic variants in a genomic region around rs17070145. In the Montreal study, we observed statistically significant interactions of rs17070145 with both AMD (P<0.03) and glaucoma (P<0.005) affecting the category verbal fluency cognitive test. Among individuals with rs17070145-T alleles, those with eye disease had worse cognitive scores compared to individuals with normal vision, while among individuals with rs17070145-C alleles, those with eye disease had similar cognitive scores compared to those

with normal vision. Similar interactions with glaucoma and AMD were found in the CLSA, although not with the same cognitive measure. Our results suggest that the *KIBRA* gene may modify the association between eye disease and cognitive function. This knowledge may shed light on the mechanism by which glaucoma and AMD are related to cognitive function.

106

Multi-ancestry Proteome-wide Association Studies Leveraging Both cis-and trans-pQTL and protein-protein Interaction Networks

Zichen Zhang[1], Jiaming Liu[2,1], Lang Wu[3], Bingxin Zhao[4], Chong Wu[1]

[1]UT MD Anderson Cancer Center, Houston, Texas, United States of America; [2]Rice University, Houston, Texas, United States of America; [3]University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America; [4]University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Proteome-wide association studies (PWAS) have emerged as a powerful approach to identify putative causal proteins for complex traits and diseases. However, Traditional proteome-wide association studies (PWAS) predominantly focus on cis-acting elements and ignore in-depth biological knowledge, limiting their ability to identify putative causal proteins for complex traits and diseases. To address these limitations, we developed a novel framework to train PWAS imputation models that integrates summary-level protein quantitative trait loci (pQTL) data from both cis- and trans-loci across the genome and leverages protein-protein interaction networks. Applying our approach to extensive pQTL data from the UK Biobank Pharma Proteomics Project (UKB-PPP; n = 46,218 for European and n = 931 for African ancestries) and deCODE genetics (n = 35,892 for European ancestry), we trained large-scale PWAS models for both ancestries and prioritized biologically relevant protein networks for 620 and 642 proteins, respectively. Compared to classic *cis*-only models, the resulting models showed significantly improved predictive performance (1,796 versus 1,267, 42% more models with estimated predictive). Applying our models to GWAS summary statistics from the FinnGen, IEU OpenGWAS, GBMI, and MVP projects, we conducted a systematic multi-ancestry PWAS analysis for over 700 phenotypes. Compared to classic *cis*-only PWAS models, the resulting models showed significantly improved predictive performance and much higher power in sequential association studies (7,270 versus 1,650, 341% more associations found), enabling the identification of numerous novel protein-trait associations. Notably, using an external dataset of 6,690 FDA-approved drugs, we demonstrated that associations identified by our method are 2.4 times and 1.3 times more likely more likely to be validated for drug targets than those identified using Mendelian Randomization and cis-only PWAS, respectively. Our PWAS framework and multi-ancestry models are freely available at https://gcbhub.org/, facilitating the discovery and characterization of the proteomic architecture of complex traits and diseases.

107

Genetic Clusters of Childhood Asthma Identification of Genetic Clusters Underlying Endotypes of Childhood-onset Asthma

Raphaël Vernet[1*], Christophe Linhard[1], Anja Estermann[1], Yuka Suzuki[2], Florence Demenais[1], Hugues Aschard[2], Hanna Julienne[2,3], Emmanuelle Bouzigon[1]

[1]INSERM, Université Paris Cité, UMR 1124, Group of Genomic Epidemiology of Multifactorial Diseases, Paris, France; [2]Inst. Pasteur, Université Paris Cité, Dept. of Computational Biology, Paris, France; [3]Inst. Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France

Asthma, the most common chronic disease in children, presents various clinical expressions reflecting the existence of distinct asthma endotypes.

We aimed to characterize the genetic structure underlying childhood-onset asthma (COA) endotypes using genetic information from multiple clinical and biological asthma-related traits.

We built a comprehensive database of GWAS summary statistics containing 34 studies of asthma and asthma subtypes (childhood-onset, adult-onset, etc.) and 217 studies of 58 asthma-related traits (blood cells counts, various cytokines, lung function, etc.). We identified 228 independent SNPs (r^2<0.1) associated with COA at $P<5x10^{-8}$ and built the matrix of 228 SNPs by 58 traits association results, with Z-scores oriented to COA risk-increasing alleles. Finally, we applied a partition around medoids (based on cosine distance between the 228 Z-score vectors) to cluster SNPs with similar patterns of trait associations.

We identified five genetic clusters of COA-associated SNPs, enriched for different pathways and cell types. Specifically, one cluster, associated with decreased levels of many cytokines (including CCL24, CXCL12, IL12) but increased levels of CCL5, was primarily enriched for innate lymphoid cells in fetal spleen and lung ($P\leq3.4x10^{-4}$) and for PD1-signaling pathway (FDR-adjusted $P=7.3x10^{-5}$). Activation of this pathway by PD1 agonist was previously shown to suppress lung inflammation in mice (PMID:32778730). Another cluster, associated with increased Lymphotoxin-α levels, was enriched for dendritic cells in lungs ($P=7.3x10^{-4}$) and for viral response-related pathways (FDR<5%), highlighting the contribution of viral infections in COA.

Thus, our clustering approach can help reveal the genetic structure underlying COA endotypes and identify potential therapeutic targets.

Funding: ANR-20-CE36-0009

108

Association Analysis of Mitochondrial Heteroplasmy and RNA-seq in the Framingham Heart Study

Mengyao Wang[1*], Roby Joehanes[2,3], Daniel Levy[2,3], Chunyu Liu[1,3]

[1]Department of Biostatistics, School of Public Health, Boston University, Boston, Massachusetts, United States of America; [2]Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [3]Framingham Heart Study, Framingham, Massachusetts, United States of America

Background: Although mitochondrial function, nuclear gene expression, and DNA methylation interrelate, the nature of their relationship remains to be elucidated.

Methods: Using multi-omics data from 1644 Framingham Heart Study (FHS) participants (mean age 57 years, 56% women), we conducted a linear mixed effects model between mitochondrial local constraint score sum (MSS) identified from whole genome sequencing and gene expression measured by whole transcriptome sequencing (RNA-seq), adjusting for age, sex, smoking status, white blood cell counts, batch, and familial structure. In addition, we investigated the association between MSS and four DNA methylation age metrics and explored whether epigenetic aging mediates the association between MSS and gene expression.

Results: We identified 20 genes whose expression levels were positively associated with MSS (false discovery rate < 0.05). Eight unique MSS-related genes were associated with PC-based epigenetic age (PCA), with 13 gene-PCA pairs observed. One of MSS-related nuclear gene expressions, TGFB-induced factor homeobox 1 (TGIF1) (), also showed positive associations with three PCAs, including the Hannum (), Horvath (), and PhenoAge (). The Hannum PCA mediated up to 23% of associations between TGIF1 expression level and MSS ($P < 0.001$).

Conclusions: We observed that mitochondrial heteroplasmy was associated with expression levels of nuclear genes and with DNA methylation age metrics. Further study is warranted to focus on the regulatory mechanism under the relationship of mitochondrial heteroplasmy, gene expression, and epigenetic age.

109

Improving Accuracy of Schizophrenia Risk Prediction by Leveraging Ancestry Information for Minority Populations

Xuexia Wang[1]*, Jingwei Gu[1], Xiaohan Liu[2], Sicong Xie[3], Samantha Gonzales[1]

[1] Xuexia Wang, Department of Biostatistics, Florida International University, Miami, Florida, United States of America; [2]Xiaohan Liu, National Day School, China; [3]Sicong Xie, Department of Statistics, University of Michigan, Ann Arber, Michigan, United States of America

Schizophrenia (SCZ) is a severe and disabling mental illness. Improving accuracy of SCZ risk prediction by using ancestry specific polygenic risk scores (PRSs) can improve early identification, intervention strategies for SCZ. Minority populations (MPs) often have less representation in genetic research, leading to less effective risk prediction models for these groups. To improve accuracy of building a risk prediction model to help reduce health disparities in SCZ, we built a risk prediction model with ridge regression using allele specific PRSs. We applied the risk prediction model to a Swedish sample (2,536 schizophrenia cases and 2,543 controls) and 664 African American. The PRSs in the risk prediction models were estimated with allele-specific ancestry method and without considering allele-specific ancestry information using the effect sizes of genetic variants of two large SCZ summary datasets from the Psychiatric Genomics Consortium (PGC). The first data is a SCZ meta-analysis genome-wise association study (GWAS) dataset (13,833 cases and 18,310 controls). The second data is a more recent study including 36,989 cases and 113,075 controls. The AUC of the risk prediction model with PRS estimated with allele-specific ancestry compared to without using allele-specific ancestry can be improved 5% in Swedish sample and 8% in the African American sample. Using the risk prediction model with PRS using allele-specific ancestry, high-risk individuals for SCZ can be estimated more accurately. Identifying high-risk individuals within MPs can lead to targeted interventions which include community-specific mental health programs, resources, and support systems.

Keywords: Schizophrenia, Minority populations, allele specific ancestry, polygenic risk scores

110

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

Siru Wang[1], Oyesola O. Ojewunmi[2], Abram Kamiza[3,4,5], Michele Ramsay[3], Andrew P Morris[6], Tinashe Chikowore[7,8,9], Segun Fatumo[2,4,10], Jennifer L Asimit[1]*

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [2]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom; [3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [4]The African Computational Genomic (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda; [5]Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi; [6]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom; [7]MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [8]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [9]Harvard Medical School, Boston, Massachusetts, United States of America; [10]Precision Healthcare University Research Institute Queen Mary University of London

Amongst diverse population groups it is likely for there to be heterogeneity in effect sizes due to ancestry, as well as environmental exposures. This allelic heterogeneity impacts the power to detect genetic associations, and in turn, refinement of sets of potential causal variants underlying genetic associations, through statistical fine-mapping. The meta-regression of multi-ethnic genetic association (MR-MEGA) adjusts for and assesses heterogeneity due to ancestry by accounting for axes of genetic variation derived from allele frequencies.

Building on the MR-MEGA framework, we propose a multi-ancestry fine-mapping approach that accounts for ancestral and environmental heterogeneity, MR-MEGA-based fine-mapping (MR-MEGAfm)/environment-adjusted MR-MEGA-based fine-mapping (env-MR-MEGAfm), allowing for multiple causal variants. We use MR-MEGA/environment-adjusted MR-MEGA together with approximate conditional analyses to identify signals and their coinciding credible sets. This requires GWAS summary statistics and linkage disequilibrium (LD) from each cohort, as well as summary-level environmental covariates (for env-MR-MEGAfm).

In simulation studies, we show that: (i) when allelic heterogeneity is correlated with ancestry alone, both methods produce nearly equivalent results; (ii) when allelic heterogeneity is more strongly correlated with environment factors compared to ancestry, env-MR-MEGAfm yields improved resolution over MR-MEGAfm.

MR-MEGAfm and env-MR-MEGAfm are efficient multi-ancestry fine-mapping methods that utilize summary-level data, allowing for multiple causal variants while adjusting for sources of heterogeneity due to ancestry (MR-MEGAfm) or ancestry and environment (env-MR-MEGAfm).

111

Identify Sex-Specific Genetic Effects on Neurological Traits and Disorders Through Genetically Regulated Gene Expression

Ting-Chen Wang[1,2], Xavier Bledsoe[1], Eric R. Gamazon[1], 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

Neurological disorders, such as Alzheimer's disease, exhibit notably higher prevalence in women than men. However, genetic effects underlying observed sex differences remain underexplored. We aim to functionally characterize genes by linking genetically regulated gene expression (GReX) to neuroimaging-derived phenotypes (NIDPs) to investigate sex-specific genetic effects on brain features.

We utilized SPrediXcan to perform transcriptome-wide association (TWAS) analyses on the 3,935 sex-stratified summary statistics of NIDPs from UK Biobank MR imaging (22,138 subjects; 11,624 genetic females) and the GReX data generated by applying the Joint Tissue Model framework to 49 GTEx tissues. Quality control included removing TWAS associations with a low prediction ($R^2 < 0.1$), NIDPs with non-significant SNP-mediated heritability, and associations involved genes in the MHC region on chromosome (chr) 6 and inversion regions on chr 8 and chr 17. We employed the Benjamini-Hochberg false discovery rate (FDR) threshold of 0.05 for multiple testing corrections.

We identified 45,719 FDR-significant GReX-NIDP TWAS associations, with 18,499 female-, 17,944 male-specific, and 9,276 shared between sexes with directionally consistent zscore estimates. These numbers excluded 4,236 associations from sex-specific tissues: ovary, vagina, uterus, testis, and prostate. Across shared and sex-specific GReX-NIDP TWAS associations, the cortical grey-white contrast NIDPs and GReX in brain tissues are enriched.

GReX in nerve-tibial, thyroid, and other tissues are also enriched in sex-specific associations.

Our study provides a rich resource linking GReX to neurophysiology in a sex-specific manner. This resource can facilitate gene function characterization influencing neurophysiology differently in males and females for forming hypotheses and prioritizing candidates for further validation.

112

Polygenic Risk of Coronary Artery Disease for Long-Term Survivors of Breast Cancer

Gordon P. Watt[1]*; Xiang Shu[1]; Anne S. Reiner[1]; Kathleen E. Malone[2]; Julia A. Knight[3,4]; Esther M. John[5,6]; Eric J. Chow[2,7]; Charles F. Lynch[8]; Lene Mellemkjær[9]; Meghan Woods[1]; Jonine L. Bernstein[1]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America; [2]Division of Public Health Sciences, Epidemiology Program, Fred Hutchison Cancer Center, Seattle, Washington, United States of America; [3]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada; [4]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada.

[5] Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, California, United States of America; [6]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford, California, United States of America; [7]Clinical Research Division, Fred Hutchison Cancer Center, Seattle, Washington, United States of America; [8]Department of Epidemiology, University of Iowa, Iowa City, IA, United States of America; [9]Danish Cancer Institute, Copenhagen, Denmark; [10]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America

Background. We evaluated whether a polygenic risk score for coronary artery disease (CAD-PRS) was associated with incident CAD for breast cancer survivors.

Methods. The study sample included 1,307 participants from the WECARE Study, an international, population-based study of >1-year breast cancer survivors diagnosed when <55 years of age. We created the CAD-PRS using variant weights for a published genome-wide PRS. Restricted to participants of European ancestry, we used Cox proportional hazards models to estimate the association between incident CAD and the CAD-PRS, adjusting for age, CAD risk factors, type of chemotherapy, receipt/laterality of radiation therapy, WECARE Study phase, subsequent cancer diagnosis and treatment, and the top 5 genetic principal components, with censoring at end of follow-up. We then explored whether the CAD-PRS modified the association of incident CAD with chemotherapy and radiation therapy receipt.

Results. Of 1,307 participants, 66 were diagnosed with incident CAD >1 year after their first breast cancer diagnosis. In multivariable models, participants with a CAD-PRS≥median had a 2.47-times increased risk of CAD (95%CI=1.45-4.22) relative to participants with CAD-PRS<median. Anthracycline-based chemotherapy was associated with CAD risk overall (HR=2.04, 95%CI=1.04-3.98), and the association did not differ significantly between CAD-PRS strata. When restricted to irradiated participants (n=882), the association between incident CAD and left-sided radiation therapy was increased for those with CAD-PRS≥median (HR=2.95, 95%CI=1.28-6.80), but not for those with CAD-PRS<median (HR=0.95, 95%CI 0.32-2.85).

Conclusion. A genome-wide CAD-PRS predicts incident CAD risk for breast cancer survivors, providing potential utility in personalized cardiovascular care for breast cancer survivors.

113

Multi-ancestral Maternal GWAS Meta Analysis of Longitudinal Trajectory of Fetal Growth

Prabhavi Wijesiriwardhana[1*], Richard J. Biedrzycki[2], Tesfa Dejenie Habtewold[1], Fasil Tekola-Ayele[1]

[1]Epidemiology Branch, Division of Population Health Research, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, United States of America; [2]Glotech, Inc., contractor for Division of Population Health Research, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, United States of America

Fetal growth sets the foundation for health across lifespan. Previous genome wide association studies (GWAS) in European ancestry populations have identified genetic variants associated with birthweight. However, the genetic architecture of longitudinal fetal growth may not overlap with that of birthweight. We aimed to identify maternal genetic loci associated with longitudinal trajectory of fetal weight throughout pregnancy by combining genome wide datasets of pregnancy cohorts in the United States across groups with African, East Asian, Hispanic, and European ancestry (total N=7,918). For each group, GWAS of the change in fetal weight trajectory was performed using a linear mixed effects-based model adjusted for fetal sex and genotype principal components, with varying slope for gestational week. The results were combined using trans-ancestral meta-regression using the MR-MEGA tool. We identified 57 SNP loci associated with fetal weight trajectory ($p < 5E-8$). Allelic effect heterogeneity was largely correlated with ancestry. Multi-ancestry meta-regression fine mapped some loci. The identified genes were enriched for disease and biological functions such as metabolism, immune response, neuropsychology, organ development, and perinatal death. A genetic risk score (GRS) of the 57 maternal loci was associated with longitudinal fetal weight trajectory, but GRS of previously known maternal loci for birthweight was not. Our study showed that the maternal genetic architecture of in-utero fetal growth trajectory is distinct from that of birthweight. Multi-ancestral cohorts with longitudinal fetal biometry measures advance insights into the biology of fetal growth.

114

A Novel Polygenic Risk Scoring Framework Integrating Common and Rare Variants for Enhanced Genetic Prediction Across Ancestries

Jacob Williams[1,*], Tony Chen[2], Xing Hua[1,3], Wendy Wong[1], Kai Yu[1], Peter Kraft[1], Xihao Li[4,5†,*], Haoyu Zhang[1,†,*]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America; [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [3]Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research Inc, Rockville, Maryland, United States of America; [4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [5]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

**Correspondence to: Jacob Williams (jacob.williams@nih.gov), Xihao Li (xihaoli@unc.edu) and Haoyu Zhang (haoyu.zhang2@nih.gov)*

†These authors jointly supervised this work: Xihao Li and Haoyu Zhang

Polygenic risk scores (PRS) predict complex diseases and traits by aggregating multiple genetic variants. Currently, existing PRS models focus on common variants, missing the potential of rare variants (minor allele frequency < 1%) to uncover the hidden heritability of complex traits. Here we introduce RICE (polygenic Risk predictions Integrating Common and rarE variants), a novel PRS framework optimized for biobank-scale sequencing data, aimed at improving genetic risk prediction accuracy across diverse ancestries. RICE is assessed using extensive simulated datasets and UK Biobank WGS data with over 740 million genetic variants from 137,012 independent individuals across African, European, South Asian, and admixed ancestries with 11 complex traits—six continuous traits (e.g. height, BMI) and five binary traits (e.g. breast cancer, type 2 diabetes). Simulation studies demonstrate RICE yields a significant 71% increase in the average effect size of PRS per standard deviation (SD) compared to top existing common variant methods. Within RICE, the inclusion of the rare variant PRS contributes 38% to the total PRS effect size. In real data analyses of six continuous traits, RICE leads to an average increase of 50% in effect size of PRS per SD, with the highest gain of 81% observed in individuals of African ancestry. For the five binary traits, we observe significant improvements of RICE, including a 51% increase of effect size of PRS per SD for type 2 diabetes in South Asian ancestry. RICE significantly advances PRS by incorporating rare variants, offering a more accurate and inclusive approach to genetic risk prediction.

115

Integration of Mitochondrial DNA Variation Calling with Nuclear Omic Profiles in a Longitudinal Cohort

Phyo W. Win[1,3*], Maya Lekhi[1], Julia Nguyen[1], Yun-Hee Choi[2], Christina A. Castellani[1,2,3]

[1]Department of Pathology and Laboratory Medicine, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario; [2]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario; [3]Children's Health Research Institute, Lawson Research Institute, London, Ontario

Mitochondrial DNA (mtDNA) variation defined as copy number (mtDNA-CN), heteroplasmy, haplotype, and haplogroup are associated with several complex age-related diseases, including cardiometabolic traits and cardiovascular disease (CVD). mtDNA crosstalk with nuclear DNA (nDNA) has emerged as a potential mechanism by which mt influences disease. However, methods for the comprehensive assessment of mtDNA variation have lagged compared to nDNA calling, making integration between mtDNA and nDNA datasets challenging. Further, matched data is often incomplete in large-scale cohorts. To address this, we used the Canadian Longitudinal Study on Aging (CLSA), a middle-aged prospective longitudinal human cohort with matched genomic (Affymetrix Axiom array, N=26,622), epigenomic (Infinium EPIC Microarray,

N=1,479), and metabolomic (Metabolon HD4, N=9,500) data to i) benchmark and improve mtDNA calling methods, ii) determine associations between mtDNA and nDNA features, and iii) establish the mediating relationship between mtDNA, metabolites, and nDNA methylation in CVD risk. We found 738 and 88 methylation sites to be significantly associated with haplogroup and haplotype, respectively (N=1,336, $P<1x10^{-7}$). Haplogroup L was found to be associated with the development of peripheral vascular disease (PVD) (N=26,622, $P=3x10^{-2}$) and a causal mediation effect was established for the nuclear methylation site cg22944890 in PVD. The guanyl-nucleotide exchange factor ($P=3x10^{-4}$) pathways were enriched in haplogroup-associated CpGs. Improvement in methods for mtDNA calling and integration will allow for better understanding of the mechanisms driving mtDNA variation on disease risk in the context of nuclear DNA dynamics.

116

Gestational Diabetes Mellitus Shares Genetic Risk Factors With Type 1 and Type 2 Diabetes and Is Predicted by Diabetes Polygenic Scores

Jessica L.G. Winters[1,2,3*], Elizabeth A. Jasper[3,4,5], Jacklyn N. Hellwege[2,4,6], Todd L. Edwards[4,5], Digna R. Velez Edwards[3,4]

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

Gestational diabetes mellitus (GDM), the most prevalent gestational metabolic disorder, can be defined as hyperglycemia or glucose intolerance arising during pregnancy. This can result in both maternal and fetal morbidity and chronic complications. Although association with type 2 diabetes mellitus (T2DM) is known, genetic connection to type 1 diabetes mellitus (T1DM) is unclear. This study examined if published polygenic risk score (PRS) models for T1DM (PGS002025) and T2DM (PGS003867) correlate with GDM risk. Data from Vanderbilt University Medical Center's BioVU included pregnant, biologically female individuals, at least 18 years old and identifying as Non-Hispanic White (NHW) or Non-Hispanic Black (NHB). GDM cases were identified using ICD codes, with controls having pregnancy related codes. Groups were created with and without preexisting diabetes, resulting in 1014 cases and 5594 controls (4613|1995; NHW|NHB) and 382 cases and 5314 controls (3950|1746; NHW|NHB). Logistic regression assessed PRS associations with GDM, adjusting for BMI, age, and top 10 principal components of ancestry. The T1DM PRS was validated in NHW individuals when diabetics were included but exhibited no significant results ($P<0.05$) upon exclusion. The T2DM PRS was validated in both NHW and NHB individuals when diabetics were included, but only displayed significant associations with GDM in NHW individuals upon exclusion (AUC=0.73; OR=14.58; SE=0.82). Significant genetic correlations were found between GDM and T2DM (RG=0.76; SE=0.06) and T1DM (RG=0.73; SE=0.12). Our findings suggest other diabetic PRS models could be useful predictors of GDM and indicate genetic correlations exist between GDM and both T1DM and T2DM.

Keywords: Diabetes, Risk Assessment, Reproductive Genetics, Complex Traits

117

Investigating the Relationship between Breast Cancer Risk Factors and Mammographic AI-generated Texture Feature

Xueyao Wu[1], Shu Jiang[2], Aaron Ge[1], Constance Turman[3], Graham Colditz[2], Rulla Tamimi[4], Peter Kraft[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America; [2]Washington University School of Medicine in St. Louis, St. Louis, Missouri, United States of America; [3]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [4]Weill Cornell Medical School, New York, New York, United States of America

Background: Mammographic risk score (MRS), an AI-driven texture feature derived from whole mammograms, is strongly associated with breast cancer risk independent of breast density. However, the mechanisms linking MRS to breast cancer remain unclear. We investigated the relationship between established breast cancer risk factors and MRS.

Methods: 383 women from the Nurses' Health Study II were included. MRS were calculated using a model previously trained on 10,126 women. Analyses evaluated associations between MRS and 10 phenotypes across childhood/adult body size, fat distribution, reproductive characteristics, and mammographic density measures through: linear regressions of MRS on each risk factor; on genetic components associated with each risk factor; and two-stage least squares regression and two-sample Mendelian randomization (TSMR) of MRS on each genetically predicted risk factor. Analyses adjusted for age, genotyping platform, and genetic principal components. Associations with P value < 0.05/10 were highlighted.

Results: Higher polygenic scores for dense-area (β = 0.07 SD difference in MRS per SD change in polygenic score; 95%CIs: 0.03-0.11) and percent density (β = 0.06; 95%CIs: 0.02-0.10) associated with increased MRS. TSMR also identified associations between genetically predicted dense-area and MRS (β = 0.82 SD difference in MRS per SD change in density; 95%CIs: 0.39-1.25), and genetically predicted percent density and MRS (β = 1.11; 95%CIs: 0.48-1.74). No evidence supported significant phenotypic/genetic associations with other risk factors.

Conclusion: Apart from mammographic density, no associations/causal effects of other breast cancer risk factors on MRS were detected, possibly due to limited sample size/power. Further validating research is warranted.

118

Identification of shared genetic etiology of cardiovascular and cerebrovascular diseases through common cardio-metabolic risk factors

Xueying Qin[1], Huairong Wang[1], Kun Wang[1], Tao Wang[2], Yiqun Wu[1]*

[1]*Department of Epidemiology and Biostatistics, School of Public Health, Peking University; Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education; Beijing, China;* [2]*Department of Epidemiology and Population Health, Yeshiva University Albert Einstein College of Medicine, Bronx, New York, United States of America*
* Corresponding author

Rational: Cardiovascular diseases (CVDs) and cerebrovascular diseases (CeVDs) are closely related vascular diseases, sharing common cardiometabolic risk factors (RFs). Pleiotropic genetic variants of two diseases have been reported, however, it is still unclear the pathological mechanisms these variants are involved in.

Objective: To unravel the shared genetic etiology of CVDs and CeVDs, by disentangling the pleiotropic variants with their common RFs.

Methods and Results: Leveraging GWAS summary data, we identified 11 colocalized loci for CVDs and CeVDs from the overlapped pleiotropic loci for the separate pairs of CVDs-RFs and CeVD-RFs conditioning on a specific RF. These loci were all related to blood pressure and lipid profiles, rather than type 2 diabetes or body mass index. The 11 loci were mapped to 12 genes, namely *CASZ1*, *CDKN1A*, *TWIST1*, *CDKN2B*, *ABO*, *SWAP70*, *SH2B3*, *LRCH1*, *FES*, *GOSR2*, *RPRML*, and *LDLR*. They were enriched in pathways related to cellular response to external stimulus and regulation of the phosphate metabolic process and were highly expressed in endothelial cells, epithelial cells, and smooth muscle cells. Multi-omics analysis revealed methylation of *CASZ1* and *LRCH1* may play a causal role in the genetic pleiotropy. Notably, these pleiotropic loci are highly enriched in the targets of antihypertensive drugs, which further emphasizes the role of the blood pressure regulation pathway in the shared etiology of CVDs and CeVDs.

Conclusions: The shared genetic components between CVDs and CeVDs are primarily linked to blood pressure and lipid traits, are likely to be regulated through epigenetic mechanisms, and may potentially serve as targets for antihypertensive drugs. This research can provide valuable insights into the underlying mechanisms of these diseases and potentially contribute to the development of more targeted interventions and treatments.

119

Revolutionizing Precision Health for T2D: AI Integration of Multimodal Medical Imaging and Genome-Wide SNP Data in a Large-Scale Biobank

Yi-Jia Huang[1], Chun-houh Chen[1], and Hsin-Chou Yang[1,2,3,4,*]
[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan;* [2]*Biomedical Translation Research Center, Academia Sinica, Taipei, Taiwan;* [3]*Institute of Public Health, National Yang-Ming Chiao-Tung University, Taipei, Taiwan;* [4]*Department of Statistics, National Cheng Kung University, Tainan, Taiwan*

Effective risk assessment and prevention improve patients' quality of life and reduce national insurance costs. This research delves into the realm of precision health for Type 2 Diabetes (T2D) by scrutinizing medical images (abdominal ultrasonography and bone density scan images) in conjunction with whole-genome SNPs to assess T2D risk in a cohort of 17,785 Han Chinese participants from Taiwan Biobank. Densely connected convolutional network deep-learning models were conducted for imaging analysis. In addition, the graphic neural network was applied to account for the intra-individual dependency of multiple images available for each individual. Bayesian statistical learning models were conducted for a polygenic risk score in genetic analysis. We fuse these modalities utilizing the eXtreme Gradient Boosting machine-learning method. Our findings from imaging analysis underscore the superiority of pixel-based analysis over feature-centric methods in terms of accuracy, ushering in automated and cost-effective processes. Furthermore, the incorporation of multi-modality analysis enhances accuracy compared to single-modality approaches. The ultimate model yields a compelling area under the Receiving Operation Curve of 0.954 (ACC = 0.875, SEN = 0.882, and SPE = 0.875) in accurately assessing the risk of T2D. Moreover, our results show that the OR of T2D and its corresponding CI revealed a positive correlation between multi-image risk score (MRS) and T2D and a positive correlation between polygenic risk score (PRS) and T2D. Based on MRS and PRS, we identified specific high-risk subgroups within the study population. This methodology and results significantly advance our comprehension of precision health in the context of T2D.

120

Revolutionizing Precision Health in Type 2 Diabetes: Unveiling the Synergy of Medical Imaging and Genetic Data Integration through Artificial Intelligence in a Large-scale Biobank

Yi-Jia Huang[1], Chun-houh Chen[1], and Hsin-Chou Yang[1,2,3,4,*]
[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan;* [2]*Biomedical Translation Research Center, Academia Sinica, Taipei, Taiwan;* [3]*Institute of Public Health, National Yang-Ming Chiao-Tung University, Taipei, Taiwan;* [4]*Department of Statistics, National Cheng Kung University, Tainan, Taiwan; *Corresponding author: Hsin-Chou Yang, Institute of Statistical Science, Academia Sinica. No. 128, Sec. 2, Academia Road, Nankang 115, Taipei, Taiwan; (Fax) 886-2-27886833; (Tel) 886-2-27875686 (E-mail) hsinchou@stat.sinica.edu.tw*

Effective risk assessment and prevention improve patients' quality of life and reduce national insurance costs. This research delves into the realm of precision health forT2D by scrutinizing medical images (abdominal ultrasonography and bone density scan images) in conjunction with whole-genome single nucleotide polymorphisms, to assess T2D risk in a cohort of 17,785 Han Chinese participants from Taiwan Biobank. Densely connected convolutional network deep-learning models were conducted for imaging analysis. In addition, the graphic neural network was applied to account for the intra-individual dependency of multiple images available for each individual. Bayesian statistical learning models were conducted for a polygenic risk score in genetic analysis. We fuse these modalities utilizing the eXtreme Gradient Boosting machine-learning method. Our findings from imaging analysis underscore the superiority of pixel-based analysis over feature-centric methods in terms of accuracy, ushering in automated and cost-effective processes. Furthermore, the incorporation of multi-modality analysis enhances accuracy compared to single-modality approaches. The ultimate model yields a compelling area under the Receiving Operation Curve of 0.954 (ACC = 0.875, SEN = 0.882, and SPE = 0.875) in accurately assessing the risk

of T2D. Moreover, our results show that the odds ratio of T2D and its corresponding confidence interval revealed a positive correlation between multi-image risk score (MRS) and T2D and a positive correlation between polygenic risk score (PRS) and T2D. Based on MRS and PRS, we identified specific high-risk subgroups within the study population. This methodology and results significantly advance our comprehension of precision health in the context of T2D.

121

Investigation of the Bias from Using Marginal effect sizes of instrumental variables in Mendelian randomization

Yihe Yang, Noah Lorincz-Comi, Xiaofeng Zhu
Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America

Mendelian Randomization (MR) is widely employed to infer causal relationships between exposures and outcomes by leveraging GWAS data. Typically, instrumental variables (IVs) are chosen as the most significant variants from GWAS loci associated with exposures and their marginal effects are used to perform either univariable or multivariable MR analysis. However, this approach can introduce correlated horizontal pleiotropy (CHP) bias when using marginal effects instead of direct effects of IVs, particularly when different causal variants in linkage disequilibrium at a locus are associated with both exposures and outcomes.

We propose selecting independent causal variants and using their estimated direct effects for MR analysis. We examined the causal contributions of HDL-C, LDL-C, triglycerides, BMI, SBP, and T2D on CAD. We estimated the direct effects of variants on CAD and the exposures across the genome using SBayesRC, selected variants with posterior inclusion probabilities larger than 0.9 as IVs, and performed clumping to exact a subset of independent IVs. Using multivariable IVW, the causal effect estimates and P-values for CAD were: HDL-C (EST=-0.025, P=0.121), LDL-C (EST=0.1109, P=2.5E-14), triglycerides (EST=-0.031, P=0.158), BMI (EST=0.035, P=0.181), SBP (EST=0.150, P=3.5E-09), and T2D (EST=0.037, P=0.074). The causal effect of HDL-C on CAD was no longer significant. However, analysis using marginal effect estimates yielded a p-value of 7.13E-13 for HDL-C on CAD, highlighting the bias of using marginal GWAS statistics.

Our findings suggest this bias is likely widespread in MR studies, emphasizing the need for careful selection of IVs based on direct effects.

122

Sex-specific Genetic Risk Factors of Coronary Heart Disease in Hispanic/Latino Populations

Yao Tu[1]*, Geetha Chittoor[2], Anne E. Justice[3], Zhe WANG[3], Alexandre Pereira[4], Andrea R.V.R. Horimoto[4], Elizabeth Frankel[5], Jennifer E. Below[5], Kari E. North[6], Misa Graff [6], Lindsay Fernandez-Rhodes[1]
[1]Department of Biobehavioral Health, Pennsylvania State University, State College, Pennsylvania, United States of America; [2]Department of Population Health Sciences, Geisinger Health System, Port Matilda, Pennsylvania, United States of America; [3] Icahn School of Medicine at Mount Siani, New York City, New York [4] Division of Aging, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [5]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, [6]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Hispanic/Latino (H/L) populations are under-represented in studies of coronary heart disease (CHD), even though they are the largest US racial/ethnic minority. Sex-stratified analyses of coronary heart disease (CHD) are lacking in H/L populations. To explore the sex-specific genetic underpinnings of CHD, we assembled over 110,000 Hispanic/Latino adults with genetic imputation or sequencing data and CHD information from ten electronic health record (EHR)-based biobanks and cohorts.
CHD cases were defined by International Classification of Diseases (ICD)-9/10 codes in EHRs, or with cohort-specific indicators. Genome-wide association results were adjusted by age and principal components, and filtered (minor allele frequency> 0.01, $r^2 > 0.7$). Analyses were run for ~58,000 females (9.8% cases) and ~29,000 males (15.6% cases) separately. Although inverse variance-weight meta-analysis did not reveal any sex-specific loci at genome-wide significance (GWS p<5e-8), 33 loci in females and 21 loci in males reached suggestive significance (p<5e-6). Only one locus was suggestively significant in both sexes (led by rs6978306 in females and rs799936 in males). A formal test of sex-differences revealed 5 loci reached suggestively significant interactions (rs12426934, rs73467552, rs116120189, rs148263928, rs2987741) and a GWS joint p-value at rs55730499 at gene *LPA*. Preliminary observations of sex differences illustrate the critical need for methodologies to account for environmental factors in sex stratified and ancestrally diverse populations. Future work on identifying sex-specific genetic effects and interactions with environmental risk factors will facilitate precise CHD prevention and health interventions in H/L populations.

Keywords: CHD, GWAS, Hispanic/Latino populations, sex-stratified

123

A Framework to Maximise Genetic Diversity in Genome-Wide Association Study Meta-Analyses

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

There have been recent efforts by the human genetics research community to increase the genetic diversity of participants contributing to genome-wide association studies (GWAS). The standard approach is to first assign participants to an "ancestry group" and then aggregate results across ancestry-specific GWAS through multi-ancestry meta-analysis. However, participants are typically assigned "continental" labels that may not reflect personal views of ethnicity/race or fully represent genetic diversity. Furthermore, some admixed participants are excluded because they cannot be assigned to a single ancestry group. Here, we present a novel pipeline for multi-ancestry meta-analysis that employs a continuous and multi-dimensional representation of ancestry by first projecting all participants onto genetic principal components (PCs) derived from the Human Genetic Diversity Project. Association analyses

are then conducted for all participants in a mixed model with adjustment for PCs as covariates. Testing for interaction with PCs enables assessment of ancestry-correlated heterogeneity. We applied the new pipeline to meta-analysis of GWAS of type 2 diabetes in up to 55,090 cases and 503,503 controls of diverse ancestry. Despite increasing the genetic diversity of participants included in the meta-analysis, we demonstrate that association summary statistics were not more inflated for the new pipeline than the standard approach (genomic control lambda 1.103 for both). The new pipeline identifies 187 loci at genome-wide significance ($P<5x10^{-8}$), compared with 178 for the standard approach. The new pipeline increases sample size and the diversity of participants included in multi-ancestry GWAS meta-analysis, offering enhanced power for discovery of loci that are relevant across global populations.

Keywords: meta-analysis, multi-ancestry, statistical genetics

124

Validation of Multi-ancestry Polygenic Scores for Lipid Levels in 3,119 Participants from Samoa and American Samoa

T.J. Yapp[1], M. Krishnan[2], S. Liu[1], S.L. Manna[3], H. Cheng[4], T. Naseri[5], M.S. Reupena[7], S. Viali[8], J. Tuitele[9], R. Deka[4], N.L. Hawley[10], S.T. McGarvey[6, 11], D.E. Weeks[1, 13], R.L. Minster[1], J.C. Carlson[1,13]

[1]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [3]Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [4]Department of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America; [5]Naseri & Associates Public Health Consultancy Firm and Family Health Clinic, Apia, Samoa; [6]International Health Institute, School of Public Health, Brown University, Providence, Rhode Island, United States of America; [7]Lutia i Puava 'ae Mapu i Fagalele, Apia, Samoa; [8]School of Medicine, National University of Samoa, Apia, Samoa; [9]Department of Public Health, Government of American Samoa, Pago Pago, United States of America; [10]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, United States of America; [11]Department of Epidemiology, School of Public Health, Brown University, Providence, Rhode Island, United States of America; [12]Department of Anthropology, Brown University, Providence, Rhode Island, United States of America; [13]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

Dyslipidemia is a major risk factor for cardiovascular disease, the leading cause of death in Samoa. This study applied polygenic scores (PGS) for LDL-C, HDL-C, triglycerides (TG), and total cholesterol (TC) derived from a multi-ancestry meta-analysis to three time-separated Samoan cohorts (1990-1991, 2002-2003, and 2010; total n=3,119). PGS performance was assessed using partial r2 from linear regression models adjusting for age and sex. The PGS for LDL-C had r2 ~8% across the Samoan cohorts, lower than in African American and Hispanic populations from the original study. HDL-C PGS had r2 ~10% for the discovery and 2002 cohort but ~5% in the 1990 cohort. TC PGS had r2 ~10% across cohorts, while TG PGS had r2 ~5-7%. These findings suggest reduced predictive power of multi-ethnic ancestry-derived PGS in Samoans and potential differences in environmental influences across traits. Further research is needed to refine and validate PGS in Samoans and assess their clinical utility for blood lipid traits.

Keywords: Cardiovascular system, polygenic scores, Complex traits, Statistical genetics

125

Multi-Trait Inference of Full Genome-Wide Associations of Type 2 Diabetes Subtypes Uncovers Distinctive Biology and Putatively Causal Genes

Satoshi Yoshiji[1,2,3], Oliver Ruebenacker[1], Patrick Smadbeck[1], Trang Nguyen[1], Ahmet Sayici[1], Chen-Yang Su[4], Edgar Alejandro Llamas Mejia[1,2,3], Jose Florez[1,3,5,6], Miriam Udler[1,3,5,6], Jason Flannick[1,2,3]

[1]Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [2]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [3]Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, United States of America; [4]Quantitative Life Sciences, McGill University, Montreal, QC, Canada; [5]Diabetes Unit, Endocrine Division, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [6]Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America

Type 2 diabetes (T2D) is a heterogeneous disease, and clustering of T2D GWAS variants has suggested multiple genetic clusters or subtypes of T2D with distinctive characteristics. However, the number of known loci for each subtype is limited, their GWAS cannot be directly conducted because they are latent endophenotypes. We propose a novel Bayesian method to infer full genetic associations of latent subtypes of complex traits using observed summary statistics of related traits. Applying it to T2D, we significantly expand our understanding of T2D subtypes.

The method consists of training and classification. During training, we used genetic variants associated with three diabetes subtypes identified through clustering of T2D GWAS variants: 83 for obesity, 42 for beta-cell, and 30 for lipodystrophy. Conditional on observed associations of these variants with 41 T2D-related traits, the model learned latent patterns for obesity (e.g., positive effect for BMI and waist-hip ratio (WHR)), lipodystrophy (negative for BMI but positive for WHR), and beta-cell (negative for fasting insulin). In classification, we inferred genetic associations of 2.3 million variants with each subtype, identifying 387 cluster-associated variants. Variant-to-gene mapping identified 375 cluster-specific causal genes, including *TGFB2* for lipodystrophy, coding an adipokine hormone. Pathway analyses showed enrichment of lipid-related pathways for obesity and lipodystrophy, and insulin secretion-related pathways for the beta-cell subtype.

In summary, this multi-trait Bayesian inference method generates full GWAS of T2D subtypes, identifying putatively causal genes and enrichment of relevant pathways. The method is applicable to any complex trait in any ancestry,

opening avenues to disentangle genetic heterogeneity of complex traits.

Keywords: Complex traits, Diabetes, Genetic clusters, Endophenotypes, Statistical method

126

DKLasso: Bridging Complexity and Interpretability in Genetic Epidemiology through Deep Kernel Learning with Feature Sparsity

Yixiao Zeng[1,2,*], Archer Yang[6], Celia Greenwood[1,2,3,4,5]

[1]PhD Program in Quantitative Life Sciences, McGill University, Montréal, Canada; [2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Canada; [3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Canada; [4]Department of Human Genetics, McGill University, Montréal, Canada; [5]Gerald Bronfman Department of Oncology, McGill University, Montréal, Canada; [6]Department of Mathematics & Statistics, McGill University, Montréal, Canada

In the rapidly evolving field of genetic epidemiology, the need for advanced statistical tools capable of unraveling complex biological phenomena is paramount. We present DKLasso, an innovative adaptation of Deep Kernel Learning (DKL) that emphasizes feature sparsity and model interpretability.

Despite its superiority in quantifying prediction uncertainty and capturing complex, nonlinear patterns in data, DKL often suffers from a lack of feature interpretability. DKLasso embeds global feature selection into the DKL framework, by introducing a linear skip connection, subject to an L1 penalty. This complementary architecture allows explicit feature selection on input variables: non-contributing features excluded by the skip connection are deactivated by the entire DKL model. In simulations, DKLasso offers robust performance, particularly in overfitting-prone scenarios, through a synergistic blend of Gaussian Process-driven regularization and L1-induced feature sparsity.

DKLasso balances the strengths of DKL with the clarity of lasso-like feature selection. Analysis of metabolomic data from the Canadian Longitudinal Study on Aging shows that DKLasso outperforms both state-of-the-art sparse neural networks (LassoNet) and classical DKL models by providing accurate quantification of prediction uncertainty and improved feature interpretability.

DKLasso provides powerful, efficient, and interpretable solutions to complex analyses. It not only expands the statistical toolbox available to researchers, but also promises more accurate, comprehensive, and understandable insights into the intricate tapestry of life science phenomena.

Keywords: Deep Kernel Learning, L1 Regularization, Feature Selection, Gaussian Process

127

Learning Sparse Gaussian Graphical Models from Correlated Data

Zeyuan Song[1,*], Sophia Gunn[2], Stefano Monti[3,4], Gina Marie Peloso[5], Ching-Ti Liu[5], Kathryn Lunetta[5], Paola Sebastiani[1,6,7]

[1]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, (Zeyuan.Song@tuftsmedicine.org; psebastiani@tuftsmedicalcenter.org); [2]The New York Genome Center, New York, New York, United States of America (sgunn@nygenome.org); [3]Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA 02218, United States of America; [4]Bioinformatics Program, Boston University, Boston, MA 02215, United States of America (smonti@bu.edu)

5: Department of Biostatistics, Boston University School of Public Health, Boston, MA (gpeloso@bu.edu, ctliu@bu.edu, klunetta@bu.edu); [6]Tufts University School of Medicine, Boston MA; [7]Data Intensive Study Program, Tufts University, Medford MA

*: Corresponding author at: Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA (Zeyuan.Song@tuftsmedicine.org)

Gaussian Graphical Models (GGM) are widely used in biomedical research to explore the relationships between biological and social factors by providing networks to describe the dependencies between factors of interest. However, contemporary biomedical studies often adopt clustered and longitudinal data collection methodologies, resulting in correlated data. Such correlations among samples or among observations within a sample can lead to false discoveries of the dependencies. Moreover, the complexity of the network analysis amplifies when analyzing high dimensional omics data within and across omics layers in addition to correlated observations. We proposed a Bootstrap-based algorithm to learn GGMs from correlated data. Here, we extend our Bootstrap algorithm to learn sparse GGMs in correlated data by expanding the hypothesis test to . We will show through simulation studies that this extended test does not inflate the false positive rate. By imposing a sequence of values of ranging from 0 to 1, we can investigate the dynamic changes of networks, revealing their contraction and dissection as edges with partial correlations below the threshold are systematically pruned. The application of this method in real data analysis uncovers clusters that are tightly connected in the dynamic changes of the Polygenic Risk Scores and metabolites networks.

Keywords: Gaussian Graphical Models, correlated data, sparse networks, omics data, Polygenic Risk Scores, metabolomics

128

Leveraging genetic similarity to investigate understudied genetic variation associated with EHR-derived phenotypes in diverse patient biobanks

David Y. Zhang[1,2*], Scott M. Damrauer[3,4], 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; [3]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA; [4]Department of Surgery, Perelman School of Medicine, University of Pennsylvania

The lack of representative population diversity in existing genetic studies both hinders our understanding of disease pathobiology and contributes to existing health disparities. We herein propose a tailored genome-first approach that leverages large-scale biobanks to investigate understudied genetic variations. The Penn Medicine Biobank includes over

11,000 individuals genetically similar to African reference populations (AFR) with whole-exome sequencing. For each sequenced variant, we extracted its gnomAD allele counts in the African (AFR) and non-Finnish European (EUR) population groups to identify variants significantly more common in the AFR population. We then filtered for EUR minor allele frequency (MAF) 5%, AFR MAF > 0.05%, and identified ~70,000 protein-altering variants with an AFR/EUR MAF ratio 200. We performed phenome-wide association studies (PheWAS) for each variant against diagnosis codes, laboratory values, and CT-derived imaging traits. We identified significant positive controls such as *HBB* p.Glu7Val with sickle cell anemia (odds ratio [OR]=103.54, p=9.80E-181) and *APOL1* p.Ser342Gly with chronic kidney disease (OR=2.04, p=2.41E-23). Additionally, we found novel associations including *SPATA5L1* p.Arg119Pro with hypocalcemia (OR=0.56, p=4.40E-08), *RAET1G* p.Thr84Arg with gastric ulcers (OR=3.90, p=5.61E-08), and *UNC45B* p.Ala881Thr with hearing loss (OR=3.06, p=8.03E-08). Numerous associations were significantly replicated. *SPATA5L1* p.Arg119Pro was associated with chronic renal failure in the Million Veteran Program (OR=1.10, p=5.63E-17), consistent with its negative association with hypocalcemia. *UNC45B* p.Ala881Thr was associated with hearing loss in the *All of Us* Research Program (OR=1.57, p=0.038). Our approach represents a unique methodological strategy for targeting underexplored variants to better understand genetic risk for disease.

129

Multi-Ancestry Analysis Identifies Susceptibility Variants and Improves Polygenic Risk Scores for Breast Cancer Subtypes

Haoyu Zhang[1,*], Xiaoyu Wang[1,2,*], Thomas U. Ahearn[1], Kyriaki Michailidou[3,4], Roger L. Milne[5,6,7], Jacques Simard[11], Paul D.P. Pharoah[12,13,14], Soo-Hwang Teo[15,16], Marjanka K. Schmidt[17,18,19], Douglas F. Easton[7,20], Peter Kraft[1], Weang Kee Ho[21,22], Montserrat Garcia-Closas[23] on behalf of the Breast Cancer Association Consortium

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America; [2]Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research; [3]Biostatistics Unit, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus; [4]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [5]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia; [6]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia; [7]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia

[11]Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval Research Center, Québec City, QC, Canada; [12]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, California, United States of America

[13]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom; [14]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [15]Breast Cancer Research Programme, Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia; [16]Department of Surgery, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia; [17]Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands; [18]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital Amsterdam, The Netherlands; [19]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands; [20]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom; [21]School of Mathematical Sciences, Faculty of Science and Engineering, University of Nottingham Malaysia, Jalan Broga, Semenyih, 43500, Selangor, Malaysia; [22]Cancer Research Malaysia, 1 Jalan SS12/1A, Subang Jaya, 47500, Selangor, Malaysia; [23]Division of Genetics and Epidemiology, Institute of Cancer Research, London, United Kingdom

*These authors equally contributed to this work

Breast cancer is a heterogeneous disease with distinct molecular subtypes that vary in clinical features and outcomes. Research predominantly focuses on European populations, potentially missing insights from non-European groups. Current polygenic risk score (PRS) models mainly address overall risk and estrogen receptor (ER) status, neglecting subtypes defined by progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and tumor grade. We propose a multi-ancestry analysis to identify genetic variants with varied associations across subtypes and develop risk prediction models including ER, PR, HER2, and tumor grade. This study uses a genome-wide association study (GWAS) including 119,962 breast cancer cases and 106,377 controls from the Breast Cancer Association Consortium (BCAC), comprising 81% European, 16% East Asian, 2% African, and 1% Hispanic populations. We used a two-stage polytomous model within each ancestry to identify subtype-associated variants, followed by fixed-effect meta-analyses. Seven PRS models were implemented, with a super learning model combining these PRSs for enhanced risk prediction. Our analysis identified 11 novel loci (P < 5 x 10-8), with two specific to East Asian populations. The best PRS model achieved an AUC of 0.664 for European, 0.621 for East Asian, 0.572 for African, and 0.621 for Hispanic populations. Subtype-specific PRSs showed higher AUCs for ER/PR+ and HER2- subtypes, particularly in European and East Asian populations. For triple-negative subtypes, PRSs increased AUC by 9.6% on average. Our study demonstrates the effectiveness of multi-ancestry GWAS and PRS approaches in identifying genetic variants and enhancing risk prediction for breast cancer subtypes.

130

Robust Fine-Mapping in the Presence of Linkage Disequilibrium Mismatch

Wenmin Zhang[1,*], Tianyuan Lu[2,3], Josée Dupuis[4], Guillaume Lettre[1,5]

[1]Montreal Heart Institute, Montreal, Quebec, Canada; [2]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [4]Department of Epidemiology, Biostatistics and Occupational

Health, McGill University, Montreal, Quebec, Canada;[5]Department of Medicine, University of Montreal, Montreal, Quebec, Canada
Summary statistics-based fine-mapping methods are widely used. However, linkage disequilibrium (LD) mismatch between the LD reference panel and the GWAS population is common and can lead to compromised accuracy.

We developed RSparsePro, an errors-in-variables model implementing efficient variational inference, to simultaneously perform LD mismatch detection and robust fine-mapping. In simulations, RSparsePro more accurately identified mismatched variants than methods designed for LD mismatch detection while substantially outperforming existing fine-mapping methods in identifying causal variants.

We applied RSparsePro to fine-map the largest GWAS meta-analyses for LDL levels. We constructed ancestry-specific LD reference panels using the UK Biobank European, South Asian, East Asian, and African populations. In 496 genome-wide significant loci, we first performed fine-mapping using ancestry-matched LD reference panels for each population, where LD mismatch was modest. Compared to SuSiE, RSparsePro identified variants in the 95% credible sets that were 1.89-fold (95% CI: 1.32-2.71) more likely to have protein-altering effects. Next, for each non-European ancestry population, we performed fine-mapping using the European LD reference panel, which artificially induces severe LD mismatch. Importantly, 62.4%, 51.9%, and 68.4% of the credible sets obtained in South Asian, East Asian, and African populations by RSparsePro using the incorrect, European LD reference panels were still consistent with those based on ancestry-matched LD reference panels. In contrast, with SuSiE, the consistency rates were 42.3% (South Asian), 22.5% (East Asian), and 16.2% (African).

RSparsePro will greatly expand the applicability of fine-mapping analyses, especially in increasingly larger GWAS involving multiple cohorts and diverse populations where in-sample LD is unlikely available.

131

scPrediXcan: A Method for Transcriptome-Wide Association Studies at Cell-type Level Using Deep Learning

Yichao Zhou[1,*], Temidayo Adeluwa[1], Saideep Gona[1], Lisha Zhu[2], Festus Nyasimi[3], Ravi Madduri[4], Mengjie Chen[2], Hae Kyung Im[2]
[1]Division of Biological Sciences, Committee of Genetic, Genomics, and Systems Biology, University of Chicago, Chicago, Illinois, United States of America; [2]Division of Biological Sciences, Department of Medicine, Section of Genetic Medicine, University of Chicago, Chicago, Illinois, United States of America; [3]Division of Biological Sciences, Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America; [4]Data Science and Learning Division, Argonne National Laboratory, Chicago, Illinois, United States of America

Transcriptome-wide association studies (TWAS) have been key in identifying genes linked to complex traits and diseases but often fail to pinpoint disease mechanisms at the cellular level. Whereas TWAS approaches use tissue-level prediction models, our association method (scPrediXcan) employs a deep learning model, scPred, trained on a reference genome and single-cell RNAseq data. scPred is a lightweight multi-layer perceptron that uses an existing sequence-to-epigenomics model (Enformer) as a feature extractor and predicts the gene expressions at single-cell pseudobulk level. scPred predicts gene expression with high accuracy: Pearson correlations R= 0.75–0.89 in all 51 cell types tested. Moreover, scPred surpasses current linear TWAS gene expression prediction models in predicting pseudobulk expression levels across individuals when comparing the absolute Pearson correlations. For TWAS, we compared the performance of scPrediXcan and the canonical TWAS method, PrediXcan, in type 2 diabetes (T2D) and systemic lupus erythematosus (SLE). We used the curated T2D gene set from Common Metabolic Diseases Knowledge Portal as a silver standard T2D gene list to calculate power; scPrediXcan (power=0.429, precision=0.107) outperformed PrediXcan trained on two datasets separately: GTEx bulk (power=0.122, precision=0.108) and T2D pseudobulk (power=0.112, precision=0.115). For SLE, scPrediXcan identified 176 candidate causal genes in T cells distributed at 23 genomic loci, whereas PrediXcan trained by GTEx whole blood only identified 54 candidates at 14 genomic loci. Overall, our results demonstrate that scPrediXcan offers a significant advance in the precision and relevance of gene nominations in TWAS, promising to deepen our understanding of the cellular mechanisms underlying complex diseases.

132

Genetically Regulated Prediction Modeling in Lipidomics and Transcriptomics in a Hispanic Cohort

W. Zhu[1*], H-H. Chen[2], A. Petty[1], J. Curran[3], P. Meikle[4,5], J. McCormick[6], S. Fisher-Hoch[6], K. North[7], E. Gamazon[1,8], 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]Academia Sinica, Taipei, Taiwan; [3]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; [5]Baker Department of Cardiometabolic Health, University of Melbourne, Parkville, State of Victoria, Australia; [6]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; [7]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [8]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom

Integrated analysis of multi-omics data has proven to provide comprehensive insights to complex biological mechanisms. Although predicted gene expressions are widely used in genetic studies, a similar approach is lacking in the human lipidome. Characterizing and contrasting the performance of predictive models in new data resources is needed to ensure rigor.

In the Cameron County Hispanic Cohort (CCHC), we trained elastic net prediction models using 10-fold cross-validation on 830 lipidome traits through genetic variants selected by GWAS p values on a training set of 1606 samples. True model performance was evaluated by Pearson r in a separate test set of 681 samples. While the overall model performance is high in the training set (mean r=0.67, std=0.07), low performance

was observed in the test set (mean r=0.11, std=0.07). Models in 40 lipid species have moderate performance (r>0.3). The same modeling method was applied in the transcriptomics data utilizing SNPs in the +/- 1MB region of each gene. With a training set of 1029 samples and a test set of 259 samples, mean model performance was similar in the training set (mean r=0.29, std=0.16) and the test set (mean r=0.18, std=0.19). Out of 25,782 genes, we identified 5,321 models with performances greater than 0.3.

Our result suggests that while both lipidome traits and transcripts can be predicted through genetic variants, modeling is more challenging in lipidopmics due to easier overfitting. Further evaluation, larger sample sizes or modified statistical approach may be needed to rigorously leverage genetically regulated predictions of the lipidome.

133

Rare Variant Analysis Pipelines: A Systematic Review

Cristian Riccio[1,2], Max L. Jansen[1,2], Felix Thalén, Hugo Solleder[1,2], Andreas Ziegler[1,2,3,4,5]

[1]Cardio-CARE, Medizincampus Davos, Herman-Burchard-Str. 1, 7265 Davos Wolfgang, Switzerland; [2]Swiss Institute of Bioinformatics, Davos, Switzerland; [3]Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [4]University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [5]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Corresponding author: Andreas Ziegler, ziegler.lit@mailbox.org

The sequencing of increasingly larger cohorts has unveiled many rare variants, presenting an opportunity for further unravelling the genetic basis of complex traits. In contrast to the analysis of common variants, rare variant analyses are more complex. Their analysis has been facilitated by computational tools, which should be flexible and easy to use. However, an overview of the available rare variant analysis pipelines and their functionalities is lacking. In this work, we thus provide a systematic review of the currently available rare variant analysis pipelines. We searched MEDLINE and Google Scholar until November 27, 2023, and included open-source rare variant pipelines that accept genotype data from cohort and case-control studies and group variants into testing units. Eligible pipelines were assessed based on functionality and usability criteria. We identified 17 rare variant pipelines that collectively support a variety of trait types, association tests, testing units, and variant weighting schemes. No single pipeline can currently handle all data types in a scalable and flexible manner. We recommend different tools to meet diverse analysis needs. STAARpipeline is suitable for beginners thanks to its built-in definitions for testing units. REGENIE is a scalable pipeline that is actively maintained, regularly updated, and well-documented. Ravages is suited for analyzing multinomial variables, while OrdinalGWAS is tailored for analyzing ordinal variables. There remains an opportunity to develop a user-friendly pipeline providing high degrees of flexibility and scalability. Such a pipeline would enable researchers to exploit the potential of rare variant analyses in uncovering the genetic basis of complex traits.

Keywords: Exome sequencing, Genome sequencing, Genome-wide association, Rare variants, Statistical genetics

134

The First Case of Alopecia-Intellectual Disability Syndrome 4 in a Filipino Newborn

J.M. Yabut, A.K. Esguerra, M.J. Racoma

St. Luke's Medical Center, Institute of Pediatrics and Child Health, Global City, Philippines

Alopecia-Intellectual Disability Syndrome (APMR) is a rare autosomal recessive neurocutaneous disorder characterized by alopecia and varying degrees of intellectual disability. As of 2023, there were only 29 reported families affected with APMR worldwide. We report a live late preterm female born to a non-consanguineous Filipino family, presented with alopecia, localized skin collodion membrane, syndactyly, and brain structural abnormalities on imaging such as ventriculomegaly and lissencephaly. Genetic testing revealed a pathogenic variant in the LSS gene consistent with APMR4. After extensive work-up, the patient was discharged stable after 10 days hospital stay with close monitoring and follow up. Notably, this case expands the phenotypic spectrum of APMR4, adding syndactyly, localized collodion membrane, and brain abnormalities on imaging (lissencephaly and ventriculomegaly) as novel features. APMR is extremely rare with limited prognostic and therapeutic data available. Documenting such cases is crucial for expanding our understanding of this condition, potentially leading to advancements in treatment strategies and improved patient outcomes.

Keywords: Alopecia in newborns, APMR, Alopecia Intellectual Disability Syndrome

135

Loci on Chromosome 20 Interact with rs16969968 to Influence Cigarettes per Day in European Ancestry Individuals

Pamela N. Romero Villela[1,2*], Luke M. Evans[1,3], Teemu Palviainen[4], Richard Border[5], Jaakko Kaprio[4], Rohan H. C. Palmer[7], Matthew C. Keller[1,2], Marissa A. Ehringer[1,5]

[1]Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, Colorado, United States of America; [2]Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, Colorado, United States of America; [3]Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, Colorado, United States of America; [4]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland; [5]Departments of Neurology and Computer Science, University of California Los Angeles, Los Angeles, California, United States of America; [6]Department of Integrative Physiology, University of Colorado Boulder, Boulder, Colorado, United States of America; [7]Behavioral Genetics of Addiction Laboratory, Department of Psychology, Emory University, Atlanta, Georgia, United States of America

Background: The understanding of the molecular genetic contributions to smoking is largely limited to the additive effects of individual single nucleotide polymorphisms (SNPs), but the underlying genetic risk is likely to also include dominance, epistatic, and gene-environment interactions.

Methods: To begin to address this complexity, we attempted to identify genetic interactions between rs16969968, the most replicated SNP associated with smoking quantity, and all SNPs and genes across the genome.

Results: Using the UK Biobank European subsample, we found one SNP, rs1892967, and two genes, *PCNA* and *TMEM230*, that showed a significant genome-wide interaction with rs16969968 for log10 CPD and raw CPD, respectively, in a sample of 116 442 smokers of European ancestry. We extended these analyses to individuals of South Asian descent and meta-analyzed the combined sample of 117 212 individuals of European and South Asian ancestry. We replicated the gene findings in a meta-analysis of five Finnish samples (N=40 140): FinHealth, FINRISK, Finnish Twin Cohort, GeneRISK, and Health-2000-2011.

Conclusions: To our knowledge, this represents the first reliable epistatic association between single nucleotide polymorphisms for smoking behaviors and provides a novel direction for possible future functional studies related to this interaction. Furthermore, this work demonstrates the feasibility of these analyses by pooling multiple datasets across various ancestries, which may be applied to other top SNPs for smoking and/or other phenotypes.

136

Evaluating Methods for Genome-Wide Associations Studies in Diverse Ancestral Populations

Julie-Alexia Dias[1], Tony Chen[1], Alex A. Rodriguez[2], Ravi K. Madduri[2], Peter Kraft[3,*], Haoyu Zhang[3,*]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois, United States of America; [3]Division of Cancer Epidemiology & Genetics, National Cancer Institute, National Institute of Health, Bethesda, Maryland, United States of America

*Correspondence to: Peter Kraft (phillip.kraft@nih.gov) Haoyu Zhang (haoyu.zhang2@nih.gov).

Ancestrally diverse biobanks enable studying complex traits across various ancestries by leveraging differences in linkage disequilibrium (LD) and allele frequencies. However, the effectiveness of different genome-wide association study (GWAS) methods in multi-ancestry groups is debated. Pooled methods avoid classifying individuals into ancestry groups, benefiting admixed populations. Conversely, meta-analysis methods may better account for finer population stratification but involve arbitrary population binning, struggling within small ancestry groups.

We compared pooled and meta-analysis methods (fixed-effects meta-analysis and MR-MEGA) by simulating traits using UK Biobank (UKB) and All of Us (AoU) data, with and without population structure, on both ancestry-group-specific and global scales. We tested these methods' ability to control type I error and evaluated power for continuous and binary traits using a large-scale simulated multi-ancestry genotype dataset with realistic LD, similar to the 1000 Genomes Project dataset with 600K subjects in five ancestral groups.

In power tests, pooled analysis outperformed both meta-analysis methods. Within meta-analysis, MR-MEGA showed better power over fixed-effects meta-analysis, particularly in scenarios with substantial non-European proportions and large sample sizes. We corroborate our simulation findings using multi-ancestry data from UKBB (N ≈ 340K) and AoU (N ≈ 310K).

Pooled analysis increases power where allele frequencies between ancestries diverge and representation is balanced. While pooled analysis may have limitations in correcting for recent population structure in some settings, it offers significant advantages in multi-ancestry cohorts. MR-MEGA outperforms fixed-effects meta-analysis under certain conditions, showing the importance of method selection based on cohort characteristics and study goals.

137

NMR Metabolomics Data as a Powerful Predictor of Mortality and Biological Age in Estonian Biobank

Māra Delesa-Vēliņa[1]*, Krista Fischer[1,2], Estonian Biobank Research Team[2]

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

There has been great interest in studying nuclear magnetic resonance (NMR) metabolomics data as a predictor of overall mortality and a proxy for biological age in biobanks data. As the number of participants having NMR data in the Estonian Biobank exceeds 200,000, we aim to develop a model for all-cause mortality and biological age estimation based on NMR data for the Estonian Biobank.

We develop the model based on the first wave of biobank participants, who have the longest follow-up time (recruitment 2002–2010, mean follow-up 13.3 ± 4.4 years). We use the second wave of data for validation (recruitment 2018 onwards, mean follow-up 4.2 years ± 0.6 years).

We employ a Cox model with age as a timescale and stepwise selection to identify biomarkers independently associated with mortality. We model an individual's survival probability parametrically, using NMR score and phenotype as covariates. We define the biological age of a particular individual as the age of an average individual with the same survival probability.

The resulting NMR score comprises 17 metabolic biomarkers and is highly associated with mortality in both the development and validation sets, with hazard ratios (per SD of NMR score) of 1.77 (95% CI 1.73–1.82) and 1.83 (95% CI 1.77–1.90), respectively. Our proposed biological age estimate is a better predictor of 5-year mortality than chronological age, with AUC (improvement compared to chronological age) of 0.890 (0.056) and 0.851 (0.043) in the discovery and validation sets, respectively.

138

Multi-ancestry Meta-analysis Identifies Genetic Modifiers of Age-at-onset of Alzheimer's Disease at Known and Novel loci

Elizabeth Blue[1*], Jai Broome[1], Diane Xue[1], Stephanie Gogarten[2], Adam Naj[2], Ellen Wijsman[1]

[1]University of Washington, Seattle, WA, United States of America; [2]University of Pennsylvania, Philadelphia, PA, United States of America

*Correspondence to: Elizabeth Blue, em27@uw.edu,

While Alzheimer's disease (AD) is highly heritable, much of AD risk is explained by age, ancestry, *APOE* genotype, and sex. Genome-wide association study (GWAS) signals can vary dramatically in strength and direction of effect when studies vary by these factors. GWAS of age-at-onset of AD (AAO) can offer increased power; despite their smaller numbers and sample sizes. We investigated the relationship between sex, *APOE*, and genetic modifiers of AAO in two diverse data sets: a discovery cohort representing 16 publicly available AD GWAS data sets and the independent Alzheimer's Disease Genetics Consortium (ADGC) data. We performed relatedness estimation, principal components (PCs) analysis, local ancestry inference, and GWAS with and without adjustment for sex and *APOE* genotype, followed by meta-analysis. Subjects with >20% non-European ancestry were well-represented in both the discovery (44% of 21736) and ADGC (36% of 19483) data. Covariate adjustment for sex and sparsifying the genetic relatedness matrix had minimal effects on GWAS effect sizes ($r^2 > 0.96$), whereas *APOE* adjustment influenced effect sizes beyond chr19 ($r^2 \leq 0.78$). Meta-analysis of GWAS adjusted for sex, PCs, and relatedness revealed 13 significant loci ($p < 5E-08$), including one novel signal (9q21.32) and 12 at AD risk loci. The *APOE*-adjusted model revealed 17 significant loci, including four novel signals (9p22.3, 9q21.33, 11q13.1, 18p11.21). Our results reinforce the power of an AAO approach by identifying novel signals and replicating several signals only recently discovered in GWAS representing hundreds of thousands of cases and controls with predominantly European ancestry.

139

Genetic Factors Associated with Depression and Cognitive Decline

Jessica K. Dennis[1*], Karanvir Singh[1], Sabine Bonnor[1], Graham Boucher[1]

[1]*Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada*

[*]presenting author

Background: Depression in dementia is common, yet poorly treated by available therapies. We hypothesize that depression in dementia is etiologically distinct from primary depression, and that genetic analyses will elucidate these differences. Since the neurodegenerative processes that lead to dementia begin decades before diagnosis, we focus on depression and cognitive decline, before dementia onset.

Methods: We used data on 11,958 participants in the Canadian Longitudinal Study on Aging who were aged 65-85 at baseline in 2010-2015, and who have been followed prospectively thereafter. Cognitive function was assessed at up to three different time points. Depressive symptoms were assessed at up to five different time points. We used growth curve modeling to quantify cognitive function and depressive symptom trajectories spanning up to nine years. We created polygenic scores for Alzheimer's disease (AD) and major depressive disorder (MDD) using the most recent GWAS summary statistics.

Results: Participants who were older at baseline experienced the fastest rate of cognitive decline, with no differences between males and females. Increasing genetic risk for AD (i.e., PGS for AD and/or *APOE* e4 genotype) associated with cognitive function trajectories, validating these trajectories as AD endophenotypes. In preliminary analyses, people who experienced new onset depression were more likely to experience cognitive decline than people with persistent or no depression. Next, we will test genetic risk for AD and PGS for MDD for association with new onset depression, while accounting for cognitive function trajectories. Our findings will help guide new treatments for depression in dementia.

140

Enhancing Antidepressant Response Prediction in Late-Life Depression Across Diverse Ancestries Using BridgePRS

Samar S. M. Elsheikh[1], Daniel Felsky[1,4,5], James L. Kennedy[1,4,5], Benoit H. Mulsant[1,4,5], Charles F. Reynolds 3rd[6], Eric J. Lenze[7,] Daniel J. Müller[1,3,4]

[1]*Campbell Family Mental Health Research Institute, Center for Addiction and Mental Health, Toronto, Ontario, Canada;* [3]*Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada;* [4]*Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada;* [5]*Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada;* [6]*Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America;* [7]*Healthy Mind Lab, Department of Psychiatry, Washington University, St. Louis, Missouri, United States of America*

Background. Antidepressant treatment response has an estimated heritability of ~30-40% (Tansey et al., 2013), varying by population and methodology. Antidepressant remission (AR) in late-life depression (LLD) is challenging due to factors like comorbidities and age-related pharmacokinetic changes. Ancestry-specific genetic factors affecting drug metabolism can introduce biases if not properly accounted for. Tools like BridgePRS optimize multi-ancestry and ancestry-specific polygenic risk scores (PRS) by leveraging shared genetic effects across ancestries.Aim. This study aims to investigate the added predictive power of AR in LLD across diverse ancestries. Specifically, we will construct PRSs using BridgePRS and assess its performance in predicting AR.

Methods. BridgePRS will use antidepressant non-remission (ANR) GWAS summary statistics from European and East Asian populations (Pain O et al., 2021) to create weighted, ancestry-corrected PRS estimates to predict AR. The performance of PRS from standard and BridgePRS methods will be compared using adjusted R-squared and AUC-ROC.

Results. The study used IRL-GRey trial data (N = 342) of older adults, ~89% European ancestry, including African, Asian, and Indian individuals, treated with venlafaxine for 12 weeks. Preliminary results using standard PRS-PCA methods showed nominal associations between AR and PRS for bipolar disorder (OR = 0.75 [0.58, 0.97], p = 0.031) and ADHD (OR = 1.34 [1.06, 1.70], p = 0.016), but not ANR (OR = 0.95 [0.74, 1.22], p = 0.70). BridgePRS for ANR only (due to available non-European GWAS) will be computed.

Conclusion. By leveraging BridgePRS for multi-ancestry samples, this study aims to enhance AR prediction in LLD, advancing personalized treatment strategies.

Keywords:Antidepressant, remission, polygenic risk scores, Ancestry