Abstracts

INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

The 2022 Annual Meeting of the International Genetic Epidemiology Society

1

Mitochondrial Genome-wide Association Study of Pancreatic Cancer

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

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

Pancreatic ductal adenocarcinoma (PDAC) remains to have a very poor prognosis. Considering that somatic mutations in the mitochondrial genome have shown to be associated with many cancers and age-related diseases, we have investigated the relationship between mitochondrial single-nucleotide genetic variants (mtSNVs) and pancreatic cancer. Hence, in this study, we applied a genome-wide mitochondrial association analysis between a total number of 5,595 mtSNVs and PDAC in a sample of 185 affected and 50 healthy individuals. The study comprises sporadic cases as well as cases from the German National Case Collection of Familial Pancreatic Cancer (FaPaCa). We achieved an average of 3,093-fold coverage of the mitochondrial genome (25% quantile = 2,272, 75% quantile = 3,949) through next-generation sequencing. We derived mitochondrial heteroplasmy through our previously established pipeline for processing and quality control of mitochondrial sequencing data. The association analysis was performed using a reverse approach with heteroplasmy at a particular mtSNV as outcome in a log-linear regression model and disease status as independent variable, including adjustment for age, sex, and coverage.

Based on the accurate heteroplasmy determined from sequencing data and after adjustment for multiple testing, we were able to identify two different mitochondrial variants negatively associated with PDAC (genes MT-HV2, MT-HV3) and a positively associated variant (gene MT-16SrRNA). Our findings provide first insights into the role of mtDNA variants regarding the pathogenesis of PDAC. They shed light on the potential of studying mitochondrial genetic variants for better understanding the development of PDAC.

2

Sex Hormones Minimally Contribute to the Effect of Body Mass Index on Reproductive Dysfunction: A Mendelian Randomization Study

Ky'Era V. Actkins[1,2]*, Nikhil Khankari[2], and Lea K. Davis[2]

[1]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Polycystic ovary syndrome (PCOS) is a highly prevalent endocrine disorder in females that increases the risk of developing cardiometabolic disorders (CMD). To date, Mendelian Randomization (MR) has provided little support for PCOS as a causal risk factor for CMD, suggesting alternative pathways and/or mediators as a potential explanation for the increased risks observed. Multivariable (MV) and bidirectional MR analyses were implemented to identify potential direct effects of PCOS on CMD, and potential reverse causation of CMD on PCOS risk, respectively. First, bidirectional MR analyses were conducted between cardiometabolic traits (i.e., body mass index (BMI), T2D, blood pressure, etc.) and PCOS. We observed BMI (β=0.65, *P*=<0.001) and T2D (β=0.09, *P*=<0.001) to be statistically significant risk factors for PCOS, but PCOS was not a significant exposure for any CMD trait. We then examined the potential effect of PCOS on CMD traits while adjusting for female testosterone (FT) and sex-hormone binding globulin (SHBG) using MVMR. After independent adjustment for FT and SHBG, PCOS was associated with a modest increase in BMI (β_{SHBG}=0.02, *P*=<0.001; β_{FT}=0.02, *P*=<0.001). When examining whether CMD are risk factors for PCOS, BMI remained strongly associated with PCOS after adjustment (β_{SHBG}=0.45, *P*=<0.001, β_{FT}=0.44, *P*=<0.001). No

significant associations were observed with T2D for either direction. Hormones may play an important role for the association between PCOS and CMD, and may influence the associations observed in epidemiologic studies. Further research is needed to elucidate these pathways and understand the independent, non-hormonal effects of PCOS genetic risk that is modifiable by BMI.

3

Benchmarking of Univariate Pleiotropy detection Methods, with an Application to Epilepsy Phenotypes

Oluyomi M. Adesoji[1*], Herbert Schulz[2], Patrick May[3], Roland Krause[3], Holger Lerche[4], Michael Nothnagel[1,5], ILAE Consortium on Complex Epilepsies

[1]*Cologne Center for Genomics, University of Cologne, Cologne, Germany;* [2]*Department of Microgravity and Translational Regenerative Medicine, Clinic of Plastic, Aesthetic and Hand Surgery, Otto von Guericke University, Magdeburg, Germany;* [3]*Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg;* [4]*Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany;* [5]*University Hospital Cologne, Cologne, Germany*

Complex traits have been found to share genetic relationships in separate association studies but the joint study of phenotypes especially through pleiotropy analysis improves the power to discover associations, drug target development and provision of effective therapy for affected individuals. Pleiotropy is the phenomenon of a hereditary unit affecting more than one trait. However, due to separate data collection and analysis of individual traits, the application of multivariate pleiotropy detection approaches is often not feasible. Therefore, to identify these switch-like or shared loci in multiple traits, we benchmarked five univariate pleiotropy detection methods namely, classical meta-analyses (MA), conditional false discovery rate (cFDR), subset-based meta-analysis (ASSET), cross-phenotype Bayes (CPBayes) and pleiotropic analysis under the composite null hypothesis (PLACO), though simulation of datasets with different underlying aetiologies, to identify the most powerful approach that keeps type 1 error low. The ASSET method which gave a good trade-off between power and false-positive rate (FPR) is then applied to summary statistics of epilepsy phenotypes provided by the ILAE Consortium on complex epilepsies, thereby identifying already known and new putative epilepsy loci. Based on our results, the classical meta-analysis approach is not recommended for pleiotropy detection.

4

Lessons From the Past: Genetic Ancestry and Demographic History of the Nicoyan Peninsula, a Longevity Hot-spot

Paola Arguello-Pascualli[1,2,3]*, Santiago G. Medina-Muñoz[4], Carmina Barberena-Jonas[4], Nicole Gladish[5], Sarah M. Merrill[1,2,3], David H. Rehkopf[5], Michael S. Kobor[1,2,3], Andrés Moreno-Estrada[4], Jessica Dennis[1,2]

[1]*Department of Medical Genetics, Faculty of Medicine, University of British Columbia, Vancouver, Canada;* [2]*British Columbia Children's Hospital Research Center, University of British Columbia, Vancouver, Canada;* [3]*Center for Molecular Medicine and Therapeutics, Vancouver, British Columbia;* [4]*Center for Research and Advanced Studies of the National Polytechnic Institute, Irapuato, Mexico;* [5]*Department of Epidemiology and Population Health and Department of Medicine, School of Medicine, Stanford University, California, United States of America.*

Little is known about the genetics of Costa Rica and its Nicoyan Peninsula, a longevity hotspot worldwide. The reasons behind the Nicoyans increased health and life span remain unclear, including the extent to which genetic ancestry contributes. European and African migrations following the colonization, brought together previously separated populations, resulting in admixture events that can be genetically recapitulated. We performed global ancestry analysis on 500 Costa Rican genomes, including 100 from the Nicoyan Peninsula. Our results revealed an average increment of 12% African and 27% Indigenous American (NAT) ancestry in Nicoyans compared to non-Nicoyans. To explain if these differences could be the result of differing demographic histories, we used genetic-based migration modelling. Earlier and singular migrations of European and African ancestry were seen in Nicoya, whereas smaller but constant African migrations were observed in the rest of Costa Rica, matching demographic and historical records of the country. Lastly, to assess the association of genetic ancestry and longevity meta-analysis, we tested 18 genetic variants previously associated to longevity in a Genome-Wide Association Study of over 26,000 participants. Preliminary data show a significant enrichment in NAT for 8 sites falling in the intronic region of CDKN2B-AS1, a gene associated to longevity and health span in several independent studies.

Taken together, our analysis show that genetic ancestry can be an incredible resource to deepen our understanding of health-related traits in modern-day populations.

5

Shared Genetic Aetiology of Osteoarthritis and Type 2 Diabetes

Ana Luiza de S. V. Arruda[1,2,4*], April Hartley[5], Konstantinos Hatzikotoulas[1], William Rayner[1], Andrei Barysenka[1], Georgia Katsoula[1,4], George Davey Smith[5], Andrew Morris[1,6], Eleftheria Zeggini[1,3]

[1]*Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany;* [2]*Munich School for Data Science, Helmholtz Zentrum Munich, Neuherberg, Germany;* [3]*TUM School of Medicine, Technical University Munich and Klinikum Rechts der Isar, Munich, Germany;* [4]*TUM School of Medicine, Technical University of Munich, Graduate School of Experimental Medicine, Munich, Germany;* [5]*MRC-Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom;* [6]*Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom*

Osteoarthritis (OA) and type 2 diabetes (T2D) are two of the most prevalent chronic health disorders worldwide. Observational studies report a positive epidemiological association between the diseases beyond their common risk factors, such as obesity and increasing age. Taking into consideration that the world's obesity rates, and average age are rising, this comorbidity pair can be considered an increasing global health challenge. Thus, in this research project, we aim to disentangle the genetic correlation between OA and T2D.

Using summary statistics of large-scale GWAS from T2D (n=898,130) and knee or hip related OA phenotypes (n=490,345), we investigate the genetic intersection between the traits by performing statistical colocalization analysis of established association signals. For colocalizing regions, we derive a set of high confidence likely effector genes based on biological lines of evidence, including colocalization with molecular QTL from disease relevant tissues. Additionally, for each of those genes, we perform Mendelian randomization analyses between expression QTL and each disease. Eighteen genome regions show robust evidence of colocalization between T2D and at least one OA phenotype. Nineteen genes were defined as high confidence likely effector genes, including *TCF7L2* and the obesity related *FTO* gene. *TCF7L2* is among the leading signals for T2D risk but had not been identified as associated with OA at genome-wide significance levels yet. We find statistical evidence that *TCF7L2* expression in pancreatic islets is causal for T2D and protective against knee related OA phenotypes. Our shared effector genes support the epidemiologically known link between BMI and the investigated comorbidity.

6

The Value of Genetic Data from 665,460 Individuals in Predicting Anemia and Suitability to Donate Blood

Jarkko Toivonen[1], Johanna Castrén[1], FinnGen, Mikko Arvas[1*]
[1]Finnish Red Cross Blood Service, Helsinki, Finland

Background/Objectives: We aim to find out whether genetic information has significant effect in predicting iron deficiency anemia or blood donation suitability.

Methods: Genetic data from FinnGen release 6 (230,000 participants), Blood Service Biobank (30,000 participants) and UK Biobank (400,000 participants) were analyzed. In addition, age, sex, weight, height, anemia status and blood donation histories were used when available.

We performed GWAS for anemia and blood donation ability and computed polygenic risk score weights for anemia, ferritin (based on Bell & al 2021 GWAS) and hemoglobin (based on Vuckovic & al 2020 GWAS). A Bayesian logistic regression for anemia was fitted on all FinnGen participants and for donation ability on blood donors, stratified into three demographic groups.

Results: A single significant SNP rs199598395 in gene RNF43 was revealed by our anemia and deferral GWAS. The meta-analysis from FinnGen and UKBB for anemia provided three more significant lead SNPs. The largest effect of the genetic data in anemia model was by the RNF43 SNP for pre-menopausal females (OR 2.9, CI 2.1 – 4.0) and in the deferral model again the RNF43 SNP for pre-menopausal females (OR 3.3, CI 2.0 – 5.3). The effect of PRSs was slightly different from zero in some models.

Conclusion: A single SNP can have a strong effect on prediction of both anemia and blood donation suitability. PRSs are likely to become more useful in the future.

7

Linking the Joint Genetic Structure of Neuroanatomical Phenotypes with Psychiatric Disorders

Antoine Auvergne*[1], Nicolas Traut[2], Hanna Julienne[1], Lucie Troubat[1], Sayeh Kazem[1], Roberto Toro[2], Hugues Aschard[1]
[1]Génétique Statistique, Département de Biologie Computationnelle, Institut Pasteur, Paris, France
[2]U5 Neuroanatomie Appliquée et Théorique, Département de Biologie Computationnelle, Institut Pasteur, Paris, France

There are increasing evidences of genetic correlations between mental disorders and brain magnetic resonance imaging (MRI) phenotypes. However, deciphering the joint genetic architecture of these outcomes has proven challenging, and new approaches are needed for inferring potential genetic structure underlying those phenotypes. Here, we investigated how clusters of genetic variants with similar neuroanatomical multi-trait associations can be linked to psychiatric disorders.

We first conducted univariate and multi-trait Genome-Wide Association Studies from nine MRI derived brain volume phenotypes in 20K UK Biobank participants and identified a total of 1,392 independent variants including 32% detected only in the multi-trait analysis. Those variants display significant enrichment for association with almost all mental disorders we considered: bipolarity (p=2.7e-13), attention-deficit/hyperactivity disorder (p=1e-9), autisms (p=1.5e-8), schizophrenia (p=2e-74), obsessive-compulsive disorder (p=0.2) and major depressive disorder (p=7.4e-5).

We next clustered those variants based on their multi-trait association with MRI phenotypes using an optimized k-medoids approach, and assess the ability of those clusters to distinguish disease-associated and non-disease-associated variants. We considered 2-15 clusters per disease, and search for clustering that maximize the variance of the enrichment. For bipolarity, this maximum was achieved with six clusters, with enrichment concentrated in two clusters (p<1e-5), one related to caudate and putamen, and a second related to accumbens and amygdala. For obsessive-compulsive disorder, maximum variance was achieved for eight clusters, with none except one cluster being significantly enriched (p=3e-3) and related to accumbens and amygdala. Both results are in agreement with the recent literature, highlighting the relevance of our data-driven approach.

Grants: This research was supported by the FRM (ECO202106013759). This research has been conducted using the UK Biobank Resource under Application Number 18584.

8

The Individual and Global Impact of Copy Number Variants on Complex Human Traits

Chiara Auwerx[1,2,3,4]*, Maarja Lepamets[5,6], Marie C. Sadler[3,4], Marion Patxot[2], Miloš Stojanov [7], David Baud[7], Reedik Mägi[6], Eleonora Porcu[1,3,4], Alexandre Reymond[1], Zoltán Kutalik[2,3,4]
[1]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4]University Center for Primary Care and Public Health, Lausanne, Switzerland; [5]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; [6]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia; [7]Materno-fetal and Obstetrics Research Unit, Department Woman-Mother-Child, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

The impact of copy number variations (CNVs) on complex human traits remains understudied. We called CNVs in 331'522 UK Biobank participants and performed genome-wide association scans (GWASs) between the copy number of CNV-proxy probes and 57 continuous traits, revealing 131 signals spanning 47 phenotypes. Our analysis recapitulated well-known associations (e.g. 1q21 and height), revealed the pleiotropy of recurrent CNVs (e.g. 26 and 16 traits for 16p11.2-BP4-BP5 and 22q11.21, respectively), and suggested gene functionalities (e.g. *MARF1* in female reproduction). Forty-eight CNV signals (38%) overlapped with single nucleotide polymorphism (SNP)-GWAS signals for the same trait. For instance, deletion of *PDZK1*, which encodes a urate transporter scaffold protein, decreased serum urate levels, and deletion of *RHD*, which encodes the Rhesus blood group D antigen, associated with hematological traits. Other signals overlapped Mendelian disorder regions, suggesting variable expressivity and broad impact of these loci in the general population, as illustrated by signals mapping to Rotor syndrome (*SLCO1B1/3*), renal cysts and diabetes syndrome (*HNF1B*), or Charcot-Marie-Tooth (*PMP22*) loci. Total CNV burden negatively impacted 35 traits, leading to increased adiposity, liver/kidney damage, and decreased intelligence and physical capacity. Thirty traits remained burden-associated after correcting for CNV-GWAS signals, pointing to a polygenic CNV-architecture. The burden negatively correlated with socio-economic indicators, parental lifespan, and age (survivorship proxy), suggesting a contribution to decreased longevity. Together, our results showcase how studying CNVs can expand biological and epidemiological insights, emphasizing the critical role of CNVs in shaping human traits and arguing in favor of a continuum between Mendelian and complex diseases.

9

Comparison of Meta-analysis and Mega-analysis for Genome-wide Association Studies

Harold Bae[1*], Anastasia Gurinovich[2], Zeyuan Song[3], Anastasia Leshchyk[4], Mengze Li[4], Cristina Giuliani[5], Marianne Nygaard[6], Paola Sebastiani[2]
[1]Biostatistics Program, College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon; [2]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA; [3]Department of Biostatistics, Boston University School of Public Health, Boston, MA; [4]Bioinformatics Program, Boston University, Boston, MA; [5]Department of Biological, Geological, and Environmental Sciences, University of Bologna, Italy; [6]Epidemiology, Biostatistics and Biodemography, Department of Public Health, University of Southern Denmark, Odense, Denmark

Meta-analysis has become a common practice for genome-wide association studies (GWAS), owing to increased power to detect novel disease loci by aggregating summary statistics across different studies. An alternative approach, mega-analysis, combines the genotypic and phenotypic data at the individual level across different studies, performs joint quality control, and models the associations in a single, joint data set. It has been hypothesized that the latter approach can have increased power to detect disease loci, particularly for rare variants, by increasing the number of minor allele counts. Some studies have found comparable power between the two methods, and argued for the use of meta-analysis given that mega-analysis is not always feasible. Yet, there is limited evidence from empirical evaluations of the differences in yield and false positive associations of the two methods using real data. A critical question remains on how much gain we can have from mega-analysis and whether combining data at the individual level is worth the effort. We report a systematic comparison of the two methods in the GWAS setting using a consortium of four studies of longevity with ~2300 cases and ~5900 controls in a combined data set. Our results suggest that the two methods yield nearly identical results for common variants, but the mega-analysis is more robust from false positive associations for rare variants, particularly when a single study is driving the false positive association in the meta-analysis. However, a drawback of mega-analysis may be a challenge of combining data from different genotype platforms and protocols.

10

Leveraging Identity by Descent to Investigate the Role of RGS16 in Obesity

James T. Baker[1*], Hung-Hsin Chen[1], Hannah G. Polikowsky[1], Kalypso Karastergiou[2], Susan K. Fried[2], Joseph B. McCormick[3], Susan P. Fisher-Hoch[4], David C. Samuels[1,5], Kari E. North[6,7], Jennifer E. Below[1]
[1]Vanderbilt Genetics Institute, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Diabetes Obesity and Metabolism Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States of America; [3]University of Texas School of Public Health Brownsville TX, United States of America; [4]Department of Epidemiology, Human Genetics and Environmental Sciences, The University of Texas Health Science Center at Houston School of Public Health, Brownsville Regional Campus, Brownsville, TX, United States of America; [5]Department of Molecular Physiology & Biophysics, Vanderbilt University School of Medicine, 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; [7]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

The US prevalence of severe obesity (SevO, body mass index [BMI] ≥40 kg/m^2) is increasing at an alarming rate. While genome-wide association studies have identified >1000 loci associated with body mass index, the function of much of this variation is unknown. Gene expression measures can illuminate the link between genetic variation and disease highlighting pathways for targeted therapeutic intervention, but to-date, only a handful of studies has examined the role of gene expression to identify molecular signatures associated with SevO. To this end, we leveraged extant whole blood (WB) RNA sequencing (RNAseq) data in 75 SevO cases and 116 controls (with BMI = 18-25) collected from randomly selected Mexican Americans in the Cameron County Hispanic Cohort (CCHC) to identify patterns associated with SevO. We used established protocols and alignment, yielding 18,565 genes after quality control. We applied DESeq2 to assess DE associated with SevO, using a negative binomial regression model with a gene-specific dispersion parameter, adjusted for sex, age, T2D, hypertension, hypercholesterolemia, and 10 probabilistic estimation of expression residual (PEER) factors. After FDR correction, 124 genes were significantly DE, including top genes *C1RL*, *IL4R, and RGS16,* a member of the regulator of G protein signaling family. We then identified clusters of individuals within Vanderbilt's biobank, BioVU, who share identical copies of *RGS16* due to recent common ancestry and identified a cluster of 8 individuals with significantly high rates of obesity. Analysis of whole exome sequencing data to identify variants causal for monogenic obesity in *RGS16* in this cluster is ongoing.

11

LDAK-GBAT - a Powerful and Efficient Tool for Gene-based Analysis of GWAS Data

Takiy Berrandou[1*], David Balding[2], Doug Speed[1]
[1]*Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark;* [2]*Melbourne Integrative Genomics (MIG), Melbourne University, Melbourne, Australia*

Genome-wide association studies (GWAS) test each SNP individually for association with the phenotype. However, it is now recognized that gene-based analyses - which jointly test sets of SNPs for association with the outcome - can complement single-SNP analysis and provide additional insights for the genetic architecture of complex traits. Here, we propose LDAK-GBAT, a new tool for gene-based association analysis that requires only summary statistics from GWAS and a reference panel.

We first evaluate the performance of LDAK-GBAT and alternative tools such as MAGMA, GCTA-fastBAT and sumFREGAT (which implemented SKAT-O, PCA and ACAT methods) using 14 traits from UK Biobank. We show that LDAK-GBAT is computationally efficient, taking approximately four minutes to analyze imputed data (>7.1 million SNPs) when using a reference panel of 404 individuals. It also produces *P* values that are well-calibrated under the null and is robust to the choice of reference panel. LDAK-GBAT finds more significant genes than the five other tools. For example, LDAK-GBAT finds on average 25% more significant genes than sumFREGAT-PCA, the second best-performing method. Next, we apply LDAK-GBAT to 18 traits from the Million Veterans Project and nine traits from the Psychiatric Genetics Consortium. In total, we find 6,083 significant genes, which is 46% more than found by single-SNP analysis, and 55% more than MAGMA the largely used tool for gene-based association analysis.

In conclusion, our proposed tool, implemented in our freely available software LDAK (www.ldak.org/), has the potential to identify additional novel disease-susceptibility genes for complex diseases from GWAS datasets.

12

Implementing Mendelian Randomization to Assess Foetal Risk from Intrauterine Prescriptive Drugs for the Treatment of Diabetes, Hypertension and Thyroidism in Pregnancy

Ciarrah-Jane S. Barry[1,2], Alexandra K. Havdahl[1,3,4], Christy Burden[5], Venexia M. Walker[1,2,6], Neil M. Davies[1,2,7]
[1]*Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, BS82BN, United Kingdom;* [2]*Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom;* [3]*Nic Waals Institute, Lovisenberg Diaconal Hospital, Spångbergveien 25, 0853, Oslo, Norway;* [4]*Department of Mental Disorders, Norwegian Institute of Public Health, Sandakerveien 24 C, 0473, Oslo, Norway;* [5]*Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom;* [6]*Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America;* [7]*K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway*

Background: Pregnant women are excluded from clinical trials for ethical and practical reasons; thus, little is known about the causal impact of intrauterine medication exposure on developing fetuses. Yet many chronic conditions necessitate maternal medication use during pregnancy, such as diabetes, hypertension, and thyroid disorders. Therefore, the objective of this study is to examine the potential adverse effects of exposure to medications for these conditions during pregnancy on neonates.

Methods: We conducted a one-sample within-family Mendelian Randomization analysis of maternal genetic drug targets on neonatal outcomes using The Norwegian Mother, Father and Child Cohort Study (MoBa). We identified genetic proxies for drug targets using DrugBank. We used data from parent-offspring trios from the MoBa study. We investigated the association of these genetic variants and gestational age, birth weight, mode of delivery and APGAR score. We used within family models that included parental genotype to assess intrauterine effects.

Results: The cohort of interest to this study contains complete genetic data on 20,183 parent-offspring trios. Preliminary summary statistics indicate analysis should be sufficiently powered, using the available linked phenotypic data. Adverse neonatal outcomes of interest to this study include those such as gestational age, birthweight, mode of delivery and Apgar score.

Conclusions: Genetic epidemiological data can provide evidence about the risks to neonates and benefits to mothers of intrauterine medication exposure. Evidence from this study may be used with existing literature, clinical trials, and alternative study types to guide physicians and mothers during pregnancy.

13

MetFLEX: A Novel Multi-tissue Transcriptomics Learning

Halima Bensmail [1,*] and Abdelkader Baggag [1]
[1]*Qatar Computing Research Institute, Hamad Bin Khalifa University*

We propose an artificial intelligence framework for learning a multivariate response model -jointly- with the error precision matrix, i.e., the tissue-tissue expression correlation, for predicting gene expression in multiple tissues simultaneously. Unlike existing methods for multi-tissue, our approach incorporates tissue-tissue expression correlation, assume non-normality of the gene expression which allows us to handle missing expression measurements more efficiently and more accurately and predict gene expression using a weighted summation of genotypes. We use a conditional penalized estimation maximization (ECM) approach to estimate the missing information related to several tissues, and this allows us to handle missing measurements more efficiently and more accurately. We applied the algorithm on simulated data and on real data from GTEx for which 44 tissues were collected by Genotype-Tissue Expression (GTEx) database and compare our method with the existing algorithms UTMOST and S-MultiXcan.

14

PRISQ: a Risk Score for Screening Prediabetes for Qatari Population

Abdelilah Arredouani [1] and Halima Bensmail [2,*]
[1]*Qatar Biomedical Research Institute, Hamad Bin Khalifa University;* [2]*Qatar Computing Research Institute, Hamad Bin Khalifa University*

Materials and methods: In this cross-sectional, case-control study we used data of 4895 controls and 2373 prediabetic adults obtained from the Qatar biobank cohort. Significant risk factors were identified by logistic regression and other machine learning methods. The receiver operating characteristic was used to calculate area under the curve, cut-off point, sensitivity, specificity, positive and negative predictive values. The prediabetes risk score was developed from data of Qatari citizens as well as long-term (≥15 years) residents.

Results: The significant risk factors for the Prediabetes Risk Score in Qatar (PRISQ) were age, gender, BMI, waist circumference, and blood pressure. The risk score ranges from 0 to 45. The AUC of the score was 80% [78%-83%], and the cutoff point 16 yielded sensitivity and specificity of respectively 86.2% [82.7% to 89.2%] and 57.9% [65.5% to 71.4%]. PRISQ performed equally in Qatari nationals and long-term residents.

Conclusions: PRISQ is the first prediabetes screening score developed in Middle Eastern population. It only uses risk factors measured non-invasively, is simple, cost-effective, and can be easily understood by the general public and health providers. PRISQ is an important tool for early detection of prediabetics and can help tremendously in curbing the diabetes epidemic in the region.

15

Multiple Tissue Learning for Transcriptomics Studies

Abdelkader Baggag[1], Halima Bensmail[1*]
[1]*Qatar Computing Research Institute*

We propose an artificial intelligence framework for learning a multivariate response model -jointly- with the error precision matrix, i.e., the tissue-tissue expression correlation, for predicting gene expression in multiple tissues simultaneously. Unlike existing methods for multi-tissue, our approach incorporates tissue-tissue expression correlation, which allows us to handle missing expression measurements more efficiently and more accurately predict gene expression using a weighted summation of genotypes. We use a regularized Bayesian method to estimate the missing information related to several tissues, and this allows us to handle missing measurements more efficiently and more accurately. We will be using 29 tissues collected by Genotype-Tissue Expression (GTEx) database and compare our method with the existing state of the art algorithms.

16

Prioritizing Functionally Relevant Lung Cancer Risk Variants Using a Novel Computational Framework

Michael J. Betti[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*

Despite the plethora of highly powered genome wide association studies (GWAS), a lingering barrier to clinical application is the functional ambiguity of identified associations. Utilizing a novel computational framework, CoRE-BED, we functionally prioritized variants from a lung cancer GWAS of 29,266 European cases and 56,450 controls. Prioritization was based on variant overlap with a promoter or enhancer across 71 lung derived biospecimens. Of the 7,673,198 variants tested, 375,101 (4.89%) fell within a regulatory element in at least one of these disease relevant biospecimens. Strikingly, the distribution of GWAS *P* values in functional variants was significantly lower than those of the remaining variants (Wilcoxon rank sum test W = 1.3475e12, *P* value < 2.2e-16), suggesting that functional status presents an effective indicator of variant importance. We next used linkage disequilibrium score regression to estimate the SNP heritability

captured by variants in lung regulatory elements. We found a higher heritability enrichment using this subset of functional variants vs. the entire set of summary statistics (9.08% vs 7.13%). To verify that this heritability enrichment was not the result of a smaller sample size, we estimated heritability in an identically sized set of 375,101 variants randomly sampled from the same summary statistics. In contrast to the >9% heritability estimate for functionally prioritized variants, the set of randomly selected SNPs had an estimated heritability of 6.71%. Our findings suggest that by limiting the set of candidate lung cancer risk variants to those within biologically relevant genomic regions, we can enrich the dataset for trait heritability.

17

Isoform-level Transcriptome-wide Association Studies Uncover Novel Mechanisms Underlying Genetic Associations with Complex Traits

Arjun Bhattacharya[1,2], Minsoo Kim[3], Jonatan L. Hervoso[4], Cindy Wen[3], Connor Jops[3], Bogdan Pasaniuc[1,5,6±], Michael J. Gandal[3,5±]
[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America; [2]Institute for Quantitative and Computational Biosciences, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America; [3]Department of Psychiatry and Biobehavioral Sciences, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America; [4]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, United States of America; [5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, United States of America; [6]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, United States of America
[±]Equal contribution

Multiple genetic associations have been discovered through integration of gene expression with genome-wide association studies (GWAS), using methods like colocalization and transcriptome-wide association studies (TWAS). However, gene expression is an imperfect measure for units of the transcriptome. Alternative splicing patterns give rise to unique isoforms of genes, whose expression are often genetically controlled. Accordingly, gene isoforms are a more fundamental unit that underlies the genetic variant to trait relationship. Here, we introduce isoTWAS, a scalable framework to integrate genetics, isoform-level expression, and phenotypic associations to not only identify genes and transcripts associated with complex traits but also potential isoform-mediated mechanisms at those prioritized genetic loci. Through extensive simulations and real data from the Genotype-Tissue Expression Project, we demonstrate that, by explicitly modeling isoform expression as a multivariate object, isoTWAS improves upon TWAS in prediction of total gene expression (more than 25% increase in prediction R^2). isoTWAS's hierarchical testing strategy also improves the power to detect gene-trait associations, especially when genetic effects vary across isoforms of the same gene (up to 40-50% increase in power in these settings). We apply isoTWAS to data from the Genotype Tissue-Expression Project to identify over 200 novel genetic associations with five neuropsychiatric traits. Specifically, we illustrate that, by focusing on isoform expression, isoTWAS detects several genetic associations that cannot be detected by TWAS. Our results show the utility of employing isoTWAS to study genetic effects on complex traits, revealing trait associations undetectable by analyses incorporating total gene expression and uncovering hidden mechanisms through gene isoforms.

18

Pleiotropic Influences of Neuropsychiatric Polygenic Risk on Common Laboratory Values in 660,000 US Veterans

Tim B. Bigdeli [1,2]*, Peter B. Barr [1,2], Roseann E. Peterson [1,3], Georgios Voloudakis [4,5,6], Bryan Gorman [7], Cooperative Studies Program (CSP) #572, Million Veteran Program (MVP), Mihaela Aslan [8,9], Philip D. Harvey [10,11], Panos Roussos [4,5,6]
[1]VA New York Harbor Healthcare System, Brooklyn, New York, United States of America; [2]Department of Psychiatry and Behavioral Sciences, SUNY Downstate Health Sciences University, Brooklyn, New York, United States of America; [3]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, United States of America; [4]James J. Peters Veterans Affairs Medical Center, Bronx, New York, United States of America; [5]Departments of Genetics and Genomic Sciences and [6]Psychiatry, Icahn School of Medicine at Mount Sinai, New York, United States of America; [7]Massachusetts Area Veterans Epidemiology, Research, and Information Center (MAVERIC), Jamaica Plain, Massachusetts, United States of America; [8]Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, Connecticut, United States of America; [9]Yale University School of Medicine, New Haven, Connecticut, United States of America; [10]Bruce W. Carter Miami Veterans Affairs (VA) Medical Center, Miami, Florida, United States of America; [11]University of Miami Miller School of Medicine, Miami, Florida, United States of America

Severe mental illness, including schizophrenia, bipolar, and major depression disorder are heritable, highly multifactorial disorders which cause significant disability, worldwide. Better understanding the pathophysiological mechanisms underlying disease risk and progression will be essential for alleviating the burden of these illnesses and comorbidities. The Million Veteran Program links genomic data, self-report survey data, and electronic health records from the Veterans Health Administration, which is the largest integrated health care system in the United States. We constructed and tested polygenic risk scores for schizophrenia, bipolar disorder, and major depression for association with median values for 70 laboratory tests in 660,000 participants. We applied genomic structural equation modeling to derive novel scores indexing shared and disorder-specific latent genetic factors. Schizophrenia polygenic scores were robustly associated (P value < 10^{-8}) with increased lipid levels, electrolyte imbalances, decreased liver and kidney function, and increased white blood cell count. Many associations were

replicable among 125,000 African Americans, albeit with lesser statistical significance (P value < 10^{-5}). We found that a cross-disorder latent genetic factor accounted for many findings, rather than any diagnosis-specific factors. These findings remained significant (P value < 10^{-6}) when restricting analyses to individuals without diagnosed psychotic or affective illness, suggesting that these relationships are not simply consequences of medication side-effects. By applying a 'reverse genetics' approach to large-scale electronic health records, we add to a growing literature demonstrating the broad pleiotropic effects of currently indexable polygenic risk. Ongoing Mendelian randomization analyses seek to further parse biological from mediated pleiotropy.

19

Bridging the Diversity Gap: Analytical and Study Design Considerations for Improving the Accuracy of Trans-ancestry Genetic Risk Prediction

Ozvan Bocher[1 *], Arthur Gilly[1], Eleftheria Zeggini[1,2,3], Andrew Morris[1,4]

[1]ITG, Helmholtz Zentrum München, Germany; [2]Technical University of Munich, Germany; [3]Klinikum Rechts der Isar, Germany; [4]University of Manchester, United Kingdom

Genetic prediction of common complex disease risk is an essential component of precision medicine. Currently, genetic risk scores (GRS) are calculated from genome-wide association study (GWAS) meta-analyses composed mostly of European-ancestry samples. These have been shown to poorly transfer to other ancestry groups for reasons including cross-ancestry heterogeneity of allelic effects. Fixed-effects meta-analysis (FETA) methods do not model heterogeneity in effects between GWAS and ancestry-specific (AS) scores suffer from low power when European ancestry individuals predominate in the discovery sample. In contrast, trans-ancestry meta-regression (TAMR) builds ancestry-aware GRS that accounts for potential heterogeneity in effect sizes between populations. Here, we examine the predictive accuracy of FETA, AS and TAMR GRS under multiple genetic architectures and ancestry configurations. We show that the predictive accuracy of FETA and TAMR scores decreases as between-ancestry effect heterogeneity increases, whereas it remains consistently low for AS scores. TAMR outperforms FETA scores when more than 10% of SNPs have heterogenous effects and conserves power in homogenous scenarios. For type 2 diabetes, where current meta-analyses reach 50% of non-European ancestry, TAMR scores explain 33% and 11% more phenotypic variance than AS and FETA methods respectively. A high proportion of non-European ancestry individuals is needed to reach an accuracy that is comparable to the one observed in European-ancestry studies. Our results highlight the need to rebalance the ancestral composition of GWAS to enable accurate prediction in non-European ancestry groups, and demonstrates the relevance of meta-regression approaches for compensating some of the current population biases in GWAS.

20

A Test of Covariance Matrix to Detect Predictors of the Gut Microbiome Variability

Christophe Boetto[1]*, Violeta Basten Romero[1], Etienne Patin[2], Marius Bredon[2], Darragh Duffy[4], Lluis Quintana-Murci[2], Harry Sokol[3,5,6,7], Hugues Aschard[1,8]

[1]Institut Pasteur, Université de Paris Cité, Department of Computational Biology, [2]Institut Pasteur, Université de Paris Cité, Department of Genomes and Genetics, Paris, France, [3]INSERM, Centre de Recherche Saine-Antoine, CRSA, AP-HP, Sorbonne Université, Gastroenterology Department, Paris France, [4]Institut Pasteur, Université de Paris Cité, Department of Immunology, Paris, France, [5]Paris Center for Microbiome Medicine, Fédération Hospitalo-Universitaire, Paris, France, [6]Carenity, Paris, France, [7]INRAE, UMR1319 Micalis & AgroParisTech, Jouy en Josas, France, [8]Harvard TH Chan School of Public Health, Department of Epidemiology, Boston, United States of America

Multivariate analysis is becoming central in studies investigating multiple omics data. However, some important characteristics of those data have been seldom explored. Here we present MANOCCA (Multivariate Analysis of Conditional CovAriance), a powerful method to test the effect of a predictor on the covariance matrix of a multivariate outcome, which we applied to study host and gut microbiome relationship. The proposed test is by construction orthogonal to tests based on mean and variance, and can capture effects missed by both approaches. We first compared the performances of MANOCCA against existing correlation methods (BoxM, Mantel) and show that the latter display severe type I error rate inflation in all simulations mimicking microbiome data. Only MANOCCA was correctly calibrated. We then applied our test to assess environmental and host genetic predictors of 16S derived gut microbiome covariance matrix in 1,000 healthy participants from the Milieu Interieur cohort, and compared results with standard univariate and multivariate (for example MANOVA) tests on means. MANOCCA strongly outperformed tests based on means, confirming associations with age and sex but with higher power (>200% power increase), and detected additional signals with smoking and diet. Host genetic genome wide association study (GWAS) identified multiple significant associations (P value < 1e-8). This includes a signal with *CSMD1*, a gene previously reported associated with alpha diversity in both human and mouse. In comparison, mean methods GWAS screening did not detect any signal. Altogether, these analysis demonstrate the capabilities of our novel approach to identify predictors of a multivariate outcome covariance matrix.

21

Identifying Genetic Determinants of Blood Pressure Variance Using a Novel Mean-Variance Test

Joseph H. Breeyear[1*], Brian S. Mautz[1,2*], Jacob M. Keaton[3], Jacklyn N. Hellwege[1,2], Eric S. Torstenson[4], Digna R. Velez Edwards[1,4,5], Chun Li[6**], Todd L. Edwards[1,2,7**]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee; [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee; [3]Medical Genomics

and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee; [5]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee; [6]Department of Population and Public Health Sciences, Kevin School of Medicine of University of Southern California, Los Angeles, CA; [7]Division of Epidemiology, Vanderbilt University Medical Center, Nashville, Tennessee

The genetic influence on the phenotypic variance of traits has often been overlooked in the field of genomics. We present the mean-variance test (MVTest), which models both mean and log-variance as functions of genotype and covariates in mutually adjusted linear models using estimating equations, that has the ability to detect effect modification when effect modifiers are unknown. We conducted simulations to demonstrate control of type I error and power compared with several alternative methods, including Joint Location-Scale test. We applied MVTest to the UK Biobank to estimate genetic effects on mean and population-level variance of systolic (SBP), diastolic (DBP), and pulse pressure (PP). GCTA joint conditional association analyses with MVTest results, conditioning on 3,800 previously reported blood pressure SNPs identified 61 previously unreported blood pressure SNPs associated with PP variance. We also evaluated the associations between blood pressure mean and variance traits and genetically predicted gene expression (GPGE) levels using S-PrediXcan and COLOC. Significant GPGE associations were identified for 257 gene-tissue pairs with mean SBP, 317 with mean DBP, 299 with mean PP, 2 with PP variance, and 1 with SBP variance. To identify phenotypes associated with the phenotypic variance of blood pressure, we developed genetic risk scores (GRSs; $p<5e^{-8}$) for each variance trait and conducted phenome-wide association studies across eMERGE and BioVU. We identified nominal associations between the DBP variance GRS and essential hypertension and hypertensive heart disease. MVTest is a novel association tool designed to examine the population-level trait variance of quantitative traits and is freely available.

22

Using Simulated Casual Data to Characterize Biases and Identify Best Practices in Electronic Health Record Research

Lindsay B. Breidenbach[1]*, Lea K. Davis[1]
[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America

Passively collected electronic health record (EHR) data is increasingly coupled with omics data to advance biological understanding of health conditions. However, retrospective studies cannot randomize treatments or exposures, and patients in the EHR may not be representative of the general population. Thus, these studies often have unaddressed confounding and selection biases. These biases, and how they affect genomic analyses, are not well characterized. Although scientists can choose from many statistical methods to control for confounding, there is little guidance on their selection. Furthermore, it is often difficult to infer which EHR variables are acting as confounders, mediators, or colliders. To mitigate these problems, we hypothesize that simulated data could model biased data and aid in method selection. To test this, we created software called "EHRomics" that creates simulated data through joint probability distributions. Here users set the effect estimate between all variables and can apply multiple confounding control methods to the resulting simulated data. For the simulation, directed acyclic graphs (DAGs) mark causal paths underlying each model. Then, the DAG and the joint distributions are combined to form a Bayesian Network (BN). In this software, we simulate selection bias by stratifying on colliders and confounding bias by not accounting for confounders. Users can also vary the simulated data's misclassification rate, prevalence, and sample size parameters. Last, we apply the confounder control methods to simulated datasets where the biases are and are not simulated. This provides insight on how these biases and other parameter inputs impact each method's effect estimate.

23

Insights Into Complex Genetic Architecture from Region-Based Association Testing of Melanoma Risk at a Locus With Allelic Heterogeneity on 16q24

Myriam Brossard[1*], Kexin Luo[1], Delnaz Roshandel[2], Fatemeh Yavartanoo[4], Yun J. Yoo[4], Andrew D. Paterson[2,3], Shelley B. Bull[1,3]
[1]*Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada; [2]Hospital for Sick Children Research Institute, Toronto, Ontario, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [4]Seoul National University, Seoul, Korea*

Region-based analysis methods are designed to improve power to detect regions with complex genetic architectures that may be missed by conventional single-SNP analysis. Such methods require specification of the regions to be analyzed and face statistical challenges arising from complex linkage disequilibrium (LD) structures. Here, we investigate application of a region-based analytical approach in 16q24 with established allelic heterogeneity for melanoma risk, characterized by multiple causal variants in *MC1R* carried by distinct risk haplotypes and an extended LD pattern. We simulated haplotypes for 107292 common SNPs on 16q in 20000 cases and 20000 controls using 1000G European-ancestry haplotypes with melanoma status generated under a log-additive model with five *MC1R* causal variants. To define regions, we partitioned 16q into 2394 non-overlapping quasi-independent LD-block regions using a haplotype block detection method adaptive to local LD structure. Within each region, we apply multiple logistic regression; and compare three tests for regional association with melanoma: the generalized Wald test, reduced *df* multiple-linear-combination (MLC) test, and regression test of SNP principal components. Among ~30 regions reaching $P_{Bonferroni} \leq 2.1 \times 10^{-5}$ for at least one region-based test, >36% showed an improved regional *P* value compared to the top single-SNP *P* value within region, which illustrates benefits of region-based over single-SNP

approaches. The causal variants were partitioned into a 321kb region, that is detected as the top region by MLC and within the top 6 regions by the other tests. Prioritization of this region is challenged by detection of other regions that include SNPs in LD with the causal variants.

24

Joint Linkage and Association Analysis with GENEHUNTER-MODSCORE

Markus Brugger[1,2,3*], Manuel Lutz[1,2,3], and Konstantin Strauch[1,2,3]
[1]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz and [2]Institute of Medical Information Processing, Biometry and Epidemiology - IBE, Chair of Genetic Epidemiology, LMU Munich, Munich and [3]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

Background: Joint linkage and association (JLA) analysis combines two disease gene mapping strategies: linkage information in families and association information in populations. JLA analysis can increase mapping power, especially when the evidence for both linkage and association is low to moderate. Similarly, an association analysis based on haplotypes instead of single markers can increase mapping power when the association pattern is complex.

Methods: We present an extension to the GENEHUNTER-MODSCORE software that enables a JLA analysis using information from arbitrary pedigree types and unrelated individuals. Our new JLA method is an extension of the MOD score approach in linkage analysis, which jointly estimates trait-model and association parameters. Association is modelled using marker-trait locus haplotypes of a single diallelic trait locus and up to three SNPs. Linkage information is extracted from additional multi-allelic flanking markers. Optimization of model parameters is achieved utilizing the derivative-free optimization algorithm COBYLA. We investigated the statistical properties of our JLA implementation using extensive simulations, and we compared our approach to the single-marker JLA test implemented in PSEUDOMARKER. Because the null distribution of our JLA test is unknown, we implemented and evaluated a simulation routine.

Results: We demonstrated the validity of our JLA analysis implementation and identified scenarios with complex association patterns, for which haplotype-based tests outperformed the single-marker tests.

Conclusion: Our new JLA-MOD score method proves to be a valuable gene mapping and characterization tool. It is particularly useful in situations where either linkage or association information alone provide insufficient power to identify disease-causing genetic variants.

25

Comparison and Integration of Single-variant and Region-based Analysis in Genome-wide Association Studies (GWAS) of Complex Traits

Shelley B. Bull[1,2*], Myriam Brossard[1], Kexin Luo[1], Delnaz Roshandel[3], Fatemeh Yavartanoo[4], Andrew D. Paterson[2,3], Yun J. Yoo[4]
[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada; [2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [3]Hospital for Sick Children Research Institute, Toronto, Ontario, Canada; [4]Seoul National University, Seoul, South Korea

Genome-wide analysis of millions of single variant tests is a standard approach for discovery of disease-/trait-associated genetic variants, leading to identification of genomic regions for subsequent fine-mapping analysis. Compared to single-variant genome-wide testing, region-based multi-variant association analysis reduces multiple testing burden and can better capture signals under complex genetic architectures. In our experience, the two approaches can be complementary and region-based tests can help prioritize sets of SNPs for further investigation, but region tests are used infrequently in GWAS involving common variants. To facilitate comparison among multi-variant region statistics and single-variant tests, we developed the *RegionScan* R package that is scalable for large datasets. In concert with a partitioning algorithm that predefines autosomal regions according to linkage disequilibrium, the implementation: (1) preprocesses high-density VCF files, with options for variant recoding, filtering and pruning; (2) applies, within each region, alternative regional association test statistics including multi-SNP linear/logistic regression tests with and without dimension reduction, SKAT-type variance component score tests, and region-level min*P* tests; and (3) visualizes signal comparisons at genome-wide and regional levels, including "locus zoom" type plots of contiguous regions. We demonstrate *RegionScan* performance in genome-wide analysis of lipid traits in nearly 18,000 participants aged 45 to 86 years in the Canadian Longitudinal Study on Aging using 5,288,020 genotyped or imputed bi-allelic SNPs (MAF 0.05) partitioned into 92,327 quasi-independent regions. In these analyses, application of permutations to estimate comparable genome-wide thresholds for region-based and single-SNP tests yields values that correspond respectively to Bonferroni-corrected and conventional *P* value criteria.

26

Power Study of Epigenetic Landscape of Prostate Cancer in Black/African Americans and Whites

Sarah G. Buxbaum[1*], Olumide M. Arigbede[1], Sara Falzarano[2], Suhn Rhie[3]
[1]Florida Agricultural and Mechanical University, College of Pharmacy, Pharmaceutical Sciences, Institute of Public Health; [2]University of Florida, College of Medicine, Department of Pathology; [3]University of Southern California, Keck School of Medicine, Department of Biochemistry and Molecular Medicine

Background: Few epigenetic studies of prostate cancer have been performed.

Methods: Power calculations were performed using PASS 2020 to simulate CpG sites in each of two groups. The number of sites, the minimum mean difference (delta) between the matched pairs' microarray signals (betas) and the SD were varied, and t-tests were performed across the simulated data.

Results: With 2200 t-tests, the power will be 80% or greater to detect 5 or more differentially methylated regions, given an SD of 0.4 or lower and a minimum difference in signal delta of 20%. Using Hotelling's two-sample T^2 test, a simulation of 100 pairs with 200 independent variables with just 2 signals with a delta of only 0.1 and 98 with no difference gave 100% power.

Discussion: This approach may be naïve because it does not take into account the complexity of the distribution of the betas. We then looked at prostate cancer epigenetic data available in cBioPortal from prostate adenocarcinoma samples in the TCGA PanCancer Atlas. This data includes only 7 African American/Black individuals and 147 whites. Comparing the groups using the methylation data, we were able to identify 14 genes among Black/African Americans that were statistically significant.

Conclusion: The simulation study showed that noise in the data was more critical than the number of sites in determining power. A global test for difference appeared very powerful. Further, we showed an example where a small number of prostate cancer specimens was sufficient to show a difference in the epigenetic landscape between Black/African Americans and whites.

27

Genetic Distance Is Highly Predictive of Vaccine Effectiveness Against Influenza A Virus

Lirong Cao[1,2*], Maggie Haitian Wang[1,2]
[1]JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China; [2]CUHK Shenzhen Research Institute, Shenzhen, China

Previously we have demonstrated the feasibility of using genetic distance on effective mutation (EM) sites of influenza viruses to evaluate vaccine effectiveness (VE). This study further investigated the optimal quantification of genetic relatedness to accurately predict the protective effect induced by vaccines, based on EM and antigenic sites (AS). Through an integrated computational framework to analyze hemagglutinin (HA) and neuraminidase (NA) protein sequences for influenza A/H1N1 and A/H3N2 in the United States, we found that not all antigenic sites of HA can contribute to VE change and genetic distance on the overlapping sites of EM and AS showed the strongest predictive power, accounting for 65.1% of the VE change for H3N2, while at most 57.7% of the variations can be explained by mismatch on EM or specific AS alone. After aggregating the genetic distance of both HA and NA by using the framework of polygenic risk score, the explained variations can attain a high level of 87.8%. Substantially improved predictive power was also detected for H1N1 by using the overlapping sites. For H3N2, we further found that NA mismatch significantly mediated the negative effect of HA mismatch on VE, accounting for 55.2% of the total effect (95%CI: 24.8 – 81.0). The findings raise awareness of NA and determine the crucial mutations for HA and NA genes related to vaccine protection, which may help candidate vaccine strain selection and optimization. Rapid evaluation of VE by using the predictive models developed herein may facilitate future vaccine deployment and medical resource allocation.

28

Functional Screening of 3'-UTR Variants Combined with Genome-wide Association Identifies Causal Genes Contributing to Alcohol Use Phenotypes

Andy B. Chen[1,2*], Kriti S. Thapa[3], Hongyu Gao[1,2,4], Jill L. Reiter[1,3], Hongmei Gu[3], Junjie Zhang[1], Xiaoling Xuei[1,4], Dongbing Lai[1], Yue Wang[1], Howard J. Edenberg[1,3], Yunlong Liu[1,2,]
[1]Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, United States of America; [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, United States of America; [3]Department of Biochemistry & Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, United States of America; [4]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN, United States of America

Genome-wide association studies (GWAS) have identified loci associated with alcohol consumption and alcohol use disorder (AUD), but many are in non-coding regulatory regions, where additional information is needed to evaluate their function. Our objective is to determine how variants and genes at regulatory loci functionally contribute to alcohol consumption and AUD.

We evaluated the activity of variants in 3' untranslated regions (3'-UTRs) of genes in loci associated with neurological disorders using a massively parallel reporter assay (MPRA) in neuroblastoma and microglia cells. Of the 13,515 variants tested, 400 (neuroblastoma) and 657 (microglia) significantly impacted gene expression. Heritability enrichment analysis found that functional variants explained a higher proportion of heritability in GWASs of alcohol phenotypes than all candidate variants.

We identified genes whose 3'-UTR are associated with alcohol consumption by aggregating variant effects from MPRA and GWAS results, using drinks per week from GSCAN as a discovery cohort and alcohol use disorders identification test-consumption (AUDIT-C) from the Million Veteran Program (MVP) as a replication cohort.

Using these identified genes, we stratified brain tissue samples using a 3'-UTR activity score calculated by combining SNP genotypes with MPRA effect values and evaluated differential expression of genes between groups with high and low 3'-UTR activity. A pathway analysis of these differentially expressed genes identified several inflammation response pathways.

By using only genotypes and MPRA effect to stratify the samples, the pathways identified are downstream of the

genetic component. This suggests that variation in response to inflammation contributes to the propensity to increase alcohol consumption.

29

Evidence of Novel Susceptibility Variants for Prostate Cancer and a Polygenic Risk Score that Improves Prediction of Aggressive Disease for Men of African Ancestry

Fei Chen[1*], Ravi K. Madduri[2], Alex A. Rodriguez[2], Burcu F. Darst[1,3], J. Michael Gaziano[4,5], Amy C. Justice[6,7] David V. Conti[1], Christopher A. Haiman[1], on behalf of AAPC and PRACTICAL/ELLIPSE Consortia and VA Million Veterans Program.
[1]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [2]Argonne National Laboratory, Lemont, Illinois, United States of America; [3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [4]VA Boston Healthcare System, Boston, Massachusetts, United States of America; [5]Division of Aging, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America; [6]Yale School of Medicine, New Haven, Connecticut, United States of America; [7]VA Connecticut Healthcare System, West Haven, Connecticut, United States of America

To identify common genetic variants contributing to prostate cancer risk in men of African ancestry, we conducted a meta-analysis of genome-wide association studies (GWAS) including 19,378 cases and 61,620 controls from the African Ancestry Prostate Cancer (AAPC) and PRACTICAL Consortia, the VA Million Veterans Program (MVP) and seven other studies and biobanks. Nine novel susceptibility loci for prostate cancer were identified, of which seven were only found or substantially more common in men of African ancestry, including a functional African-specific variant (rs60985508) that introduces a stop codon in exon 24 of the prostate-specific gene anoctamin 7 (*ANO7*). A polygenic risk score (PRS) of 278 risk variants conferred strong associations with prostate cancer risk in studies of African ancestry, with OR being > 3 and > 5 for men in the top 10% and 1% PRS categories versus the 40%-60% average risk category of the PRS, respectively. More importantly, men in the top PRS decile were 1.23-times (95%CI=1.10-1.38) and 1.65-times (95%CI=1.15-2.35) more likely to be diagnosed with aggressive and metastatic disease compared to non-aggressive prostate cancer, respectively. Our findings demonstrate the importance of large-scale genetic studies in men of African ancestry for a better understanding of prostate cancer susceptibility in this high-risk population. This study also provided strong evidence of clinical utility for PRS in differentiating risk of developing aggressive versus non-aggressive disease for men of African ancestry.

30

Genomic Shared Segments Enable Identification of At-risk Patients in Biobanks

Hung-Hsin Chen[1*†], Megan C Lancaster[2†], Matthew R Fleming[2], James T Baker[1], David C Samuels[1], Chad Huff[3], Dan M Roden[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]Division of Cardiovascular Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America
[†]These authors contributed equally to this work

Most biobanks recruit participants from local populations, resulting in significant undocumented (cryptic) relationships between participants. While cryptic relatedness can introduce bias in traditional association studies, shared genomic regions create opportunities for identifying undiagnosed/misdiagnosed carriers of pathogenic variants. Here, we demonstrate a powerful approach to identify patients at risk of long QT syndrome through identity-by-descent (IBD)-based genotype inference. Long QT syndrome is a potentially lethal arrythmia disease, but 39% of patients experienced diagnostic delay after first symptom and misdiagnosis is common. BioVU comprises DNA samples from 245,000 individuals and their linked electronic health record (EHR), including 69,817 genotyped individuals of European ancestry. We utilized twelve long QT syndrome clinical patients from four cryptically related families, who are confirmed *KCNE1* causal mutation carriers (D76N, rs74315445). We then identified BioVU subjects with IBD (via hap-IBD with SHAPEIT4 phasing) with any proband across *KCNE1* (chr21:35,818,986-35,884,508). Fourteen BioVU Europeans share the same IBD segment(≥3cM) with seven probands, and thirteen share with two other probands. Only 23/27 identified subjects were array genotyped as mutation carriers suggesting potentially erroneous genotyping, and 68.5% of all mutation carriers in BioVU were captured by our IBD-based approach. Confirmation sequencing is on-going. EHR analysis revealed only one identified mutation carrier has been diagnosed with long QT syndrome, but, among the sixteen subjects with available electrocardiogram in EHR, three have prolonged QTc interval and another five have pathologically long QTc intervals(>490msec), illustrating the utility of our shared segment approach for identifying undiagnosed/misdiagnosed patients carrying pathogenic causal variants in a large biobank.

31

Sex-Specific Genetic Variation of Weight and Waist Circumference Change: A Multi-Ancestry Meta-Analysis of Longitudinal Data

Geetha Chittoor[1*], Tugce Karaderi[2,3,4], Misa Graff[5], Germán D. Carrasquilla[3], Thorkild I.A. Sørensen[3], Anne E. Justice[1] on behalf of the Adiposity Change Working Group.

[1]Department of Population Health Sciences, Geisinger, Danville, Pennsylvania, United States of America; [2]Center for Health Data Science, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark; [3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark; [4]DTU Health Tech, Technical University of Denmark, Copenhagen, Denmark; [5]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Majority of genetic studies on adiposity are coming from cross sectional measures of body mass. Hence, we aim to uncover sex-specific genetic loci associated with longitudinal adiposity change using a genome-wide meta-analyses. Data included 156,749 multi-ancestry individuals (86% European, 7% African, 7% Hispanic/Latino, and <1% Asian) aged 20-65 years, 59% women, from 18 cohorts. Measures included weight change (WTC), weight gain (WTG), waist circumference change (WCC), and waist circumference gain (WCG). Using a linear mixed model, weight and WC were regressed on age as both fixed and random effects to derive WTC and WCC stratified by sex/ancestry. Slopes were adjusted for BMI, age, age^2, height, follow-up-years, principal components, and then residuals were inverse-normal transformed. A positive residual slope indicated WTG or WCG, where 48% were gainers. Study-specific analyses were performed using transformed slope residuals allowing for heterogeneity of effect by follow-up-years. Sex stratified results were combined using fixed-effect meta-analysis. We identified 29 suggestively significant ($P<5x10^{-7}$) loci for WTC or WTG, including four genome-wide significant ($P<5x10^{-8}$) loci near *FTO* for WTG in men and women; and *CHCHD3* for WTC, *C2CD4B* for WTG only in women. For WCC or WCG, we identified 22 suggestively significant loci including two genome-wide significant loci for WCC (women: *NSMCE2*; men: *PTPRD*). *FTO*'s association with weight gain is noteworthy here; *C2CD4B*, *NSMCE2*, and *PTPRD* associated with adiposity here were reported previously to be associated with metabolic diseases including diabetes and cancer. These results highlight the sex-specific genetic loci on adult longitudinal adiposity change and obesity related traits.

32

Genetic Basis of Resistance to Infection by *Mycobacterium Tuberculosis*

Clément Conil[1,2]*, Jérémy Manry[1,2], Elouise E. Kroon[3], Marc A. Jean-Juste[4], Marlo Möller[3], Jean-Laurent Casanova[1,2,5,6], Eileen G. Hoal[3], Erwin Schurr[7,8,9], Laurent Abel[1,2,5], Aurélie Cobat[1,2,5]
[1]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, Paris, France; [2]Université Paris Cité, Imagine Institute, Paris, France; [3]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; [4]Haitian Study Group for Kaposi's Sarcoma and Opportunistic Infections (GHESKIO), Port-au-Prince, Haiti; [5]St Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, New York, United States of America; [6]Howard Hughes Medical Institute, New York, New York, United States of America; [7]Department of Biochemistry, Faculty of Medicine, McGill University, Montreal, Quebec, Canada; [8]Program in Infectious Diseases and Global Health, The Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada; [9]McGill International TB Centre, Department of Medicine, Faculty of Medicine, McGill University, Montreal, Quebec, Canada

Despite intense exposure to *Mycobacterium tuberculosis* (*Mtb)*, some individuals seem resistant to infection, as inferred by tuberculin skin test (TST) or interferon-gamma release assays (IGRAs). Few studies have attempted to unravel the genetic factors underlying resistance to *Mtb* infection and the polymorphisms identified have only a modest effect. We performed a whole genome sequencing study of individuals with extreme phenotypes of resistance who remained TST and IGRA negative (so-called resisters) despite high vulnerability due to HIV infection and high levels of exposure to *Mtb*. We enrolled 55 resisters and 100 *Mtb* infected HIV+ individuals from South Africa and 66 resisters and 57 *Mtb* infected HIV+ individuals from Haiti. Single variant analysis identified a locus on chromosome 12q15 associated with resistance to infection in South Africa with replication in Haiti, leading to a combined *P value* of $1.1x10^{-6}$ (OR[95%CI]=0.23[0.12–0.44]). The lead SNP, rs11286051, is an eQTL for *LYZ* which encodes the antimicrobial enzyme lysozyme. The allele associated with resistance to infection is associated with an increased expression of *LYZ* in whole blood and lung in the GTEx dataset. Gene-based analysis for rare (minor allele frequency<5%) nonsynonymous variants identified *PPM1H,* a promising candidate gene with 17.4% of carriers among resisters vs. 4.5% among infected individuals in the combined sample ($P=1.7x10^{-4}$, OR=0.20[0.08–0.50]). *PPM1H* encodes a phosphatase which counteracts LRRK2, an inhibitor of BCG-triggered apoptosis and a negative regulator of *Mtb* phagosome maturation in macrophages. Our results will guide the understanding of the molecular mechanisms involved in resistance to *Mtb* infection.

33

Circulating Immune Cell Count and Colorectal Cancer Risk: A Mendelian Randomization Study

Andrei-Emil Constantinescu[1,2]*, Caroline J. Bull[1,2], Jeroen R Huyghe[3], Marc J. Gunter[4], Neil Murphy[4], Nicholas J. Timpson[1], Emma E. Vincent[1,2]
[1]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom; [2]School of Translational Health Sciences, University of Bristol, Bristol, United Kingdom; [3]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [4]Nutrition and Metabolism Branch, International Agency for Research on Cancer, World Health Organization, Lyon, France

Colorectal cancer (CRC) is one of the most common cancers in the UK and accounts for around 10% of cancer deaths worldwide. Previous studies have suggested a role for immune cell subtypes in colorectal cancer, with eosinophils having a protective effect and neutrophils having

a detrimental effect. Here, we aimed to investigate the effect of circulating immune cell counts (ICCs) on CRC risk using Mendelian randomization (MR). Genome-wide association study (GWAS) summary statistics for ICCs were accessed from a comprehensive meta-analysis (N=562,132 Europeans), and for CRC overall and by site (colon, proximal colon, distal colon and rectal) through a large meta-analysis (58,221 cases and 67,694 controls in the Genetics and Epidemiology of Colorectal Cancer Consortium, Colorectal Cancer Transdisciplinary Study, and Colon Cancer Family Registry). We performed univariable (UV) and multivariable (MV) MR analyses to assess the effect of ICCs on CRC risk. The inverse-variance weighted UVMR analysis showed evidence of an effect on CRC risk for basophils (overall CRC - OR: 0.88, CI(95%): 0.78-0.99, *P*=0.04), eosinophils (overall CRC - OR: 0.93, CI(95%): 0.88-0.98, *P*=0.01), and overall ICCs (colon - OR: 0.91, CI(95%): 0.85-0.99, *P*=0.02). These results were corroborated by sensitivity UVMR analyses (MR-PRESSO, Cochran's Q test, MR-Egger). The MVMR method provided evidence of an effect for eosinophils (Overall CRC - OR: 0.88, CI(95%): 0.80-0.97, *P*=0.01) and lymphocytes (Overall CRC - OR: 0.84, CI(95%): 0.76-0.93, *P*=0.0007) on overall CRC risk. Our study provides evidence that circulating immune cells play a role in CRC aetiology, laying the path for targeted mechanistic studies.

34

Multi-stage Germline Exome Sequencing Study of 17,546 Men with Aggressive and Non-aggressive Prostate Cancer Identifies Genes for Gene Panel Testing

Burcu F. Darst[1*], Ed Saunders[2], Tokhir Dadaev[2], Xin Sheng[1], Peggy Wan[1], Rosalind A. Eeles[2,3], Fredrik Wiklund[4], Zsofia Kote-Jarai[2], David V. Conti[1], Christopher A. Haiman[1]

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [2]The Institute of Cancer Research, London, United Kingdom; [3]Royal Marsden NHS Foundation Trust, Fulham Road, London, United Kingdom; [4]Karolinska Institute, Solna, Sweden

While gene panel testing has been used to identify men at high risk of aggressive prostate cancer (PCa), genes tested are based on limited evidence due to restrictive sample sizes of previous sequencing investigations and the focus on candidate genes. We conducted a large multi-stage case-only exome sequencing study of 9,185 aggressive and 8,361 non-aggressive cases of European ancestry from 19 international studies. Stage 1 samples (n=5,545) had whole-exome sequencing, and stage 2 samples (n=12,001) had targeted exome sequencing for 1,459 genes selected based on stage 1 and previous evidence. Gene burden analyses of rare (MAF<1%) deleterious variants validated previously associated genes *BRCA2*, *ATM*, and *NBN*. Among potentially novel genes with strong but not exome-wide significant statistical evidence were *PKD2L2*, involved in fertility (PCa death OR=3.5, 95% CI=1.76-7.04, P=$5x10^{-4}$); *MMP19*, involved in reproduction and metastasis (PCa death OR=2.8, 95% CI=1.53-5.05, P=$8x10^{-4}$); *SMPD1*, involved in converting sphingomyelin to ceramide (metastatic OR=5.3, 95% CI=1.85-14.98, P=0.002); and *RIMBP3*, involved in sperm head morphogenesis (aggressive OR=1.8, 95% CI=1.24-2.58, P=0.002). Interestingly, *SMPD1* had the strongest association with aggressive and metastatic disease in gene burden analyses of rare missense variants. Among 24 PCa panel and DNA repair genes with evidence of association with aggressive disease, rare deleterious variants were carried by 5.6% of non-aggressive cases compared to 10.0% of aggressive (OR=1.98, 95% CI=1.75-2.23, P=$9.2x10^{-28}$) and 12.3% of metastatic (OR=2.66, 95% CI=2.21-3.19, P=$1.4x10^{-25}$) cases. These findings provide evidence of association of known and novel aggressive PC genes that should be considered in gene testing panels for PCa.

35

A Data Integration Tool for Identifying Proteomic and Transcriptomic Biomarkers: An Application in TCGA Breast Cancer

Sarmistha Das* and Deo Kumar Srivastava

Department of Biostatistics, St. Jude Children's Research Hospital

Biomarkers are important predictors of disease onset/progression and hence play a vital role in the prediction of patient survival and/or response to therapy. But the key challenge in finding biomarkers for complex diseases is, decoding the intricate interplay of multiple omics data. Although information from multiple omics data intuitively provides substantial information on the disease compared to a single source of data, it is difficult to analyze such correlated information from multiple omics together due to 1) the differences in the data structure emerging from different assays, 2) less sample size than the number of features under study etc.

Protein misfolding has been held responsible for disruption in the normal mechanisms of individuals. Moreover, it is well-known that aberrations in mRNA may disrupt downstream protein formation. Thus, intuitively studying the alteration in protein may provide substantial insight into disease progression, when integrated with the transcriptomic data. But most analytical procedures are unable to integrate subject-specific multi-omics data for a disease outcome because of the high dimensionality and diversity in the nature of multi-omics data.

We propose an algorithm, to identify novel omics signatures that may differentiate between two subtypes or groups, in a single data integration framework. Our method is based on principal components that select important transcriptomic and metabolomic features associated with disease and explores the inter-relation of the most informative features using a multivariate model, to identify disease-associated multi-omics biomarkers. Application of our method to TCGA breast cancer data identified functionally relevant genes and proteins in important pathways.

36

Polygenic Scores are Correlated with Year-of-Birth in Large Biobanks

Maria Niarchou[1], Younga H. Lee[2], Loic Yengo[3], Georgios Voloudakis[4], Sylvanus Toikumo[5], Dan Zhou[1], Peter Straub[1],

Rachel L. Kember[5], Anna R. Doherty[6], Jordan W. Smoller[2], Peter Visscher[3], Naomi Wray[3], Nancy Cox[1], Guanhua Chen[8], Lea K. Davis*[1,9,10,11]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Tennessee, United States of America; [2]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, United States of America; [3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; [4]Icahn School of Medicine at Mount Sinai, NY, United States of America, [5]University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America; [6]The University of Utah, Utah, United States of America; [7]Department of Medical Genetics, Faculty of Medicine, University of British Columbia, Canada; [8]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Wisconsin, United States of America; [9]Department of Biomedical Informatics, Vanderbilt University Medical Center, Tennessee, United States of America; [10]Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Tennessee, United States of America; [11]Department of Molecular Physiology and Biophysics, Vanderbilt University, Tennessee, United States of America

Polygenic scores (PGS) are already being used in the clinical context and may eventually be implemented in routine clinical practice. However, it is important to understand how factors that are not directly related to disease biology, may affect the value of the PGS. For example, we and others observed a correlation between Year of Birth (YoB) and PGS for different health conditions and quantitative traits. We further investigated this phenomenon in the Vanderbilt Biobank (N=61,778) and replicated the finding in biobanks in the PsycheMERGE network (N=36,139). We tested PGSs for coronary artery disease, multiple neuropsychiatric disorders, behavioral traits, anthropometric traits, and blood cell traits. After accounting for sex and population stratification, all PGS – except for blood cell trait PGS – were significantly associated with YoB. To understand the mechanisms underlying these associations, we tested hypotheses related to changes in 1) population structure; 2) relatedness; and 3) ascertainment bias over time. All factors investigated partially explained the association of PGS with YoB across the biobanks investigated, but none fully accounted for the trends. Overall, our findings indicate that PGS can be systematically biased by differences in year of birth cohort, and suggest a complex interplay of time-dependent factors mediate this bias. Accordingly, further studies are needed to ensure that PGS modeling is not biased by age differences or other possible confounders.

37

Investigation of Circulating Proteins and Risk of Overall, Advanced, and Early Onset Prostate Cancer: A Mendelian Randomization Study

Trishna A. Desai*[1], Åsa K. Hedman[2,4], Eleanor L. Watts[3], Mattias Johansson[5], Anders Mälarstig[2,4], Timothy J Key[1], Ruth C. Travis[1], Karl Smith-Byrne[1].

[1]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, England; [2]Department of Medicine, Karolinska Institute, Stockholm, Sweden; [3]Metabolic Epidemiology Branch, National Cancer Institute, Bethesda, Maryland, United States of America; [4]Emerging Science & Innovation, Pfizer Worldwide Research & Development, Cambridge, United States of America; [5]Genomic Epidemiology Branch, International Agency for Research on Cancer, Lyon, France

Introduction: Prostate cancer (PC) is the second most diagnosed malignancy in men worldwide, with few established risk factors and poor prognosis for advanced stage disease. To elucidate the potential molecular underpinnings of PC, we leveraged recent advances in -omics technology to investigate the relationship between genetically predicted circulating protein concentrations and PC risk, including advanced (metastatic disease or Gleason score >=8 or Prostate Specific Antigen >100 ng/mL or PC death) and early onset (diagnosis <55 years) subtypes.

Methods: We conducted a two-sample Mendelian randomization study collating genetic instruments for 2,329 proteins using 6,476 *cis* SNPs and outcome data for overall (N_{cases} = 85,554), advanced (N_{cases} = 15,167) and early onset (N_{cases} = 6,988) PC from the PRACTICAL consortium. We considered a protein robustly associated with prostate cancer risk where a Wald ratio was Bonferroni significant (0.05 divided by the number of proteins investigated) and there was evidence of a shared genetic signal from colocalization (Posterior Probability 4 > 0.7).

Results: Seven hundred and fifty-one unique proteins were associated with at least one PC endpoint. Four hundred and sixty-seven associations were observed for overall PC, 313 for advanced PC, and 332 for early onset PC. Of these 751 unique proteins, 66 passed the Bonferroni correction of significance, and 9 further showed evidence in support of colocalization.

Discussion: These findings introduce a suite of credible molecular biomarkers that may be implicated in the aetiology of PC and highlight an opportunity to further investigate related metabolic pathways and potential drug targets.

38

Identifying Genes Contributing to Dementia and Cardiovascular Risk Differently in Men and Women: Extending ExPheWAS to Test Sex-by-gene Interactions

Tatiana Dessy[1,2,3,5]*, Sarah Gagliano Taliun[1,4,6], Louis-Philippe Lemieux-Perreault[1,2], Amina Barhdadi[1,2], Marie-Pierre Sylvestre[1,3,5], Marie-Pierre Dubé[1,2,3,4]

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

Background: Health conditions may develop differently by sex. Gene discovery methods that include sex-specific analyses have the potential to inform on the sex-specific

nature of disease etiology. Here, we explore genetic variability at the *APOE* gene as a risk factor for both cognitive and cardiovascular conditions using a sex-specific approach.

Methods: Using ExPheWAS, a gene-based genome and phenome-wide association tool that models variation in genes by principal component analysis, we tested for interaction between sex and gene by analysis of deviance. Three nested generalized linear models are estimated for each gene-outcome pair, to test (1) the significance of the gene's principal components and (2) of the sex-by-gene interaction. We applied the interaction test in the UK Biobank to study *APOE* using carotid intima-media thickness (cIMT, mean and max; N=21,000), fluid intelligence score (N=123,397) and mean time to correctly identify matches (N=381,850) as endophenotypes for cardiovascular and cognitive health conditions.

Results: In generalized linear models adjusted for age, sex and principal components for ancestry, the *APOE* gene was significantly associated with mean cIMT (P=3.21x10^{-11}) and max cIMT (P=9.91x10^{-10}). The interaction test provided evidence of heterogeneity between sexes for cIMT mean only (P=0.04). No association was significant for cognitive endophenotypes.

Discussion: *APOE* may contribute to cardiovascular risk differently in men and women. Validation in diverse populations is needed to inform on the generalizability of findings. Results from the ExPheWAS sex-by-gene interaction test applied genome-wide is expected to provide additional insight into the sex-specific genetic architecture of diseases and traits.

39

A Versatile, Fast, and Unbiased Method for Estimation of Gene-by-environment Interaction Effects on Biobank-Scale Datasets

Matteo Di Scipio[1,2*], Mohammad Khan[1,2 *], Shihong Mao[1], Michael Chong[1,3,4], Nazia Pathan[1,2], Conor Judge[1], Nicolas Perrot[1], Walter Nelson[5,6], Shuang Di[5,7], Jeremy Petch[1,2,5,8], Guillaume Paré[1,3,4,9]

* Contributed equally

[1]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada; [2]Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada; [3]Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada; [4]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Canada; [5]Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada; [6]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada; [7]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; [8]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada; [9]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada.

Current methods to evaluate gene-by-environment interactions (GxE) on biobank-scale datasets are limited due to constraints such as high data dimensionality (p>>n), weaker signals than marginal genetic and environmental effects, computational burden, and common sources of heritability biases. Our method, MonsterLM, enables multiple linear regression on blocks of up to 30,000 SNPs and their interactions to estimate GxE on genome-wide datasets without requiring assumptions about genetic architecture. We show that MonsterLM provides unbiased estimates of variance explained by GxE effects across a range of MAF and LD quantiles (MAF>0.01; LD $r^2 < 0.9$) and is robust to common biases including varying SNP effect sizes and collider bias. We applied MonsterLM to the UK Biobank to test for GxE of waist-to-hip ratio (WHR) with eleven complex traits, including ten blood biomarkers and height (N=297,529-325,989), and two dichotomous disease traits (N=324,858-325,989). We identified significant genome wide GxE with WHR for eight biomarkers and both dichotomous diseases, with variance explained by interactions ranging from 0.009 to 0.071. Generally, >50% of GxE was attributed to variants without significant marginal association with the phenotype of interest. Conversely, 5% or less of variants contributed to >50% of GxE. We observed modest improvements in polygenic score prediction by additionally incorporating GxE for some biomarkers. Our results imply an important contribution of GxE to complex trait variance, driven largely by a restricted set of variants distinct from loci with strong marginal effects.

40

Tofu Intake and Risk of Neuropsychiatric Diseases: A Two-sample Mendelian Randomization Study

Ziyi Ding [1*], Yuxuan Sun [2], Maggie H. Wang [3], Rui Sun [2,3]

[1]The Herbert Wertheim School of Public Health and Longevity, University of California San Diego, La Jolla, California, United States of America; [2]Scientific Research Center, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518107, P. R. China; [3]The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China

Epidemiological studies report associations of tofu intake with neuropsychiatric diseases, but the results were inconsistent and whether these associations are casual remained unclear. This study aims to explore the potential causality between tofu intake and risk of three common neuropsychiatric disorders, including Alzheimer's disease (AD), schizophrenia, and major depressive disorder (MDD) using two-sample univariable and multivariable Mendelian randomization (MR) analyses. The causal inference was investigated using summary data from genome-wide association studies (GWASs) in Europeans with tofu intake (N = 64,945), AD (N = 63,926), schizophrenia (N = 77,096), and MDD (N = 173,005). Independent GWAS summary statistics from BioBank Japan Project on East Asians were used to evaluate the causality in a different population. In Europeans, the OR of tofu intake was 29.04 in AD (P value = 0.026) after adjusting for smoking and alcohol consumption and the result was consistent in univariable analysis (P value = 0.006). In East Asians, genetically predicted tofu intake was

negatively associated with risk of AD with an OR to be 0.34 after controlling covariates (*P* value = 0.050). In all the analyses, there was no evidence of causal association of tofu intake with schizophrenia and MDD. Besides, multivariable MR analyses indicated smoking and alcohol intake contributed to increased risk of schizophrenia (smoking: OR = 1.30, *P* value = 1.48×10^{-5}; alcohol: OR = 1.33, *P* value = 0.037). Our results provided evidence
in support of casual association between tofu intake and AD and the effects may be population-specific.

41

Polygenic Risk Score Improves the Predictive Ability of Subsite-specific Colorectal Cancer: Evidence from Two Large-scale Prospective Cohorts

Junyi Xin[1,] Xia Jiang[2], Ni Li[3], Qianyu Yuan[4], Xuesi Dong[3], Zhengdong Zhang[1,5], David C. Christiani[4,6], Mulong Du[4,7,*], Meilin Wang[1,5]

[1]Department of Environmental Genomics, Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China; [2]Department of Clinical Neuroscience, Centre for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden; [3]Office of Cancer Screening National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; [4]Departments of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America; [5]Department of Genetic Toxicology, The Key Laboratory of Modern Toxicology of Ministry of Education, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China; [6]Department of Medicine, Massachusetts General Hospital, Boston, MA, United States of America; [7]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

Colorectal cancer (CRC) is a heterogeneous disease presenting subsite-specific characteristics. While the general polygenic risk score (PRS) built based on previously identified CRC risk loci has been demonstrated as an effective tool to stratify high-risk population, its subsite-specific predictive performance remains undetermined.

Here, leveraging data from two large-scale prospective cohorts of European ancestry, including the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial involving 13,236 individuals and the UK Biobank involving 355,543 individuals, we found that carriers of high PRS had an increased risk of developing left-sided (PLCO: HR per SD = 1.36, $P = 2.65 \times 10^{-8}$; UK Biobank: HR = 1.60, $P = 3.45 \times 10^{-82}$) and proximal CRC (PLCO: HR = 1.25, $P = 7.31 \times 10^{-5}$; UK Biobank: HR = 1.56, $P = 2.28 \times 10^{-36}$), both in dose-response manner without subsite heterogeneity (PLCO: $P_{heterogeneity} = 0.296$; UK Biobank: $P_{heterogeneity} = 0.595$).

Notably, receiver operating characteristic (ROC) curve revealed that PRS could significantly improve the discriminatory ability of traditional risk prediction model (including sex, age, body mass index [BMI], smoking and family history of CRC) across different sites, particularly for individuals of left-sided CRC where an improvement of over 5% was observed (PLCO: 7.02%; UK Biobank: 6.08%).

To conclude, as evidenced by two large-scale cohorts, the general PRS is effective in identifying high-risk individuals susceptible to subsite-specific CRC, providing clinical relevance by aiding individualized prevention.

42

Network Analysis of Multi-omics Data Identifies Shared Genes and Pathways Underlying the Risk of Allergic Diseases and IgE Production

Pradeep Eranti[1]*, Raphael Vernet[1], Emmanuelle Bouzigon[1], Florence Demenais[1] on behalf of the EGEA, EVADA and RESET-AID Consortia

[1]Université Paris Cité, Inserm, UMRS-1124, Group of Genomic Epidemiology and Multifactorial Diseases, Paris, France

Both genetic and epigenetic mechanisms influence risk of allergic diseases (AD) and Immunoglobulin E (IgE) levels, a key intermediate phenotype in allergy. Integration of omics data related to AD and IgE can provide more insight into allergy-related mechanisms.

We applied network analysis to summary-level data from a publicly available genome-wide association study (GWAS) of AD (242,569 subjects, GWAS catalog:GCST005038), and from epigenome-wide association studies (EWAS) of IgE levels based on targeted bisulfite sequencing and conducted separately in 343 asthmatics and 370 non-asthmatics from the EGEA study. We used the STRING protein-protein interaction network as background and the SigMod method to identify modules (sub-networks) enriched in trait-associated genes. We then investigated the relationships between the AD-GWAS gene module and each IgE-EWAS gene module.

We identified three modules of (i) 292 genes and 1,233 interactions derived from AD-GWAS; (ii) 251 genes and 720 interactions derived from IgE-EWAS in asthmatics; (iii) 205 genes and 534 interactions derived from IgE-EWAS in non-asthmatics. More than 99% of the genes from each module were nominally associated ($P \leq 0.05$) with their respective trait. The AD-GWAS module shared eight and seven genes, respectively, with IgE-EWAS modules (asthmatics and non-asthmatics). A significant proportion of AD-GWAS module genes directly interacts with genes from each IgE-EWAS module (more than 70%, $P < 10^{-5}$). All three modules were enriched for pathways related to antigen presentation in the context of HLA complex. This study shows shared biological mechanisms underlying AD-GWAS and IgE-EWAS gene modules. Further functional characterization of these modules is underway.

Funding: H2020-MSCA-ITN-2018 (Grant-ID 813533)

43

Genome-wide Association Study of Asthma in 1,587 French-Canadian Subjects

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

[1]*Institut universitaire de cardiologie et de pneumologie de*

Québec, Université Laval, Quebec, QC, Canada; [2]Department of Social and Preventive Medicine, Laval University, Quebec, QC, Canada; [3] Department of Molecular Biology, Medical Biochemistry and Pathology, Laval University, Quebec, QC, Canada; [4]Department of Molecular Medicine, Laval University, Quebec, QC, Canada

Context: Asthma is a common respiratory disease with both genetic and environmental risk factors. Heritability estimates for asthma range from 0.55 to 0.90. Numerous genome-wide association studies (GWASs) have been completed and have identified several single nucleotide polymorphisms (SNPs) associated with asthma.

Aims: Our research aims to perform a GWAS in the Quebec City Case-Control Asthma Cohort (QCCCAC) which consists of 1,587 French-Canadian subjects (1,056 asthmatics and 531 healthy controls). We also create a genome-wide Polygenic Risk Score (PRS) to estimate an individual's genetic predisposition for asthma in our cohort.

Methodology: The genetic association analysis was performed using SAIGE (Scalable and Accurate Implementation of Generalized mixed model) adjusting for age, sex, and the first 20 ancestry-based principal components. We used summary statistics reported by the Trans-National Asthma Genetic Consortium meta-analysis (23,948 cases and 118,538 controls, 1,991,789 SNPs) to calculate the PRS in our cohort by applying the LDpred2 method, a Bayesian shrinkage approach. The prediction accuracy of models was evaluated by the adjusted area under the receiver operating characteristics curve (AUC).

Results: Seven distinct genetic loci were identified (a suggestive association threshold of P value $< 1 \times 10^{-6}$) in the GWAS analysis, including asthma-associated SNPs located in *KCNQ5* and *MYL10*. The model with only sex, age, and the first 20 principal components showed an AUC of 0.588 [95% CI: (0.533-0.643)]. After adding PRS to the model the AUC increased to 0.610 [95% CI: (0.556-0.665)].

Conclusion: These results suggest that individuals' genetic background may improve asthma risk prediction.

44

IBDMap: Biobank Scale Shared Segments Analysis

Grahame F. Evans[1]*, Ryan J. Bohlender[2], Alexander S. Petty[3], Lauren E. Petty[3], Hung-Hsin Chen[3], James T. Baker[1], Jennifer E. Below[1,3], Chad D. Huff[2]

[1]Department of Human Genetics, Vanderbilt University, Nashville, Tennessee, United States of America; [2]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

The recent increase in available biobank genetic datasets has created opportunities for novel disease gene discovery approaches. Identity by descent (IBD) mapping evaluates IBD segment sharing among cases and controls to identify genomic regions harboring rare, disease causing variation undetectable through traditional genome wide association studies. However, IBD mapping in biobank scale data is a substantial computational challenge. To this end, we developed IBDMap, a multithreaded, scalable C++ tool which employs a map/reduce parallelized permutation strategy to determine IBD sharing enrichment by comparing genome wide sharing rates among cases and controls. IBDMap includes a novel, optimized implementation of the Benjamini and Yekutieli false discovery rate (FDR) procedure, offering a substantial statistical power increase given the highly correlated nature of IBD segment data. Using hapIBD and iLASH, we established a consensus IBD segment repository in Vanderbilt University's biobank, BioVU for 69,819 individuals genotyped on the MEGA^{EX} array. We applied IBDMap to identify two statistically significant regions on chromosomes 2 and 13 with genome wide significant enrichment of IBD sharing in atherosclerosis cases (FDR Y and Z, respectively). The chromosome 13 region would not have been identified using standard multiple testing correction approaches. The chromosome 2 region contains gene GALNT13, previously linked to atherosclerosis related conditions in diabetic individuals. We are following up on these findings by sequencing the top four haplotype clusters enriched for cases in each region of interest and across additional cardiovascular phenotypes. Our results demonstrate the potential of IBDMap as a novel approach to gene discovery.

45

The X Chromosome in Neurodegeneration

Patrick D. Evans*[1], Eric R. Gamazon[1,2]

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

Neurodegeneration is a key phenotype in aging and in diseases such as Parkinson's and Alzheimer's disease. The X chromosome is enriched for genes functioning in the brain and has many links to brain physiology and development. However, the X chromosome is also an often overlooked chromosome in analysis of the genetic etiology of diseases. We performed analysis of the X chromosome in the Vanderbilt University Medical Center biobank samples, called BioVU, across neurodegeneration phenotypes. Genome-wide association analysis (GWAS) was performed for each neurodegeneration phenotype within BioVU. GWAS results were then used in a combined polygenic risk score analysis within BioVU for neurodegenerative traits using existing brain phenotype GWAS as training sets. Finally, predicted expression of each gene on the X chromosome was calculated from models using elastic net from the Genotype-Tissue Expression (GTEx) consortium data. The models were then applied to all samples in BioVU and transcriptome-wide association analysis (TWAS) was performed on all neurodegenerative phenotypes. Results of these analyses will be presented and the function of implicated genes will be discussed.

46

Immunogenetics of Healthy Aging at the Single Cell Level

Elyssa Bader[1,2], Marie-Julie Favé[1]*, Mawusse Agbessi[1], Jasmina Uzunovic[1], Nicholas Cheng[1,2], Elias Gbeha[1], Vanessa Bruat[1], Kim Skead[1,2], David Soave[3], and Philip Awadalla[1,2,4]
[1]Ontario Institute for Cancer Research; [2]Department of Molecular Genetics, University of Toronto; [3]Wilfried Laurier University; [4]Dalla Lana School of Public Health

During aging, hematopoiesis is dysregulated leading to deleterious effects on health and contributes to increased risk for developing infections, cancer, chronic diseases, and death. However, the direct effects of aging on hematopoietic cells have been difficult to dissect because of the genetic and environmental contributions, and further confounded by age-associated diseases. Here, we profiled with multi-omics 400 participants free of diagnosed conditions selected from the extremes of the age and health risk spectrum to identify cellular, transcriptomic, and epigenomic features that contributes to healthy aging. We performed whole-genome, ATAC-seq and single-cell RNA sequencing on over 500,000 cells, and although we find limited changes in cell type counts, we found cell type specific differential gene expression associated with immune health, suggesting that functions rather than counts are impacted by health as one age. We developed cell-type specific transcriptional risk scores and were able to classify individuals based on immune health with consistent accuracy. We identified 1105 eQTLs associated with immune health, with higher numbers and effects sizes across innate immune cell types. We show that genetic contribution to gene expression decreases in healthy aged individuals in innate cell types, specifically in stress related pathways. Our results show that healthy aging of the hematopoietic and immune system happens through the modification of cell type specific functions, particularly of the innate immune cell types. Our results suggest that healthy agers might be able to delay or escape immunoscenescence through cell type specific regulation and decreased activation of stress and pro-oncogenic genetic regulatory mechanisms.

47

Deciphering the Causal Effects of Cytokines on Human Health: A Phenome-wide Mendelian Randomization Study

Alba Fernández-Sanlés[1,2]*, Nancy McBride[1,2], Marwa Al Arab[1,2], Louise A.C. Millard[1,2], Jie Zheng[1,2], Tom Gaunt[1,2], Deborah A. Lawlor[1,2], Maria C. Borges[1,2]
[1]MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom; [2]Population Health Science, Bristol Medical School, University of Bristol, Bristol, United Kingdom

Cytokines are signaling proteins involved in multiple processes such as inflammation and immune responses, which have been implicated in many diseases. Understanding the causal effect of cytokines on unselected outcomes in a hypothesis-free approach could provide insights for management of multimorbidity. We conducted a *phenome-wide two-sample Mendelian Randomization study (MR-PheWAS)* to probe the potential effect of 133 cytokines and downstream proteins on 1,235 traits from the IEU OpenGWAS project. Those proteins, measured in 35,559 Icelandic individuals using an aptamer-based proteomic technology, were instrumented with 444 cis-acting protein Quantitative Trait Loci (pQTLs). We found that 102 proteins are related to at least one of 188 traits (false discovery rate <0.05), including cancer, cardiovascular, endocrine, respiratory, immunological, infection, anthropometric, lifestyle and hematological traits. 53% of proteins were related to more than one trait (up to 24), being 12% associated with two or more diseases (up to six). For example, higher IL6R levels were associated with lower risk of coronary artery disease (OR=0.95[95%CI:0.93-0.97]), rheumatoid arthritis (OR=0.93[95%CI:0.90-0.96]) and inflammatory bowel syndrome (0.96[95%CI:0.94-0.98]), and higher risk of eczema (OR=1.07[95%CI:1.04-1.11]). 8% of traits were modulated by at least two proteins (up to 91). In follow-up analyses, we are seeking replication in independent samples (UK Biobank and FinnGen) and conducting sensitivity analyses to explore potential bias due to confounding by linkage disequilibrium (genetic colocalization), horizontal pleiotropy and epitope effects on pQTLs. Our results highlight the role of cytokines in potentially influencing a wide range of diseases, with implications for drug development, especially in the context of multimorbidity.

48

Leveraging Consanguinity in the UK Biobank Cohort to Identify Rare Recessive Variants Involved in Complex Traits

Sidonie Foulon[1,*], Margot Derouin[1], Marie-Sophie Ogloblinsky[2], Steven Gazal[3], Hervé Perdy[4] and Anne-Louise Leutenegger[1]
[1]Institut national de la santé et de la recherche médicale (Inserm), Université Paris Cité, NeuroDiderot, Paris, France; [2]Inserm, Université de Bretagne Occidentale, Génétique Génomique fonctionnelle et Biotechnologies, Brest, France; [3]University of Southern California, Los Angeles, CA, United States of America; [4]Université Paris Saclay, Inserm, Centre de recherche en épidémiologie et santé des populations, Villejuif, France

Mating between relatives is frequent and encouraged for social or economic reasons in many human populations (about 10% of the world population). Offspring of relatives are consanguineous and their genome carries homozygous-by-descents (HBD) segments. The genetic component of common complex traits could be partly due to the contribution of rare variants with recessive effects. These types of variants can be found in HBD-segments. We propose here an approach that relies on an excess of HBD segments shared among cases compared to what is expected among controls. We have implemented it in the R package Fantasio. We illustrate its performance and adjust the model on the UK Biobank prospective cohort (~500,000 individuals living in the United-Kingdom; https://www.ukbiobank.ac.uk/). We focus on the diabetes phenotype constructed from available fields of the biobank. We inferred the genetic ancestry of the individuals using a random forest classifier based on the 1000 Genomes and HGDP-CEPH reference panels. We found consanguineous individuals in all ancestries with the largest proportions observed in the Central South Asian (42%) and

Middle Eastern (40%) ancestries. Moreover, we noticed an excess of consanguineous individuals in the diabetes cases in all ancestries but most significantly in the African (P=0.004), Central South Asian (P=0.047) and European ancestries (P=0.002). We finally present the regions of the genome found associated with diabetes and likely carrying rare recessive variants in the different population ancestries.

This research was conducted using the UK Biobank Resource under Application #59366.

49

A Hybrid Random Forest Variable Selection Approach for Omics Data

Césaire J. K. Fouodo, Inke R. König Silke Szymczak*
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Identifying important variables for prediction is challenging in the context of omics data sets such as genetics or transcriptomics. Often, only a few of the numerous predictor variables are relevant, and applying standard approaches might be computationally demanding. As the usage of random forests (RF) for classification and regression problems in high-dimensional settings increases, RF based testing procedures for variable selection become more popular. The Vita RF testing procedure have been showed to be computational fast and effective for identifying relevant predictor variables in omics data analysis. However, depending on the chosen significance threshold, results can be unstable or contain an increased number of false-positive findings. An alternative is the iterative Boruta approach, which is more powerful but time consuming, especially in high dimensional settings.

The hybrid approach we propose combines the Vita and Boruta methods by using the idea of Vita at each iteration of the Boruta method. We conduct simulation studies based on both theoretical and experimental expression datasets to compare the three procedures using different evaluation criteria. We also applied the three methods on real omics experimental data.

Results show that the hybrid approach is a good compromise between the two underlying methods. It is more stable than Vita, considerably faster than Boruta, and leads to fewer false positive findings. The application of the three methods on experimental data sets confirms our findings. We thus recommend our hybrid approach for variable selection on omics data sets. We anticipate that it will be particularly beneficial for genetic studies.

50

A Novel Multi-Components Mixed Model Based Bacterial-Gwas Method and Its Application to Listeria Monocytogenes

Arthur Frouin[1]*, Fabien Laporte[1], Lukas Hafner[2], Marc Lecuit[2], Mylène Maury[2], Rayan Chikhi[1], Hugues Aschard[1]
[1]Institut Pasteur, Université de Paris, Department of Computational Biology, Paris, France; [2]Institut Pasteur, Université de Paris, Department of Cell Biology & Infection, Paris, France

Modern human cohorts now provide access to an overwhelming amount of omics data, including molecular phenotypes pending to the host, but also an increasing amount of genetic and omics data from organisms involved in human health. Here we investigate methods for genome-wide association studies (GWAS) of pathogenic bacterial genome, focusing on the role of Listeria Monocytogenes genetics on the severity of listeriosis infection. Performing bacterial GWAS requires to address exacerbated population structure, which has proven to be extremely challenging. Diverse and heterogeneous methods have been proposed, but fundamental genetic modelling aspects have been surprisingly seldom discussed.

We used real bacterial sequence data and simulated phenotypes to conduct a formal comparison of alternative methods for modelling genetic structure, estimating heritability, and testing association. We leverage those results to develop a robust and powerful approach which we applied to the MONALISA cohort, a unique large scale national prospective repository that systematically collects Listeria Monocytogenes strains in France, currently including whole genome sequencing of 3718 strains.

We first demonstrate that the classic human heritability model, commonly assumed in existing bacterial GWAS methods, is strongly unadapted to study highly structured organisms such as Listeria. The most efficient approach consists in a multicomponent linear mixed model applied to unitigs, where components are inferred from a hierarchical clustering of the genetic relatedness matrix. Virulence's heritability analyses showed shared and specific genetic effect across clonal complexes (CC), in agreement with clinical results. The GWAS indicated suggestive associations among unitigs shared across CC.

51

Glucose, Insulin, and Brain Health in the UK Biobank: A Mendelian Randomization Study

Victoria Garfield*[1], Christopher T. Rentsch[2], Liam Smeeth[2], Krishnan Bhaskaran[2], Nishi Chaturvedi[1]
[1]MRC Unit for Lifelong Health and Ageing, Institute of Cardiovascular Science, University College London, United Kingdom
[2]Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, United Kingdom

Previous Mendelian randomization (MR) studies have not demonstrated a causal association between higher glycated hemoglobin (HbA_{1c}) or type-2 diabetes and worse brain health. It is, however, plausible that other unexplored glycemic traits causally relate to poorer cognition, lower brain volumes and excess dementia risk. We used a two-sample MR design in the UK Biobank to understand whether fasting glucose (FG-109 SNPs), fasting insulin (FI-48 SNPs) and post-load glucose (2hPG-15 SNPs) causally relate to cognitive function (reaction time/visual memory, n=329,288), brain volumes (total/hippocampal/white matter hyperintensity

volume - WMHV, n~32,000) and all-cause dementia. We used genetic variants from the latest genome-wide association study (GWAS) and performed inverse-variance weighted (IVW) MR analyses, alongside tests for horizontal pleiotropy (Weighted median and MR-Egger). Results are expressed as exponentiated beta coefficients for cognition and WMHV due to log-transformations, betas(cm^3) for total brain/hippocampal volumes and ORs for dementia. None of our glycemic exposures was associated with reaction time [FG exp(ß)=1.00 (95%CI=0.99;1.01); FI exp(ß)=0.99 (95%CI=0.98;1.00; 2hPG exp(ß)=1.00 (95%CI=0.99;1.01]; or total brain volume [FG ß=-4.82cm^3 (95%CI=-12.94;3.30); FI ß=3.02cm^3 (95%CI=-12.30;18.34); 2hPG ß=-1.21cm^3 (95%CI=-6.13;3.71);]. Other results in the cognitive and neuroimaging domains followed a similar pattern. We also found no associations with all-cause dementia [FG OR=0.92 (95%CI=0.66;1.30); FI OR=0.68 (95%CI=0.36;1.31); 2hPG OR=1.27 (95%CI=0.97;1.67)]. MR-Egger intercept *P* values indicated no concerns with horizontal pleiotropy. Using MR, we did not find evidence that fasting glucose, fasting insulin and 2-hour post-load glucose are causally associated with brain health outcomes. It is likely that other factors related to hyperglycemia should be investigated to understand the observed association with brain health.

52

PRS: A Misled Interpretation of Genetic Variation Used for Complex Disease Risk Prediction

Anthony F. Herzig[1*], Emmanuelle Génin[1], Françoise Clerget-Darpoux[2]

[1]Inserm, Université de Brest, EFS, CHU Brest, UMR 1078, GGB, Brest, France; [2]Emeritus Research Director at INSERM, Paris, France

Polygenic Risk Scores (PRS) are described as promising tools that will soon be ready to enter the clinics to help disease risk prediction and refine diagnosis. For a given disease, a PRS is constructed by combining information at different SNPs to maximize the contrast between cases affected by the disease and controls. Leveraging on the numerous GWAS that have been conducted, PRSs are now available for many diseases and used at the population level to classify individuals according to their relative genetic risk and at the individual level to estimate their absolute risk. Behind PRS is the strong assumption that all multifactorial diseases can be explained by an underlying quantitative trait, the liability, explained by the additive and small effects of many independently acting genetic and environmental factors. This is the so-called Polygenic Additive Liability (PAL) model proposed 60 years ago to explain the familial segregation of diseases not compatible with a monogenic transmission model.

We show how the PAL model assumptions of genetic and phenotypic homogeneity and lack of gene-gene and gene-environment interactions are far from the complex reality of pathophysiological processes and could lead to false disease risk estimates. We warn against the general adoption of the PAL model that accredits the idea that our diseases are genetically determined and that our genetic risks of contracting a disease are known at birth and call for alternative strategies to better understand disease complexity.

53

Appraising the Causal Role of Risk Factors in Coronary Artery Disease and Stroke: A Systematic Review of Mendelian Randomization Studies

Andrea N. Georgiou[1*], Loukas Zagkos[2], Wei Xu[3], Lijuan Wang[3], Georgios Markozannes[1,2], Eleni M. Loizidou[1,5], Evropi Theodoratou[3,4], Evangelos Evangelou[1,2,6], Konstantinos K Tsilidis[1,2,6], Ioanna Tzoulaki[1,2,6,7]

[1]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; [2]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [3]Centre for Global Health, Usher Institute, The University of Edinburgh, Edinburgh, United Kingdom; [4]CRUK Edinburgh Centre, Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, United Kingdom; [5]Biobank.cy Center of Excellence in Biobanking and Biomedical Research, University of Cyprus; [6]Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, Ioannina, Greece; [7]BHF Centre of Research Excellence, Imperial College London, London, United Kingdom

Background: Mendelian randomization (MR) offers a powerful approach to study potential causal associations between exposures and health outcomes, by using genetic variants associated with an exposure as instrumental variables. In this systematic review, we aimed to summarize previous MR studies and evaluate the evidence for causality for a broad range of exposures and two principal subtypes of cardiovascular diseases, coronary artery disease (CAD) and stroke.

Methods: MR studies investigating the association of any genetically predicted exposure with CAD or stroke were identified in PubMed. Studies were classified into four categories, namely robust, probable, suggestive, insufficient, based on the significance of the main MR analysis results and the concordance with the sensitivity analyses (MR-Egger, weighed median and MR-PRESSO) based on data extracted from each study. Associations that did not perform any sensitivity analysis were classified as non-evaluable.

Findings: We identified 810 associations eligible for evaluation examining 340 distinct exposures. Of them, 87 were robust, 197 were probable,48 were suggestive and 478 had insufficient evidence. The most prominent *robust* associations were observed for anthropometric traits with CAD, clinical measurements with CAD and stroke, lipids and lipoproteins with CAD and thrombotic factors with stroke.

Conclusion: In this large-scale systematic review, we summarized and evaluated the evidence of association between genetically determined exposures and CVD risk. Only a limited number of associations were robust, whereas most associations were either insufficient or non-evaluable. We provide an overview of future avenues of research and approaches for a more systematic assessment of main and sensitivity MR analyses.

54

Can Individual Studies Add Value to Characterize Loci from Genome-wide Association Meta-analyses? A Case Study.

Dariush Ghasemi-Semeskandeh[1,2*], David Emmert[1], Eva König[1], Luisa Foco[1], Martin Gögele[1], Laura Barin[1], Christian Fuchsberger[1], Dorien J.M. Peters[2], Peter P. Pramstaller[1], Cristian Pattaro[1]

[1]Institute for Biomedicine, Eurac Research, Bolzano, Italy; [2]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

Genome-wide association study (GWAS) meta-analyses identified thousands of loci. Given genetic variant effect sizes vary across contributing studies due to environmental or genetic differences, study-specific features might contribute to biological mechanism elucidation. We contrasted a GWAS of estimated glomerular filtration rate (eGFR) in a population-based study from the Alps, where thyroid disease is common, against a meta-analysis from the CKDGen Consortium (n>1 million) that previously identified 147 kidney function relevant loci.

In the Cooperative Health Research in South Tyrol (CHRIS) study (n=10,146), we fitted mixed models on ln(eGFR), accounting for relatedness (genomic inflation λ=1.02), to assess direction-consistent replication of CKDGen results. Replicated variants underwent phenome-wide mediation analysis across 72 biomarkers.

We replicated 10 loci, totaling 162 variants correlated with the locus lead variant ($r^2 \geq 0.8$). In CHRIS, replicated variants had 1.3-to-5.8 times larger effects than in CKDGen, at similar minor allele frequencies. Free triiodothyronine and thyroxine resulted as potential mediators for most loci. Variant-by-hyperthyroidism interaction testing P-values were shifted towards small values, compatibly with a systematic effect modification.

The potential modifier role of hyperthyroidism on kidney function genetic associations warrants independent replication.

55

An Exploration of Linkage Fine-Mapping on Sequences from Case-Control Studies

Payman Nickchi[1], Charith Karunarathna[1,2], Jinko Graham[1,*]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada

[2]Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada

Linkage analysis maps genetic loci for a heritable trait by identifying genomic regions with excess relatedness among individuals with similar trait values. Analysis may be conducted on related individuals from families, or on samples of unrelated individuals from a population. For allelically heterogeneous traits, population-based linkage analysis can be more powerful than genotypic-association analysis. Here, we focus on linkage analysis in a population sample, but use sequences rather than individuals as our unit of observation. Earlier investigations of sequence-based linkage mapping relied on known sequence relatedness, whereas we infer relatedness from the sequence data. We propose two ways to associate similarity in relatedness of sequences with similarity in their trait values and compare the resulting linkage methods to two genotypic-association methods. We also introduce a procedure to label case sequences as potential carriers or non-carriers of causal variants after an association has been found. This post-hoc labeling of case sequences is based on inferred relatedness to other case sequences. Our simulation results indicate that methods based on sequence-relatedness improve localization and perform as well as genotypic-association methods for detecting rare causal variants. Sequence-based linkage analysis therefore has potential to fine-map allelically heterogeneous disease traits.

56

Fluorescent Signal Probe Patterns Strongly Influence Performance of Methods to Estimate Covariate Effects on DNA Methylation Levels

Lai Jiang[1], Keelin Greenlaw[1], Antonio Ciampi[2], Angelo J. Canty[3], Jeffrey Gross[4], Gustavo Turecki[4], Celia M.T. Greenwood[1,2,5,6,*]

[1]Lady Davis Institute, Jewish General Hospital, Montreal, Canada; [2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Canada; [3]Department of Mathematics and Statistics, McMaster University, Canada; [4]Douglas Mental Health University Institute, Montreal, QC, Canada; [5]Gerald Bronfman Department of Oncology, McGill University, Montreal, QC, Canada; [6]Department of Human Genetics, McGill University, Montreal, QC, Canada

Introduction: 5-hydroxymethylcytosine (5hmC) is a methylation state linked with gene regulation, commonly found in cells of the central nervous system. 5hmC is associated with demethylation of cytosines from 5-methylcytosine (5mC) to the unmethylated state. The presence of 5hmC can be inferred by a paired experiment involving bisulfite and oxidation-bisulfite treatments on the same sample, followed by a methylation assay using a platform such as the Illumina Infinium MethylationEPIC BeadChip (EPIC).

Methods and Results: We developed a Bayesian hierarchical model for estimating the presence of 5hmC, and whether its levels depend on covariates (previously presented at IGES 2020). To evaluate performance, we designed simulations that started from observed fluorescent signal patterns in data from brain tissues. In fact, we designed 7 different sets of simulations, each representative of a different kind of pattern of fluorescent signals among the EPIC probes. The sensitivity and false positive rate for detection of 5hmC, and for inferring covariate effects, varied importantly across the 7 sets of simulations.

Conclusions: It may not be true that "one size fits all"! Careful simulation design prevented false conclusions about the performance of methods for inferring 5hmC.

57

Predicting Cancer Risk from Germline Next-generation Sequencing Data Using a Novel Context-based Variant Aggregation Approach

Zoe Guan[1]*, Colin B. Begg[1], Ronglai Shen[1]
[1]*Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, United States*

Twin studies have shown that most common cancer types have a substantial heritable component. However, known risk variants (SNPs from GWAS and pathogenic variants in known cancer predisposition genes) explain only a limited proportion of the estimated heritability of cancer. It has been hypothesized that much missing heritability lies in rare variants not captured by SNP arrays. Rare variants need to be aggregated to achieve sufficient statistical power for detection. We propose a novel context-based variant aggregation approach for extracting signals from rare variants detected through germline whole-exome and whole-genome sequencing. Many studies have shown that the distributions of the genomic, nucleotide, and epigenetic contexts of somatic variants in tumors are informative of cancer etiology and site of origin (Alexandrov et al. 2014, Alexandrov et al. 2020, Chakraborty et al. 2021). Recently, a new direction of research has focused on extracting signals from the contexts of germline variants (Seplyarskiy et al. 2020, Xu et al. 2020) and evidence has emerged that patterns defined by the nucleotide contexts of germline variants are associated with oncogenic pathways, histological subtypes, and prognosis (Xu et al. 2020). It remains an open question whether aggregating germline variants based on these contexts can improve cancer risk prediction. Using germline whole-exome sequencing data from over 200,000 individuals from the UK Biobank and whole-genome sequencing data from over 2,000 individuals from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, we investigate the predictive value of meta-features aggregating rare variants based on their genomic, nucleotide, and epigenetic contexts for distinguishing cancer cases from controls and for predicting tumor subtypes (such as homologous recombination deficiency). We compare the performance of risk models based on known risk variants and risk models that additionally include the meta-features.

58

Evaluation of Tools for GWAS of Binary Traits in Correlated Data

Anastasia Gurinovich[1*], Mengze Li[2], Anastasia Leshchyk[2], Harold Bae[3], Zeyuan Song[4], Thomas T. Perls[5], Paola Sebastiani[1]
[1]*Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, United States of America;* [2]*Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America;* [3]*Biostatistics Program, College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon, United States of America;* [4]*Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America;* [5]*Department of Medicine, Geriatrics Section, Boston University School of Medicine, Boston, Massachusetts, United States of America*

Performing a genome-wide association study (GWAS) with a binary phenotype using family data is a challenging task. Using linear mixed effects models is typically unsuitable for binary traits, and numerical approximations of the likelihood function may not work well with rare genetic variants with small counts. Additionally, imbalance in the case-control ratios poses challenges as traditional statistical methods such as the Score test or Wald test perform poorly in this setting. In the last couple of years, several methods have been proposed to better approximate the likelihood function of a mixed effects logistic regression model that uses Saddle Point Approximation (SPA). SPA is implemented in both GENetic EStimation and Inference in Structured samples (GENESIS) and Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) software, two increasingly popular tools that were developed to perform GWAS of binary traits in family data. A few studies have examined the performance of these tools in simulated data. We compared Score and SPA tests as implemented in GENESIS and SAIGE using real datasets to evaluate computational efficiency and the agreement of the results. We used the New England Centenarian Study imputed genotype data and the binary phenotype of human extreme longevity to compare the agreement of the results and tools' computational performance. The evaluation suggests that SAIGE and GENESIS produce similar, but not equivalent, results, with SPA adjustment of the Score test with full genetic relationship matrix performing well in small real correlated data, and SAIGE is more computationally efficient than GENESIS.

59

Genetic Adaptation to Climate in Human Populations

Elena S. Gusareva[1,2,3], Vladimir N. Kharkov[3,4], Tatiana Tatarinova[3,5], Ghosh Amit Gourav[1,3], Namrata Kalsi[1,3], Aleksei A. Zarubin[3,4], Daniela I. Drautz-Moses[2,3], Stephan C. Schuster[2,3], Vadim A. Stepanov[3,4], and Hie Lim Kim[1,2,3]
[1]*The Asian School of the Environment, Nanyang Technological University, Singapore;* [2]*Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore;* [3]*GenomeAsia 100 consortium;* [4]*Research Institute of Medical Genetics, Tomsk National Research Medical Center, Russian Academy of Sciences, Tomsk, Russian Federation;* [5]*University of La Verne, La Verne, California, United States of America*

The adaptation pressure shapes the genomic landscape and leave signatures in genomes that are evident in patterns of allelic frequencies associated with human physiological traits. Cold is one of the important unavoidable stressors for human populations of the Arctic. We performed genome-wide screening for footprints of positive selection in Far Eastern Eskimo, Chukchi, and Koryaks adopted to live in the cold polar climate and specific food system with low carbohydrate intake. We used the cross-population extended haplotype homozygosity (XP-EHH) analysis to compare the Far Easterners

with reference populations from East Siberia (Buriats, Evens, and Evenks) and East Asia (Japanese). Assuming the significance level of 10^{-6}, the cut-off for the value of std. XP-EHH was set to |5|. The strongest selection signals both in Koryaks and Eskimo/Chukchi groups (std. XP-EHH = 7.7 and 7.4, resp.) were identified in the locus of the carnitine palmitoyltransferase 1 A (*CPT1A*) gene. This selection signal was also confirmed by the Integrated Haplotype Score (iHS) and integrated Selection of Allele Favoured by Evolution (iSAFE) analyses. The CPT1A is involved in lipids metabolism and is a key regulatory enzyme that imports long-chain fatty acids into the mitochondria for fatty acid oxidation. We identified a specific mutation Pro479Leu in this gene that was enriched in Koryaks, Eskimo, and Chukchi, while been absent in reference populations. This variant is known to be deleterious in many populations, however, it seems to be benign in Arctic populations in which it likely helps the carriers to maintain body heat by keeping certain fats unmetabolized.

60

Proteomics Biomarker Discovery for Individualized Prevention of Familial Pancreatic Cancer Using Statistical Learning

Chung Shing Rex Ha[1,2,6*], Martina Müller-Nurasyid[1,2,3,7], Agnese Petrera[4], Stefanie M. Hauck[4], Detlef K. Bartsch[5], Emily P. Slater[5], Konstantin Strauch[1,2,6]

[1]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany; [2]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [3]Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Munich, Germany; [4]Research Unit Protein Science and Metabolomics and Proteomics Core Facility, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [5]Department of Visceral-, Thoracic- and Vascular Surgery, Philipps University, Marburg, Germany; [6]Chair of Genetic Epidemiology, Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Munich, Germany; [7]Pettenkofer School of Public Health Munich, Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Munich, Germany

Pancreatic ductal adenocarcinoma (PDAC) is a tumour entity with a dismal prognosis. Familial pancreatic cancer (FPC) refers to families with an inherited predisposition and hence higher risk to develop PDAC. These families are characterized by two or more first-degree relatives with PDAC that do not fulfil the criteria for another inherited tumor syndrome. In this study, we focus on discovering potential biomarkers to improve the current diagnostic tools in established FPC screening procedures. To this end, we applied high-throughput proteomics to obtain comprehensive serum protein profiles for individuals at risk from the German National Case Collection of Familial Pancreatic Cancer (FaPaCa) in different potential pre-cancer stages.

Data analysis in this study encountered two major difficulties: a small sample size and an unbalanced data structure. High dimensional proteomics data and the above-mentioned difficulties challenge traditional statistical analysis tools. Therefore, we applied advanced statistical learning methods to enhance the quality of statistical analysis and the interpretability of the results. First, we performed variable selection and model fitting via model-based gradient boosting. Next, we applied stability selection to discover stable subsets of selected biomarkers.

We were able to show that model-based gradient boosting could handle the unbalanced data structure in a high-risk screening program. We identified a relevant subset of biomarkers in the context of high-throughput proteomics with good prediction performance. Stability selection helped us further sharpen the sets of stable biomarkers. Our discoveries can potentially improve the current screening procedure of FPC and lead to optimized strategies for its early detection.

61

Unsupervised Outlier Detection Applied to SARS-CoV-2 Nucleotide Sequences Identifies Sequences of the Omicron Variant and Other Variants of Interest

Georg Hahn, Christoph Lange

Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

As of January 2022, the GISAID database contains more than one million SARS-CoV-2 genomes, including more than ten thousand nucleotide sequences of the recently discovered omicron variant. These SARS-CoV-2 strains have been collected from patients around the world since the beginning of the pandemic. We start by assessing the similarity of all pairs of nucleotide sequences using the Jaccard index and principal component analysis. As shown previously in the literature, an unsupervised cluster analysis applied to the SARS-CoV-2 genomes results in clusters of sequences according to certain characteristics such as their strain or their clade. Importantly, we observe that nucleotide sequences of the omicron variant are outliers in clusters of sequences stemming from variants identified earlier on during the pandemic. Motivated by this finding, we are interested in applying outlier detection to nucleotide sequences, and demonstrate that nucleotide sequences of the omicron variant can be identified solely based on a statistical outlier criterion. We argue that outlier detection might be a useful tool to identify emerging variants in real time as the pandemic progresses.

62

Polyexposure Risk Score Offers Greater Predictive Performance for Chronic Obstructive Pulmonary Disease than Polygenic Risk Score and Smoking Alone

Yixuan He[1,2*], Chirag Patel[1]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, United States of America; [1]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, Massachusetts, United States of America

Background/Objective: Chronic obstructive pulmonary disease (COPD) is associated with a combination of genetic and environmental risk factors. While the polygenic risk score (PGS), a weighted sum of genetic variants, exists for COPD, the collective effect of multiple exposures is unknown.

Methods: From unrelated individuals of white European ancestry in the UK Biobank, we identified 269506 individuals with no prior diagnosis of COPD and FEV1/FVC≥0.7 at baseline. We used machine learning feature selection methods to extract non-redundant and independent exposures from 109 exposure and lifestyle factors for the PXS. We adjusted for PRS, constructed from previously published weights of approximately 2.5 million genetic variants.

Results: The PXS for COPD consists of eight exposures (employment, qualifications, physical activity, homeownership, TV, alcohol, and smoking status) and has greater prediction for COPD incidence (C index: 0.782, 95% CI 0.766-0.798) than PGS (0.712, 95% CI 0.696-0.728) alone or PGS and smoking (0.746, 95% CI 0.730-0.762). After adjusting for PRS, TV screentime was no longer selected in the PXS. Compared to the remaining population, individuals in the top 5% of PXS and PRS had hazard ratios of 4.95 (95% CI 4.19-5.84, p<0.001) and 1.47 (95% CI 1.12-1.93, p<0.001) for COPD, respectively.

Conclusion: PXS predicts incident COPD better than PGS or smoking alone. PXS is independent of PGS and identifies individuals with the greatest risk for incident disease.

63

Genetic Links Between Cardiometabolic Traits and Risk for Uterine Fibroids

Joséphine Henry[1*], Takiy Berrandou[1,2], Lizzy M. Brewster[3,4], Nabila Bouatia-Naji[1]

[1]PARCC, INSERM, Université de Paris, Paris, France; [2]Quantitative Genetics and Genomics (QGG), Aarhus University, Denmark; [3]Amsterdam Institute for Global Health and Development (AIGHD), The Netherlands; [4]CK Foundation, Amsterdam, The Netherlands

Uterine Fibroids (UFs) are common neoplasms in the uterus, mainly affecting women of reproductive age. UFs can cause abnormal uterine bleeding and other symptoms, including outside the uterus. UFs generally result in hysterectomy for most severe cases and clinical series suggest impaired cardiometabolic features in patients. This study aims to uncover genetic links between cardiometabolic traits and risk for UFs.

We used existing GWAS summary statistics from the UK Biobank including 5,514 UFs cases and 188,639 controls, along with traits mainly related to blood pressure, adiposity, and lipids. We used LD score regression to estimate the genetic correlations between UFs and cardiometabolic traits, GCTA-mtCOJO to perform conditioned correlations, and Mendelian Randomization (MR) to determine potential causal associations between traits and risk for UFs.

We found a significant positive genetic correlation between UFs and systolic blood pressure (SBP, r_g=0.13, P=4×10^{-5}). We also found significant positive correlations with body mass index (BMI, r_g=0.21, P=1×10^{-8}), triglycerides (r_g=0.27, P=2×10^{-11}), and hemoglobin levels (r_g=0;16, P=1×10^{-5}). Conditioning on SBP or BMI genetics marginally influenced the correlations with lipids and hemoglobin levels. MR analyses indicated that genetically determined higher SBP did not associate with increasing risk of UFs (P=0.23). However, genetically predicted higher BMI (P=1.4×10^{-3}), triglycerides (P=1.1×10^{-5}) and hemoglobin levels (P=1.3×10^{-2}) associated with increased risk for UFs, suggesting them as potential genetic risk factors.

We report genetic links between UFs and SBP, BMI and triglycerides. Ongoing work includes confirming these results using a larger GWAS meta-analysis for UFs, exploring the genetic link locally in specific shared loci.

64

Epigenome-wide Association Study of Pediatric Asthma in Latinos

Esther Herrera-Luis[1*], Carlos de la Rosa-Baez[1], Celeste Eng[2], Kenneth B. Beckman[3], Luisa N. Borell[4], Esteban G. Burchard[2,5], María Pino-Yanes[1,6,7]

[1]Genomics and Health Group, Department of Biochemistry, Microbiology, Cell Biology and Genetics, Universidad de La Laguna (ULL), San Cristóbal de La Laguna, Tenerife, Spain. [2]Department of Medicine, University of California San Francisco, San Francisco, California, United States of America; [3]University of Minnesota (UMN) Genomics Center, Minneapolis, Minnesota, United States of America; [4]Department of Epidemiology & Biostatistics, Graduate School of Public Health & Health Policy, City University of New York, New York, New York, United States of America; [5]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, United States of America; [6]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [7]Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna (ULL), Santa Cruz de Tenerife, Spain

The epigenetic mechanisms of asthma remain largely understudied in Latinos, a population disproportionately affected by asthma. Here, we aimed to identify epigenetic loci and processes associated with pediatric asthma in Latinos. An epigenome-wide association study of asthma was performed in 765 children (410 with asthma and 356 without asthma) from the Genes-Environments and Admixture in Latino Asthmatics (GALA II). DNA methylation levels in whole-blood were profiled with the Infinium EPIC or the HumanMethylation450 BeadChip arrays (Illumina, San Diego, CA, United States of America). The association between 824,938 CpG probes and asthma was evaluated using linear regression models adjusted for age, sex, genetic ancestry, *in utero* exposure to maternal smoking, and cell-type heterogeneity. Probes that exceeded the significance threshold (Q value≤0.05) were evaluated for enrichment analysis in gene ontology terms and previous epigenetic signals. A total of 127 probes were significantly associated with asthma. The most significant association was located at cg04873169 (*GRAMD4*, P value=2.05x10^{-8}), which has been previously associated with asthma in airway epithelial cells

from European adults. Moreover, significant probes were enriched in biological processes related to the regulation of immune response, response to stimulus, and cell activation (*Q* value≤0.05), and in previous epigenetic associations, including fractional exhaled nitric oxide, nitrogen dioxide, smoking, asthma, allergic sensitization, and atopy (*P* value ≤0.05). Our findings from Latinos highlight the transferability of asthma-related epigenetic markers across tissues and the epigenetic overlap with similar traits.

Supported by NIH R01HL155024, R01MD010443, R56MD013312 and MCIN/AEI/10.13039/501100011033 PID2020-116274RB-I00.

65

Constructing a SURrogate-Family Based Association Test (SURFBAT) with Genotype Imputation Algorithms

Anthony F. Herzig[1*], Gaëlle Marenne[1], Hervé Perdry[2], FranceGenRef Consortium[3], Emmanuelle Génin[1,4]
[1]Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest, France; [2]Université Paris-Saclay, Université Paris-Sud, Inserm, CESP, Villejuif, France; [3]LABEX GENMED, Centre National de Recherche en Génomique Humaine, Evry, Paris; [4]Centre Hospitalier Régional et Universitaire de Brest, France

Genotype-phenotype association tests are typically adjusted for population stratification using principal components that are estimated genome-wide. This lacks resolution when analyzing populations with fine structure and/or individuals with fine levels of admixture. This can affect power and precision, and is a particularly relevant consideration when control individuals are recruited using geographical selection criteria. Such is the case in France where we have recently created reference panels of individuals anchored to different geographical regions. To make correct tests against case groups, that would likely be gathered from large urban areas, new methods are needed.

We present SURFBAT (a SURrogate Family Based Association Test) which performs an approximation of the transmission-disequilibrium test. Our method hinges on the application of genotype imputation algorithms to match similar haplotypes between the case and control groups. This permits us to estimate local ancestry informed posterior probabilities of the pseudo-untransmitted parental alleles of each case individual. The method is suitable when the control panel spans the local ancestry spectrum of the case-group population and each control has similar paternal and maternal ancestries. This is the case in our reference panels where individuals have their four grandparents born in the same geographic area.

SURFBAT provides an association test that is inherently robust to fine-scale population stratification. We demonstrate the interest of our tool on simulated datasets created from the 1000 Genomes Project and the FranceGenRef project.

66

Linear Regression on Martingale Residuals Enables Fast and Accurate Recurrent Event Analysis for Genome-wide Association Studies

Jasper P. Hof[1]*, Sita H. Vermeulen[1], Anthony C.C. Coolen[2], Tessel E. Galesloot[1]
[1]Department for Health Evidence, Radboud university medical center, Nijmegen, The Netherlands; [2]Department of Biophysics, Radboud University, Nijmegen, The Netherlands

Many diseases are characterized by recurrences after recovery, e.g. recurrences in cancer and infections. The Cox proportional hazards frailty model is the current state-of-the-art model for recurrent event analysis, however this model is too statistically complex for efficient application in genome-wide association studies (GWAS). Here, we developed a novel method for recurrent events analysis in GWAS.

In our method, every DNA variant is tested univariably for association with recurrences using a linear regression on martingale residuals (LRM). The martingale residuals are computed once from a null model and subsequently a saddle-point approximation is implemented to achieve accurate statistical inference for low-frequency variants. We compared the statistical performance (type I error, power, run time) of LRM with established recurrent event models over a range of relevant and realistic parameters for recurrent event GWASs. In the next months, we will apply LRM to our bladder cancer recurrence dataset (N=1,443) to demonstrate the applicability of our method.

Our simulations showed that the novel method controls the type I error and that the statistical power is similar to state-of-the-art recurrent event models in all simulation scenarios. Additionally, LRM is significantly faster than existing recurrent event models: a recurrent event GWAS for 1 million DNA variants and 1,500 individuals requires 2-3 minutes using LRM, whereas existing recurrent event methods based on a Wald test would require 3-8 days, depending on recurrence rate. Thus, LRM enables a fast and accurate recurrent event analysis on genome-wide scale.

67

The Robustness of Bayesian Network Analysis with Respect to Data Measurement Error

Richard A. J. Howey, Heather J. Cordell
Population Health Sciences Institute, Newcastle University, United Kingdom

Bayesian networks have been proposed to identify possible causal relationships between measured variables based on their conditional dependencies and independencies, especially in complex scenarios with many variables. The quality of the data can have an impact on the accuracy of any best fit Bayesian network. In previous work we have considered data sets with missing values and now we consider complete data sets with measurement error.

Measurement error is the deviation of the recorded data from the true data values, which can manifest itself for a number of reasons, but here we consider when recorded data

is subjected to random (normally distributed) error, or "noise." This is perhaps the most typical type of measurement error expected.

We show that measurement error can lead to a decrease in the quality of the fitted Bayesian network, and that the impact on edge detection varies somewhat unpredictably, dependent upon the nearby true network structure.

Our software, BayesNetty, has been adapted to add options to introduce measurement error, and we use it to vary the amount of measurement error in simulated data. The software is designed to be fast, easy to use and practical for large data sets, especially those containing genetic data. As with any statistical method, Bayesian network analysis is dependent on the quality of data, emphasising that data should be measured as accurately as possible and the likelihood of measurement error should be taken into account when interpreting results.

68

Evaluation of Network-guided Random Forest for Disease Gene Discovery

Jianchang Hu*, Silke Szymczak
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Schleswig-Holstein, Germany

Identification of biomarkers associated with complex diseases can improve patient risk prediction and foster understanding of underlying molecular pathomechanisms. However, due to the functional interdependencies between molecular components, a complex disease such as cancer rarely occurs because of an abnormality of a single gene. Network information is believed to be beneficial for disease module and pathway identification. In this simulation study, we investigate the performance of a network-guided random forest (RF) where the network information is summarized into a sampling probability of predictor variables which is further used in the construction of RF. The identification of important genes is based on standard variable importance measures from RF. For comparison, we also consider several different constructions of this sampling probability including uniform probability and marginal association test based construction. We simulate synthetic RNASeq data along with the underlying network structure using the R package SeqNet. Performance of disease gene identification and model prediction accuracy is investigated. Our results suggest that network-guided RF tends to select hub nodes more frequently in all scenarios. When causal genes are randomly distributed within the network, network information only deteriorates the gene selection, but if they form a module, network-guided RF identifies causal genes more accurately. When effect sizes of causal genes are large, network-guided RF does not show significant improvements on prediction accuracy over standard RF. More simulation scenarios including various effect sizes and different topological structures of causal genes are under investigation, and complete results will be presented during the conference.

69

A Fast Bayesian Screen to Identify Pleiotropic Loci and Describe Pleiotropic Profiles

Sijia Huo[1]*, Sara Lindström[2,3], Lu Wang[4], Peter Kraft[1,5]
[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Department of Epidemiology, University of Washington, Seattle, Washington, United States of America; [3]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [4]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, Washington, United States of America; [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

Pleiotropy occurs when a genetic variant is associated with more than one trait. Multi-trait test statistics have been proposed that leverage pleiotropy to increase power identifying trait associated loci. However, these methods only test the global null that none of the tested traits are associated with a variant but do not provide any information regarding whether the variant is associated with more than one trait, or with which traits the variant is associated. In this paper, we propose a new fast screening approach based on the Bayesian support region to overcome these restrictions. Our approach accounts for correlation among test statistics due to sample overlap and leverages cross-trait heritability. Computation scales linearly in the number of traits. We compare our approach to widely used alternatives (including Bonferroni correction for the number of traits and ASSET) via simulation. Our approach shows both high sensitivity and high specificity and outperforms the alternatives under most scenarios. For example, our approach can correctly detect up to 67.8% more pleiotropic SNPs than Bonferroni correction. We applied our approach to GWAS summary statistics from 12 different cancers and identified 82 independent regions exhibiting pleiotropy, including TERT and ABO, each associated with three cancers. We hope that this new method can facilitate biological discoveries in the future.

70

Association of Genetic Loci for Human Plasma Proteins with Response to Treatment in People with Rheumatoid Arthritis

Andrii Iakovliev[1*], Stephanie F. Ling[2,3], Marco Colombo[4], Darren Plant[2], Myles Lewis[5], Costantino Pitzalis[5], Anne Barton[2,3], Paul McKeigue[1], Athina Spiliopoulou[1,6]
[1]Usher institute of Population Health sciences and informatics, University of Edinburgh, United Kingdom; [2]Versus Arthritis Centre for Genetics and Genomics, Division of Musculoskeletal Sciences, The University of Manchester, United Kingdom; [3]National Institute for Health Research Manchester Biomedical Research Centre, Manchester University National Health Service Foundation Trust, Manchester Academic Health Sciences Centre, United Kingdom; [4]Centre for Paediatric Research, University of Leipzig, Germany; [5]Centre for Experimental Medicine & Rheumatology, Queen Mary University of London, London, United Kingdom; [6]Institute of Genetics and Cancer, University of Edinburgh, United Kingdom

We used locus-specific genotypic scores for human plasma proteins (pQTLs), gene expression (eQTLs), and gene methylation (mQTLs) to detect effects of these intermediate variables on response to Tumor Necrosis Factor inhibitors (TNFi) in over 3,000 people with Rheumatoid Arthritis (RA). Change in erythrocyte sedimentation rate (**Δ**ESR) and in Swollen 28-Joint Count (**Δ**SJC) were used to quantify TNFi response. Associations were adjusted for false discovery rate. Significant loci were further explored by Mendelian randomisation (MR) and colocalization analyses.

One *cis* pQTL score on chromosome 10 for the *ENTPD1* protein and five *trans* pQTL scores on chromosome 3 for the *ARHGAP30*, *APOM*, *EHMT2*, *BGN*, and *FURIN* proteins were associated with TNFi response. Validation using SWATH-MS protein expression data in 180 people confirmed that *APOM* expression was significantly associated with TNFi response.

Signals on chromosome 3 occurred within butyrylcholinesterase (*BCHE*) locus. The eQTL and mQTL *cis* scores for the *BCHE* gene were associated with TNFi response ($P < 0.05$) and the genetic signals colocalized with the *trans* pQTLs suggesting common causal variants. MR analysis using two independent instruments for the butyrylcholinesterase enzyme indicated a positive causal effect on TNFi response ($\beta = 0.13$, $P = 0.033$).

We have detected a strong genetic signal for TNFi response at the *BCHE* locus and provided evidence of a weak causal effect of the butyrylcholinesterase enzyme. *BCHE* has not been previously associated with TNFi response. The apparent pleiotropy in this locus, indicated by five *trans* pQTLs makes it difficult to identify the most likely causal pathway.

71

Multi-ethnic Polygenic Risk Scores for Venous Thromboembolism

Yon Ho Jee[1*], Florian Thibord[2,3], Christopher Kabrhel[4-6], Nicholas Smith[7-9], Peter Kraft[1,10]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, Maryland, United States of America; [3]National Heart Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, Massachusetts, United States of America; [4]Center for Vascular Emergencies, Department of Emergency Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [5]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [6]Department of Emergency Medicine, Harvard Medical School, Boston, Massachusetts, United States of America [7]Department of Epidemiology, University of Washington, Seattle, Washington, United States of America; [8]Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, Washington, United States of America; [9]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, Washington, United States of America [10]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

Venous thromboembolism (VTE) is a significant contributor to morbidity and mortality, with large disparities in incidence rates across ancestry populations. Polygenic risk scores (PRSs) comprised of genome-wide significant variants have been demonstrated to identify European individuals at the highest risk of VTE. However, there is limited evidence on whether high-dimensional PRS constructed using more sophisticated methods can enhance the predictive ability and their utility in populations of non-European ancestry. We developed PRSs for VTE using summary statistics from the International Network on Venous Thrombosis consortium GWAS meta-analyses of European (71,771 cases and 1,059,740 controls) and African ancestry samples (7,482 cases and 129,975 controls). We used LDPred2, stacked clumping and thresholding, and PRS-CSx to construct PRS and evaluated the performance of these PRSs in a European ancestry population (2,222 cases and 2,201 controls). LDpred2 trained using European ancestry summary statistics performed the best with OR of 1.51 (95% CI 1.38-1.67) and area under the curve (AUC) of 0.62 (0.60-0.65). In this European ancestry test set, combined PRS-CSx of European and African ancestry population (AUC=0.60, 0.58-0.63) did not perform any better than PRS-CSx of European alone (AUC=0.60, 0.58-0.63) or African ancestry alone (AUC= 0.58, 0.55-0.60). The highest fifth percentile of the LDpred2 distribution was associated with twofold increased risk for VTE (OR=2.04, 1.73-2.40). These findings suggest that PRS may be used to identify individuals at highest risk for VTE event and provide guidance for the most effective treatment strategy. We are validating these PRSs in African-ancestry population.

72

GWAS of Longitudinal Trajectories at Biobank Scale

Seyoon Ko[1,2], Christopher A. German[2], Aubrey E. Jensen[2*], Judong Shen[3], Anran Wang[3], Devan V. Mehrotra[3], Yan V. Sun[4], Janet S. Sinsheimer[1,2,5], Hua Zhou[1,2], Jin J. Zhou[2,6,7]

[1]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, California, United States of America; [2]Department of Biostatistics, University of California, Los Angeles, Los Angeles, California, United States of America; [3]Biostatistics and Research Decision Sciences, Merck & Co., Inc., Kenilworth, New Jersey, United States of America; [4]Department of Epidemiology, Emory University, Atlanta, Georgia, United States of America; [5]Department of Human Genetics, University of California, Los Angeles, Los Angeles, California, United States of America; [6]Department of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America; [7]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, Arizona, United States of America

Biobanks linked to massive, longitudinal electronic health record (EHR) data make numerous new genetic research questions feasible, including the study of biomarker trajectories. For example, high blood pressure measurements over multiple visits strongly predict stroke onset, and consistently high fasting glucose and Hb1Ac levels define

diabetes. Recent research reveals that not only the mean level of biomarker trajectories but also their fluctuations, or within-subject (WS) variability, are risk factors for many diseases. Glycemic variation, for instance, is considered an important clinical metric in diabetes management. It is crucial to identify genetic factors that shift the mean or alter the WS variability of a biomarker trajectory. Compared to traditional cross-sectional studies, trajectory analysis utilizes more data points and captures a complete picture of the impact of time-varying factors, including medication history and lifestyle. Currently, there are no efficient tools for genome-wide association studies (GWASs) of biomarker trajectories at the biobank scale, even for mean effects. We propose TrajGWAS, a linear mixed effect model-based method for testing genetic effects that shift the mean or alter the WS variability of biomarker trajectories. It is scalable to biobank data with 100,000 to 1,000,000 individuals and many longitudinal measurements, and robust to distributional assumptions. Simulation studies corroborate that TrajGWAS controls the type I error rate and is powerful. Analysis of 11 biomarkers measured longitudinally and extracted from UK Biobank primary care data for more than 150,000 participants with 1,800,000 observations reveals loci that significantly alter the mean or WS variability.

73

Body Mass Index and Incidence of Lung Cancer in The HUNT Study: Using Observational and Mendelian Randomization Approaches

Lin Jiang[1]*, Yi-Qian Sun[2,3,4], Ben M. Brumpton[5,6,7], Arnulf Langhammer[7,8], Yue Chen[9], Xiao-Mei Mai[1]

[1]Department of Public Health and Nursing, Faculty of Medicine and Health Science, Norwegian University of Science and Technology, Trondheim, Norway; [2]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Science, Norwegian University of Science and Technology, Trondheim, Norway; [3]Department of Pathology, Clinic of Laboratory Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway; [4]TkMidt-Center for Oral Health Services and Research, Mid-Norway, Trondheim, Norway; [5]Clinic of Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway; [6]K.G. Jebsen Centre for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Norway; [7]HUNT Research Centre, Department of Public Health and Nursing, Norwegian University of Science and Technology, Levanger, Norway; [8]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway; [9]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada

Background: Observational studies have shown an inverse association between body mass index (BMI) and lung cancer risk. Mendelian randomization (MR) analysis using genetic variants as instruments for BMI may clarify the nature of the association.

Aims: We studied the causal association between BMI and lung cancer incidence using observational and MR approaches.

Methods: We followed up 62,453 cancer-free Norwegian adults from 1995–97 (HUNT2) until 2017. BMI in HUNT2 was classified as <25.0, 25.0–29.9 and ≥30.0 kg/m^2. Seventy-five genetic variants were included as instruments for BMI (among which 14 also associated with smoking behavior). Incident lung cancer cases were ascertained from the Cancer Registry of Norway. Cox regression models were used to estimate HRs with 95% CIs. Multivariable MR was used to examine the effect of BMI after genetically controlling for smoking.

Results: During a median follow-up of 21.1 years, 1009 participants developed lung cancer (327 adenocarcinoma). The HRs and 95% CIs for incidence of adenocarcinoma were 0.73 (0.58–0.92) for BMI 25.0–29.9 kg/m^2 and 0.53 (0.37–0.76) for BMI ≥ 30 kg/m^2 compared with BMI <25.0 kg/m^2 (P for trend <0.001). However, multivariable MR suggested a positive association between genetically determined 1 kg/m^2 increase in BMI and the incidence of adenocarcinoma (HR 1.28, 95% CI 1.03–1.58). No associations were found with other lung cancer subtypes.

Conclusions: Our study suggests that the inverse association between BMI and adenocarcinoma in observational analysis may not be causal. More MR studies are needed to conform our positive finding.

74

Association and Performance of Polygenic Risk Scores for Breast Cancer Among French Women Presenting or Not a Hereditary Predisposition to the Disease

Yue Jiao[1] *, Thérèse Truong[2], Séverine Eon-Marchais[1], Anne Boland-Augé[3], Jean-François Deleuze[3], Pascal Guénel[2], Dominique Stoppa-Lyonnet[4], Nadine Andrieu[1], Fabienne Lesueur[1]

[1]Inserm, U900, Team Genetic Epidemiology of Cancers, Institut Curie, PSL Research University, Mines ParisTech, Paris, France; [2]Paris-Saclay University, UVSQ, Gustave Roussy, Inserm, CESP, Team Exposome and Heredity, Villejuif, France; [3]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, Evry, France; [4]Department of Genetics, Institut Curie, Paris University, Paris, France

Three breast cancer polygenic risk scores (PRS) comprising 77, 179 and 313 SNPs (with partially overlapping SNPs) have been proposed for European-ancestry women by the Breast Cancer Association Consortium (BCAC) for predicting risk in the general population. However, even within this population, the effect of these SNPs may vary from one country to another because of other factors, which may modify performance of PRS. We assessed risk and predictive performance associated to these PRS in French women from (1) the CECILE population-based study (N_{cases}=1015, $N_{controls}$=996), (2) *BRCA1* or *BRCA2* (*BRCA1/2*) pathogenic variant (PV) carriers from the GEMO study (N=2639), and (3) familial breast cancer cases with no *BRCA1*/2 PV (N=1257) and controls (N=1266) from the GENESIS study. We showed that the three PRS were significantly associated with breast cancer in all studies, with OR per SD varying from 1.7 to 2.0 in CECILE and GENESIS, and hazard ratios varying from 1.1 to 1.4 in GEMO. The predictive performance of PRS_{313} in CECILE was similar

to that reported in BCAC and lower to that in GENESIS (area under the receiver operating characteristic curve (AUC)=0.67 and 0.75, respectively). These PRS were less performant in *BRCA2* and *BRCA1* PV carriers (AUC 0.58 and 0.54 respectively). Additionally, breast cancer risk associated with PRS may vary according to the presence of a predicted PV in a moderate-risk gene like *ATM*, *CHEK2* or *PALB2*. Work to develop more specific PRS is underway to envisage better risk stratification of women to improve screening and clinical management.

75

Multi-omics Predictive Model for Asthma-related Phenotypes

Heejin Jin[1*], Sungho Won[1]
[1]*Department of Public Health Science, Seoul National University, Seoul, Korea*

Large-scale omics datasets have provided a comprehensive biological understanding of disease. However, there have been few studies of predicting diseases using integration of multi-omics datasets. The aim of this study is to evaluate the relative importance of each omics (genome, transcriptome, proteome, metabolome, and epigenome) and build a multi-omics predictive model for asthma-related phenotypes. All omics were collected from asthmatic patients and individuals were slightly different for each omics, of which 282 had all types of omics. To avoid overestimation, feature selection of each omics was conducted using elastic-net regression with non-overlapping samples, and all selected markers were used to estimate the relative importance of each omics using McFadden's R-square () with overlapping samples. Prediction models were built with all omics and clinical variables, and we focused on phenotype with high balanced accuracy (BA). The final predictive model was built using a stepwise multivariable logistic regression and area under the curve (AUC) was estimated to evaluate the performance of the model. We have found that the classification accuracy of the blood eosinophil group (greater or less than 300 cells/μL) was very high (BA 0.964, *P* 7.70) when all omics and clinical variables were in the prediction model. Here, the proteome (51.18%) accounted for the largest portion of the variances, followed by clinical variables (24.15%), epigenome (12.81%), transcriptome (8.48%), metabolome (2.49%), and genome (1.89%). The AUC of final predictive model with proteome and metabolome was 0.977, which was very higher than that of the model with a single omics marker.

76

Random Glucose GWAS Trans-ethnic Meta-analysis Provides Insights into Diabetes Pathophysiology, Complications, and Treatment Stratification

Vasiliki Lagou[1], Longda Jiang[2], Liudmila Zudina[3], Zhanna Balkhiyarova[3], Jared Maina[4], Ayse Demirkan[3], Marika A. Kaakinen[3,5*], Ben Jones[5], Inga Prokopenko[3, 4], The Magic Investigators[6]
[1]*Wellcome Sanger Institute, Hinxton, United Kingdom;* [2]*The University of Queensland, Brisbane, Australia;* [3]*University of Surrey, Clinical and Experimental Medicine, Guildford, United Kingdom;* [4]*Institut Pasteur de Lille, CNRS, University of Lille, UMR 8199 - EGID, Lille, France;* [5]*Imperial College London, Section of Endocrinology and Investigative Medicine, London, United Kingdom;* [6]*Meta-Analysis of Glucose and Insulin-related Traits Consortium, https://magicinvestigators.org, United Kingdom*

Introduction: Conventional measurements of fasting/postprandial blood glucose levels investigated in genome-wide association studies (GWAS) cannot capture the effects of DNA variability on "around the clock" glucoregulatory processes. We performed GWAS meta-analysis of glucose measurements under non-standardized conditions (random glucose; RG) in 493,036 individuals of diverse ethnicities and without diabetes, enabling powerful locus discovery and innovative pathophysiological observations.

Methods: We dissected associations between HRC-imputed SNPs and RG, adjusted for age, sex, population structure, time since last meal (where available) in 17 studies, including UK Biobank. We investigated genetic (LD score regression/PRSs/hierarchical clustering) and causal (MR-Base) relationships with other phenotypes, and gene expression (metaXscan, DEPICT).

Results: We discovered 142 RG loci (185 distinct signals), including 84 novel signals for glycaemia, 14 with sex-dimorphic effects, 9 identified through trans-ethnic analysis and 25 rare/low-frequency signals. Regulatory, glycosylation, and metagenomic annotations highlighted ileum and colon tissues, indicating an underappreciated role of gastrointestinal tract in the control of blood glucose. Functional follow-up and molecular dynamics simulations of low-frequency coding variants in *GLP1R*, a type 2 diabetes (T2D) treatment target, revealed that optimal selection of GLP-1R agonist therapy in the clinic will benefit from a tailored genetic stratification. We provide novel compelling evidence from Mendelian randomization that lung function is modulated by blood glucose levels ($\beta_{MR\text{-}RG}$=-0.61, P=3.5x10^{-4}; $\beta_{MR\text{-}T2D}$=-0.062, P=1.42x10^{-21}), and settle the longstanding controversy that pulmonary dysfunction is a diabetes complication.

Conclusion: Our investigation yields wide-ranging insights into the biology of glucose regulation, diabetes complications and pathways for treatment stratification.

Funding: H2020-SC1-HBC-28-2019-LONGITOOLS, WCRF-2017/1641, Diabetes UK(BDA number:20/0006307), PreciDIAB(ANR-18-IBHU-0001).

77

Plasma Circulating MicroRNA Signature of Alcohol Consumption: The Rotterdam Study

Irma Karabegović[1*], Yasir Abozaid[1], Silvana C.E. Maas[1, 2], Jeremy Labrecque[1], Daniel Bos[1, 3], Robert J. De Knegt[4], M. Arfan Ikram[1], Trudy Voortman[1, 5], Mohsen Ghanbari[1]
[1]*Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands;* [2]*Vall d´Hebron Institute of Oncology (VHIO), Barcelona, Spain;* [3]*Department of Radiology and Nuclear Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands;* [4]*Department of Gastroenterology, Erasmus University Medical Center, Rotterdam, The Netherlands;* [5]*Division of Human Nutrition and Health, Wageningen University*

& Research, Wageningen, The Netherlands

Background: MicroRNAs (miRNAs) represent a class of small non-coding RNAs that regulate gene expression post-transcriptionally and are implicated in the pathogenesis of different diseases. Alcohol consumption might affect the expression of miRNAs, which in turn could play a role in the development of complex diseases.

Objective: We investigated the association between alcohol consumption and circulating miRNAs and explored the potential mediatory role of the identified miRNAs in the association between alcohol consumption and liver related traits, using data from the population-based Rotterdam Study cohort.

Design: Profiling of plasma circulating miRNAs (n=2083) was conducted using HTG EdgeSeq miRNA Whole Transcriptome Assay in 1933 participants of the Rotterdam Study. Adjusted linear regression was used to examine the association between alcohol consumption (glasses/day) and plasma miRNA levels in a hypothesis free approach. Sensitivity analysis for alcohol categories (non, light and heavy drinkers) was performed. Our secondary analysis explored whether identified miRNAs mediate the association between alcohol consumption and liver related traits.

Results: Plasma levels of four miRNAs (miR-193b-3p, miR-122-5p, miR-3937 and miR-4507) were significantly associated with alcohol consumption surpassing the Bonferroni corrected *P* value< 8.46×10^{-5}. The most prominent association was observed for miR-193b-3p (β=0.087, *P* value=2.90×10^{-5}). A potential mediatory role of miR-3937 and miR-122-5p was observed between alcohol consumption and liver traits.

Conclusions: This study indicates that plasma levels of four miRNAs are associated with alcohol consumption in a population-based setting, among them miR-3937 and miR-122-5p show a mediatory role between alcohol consumption and liver health.

78

Statistical Methods and Approaches for the Analysis of Single Cell Composition Data

Tanya Karagiannis[1], Eric Reed[2], Stefano Monti[3-5], Paola Sebastiani[1,2]

[1]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, United States of America; [2]Data Intensive Study Center, Tufts University, Boston, Massachusetts, United States of America; [3]Division of Computational Biomedicine, Boston University School of Medicine, Boston, Massachusetts, United States of America; [4]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [5]Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America

Changes of cell type composition across samples can have biological significance and provide insight into disease and other conditions. Single cell transcriptomics has made it possible to study cell type composition at a fine resolution. Most single cell studies investigate compositional changes between samples for each cell type independently, not accounting for the fixed number of cells per sample in sequencing data. To account for the constrained proportions of cell types in single cell composition data, we developed the cell type diversity statistic, a method to measure and compare the overall cell type composition of a sample. The diversity statistic is an entropy-based metric that allows for the investigation of global cell type compositional changes across multiple samples and biological conditions at the single cell level. In addition, methods to analyze cell type specific compositional changes based on the total number of cells in samples are emerging. However, this analysis provides limited insight into cell type compositional changes within cell compartments and lineages that may provide more biological meaning to cell type changes. We developed a new approach to characterize cell type compositional changes through conditional analyses within cell compartments based on a data-driven hierarchical approach. We applied these methods to assess compositional changes in peripheral blood mononuclear cells (PBMCs) related to aging and extreme old age using multiple single cell RNA-sequencing (scRNA-seq) datasets from individuals of four age groups across the human lifespan.

79

Prevalence of the *MC4R* Gene Polymorphism in Relation to Obesity and the Metabolic Syndrome in a North Lebanese Population

Racha A. Kerek[1*], Rayan A. El-Achrafi[1], Haneen S. Shami[1]

[1]*Department of Public Health Genetics, Jinan University, Tripoli, Lebanon*

The metabolic syndrome (MetS) refers to the presence of a set of physiological signs, which, when combined, greatly increase the risk of type-II diabetes, cardiovascular diseases, stroke and obesity. The diagnostic criteria are related to central obesity as an essential component, combined to factors like raised cholesterol, triglycerides, and blood pressure. Focusing on weight-gain, we are interested in studying the melanocortin-4 receptor (*MC4R*), a gene regulating appetite and energy homeostasis, and its polymorphism in developing the MetS. Our aims are to study the prevalence of the *MC4R* rs17782313 T>C polymorphism in North Lebanon. We genotyped DNA of 120 unrelated participants using the Restriction Fragment-Length-Polymorphism technique. We also evaluated the participants' anthropometric measurements, physical activity levels using the international physical activity questionnaire, excessive overeating behavior, and lipid blood levels. Results show high prevalence of the polymorphism with an allele frequency of 0.61 for the C allele. Statistical analysis shows that MC4R polymorphism is significantly linked to an increased body-mass index and Waist-hip-ratio, but not with other parameters related to weight gain (body fat percentage and fat mass). While studying the MetS components demonstrate no significant association, but preliminary results show a tendency for a disturbed lipid profile. In conclusion, genetic components should be considered as risk factors to non-communicable diseases in Lebanon.

80

Prevalence of the *TCF7L2* Gene Polymorphism in Relation to Diabetes and the Metabolic Syndrome in the Lebanese Population

Racha A. Kerek[1*], Alaa Harmoush[1]
[1]*Department of Public Health Genetics, Jinan University, Tripoli, Lebanon*

Diabetes is a worldwide public health concern and in Lebanon, it is considered a major burden with 15% of the population affected. The search for risk factors is ongoing as a strategy for disease management. Among these, genetic polymorphisms in the transcription factor 7-like 2 (*TCF7L2*) gene, a key agent in the aetiology and the development of type 2 diabetes are considered. In this study we aimed to assess the prevalence of rs290481 and rs7903146 polymorphisms of *TCF7L2* in the Lebanese population and their possible relationship with diabetes and the metabolic syndrome. For this case-control study, we recruited 100 participants of Lebanese origins from North Lebanon. The DNA extracted was genotyped using the Restriction Fragment Length Polymorphism technique and distributions and associations were analysed using SPSS. The results show a high prevalence of the *TCF7L2* rs7903146 polymorphism in 57 % of the participants, while rs290481 was present among 42 % of the studied population. Statistical tests have shown significant associations for the *TCF7L2* rs7903146 polymorphism with diabetes and cholesterol. On the other hand, rs290481 didn't indicate any association with the components of the metabolic syndrome. In conclusion, further exploration of the genetic components in the Lebanese population should be considered when looking for biomarkers of diabetes and other non-communicable diseases.

81

Integrating Genetics and Clinical Factors to Classify Non-alcoholic Fatty Liver Patients by Different Machine Learning Algorithms in UK Biobank Cohort

Jiayin Chen[1], Divya Sharma[1,2], Mei Dong[1], Mamatha Bhat[3], Sareh Keshavarzi[1,2*]
[1]*Biostatistics Division, University of Toronto, Toronto, Canada;* [2]*Department of Biostatistics, Princess Margaret Cancer Center, Toronto, Canada;* [3]*Division of Gastroenterology and Hepatology, University Health Network, Toronto, Canada*

In this study, we developed a predictive model combining clinical, behavioral, and genetic risk factors to identify patients suffering from non-alcoholic fatty liver disease (NAFLD), which is the most prevalent liver disease in the world.

We used a cohort of 2088 NAFLD cases and 428,064 non-NAFLD controls from the British ancestry in UK Biobank, defined according to ICD10. To identify patients with NAFLD, we evaluated six different supervised machine learning (ML) approaches including random forest and boosting algorithms such as CatBoost and Gradient boosting method based on potential clinical, behavioral, and 12 previously identified genetic variables. We trained and tested these models using a 70% and 30% train-test ratio.

As observed by all models, alanine transaminase, body mass index, and glutamyl transferase were the most significant clinical and lifestyle variables. SNPs in the genes for HSD17B13 and PNPLA3, which encode enzymes in hepatocytes that cause liver disease, were the most significant genetic factors associated with NAFLD. When comparing six ML approaches for NAFLD prediction, the Gradient Boosting Model achieved the highest area under the curve (AUC=0.83 95% CI: 0.82-0.84), and the logistic regression achieved the lowest (AUC= 0.74 95% CI: 0.73-0.75).The models including genetic and clinical factors slightly increased the AUC (~2%) compare to model including clinical variables only.

We conclude that lifestyle and clinical habits are more influential in NAFLD prediction than genetics. NAFLD prediction models may perform better if polygenic risk scores are included rather than single genetic factors.

82

Validation and Assessment of Predictive Ability for a Polygenic Risk Score on Parkinson's Disease

Sebastian Koch[1*], Björn-Hergen Laabs[2], Meike Kasten[3,4], Eva-Juliane Vollstedt[4], Christine Klein[4], Inke R. König[2], Katja Lohmann[4], Michael Krawczak[1], Amke Caliebe[1]
[1]*Institute of Medical Informatics and Statistics, Kiel University, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany;* [2]*Institute of Medical Biometry and Statistics, University of Luebeck, University Medical Center Schleswig-Holstein, Campus Luebeck, Luebeck, Germany;* [3]*Department of Psychiatry, University of Luebeck, Luebeck, Germany;* [4]*Institute of Neurogenetics, University of Luebeck, University Medical Center Schleswig-Holstein,Campus Luebeck, Luebeck, Germany*

Idiopathic Parkinson's disease (*PD*) is a complex multifactorial disorder with a heritability of ~ 22% for which about 90 genome-wide significant SNPs have been identified recently. However, also non-genome-wide significant SNPs can contribute to the heritability of a disease. This is taken into account by polygenic risk scores (*PRSs*), which aggregate the effects of a large number of genetic variants upon the risk for a disease like PD in a single quantity.

Before an existing PRS can be established as a valid research instrument, an assessment of its performance in independent datasets is a necessary step. We examined a previously proposed PRS for PD of 1805 SNPs for its ability to differentiate between cases and controls in an independent genetic dataset, comprising 1914 PD cases and 4464 controls and were able to replicate the results. Furthermore, we evaluated theoretically its ability to predict the development of PD in later life for healthy individuals. Here, our main objective were age-stratified predictive values because these determine the usage of this PD-PRS as a prognostic tool.

We concluded that although this proposed PRS for PD is a promising tool for research, its ability to predict PD on an individual level based on this PRS alone is not feasible.

83

Understanding Disease Mechanisms: From Genome to Phenome via the Proteome

Mine Koprulu[1*], Julia Carrasco-Zanini[1], Eleanor Wheeler[1], Nicola Kerrison[1], Nicholas Wareham[1], Maik Pietzner[1,2], Claudia Langenberg[1,2]

[1]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge, UK; [2]Computational Medicine, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Berlin, Germany

Studying the plasma proteome as the intermediate layer between the genome and the phenome has the potential to identify disease causing genes and variants and improve our understanding of the underlying mechanisms. Here, we conduct a *cis*-focused proteogenomic analysis of 2,923 proteins measured by the Olink Explore 1536 and Olink Expansion platforms in 1,180 individuals to identify disease causing genes across the human phenome and systematically refine causal genes at previously reported GWAS loci. We identify 1,553 distinct credible sets of protein quantitative trait loci (pQTL), a third of which (n=531) contained cis-pQTLs not previously reported. Of these, 182 signals were seen for 117 proteins never studied before, and 349 were detected despite proteins having been targeted in other much larger studies. We identified 224 proteins with robust evidence to contribute to the aetiology of 578 unique health outcomes using statistical colocalization, including putative drug targets for metabolic diseases such as type 2 diabetes (FGFR4). We demonstrate convergence between our colocalization results and the rare variant gene-burden associations for 25 proteins, including *TIMD4* for cholesterol metabolism providing genetic evidence for drug target consideration. Finally, we show that 481 of the credible sets overlap with reported GWAS loci and highlight novel candidate causal genes for 40.1% of these, including *TIMP4* at the intersection between hypothyroidism and creatinine metabolism.

Our findings demonstrate the ability of broad capture, high-throughput proteomic technologies to robustly identify new gene-protein-disease links, provide mechanistic insight, and add value to existing GWASs by enabling and refining causal gene assignment.

84

Development of a Breast Cancer Risk Prediction Model with Carrier Status, a Polygenic Risk Score, and Epidemiologic Risk Score

Sarah S. Kalia[1], Nicholas J. Boddicker[2], Siddhartha Yadav[3], Fergus J. Couch[4], Peter Kraft[1,5] * on behalf of the CARRIERS Consortium

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America; [3]Department of Oncology, Mayo Clinic, Rochester, Minnesota, United States of America; [4]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America; [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

Breast cancer has been associated with monogenic, polygenic, and epidemiologic (clinical, reproductive and lifestyle) risk factors, but direct empirical data on their joint effects is limited. We extended a risk model incorporating pathogenic variants (PV) in six breast cancer predisposition genes and a 105-SNP polygenic risk score (PRS) to include an epidemiologic risk score (ERS). This study was performed in a population-based sample of more than 22,700 cases and a similar number of age-matched controls from the Cancer Risk Estimates Related to Susceptibility (CARRIERS) Consortium. We assessed effect measure modification among the modeled factors, age and family history: interaction terms were chosen via penalized logistic regression with cross validation. Our final model includes a synergistic interaction between the ERS and PRS, and antagonistic interactions among age, PRS and *BRCA1/2* status. Adding the ERS increases risk discrimination. A 50-year-old female at the median PRS with the highest ERS has an odds of breast cancer nearly five times that of an age-matched female at the median PRS but with the lowest ERS. Our results illustrate that the ERS, alone and in combination with the PRS, can contribute to clinically meaningful risk stratification across high-risk thresholds for five-year and lifetime risk, especially for carriers of a PV in a moderate penetrance gene such as *CHEK2*. Appropriately integrating monogenic, polygenic, and epidemiologic risk factors to improve breast cancer risk prediction models may inform personalized screening protocols and prevention efforts.

85

Investigating the Impact of C4 Copy Number Variation on Immune Function in Schizophrenia

Allison M. Lake[1,2,3*], Lea K. Davis[2,3]

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

Schizophrenia is a debilitating neuropsychiatric disorder that is highly heritable but arises via unknown mechanisms. Immune processes are implicated in its etiology but remain poorly understood. The strongest schizophrenia GWAS signal lies in the major histocompatibility complex and is in part driven by copy number variation (CNV) in the complement component 4A (C4A) gene. Observational studies have demonstrated increased inflammation in schizophrenia, but it is unclear whether these associations are influenced by C4A CNV. In this study, we first leveraged diagnosis codes and quality-controlled laboratory measurements extracted from a large hospital electronic health record (EHR) system to investigate alterations in six immune biomarkers in patients with schizophrenia vs. individuals with no psychiatric diagnoses. Analysis of median outpatient lab values revealed significant elevations in C-reactive protein (CRP) (OR=1.52; 95% CI=1.37,1.69) and total WBC (OR=1.17; 95% CI=1.12,1.22) in cases. We next examined the role of C4

CNV in immune biomarker elevations in schizophrenia. In a subset of European-ancestry patients with SNP genotyping data (N>69,000), we imputed C4A CNV using a published reference panel and tested the association between C4A CNV and schizophrenia-associated biomarkers WBC and CRP. Only WBC was associated with C4A CNV (OR=1.04, 95% CI=1.03,1.05). When repeating the case-control biomarker analysis in the genotyped individuals and controlling for C4A CNV, the schizophrenia-WBC association persisted (OR=1.25, 95% CI=1.08,1.44). Taken together, these findings suggest that biomarker changes in schizophrenia may be independent of C4A CNV and highlight the need for detailed studies examining the joint role of genetics and immune biomarkers in schizophrenia.

86

Investigating Relationships between Negative Peer Interactions, Polygenic Influences on Pubertal Timing, and Depression Symptoms Across Sexes

Eva Lancaster[1]*, William Copeland[2], Hermine H. Maes[3,1], Roseann Peterson[1]

[1]*Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States of America;* [2]*Department of Psychiatry, University of Vermont, Burlington, Vermont, United States of America;* [3]*Department of Human & Molecular Genetics, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States of America*

Background: Sex-specific differences in depression prevalence emerge during adolescence, in the years coinciding with the onset of puberty. Associations between depression and pubertal timing have been observed, and it is likely that these associations are driven by both biological (e.g., shared genetic influences) and psychosocial factors (e.g., lack of peer social support). However, the specific underlying mechanisms remain unknown. This study examined relationships between pubertal timing, peer interactions, and depression in the Great Smoky Mountains Study (n = 1420) to provide insight into the processes underlying depression etiology and heterogeneity.

Methods: The Child and Adolescent Psychiatric Assessment, which establishes DSM-IV diagnoses and symptoms, was first administered to 8-17 year-olds with multiple assessments completed throughout development. Linear regression tested relationships between maximum lifetime depressive symptoms, three measures of peer interactions (bullying, conflicts, and shyness), and polygenic risk scores for age at menarche (PRS-AAM; females) and age at voice break (PRS-AAVB; males).

Conclusions: A significant association was identified between PRS-AAM and depressive symptoms in women ($p = 0.004$), but not between the PRS-AAVB and depressive symptoms in males ($p = 0.988$). All measures of peer interactions were significantly associated with depressive symptoms ($p < 0.008$) and improved model prediction. Shyness with peers was the strongest environmental predictor of depressive symptom load (adjusted r-squared = 0.057 [females] and 0.076 [males]). In the full models, the overall proportion of variation explained was similar across sexes (adjusted r-squared = 0.135 [females] and 0.146 [males]). Further analyses will test for gene-environment interactions.

88

Study of Effect Modifiers of Genetically Predicted CETP Reduction

Marc-André Legault[1,2,3,*], Amina Barhdadi[1,2], Isabel Gamache[1,3], Audrey Lemaçon[1,2], Louis-Philippe Lemieux Perreault[1,2], Jean-Christophe Grenier[1], Marie-Pierre Sylvestre[4,5], Julie G. Hussin[1,6], David Rhainds[1], Jean-Claude Tardif[1,6], Marie-Pierre Dubé[1,2,6]

[1]Montreal Heart Institute, Montreal, Canada; [2]Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada; [3]Department of Biochemistry and Molecular Medicine, Université de Montréal, Montreal, Canada; [4]Research Centre of the University of Montreal Hospital Centre, Montreal, Canada; [5]Department of Social and Preventive Medicine, Université de Montréal, Montréal, Canada; [6]Department of Medicine, Université de Montréal, Montreal, Canada

Genetic variants in drug targets can be used to predict the effect of drugs. Here, we use this approach to assess *effect modification* in which the magnitude of the effect of an exposure on an outcome varies across strata of a third variable. We aim to estimate changes in the effect of a genetically predicted reduction in cholesteryl ester transfer protein (CETP) on clinically meaningful phenotypes across strata of sex and/or body mass index (BMI).

We used linear and logistic regression with interaction terms to model effect modification. We report the estimated effects of a reduction in genetically predicted CETP concentration on phenotypes at representative values of sex and BMI using the R "margins" package. Interactions on the additive scale for binary outcomes were also estimated using interaction contrast and the Relative Excess Risk due to Interaction (RERI).

Both sex and BMI modified the association between a genetically predicted lower CETP and lipid biomarkers in the UK Biobank. Female sex and lower BMI were associated with higher HDL cholesterol and lower LDL cholesterol, for a same genetically predicted reduction in CETP concentration. Sex also modulated the effect of a genetically lower CETP on cholesterol efflux capacity, an important measurement associated with atherosclerotic plaque development. Our results provide insight on the clinical effects of CETP inhibitors. The investigation of effect modification by using genetic variants as proxies for drug target activity is a promising approach to identify subgroups of individuals susceptible to derive a greater benefit from a pharmacological treatment.

89

Integrated Multi-omics Analysis to Reveal Underlying Protective Mechanisms of Delaying Cognitive Decline in Centenarians

Anastasia Leshchyk[1,2*], Stefano Monti[2], Stacy Andersen[3], Tomas T. Perls[4], Paola Sebastiani[5]

[1]Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America; [2]Department of Medicine, Computational Biomedicine Section, Boston University, Boston, Massachusetts, United States of America; [3]Department of Biostatistics, Boston University, Boston, Massachusetts, United States of America; [4]Department of Medicine, Geriatric Section, Boston University, Boston, Massachusetts, United States of America; [5]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, United States of America

Multiple studies of long-lived people have shown that some individuals can delay the onset of aging-related diseases such as Alzheimer's or cardiovascular diseases to the very end of their lives. Centenarians' offspring have also demonstrated the ability to age more healthily compared to the controls without familial longevity. The compression of morbidity in centenarians and their offspring is illustrated in studies where more than 90% of centenarians remain functionally independent and postpone their disability at very old age. In addition, centenarians who experience diseases such as dementia or Alzheimer's often delay the onset to the very end of their long lives. Previous studies of centenarian cohorts showed that APOE e2 allele carriers have increased odds of reaching longevity and decreased mortality risk. In other genetic studies, subjects with the e4 allele demonstrated an increased risk of developing Alzheimer's disease. In this research project, we investigate how APOE genotypes affect cognitive decline in a cohort of centenarians, their offspring, and controls. We are developing a novel Bayesian network-based methodology that integrates multiple omics data to recognize the shared molecular profiles among subjects with familial longevity that leads to postponing cognitive decline and the onset of dementia. A Bayesian network-based approach will help to decipher the risk factors and pathways contributing to the prolonged subjects' health span. It will also help to understand how the genetic effect of APOE genotype propagates to the molecular level and eventually to the phenotype.

90

Disentangling the Aetiological Pathways Between Body Mass Index and Site-specific Cancer Risk Using Tissue-Partitioned Mendelian Randomization

Genevieve M. Leyden[1,2*], Michael P. Greenwood[2], David Murphy[2], George Davey Smith[1], Tom G. Richardson[1,3]

[*]Presenting author

[1]MRC Integrative Epidemiology Unit, Bristol Population Health Science Institute, University of Bristol, Bristol, United Kingdom; [2]Bristol Medical School: Translational Health Sciences, Dorothy Hodgkin Building, University of Bristol, Bristol, United Kingdom; [3]Novo Nordisk Research Centre, Headington, Oxford, United Kingdom

Body mass index (BMI) influences risk of various site-specific cancers, although dissecting the subcomponents of this heterogenous lifestyle factor responsible for driving cancer risk has proven difficult to establish. In this study, we have leveraged tissue-specific gene expression to separate and estimate the independent effects of distinct phenotypes underlying BMI on risk of 6 site-specific cancers. We recently developed methodology to leverage BMI-associated variants as instrumental variables within a multivariable Mendelian randomization (MR) framework weighted by their evidence of genetic colocalization with subcutaneous adipose- and brain-tissue derived gene expression. Here, we extend this approach to a two-sample setting to harness findings from large-scale consortia. Our results provide evidence that brain-tissue colocalizing variants are predominantly responsible for driving the genetically predicted effect of BMI on lung cancer (OR:1.17; 95%-CI=1.01-1.36; P=0.03). Similar findings were identified when analyzing cigarettes per day as an outcome (Beta=0.44; 95%-CI=0.26-0.61; P=1.62x10-6), suggesting that neurobiological pathways may underlie the relationship between BMI and increased lung cancer risk. Our findings also suggest that adipose-tissue colocalizing variants predominantly drive the effect of BMI and increased risk for endometrial cancer (OR:1.71; 95%-CI=1.07-2.74; P=0.02) highlighting the putatively important role of adipogenesis in the aetiology of this outcome. Our novel extension to multivariable MR provides valuable insight into the divergent underlying pathways between BMI and risk of site-specific cancers. Conducting this approach in a two-sample setting has wide potential applicability to disentangle mechanisms between adiposity and a spectrum of disease outcomes.

Funded by the BHF (FS/17/60/33474) and MRC (MC_UU_00011/).

91

Development of a yQTL Discovery Pipeline Applicable for Both Unrelated and Related Individuals

Mengze Li[1,2*], Zeyuan Song[3], Anastasia Gurinovich[4], Paola Sebastiani[4], Stefano Monti[1,2,3]

[1]Section of Computational Biomedicine, Boston University School of Medicine, Boston, Massachusetts, United States of America; [2]Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America; [3]Department of Biostatistics, Boston University, Boston, Massachusetts, United States of America; [4]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, United States of America

Quantitative trait loci (QTL) are DNA sequence variants, such as SNPs, that influence the level of a quantitative trait, for example, gene expression. QTL discovery analysis consists of multiple steps, including genome-wide principal component analysis (PCA), genome-wide association test, as well as downstream analyses such as plotting and result annotations. In order to facilitate and automate the process, we developed *yQTL*, a pipeline that is agnostic to the nature

of the dependent variable (y) to be modeled. In the genome-wide association step, the pipeline supports two different analysis modalities: i) one using standard linear models based on the use of the R package *matrixeQTL*, which is optimized to process hundreds of phenotypes at once and yields QTL results at a high speed, but does not support the incorporation of family structure information, thus adequate only when unrelated subjects are analyzed; ii) one based on the R package *GENESIS* that supports the estimation of genetic relationship matrix (GRM) and include it in linear mixed effect models, thus able to analyze related subjects. Both modalities include the genome-wide PCA and the incorporation of user-specified covariates. Through the adoption of the workflow management tool *Nextflow*, the pipeline parallelizes the analysis steps. We have tested the pipeline using proteomics and metabolomics data from the New England Centenarian Study and publicly available multi-omics datasets.

92

Investigating Genetic Inheritability of RNA Alternative Splicing in Alcohol Use Disorder

Rudong Li[1,2], Andy B. Chen[1,2], Steven X. Chen[1,2], Jill L. Reiter[1,2], Dongbing Lai[2], Tatiana Foroud[2], Howard J. Edenberg[2,3], *Yunlong Liu*[1,2*]

[1]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [2]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, United States of America; [3]Departments of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana, United States of America

Alcohol use disorder (AUD) is a genetically heritable psychiatric disorder characterized by excessive and problematic alcohol consumption. Although significant efforts focus on identifying risk loci through genome-wide association studies (GWAS), less is known about the roles of pre-mRNA splicing in AUD. We designed a Mendelian Randomization-based approach for identifying the transcripts whose splicing variants may contribute to alcoholic traits. For each alternative splicing event, we first developed an Elastic Net-based predictive model for inferring the splicing outcomes based on the genotypes of common SNPs within the gene region, using the RNA-seq data from the CommonMind Consortium. We applied these splicing models to the genotypes of the subjects in the Collaborative Studies on Genetics of Alcoholism study and examined the association between the inferred splicing outcome with AUD-related traits using the Generalized Estimating Equation method. This analysis identified 27 events whose inclusion status may contribute to AUD. We further conducted the same analysis in the Australian Twin-family Study of Alcohol Use Disorder dataset. Six of the 27 splicing events were replicated (FDR<0.05). Further analysis suggested that pathways related to immune response and neurodegeneration are enriched for the downstream genes of these splicing events. Moreover, impact of the splicing event in gene *ELOVL7* was supported in four additional large GWAS datasets with summary statistics. In conclusion, this work investigated how RNA splicing might impact the risk for AUD, which highlighted neuro-immunological functions. This work also establishes a framework for studying the impact of RNA splicing in the genetics of other complex diseases.

93

Multi-population Analysis Identified Novel Variants in Ever- and Never-smoking Lung Cancer

Yafang Li[1-3*], Xiangjun Xiao[1], Jun Xia[1,4], Meng Zhu[5], Gail F Fernandes[4], Shannon E Slewitzke[4], Susan M Rosenberg[4], Jinyoung Byun[1-3], Younghun Han[1], Chris Amos[1-3]. INTEGRAL-ILCCO Lung Cancer Consortium

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, Texas, United States of America; [2]Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, Houston, Texas, United States of America; [3]Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America; [4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America; [5]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

Leveraging the genotype data from INTEGRAL (Integrative Analysis of Lung Cancer Etiology and Risk)-ILCCO (International Lung Cancer Consortium) lung cancer consortium, we conducted a multi-population Genome-wide Association Study on ~20,000,000 high-quality imputed SNPs (information score >= 0.8) from a total of 64,897 individuals, including 44,823 ever-smokers and 20,074 never-smokers. 72.1% of the individuals have European ancestry (CEU, N=46,786), 19.1% with Asian ancestry (CHB, N=12,423) and 8.8% with African-American ancestry (YRI, N=5,688). We identified six novel variants in ever- and never- smoking lung cancer. For example, rs62303696, located in 3' untranslated region (UTR) of *GABRA4*, was identified in ever-smoking lung cancer with a joint p value of 1.22×10^{-9} and OR of 1.18. The risk effect was detected in all three continental populations with p values of 2.71×10^{-7} (CEU), 4.81×10^{-3} (CHB) and 6.08×10^{-2} (YRI). rs968516 (8.19×10^{-10}, OR=0.34), located at 5' UTR of *LCNL1*, was identified in never-smoking squamous lung cancer. We further conducted functional analysis to infer the functional impact of the identified variants. The association between GWAS signals and expression quantitative trait loci (eQTL) signals suggested rs968516 could affect lung cancer risk in never-smokers through cis-regulation of *LCNL1* gene expression. Overproduction of *GABRA4* in lung fibroblast cell line MRC5-SV40 was found to increase DNA damage level in the cell implying rs62303696 may increase lung cancer risk through regulation of DNA damage. Besides the novel findings, we also validated the lung cancer risk effect for *VTI1A* and *ACVR1B* in never-smoking women in African-American for the first time.

94

Sex-stratified vs. Sex-combined Analysis in the Presence of Genetic Effect Heterogeneity

Boxi Lin[1*], Lei Sun[1,2]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Department of Statistical Sciences, University of Toronto, Toronto, Canada

The effect of a genetic variant on a complex trait may differ between male and female, e.g. genetic effects may be sex-specific for testosterone levels. In the presence of genetic effect heterogeneity between female and male, sex-stratified analysis is often used, which provides easy-to-interpret sex-specific effect size estimates. However, from power of association testing perspective, sex-stratified analysis may not be the best approach. As sex-specific genetic effect implies SNP-sex interaction effect, jointly testing SNP main and SNP-sex interaction effects may be more powerful than sex-stratified analysis or the standard main-effect testing approach. When individual data are not available, it is then of interest to study if the interaction analysis can be derived from sex-stratified summary statistics. We considered several different sex-combined methods and evaluated them through extensive simulation studies. We observed that a) the joint SNP main and SNP-sex interaction analysis is most robust to a wide range of genetic models, and b) this joint interaction testing result can be obtained by quadratically combining sex-stratified summary statistics (i.e. squared sex-stratified summary statistics). We then provide theoretical justification for the equivalence between the joint interaction test and the quadratically combined omnibus test. Finally, we provide additional supporting evidence by utilizing the publicly available sex-stratified GWAS summary statistics of testosterone levels of the UK Biobank data.

95

Phenome-wide PGS Portability in the Colorado Center for Personalized Medicine Biobank Suggests Overlooked Challenges in Diverse Populations

Meng Lin[1]*, Christopher H. Arehart[1], Nicholas Rafaels[1], Kristy R. Crooks[1], Nikita Pozdeyev[1], Audrey Hendricks[1], Sridharan Raghavan[1], Christopher R. Gignoux[1], on behalf of the CCPM Clinical PRSUIT Working Group

[1]Colorado Center for Personalized Medicine (CCPM), University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America

The potential of polygenic scores (PGS) to improve health-related personalized risk prediction is highly promising. Yet, their transportability across populations remains largely poor. Limited studies to date have empirically and comprehensively examined the landscape of current PGS transportability across a phenome-wide range of conditions. Here, we leverage the PGS Catalog with >600 scores to systematically assess the performance of PGS prediction on >1k electronic health record (EHR) phenotypes in the Colorado Center for Personalized Medicine Biobank, totaling 661,265 predictions. To analyze heterogeneity across populations we stratified 33,863 individuals from six genetically defined ancestry groups. Although some diseases are reasonably well predicted by PGSs together with demographic covariates, such as type 2 diabetes and hypertension (P=2.7e-170 and 1.1e-164, AUC=0.77 and 0.81, respectively), the majority of predictions have considerable cross-group heterogeneity in performance (average I^2=0.18 in phecode~PGS pairs with false discovery rate <0.1). Additionally, comparison of measures in cross-group heterogeneity, when the PGS unit was per SD vs. top decile cutoff, yielded alarming discordance (r=0.29). Using multilevel nested mixed models, we found primary influences on heterogeneity include the distribution of disease prevalence between ancestry groups in the test cohort, in addition to features of the training set. Our results suggest that there are overlooked barriers to PGS transportability that can be determined empirically and that are due to both characteristics of the training and test settings. This provides a snapshot of hopes and pitfalls in the current efforts of applying PGS resources to diverse populations.

96

Inference of Causal Networks Using Bi-directional Mendelian Randomization and Network Deconvolution with GWAS Summary Data

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

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

Inferring causal relationships among potential risk factors and diseases from observational data is both important and challenging, e.g. due to hidden confounding. Emerging as a powerful tool, Mendelian randomization (MR) has been increasingly applied for causal inference with observational data by using genetic variants as instrumental variables (IVs). However, the current practice of MR has been largely restricted to investigating the total causal effect between two traits, while it would be more useful to infer the direct causal effect between any two of many traits (by accounting for mediating effects through other traits). In this work, we first extend bi-directional MR-cML, a robust MR method based on constrained maximum likelihood, to overlapping-sample MR set-up, then apply it to infer a causal network of total effects among multiple traits. Finally we apply graph deconvolution to infer a causal network of direct effects. Simulation studies showed much better performance of the extended MR-cML with sample overlap. We applied the method to 17 large-scale GWAS summary datasets to infer the causal networks of both total and direct effects among 11 common cardiometabolic risk factors, 4 cardiometabolic diseases (coronary artery disease, stroke, type 2 diabetes, atrial fibrillation), Alzheimer's disease and asthma. The inferred total causal graph identified many well-accepted risk factor-disease pairs while the direct causal graph provided more interesting insights into the mechanisms. We also provide an R Shiny app (https://zhaotongl.shinyapps.io/cMLgraph/) for users to explore any subset of the 17 traits of interest.

97

Genetic Analysis of Preserved Ratio Impaired Spirometry Using a Population-based Cohort

Alvin E. Lirio[1]*, Daniel H. Higbee[2, 3], Catherine John[1], Alexander T. Williams[1], Nick Shrine[1], James W. Dodd[2, 3], Martin D. Tobin[1,4], Anna L. Guyatt[1]

[1]*Department of Health Sciences, University of Leicester, Leicester, United Kingdom;* [2]*Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom;* [3]*Academic Respiratory Unit, University of Bristol, Southmead Hospital, Bristol, United Kingdom;* [4]*Leicester NIHR Biomedical Research Centre, Leicester, United Kingdom*

Preserved Ratio Impaired Spirometry (PRISm), defined as Forced Expiratory Volume in one second (FEV_1) <80% predicted with FEV_1/Forced Vital Capacity (FVC) ratio ≥0.70, is associated with respiratory symptoms, extra-pulmonary co-morbidities, and increased mortality. Identification of shared genetic architecture of PRISm and other respiratory conditions could inform understanding of this spirometric pattern.

In a genome-wide association study (GWAS) of PRISm in UK Biobank (38,639 PRISm cases and 257,643 controls), 27 SNPs reached genome-wide significance ($P<5x10^{-8}$) and were followed up in an association study of PRISm in the Extended Cohort for E-health, Environment and DNA (EXCEED) cohort (256 cases and 803 controls). None of the 27 SNPs met a Bonferroni-corrected threshold ($P<0.002$) for independent replication in EXCEED.

We then utilized GWAS summary statistics from UK Biobank to assess genetic overlap of PRISm and other respiratory traits via genetic correlation analysis (r_g) using bivariate linkage disequilibrium (LD)-score regression. We found genetic correlations between PRISm and: childhood-onset asthma ($r_g=0.131$, $P<0.001$); adult-onset asthma ($r_g=0.202$, $P<0.001$); moderate-to-severe-asthma ($r_g=0.310$, $P<0.001$); asthma-Chronic Obstructive Pulmonary Disease (COPD) overlap syndrome ($r_g=0.517$, $P<0.001$) and; COPD ($r_g=0.623$, $P<0.001$). The magnitude of genetic correlation with PRISm corresponds to the degree of fixed airflow obstruction in each of these traits. In addition, we found a positive genetic correlation between respiratory infections and PRISm ($r_g=0.185$, $P=0.003$), suggesting shared aetiology.

Our findings provide insights into the genetic architecture of PRISm in relation to other respiratory and complex diseases. A two-stage GWAS meta-analysis of PRISm, combining results of UK Biobank, EXCEED, and other studies is ongoing.

98

Germline Cancer Gene Expression Quantitative Trait Loci Influence Local and Global Tumor Mutations

Yuxi Liu[1,2 *], Alexander Gusev[3], Peter Kraft[1,2,4]

[1]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America;* [2]*Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America;* [3]*Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts, United States of America;* [4]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America*

Somatic mutations drive cancer development and are relevant to patients' response to treatment. Emerging evidence show that somatic mutations are influenced by germline variants. However, the mechanisms underlying these germline-somatic associations remain largely obscure. We hypothesize that germline variants can influence somatic mutations in a nearby cancer gene ("local impact") or a set of recurrently mutated cancer genes across the genome ("global impact") through their regulatory effect on gene expression. Here, by integrating tumor targeted sequencing data from 12,413 patients across 11 cancer types in the Dana-Farber Profile cohort with germline cancer gene expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression Project, we identified novel associations between eQTL and tumor mutations. For local impact, we found variants that upregulate *ATM* expression which are also associated with a decreased risk of having somatic *ATM* mutations across 8 cancer types ($P = 3.43 \times 10^{-5}$). For global impact, we identified *GLI2*, *WRN*, and *CBFB* eQTL that are associated with tumor mutational burden of cancer genes in ovarian cancer, glioma, and esophagogastric carcinoma, respectively, with $P < 3.45 \times 10^{-6}$. An *EPHA5* eQTL was associated with tumor mutation count (TMC) of cancer genes in colorectal cancer. eQTL associated with expression of *APC*, *WRN*, *GLI1*, *FANCA*, and *TP53* were associated with TMC in endometrial cancer ($P < 1.73 \times 10^{-5}$). Our findings provide evidence for the germline-somatic associations mediated through expression of specific cancer genes and open avenues for research on the underlying biological processes, especially those related to immunotherapy responses.

99

An Eigenvalue Ratio Approach to Inferring Population Structure from Whole Genome Sequencing Data

Zhonghua Liu

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

Inference of population structure from genetic data plays an important role in population and medical genetics studies. With the advancement and decreasing cost of sequencing technology, the increasingly available whole genome sequencing data provide much richer information about the underlying population structure. The traditional method (Patterson, Price, and Reich, 2006) originally developed for array-based genotype data for computing and selecting top principal components that capture population structure may not perform well on sequencing data for two reasons. First, the number of genetic variants p is much larger than the sample size n in sequencing data such that the sample-to-marker ratio n/p is nearly zero, violating the assumption of the Tracy–Widom test used in their method.

Second, their method might not be able to handle the linkage disequilibrium well in sequencing data. To resolve those two practical issues, we propose a new method called ERStruct to determine the number of top informative principal

components based on sequencing data. More specifically, we propose to use the ratio of consecutive eigenvalues as a more robust test statistic, and then we approximate its null distribution using modern random matrix theory. Both simulation studies and applications to two public data sets from the HapMap 3 and the 1000 Genomes Projects demonstrate the empirical performance of our ERStruct method.

100

Mendelian Randomization for Multiple Exposures and Outcomes

Noah J. Lorincz-Comi[1]*, Yihe Yang[1], Xiaofeng Zhu[1]
[1]*Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America*

Mendelian Randomization (MR) is typically used to estimate the causal effect of an exposure on an outcome using data from genome-wide association studies (GWAS). Including multiple exposures in MR estimation ('multivariable' methods) can reduce the potential for confounding and pleiotropy biases in causal effect(s) estimation. Yet, the power in multivariable methods could be increased by including multiple correlated outcome traits simultaneously ('multivariate' methods). However, no such MR estimator currently exists.

Here, we propose the first fully flexible MR estimator that can be used to model causal relationships between multiple exposures and outcomes simultaneously. Our proposed method uses a set of distribution-free estimating equations and corrects for measurement error in the MR instrumental variables, which popular methods do not explicitly do. The proposed estimator needs only GWAS summary statistics, which are frequently publicly available. We demonstrate in theory and simulations that our method is unbiased when GWAS estimates are unbiased and its variance can approach the Cramer-Rao lower boundary in causal effect estimation. We also compare the robustness of our method to existing alternatives in reducing confounding and pleiotropy biases and introduce a new pleiotropy test for multiple traits. This test is primarily used to ensure MR model validity but can also be used in genome-wide searches for pleiotropic SNPs. Finally, we used our method to estimate the causal effects of three lipid traits (HDL, LDL, triglycerides) on coronary artery disease (CAD) risk. These results indicate causal relationships between each lipid trait and CAD risk, and substantial pleiotropy across the genome.

101

Reimagining Gene-environment Interaction Analysis for Human Complex Traits

Jiacheng Miao[1], Gefei Song[1], Lauren L. Schmitz[2], Jason M. Fletcher[2,3], Qiongshi Lu[1]*
[1]*Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 610 Walnut Street, Madison, WI 53726, United States of America;* [2]*Robert M. La Follette School of Public Affairs, University of Wisconsin-Madison, 1225 Observatory Dr, Madison, WI 53706, United States of America;* [3]*Department of Sociology, University of Wisconsin-Madison, 1180 Observatory Dr, Madison, WI 53706, United States of America*
*Denotes presenting author

The environments are often ignored or treated as nuisance parameters in human complex trait genetics research. However, in epidemiology, social sciences, and clinical research, there is a great interest in quantifying the heterogeneity of the effect of an exposure (e.g., a treatment, a policy change, a natural experiment), and more specifically, how it interacts with genetics. Given the 'omnigenic' genetic architecture of most human traits, it is a common practice to summarize genetic propensities into polygenic risk scores (PRS) and test the interaction between PRS and the environment. However, the typical statistical methodology used in these analyses (i.e., linear models with main effects of PRS and E and their interaction PRSxE) always produces biased estimates of interactions, requires an external GWAS independent from GxE samples, and is sensitive to PRS-E correlation. We introduce an innovative statistical framework named PIGEON which links two seemingly unrelated topics: gene-environment interaction and genetic correlation estimation. We demonstrate that genetic correlations between typical GWAS and SNPxE interaction statistics provide a superior strategy for quantifying PRSxE interactions. More specifically, PIGEON provides unbiased estimates for PRSxE interaction with improved statistical power without any assumption on G-E independence. It also allows GWAS and GxE study samples to have arbitrary overlaps. We will show plenty of empirical examples that involve gene-by-education-reform interaction in the UK and gene-by-sex interactions of more than 500 complex traits to showcase its performance. Overall, PIGEON address critical limitations in existing methodologies and will have broad applications in future GxE studies.

102

Identifying Monogenic Causes for Improved Polygenic Prediction of Osteoporosis

Tianyuan Lu[1,2]*, Vincenzo Forgetta[1], Sirui Zhou[1,3,4], J Brent Richards[1,3,4,5], and Celia MT Greenwood[1,3,4,6]
[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2]Quantitative Life Sciences Program, McGill University, Montreal, Canada; [3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada; [4]Department of Human Genetics, McGill University, Montreal, Canada; [5]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom; [6]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada

More than 200 million people have osteoporosis worldwide. Predicting bone density by polygenic risk score (PRS) may improve osteoporosis and fracture risk screening. However, existing PRSs do not include rare variants that can confer strong effects.

Leveraging UK Biobank whole-exome sequencing data, we developed a PRS including both rare and common variants,

called ggSOS, for heel ultrasound speed of sound (SOS), a quantitative measure of bone density. We split European ancestry individuals into a training dataset (N = 317,434) and a test dataset (N = 74,825). In the training dataset, we regressed measured SOS on a common variant-based PRS for SOS, which aggregated variants with a minor allele frequency (MAF) >0.001. Using the residualized SOS, we conducted burden testing for 19,308 genes with computationally predicted rare pathogenic variants with a MAF ≤0.001.

Fourteen genes harbored rare pathogenic variants associated with residualized SOS (*P* value $<2.5x10^{-6}$). These variants cumulatively affected 6.6% of individuals. Amongst 4,949 carriers in the test dataset, one SD decrease in ggSOS was associated with 1.52 (1.27-1.81)-fold increased odds of osteoporosis and 1.45 (1.25-1.69)-fold increased hazard of major osteoporotic fracture. Compared to a common variant-based PRS, ggSOS had increased C-indices for predicting osteoporosis (0.625 vs. 0.619) and fracture (0.620 vs. 0.613) risks. No prominent improvement was observed in non-European ancestry populations.

In conclusion, identifying monogenic causes may assist polygenic prediction for osteoporosis. Nonetheless, the relatively high cost of sequencing, limited proportion of carriers, and small magnitude of predictive power gain entail careful considerations in research and in clinics.

103

Phenotype Prediction in Diverse Populations Using Identity by Descent Clustering in the Colorado Center for Personalized Medicine Biobank

Betzaida L. Maldonado[1*], Jonathan A. Shortt[1], Meng Lin[1], Chris R. Gignoux[1],

[1]*Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America*

The Colorado Center for Personalized Medicine (CCPM) Biobank has over 79,000 biological samples with either array or whole exome sequence data, which combined with Electronic Health Records data, provide a unique resource to study health disparities in a large scale and genome wide context. Here, we leverage inherent population structures to characterize patterns of risk in clusters of homogeneous ancestry. Identification of individuals that share genetic homology is possible by detecting haplotypes inherited from a common ancestor without intervening recombination. These identical by descent (IBD) segments persist within populations for generations, therefore, shared IBD segments between individuals can be used to identify genetic communities and gain insights into recent population history and investigate correlations of disease risk and prevalence. Here, we evaluate CCPM's current data freeze with 40 IBD informed clusters in a data resource of 605 polygenic scores. Because IBD clustering captures both genetic and environmental factors that impact risk prediction, this method may be an effective tool to overcome current phenotype and risk prediction limitations. We evaluate phenotype prediction accuracy in diverse populations using IBD clustering and identify scores for which adding an IBD cluster term improves prediction accuracy. We test the hypothesis that prediction accuracy will increase when incorporating polygenic risk scores (PRS) developed from a population with similar genetic ancestry as the target IBD cluster. Potential environmental factors captured by IBD clusters will improve phenotypic prediction beyond polygenic risk scores, particularly for admixed and understudied populations, thereby improving the potential of personalized medicine to benefit all.

104

RetroFun-RVS: A Family-based Retrospective Association Test Integrating Functional Annotations

Loïc Mangnier[*1,2], Alexandre Bureau[1,2]

Département de médecine sociale et préventive, Université Laval, Québec, QC, Canada; CERVO Brain Research Centre, Québec, QC, Canada

Over the last few years progress has been made to efficiently annotate genetic variations located within noncoding regions. Methods have already been proposed to incorporate functional annotations in rare-variant association tests, increasing the power of detection by leveraging external information. However, these models have not been extended to pedigree studies. Moreover, among functional annotations, regions based on 3D genome contacts are promising to detect causal variants in non-coding sequence. Here we are proposing a rare-variant-sharing retrospective family-based model, called RetroFun-RVS, that includes simultaneously functional annotations for extended pedigrees in a computation-efficient manner. In addition to weighting variants based on functional scores, our model exploits the affected-only design to detect shared causal variants in pedigrees. Through extensive simulations studies in 52 families totaling 270 affected individuals, we have demonstrated that incorporating 3D-based networks as functional annotations performs better to detect rare causal variants compared to not using annotations or applying RVS tests, while showing a good control of the Type-I error a. Indeed, considering three different scenarios at 2% of causal variants (relative risk = 20), where 100%, 75%, and 50% are located within the region corresponding to one network, we have shown that power is increased when at least one score is predictive for the trait (original power = 86%, power with annotations = 92% at a=1.66e-5) with a minimal loss of power when the signal is shared across different regions. Finally, we argue that RetroFun-RVS can help locating variant sets with a common functional role in disease.

105

How Should QC of Sequencing Data be Performed for Rare Variant Association Testing with External Controls?

Gaëlle Marenne[1,*], Thomas E Ludwig[1,2], Ozvan Bocher[1], Anthony F Herzig[1], Chaker Aloui[3], Elisabeth Tournier-Lasserve[3,4] and Emmanuelle Génin[1,2]

[1]*Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200 Brest, France*; [2]*CHU Brest, F-29200 Brest, France*; [3]*Université de Paris, NeuroDiderot, Inserm UMR 1141, F-75019 Paris, France*; [4]*AP-HP,*

Service de Génétique Moléculaire Neurovasculaire, Hôpital Saint-Louis, F-75010 Paris, France

Testing the excess burden of rare variants in patients affected by a given disease requires sequencing a large number of cases and controls. To reduce study cost and increase sample sizes, external controls can be used. But this strategy requires a specific and stringent quality control (QC) to remove batch effects and ensure comparability between cases and controls data.

In this work, we analyzed two real datasets of whole exome and whole genome sequencing data to show how the QC impacts on the gene-based rare-variants association test results (WES: 96 cases of moyamoya disease and 568 controls; WGS: 78 samples from 39 duplicated individuals with blood and saliva samples). We applied two QC strategies: the widely used Variant Quality Score Recalibration (VQSR) method, and the QC implemented in the RAVAQ R package.

We show that both QC strategies selected similar numbers of QCed variants, although the overlap was low in the case-control WES data. We noticed that the difference especially impacted rare variants and singletons. In both datasets, the RAVAQ QC strategy improved rare variant association test results by removing inflation due to spurious signals. On the WES dataset, RAVAQ showed a better power by finding the already-known associated gene as the top signal.
In conclusion, the QC implemented in RAVAQ is accurate and more appropriate for rare-variant association testing. The RAVAQ all-in-one pipeline is an interesting and well documented tool for researchers to conduct all the analysis steps.

106

Body Size at Different Ages and Risk of Six Cancers: A Mendelian Randomization and Prospective Cohort Study

Daniela Mariosa[1]*, Karl Smith-Byrne[1], Tom G. Richardson[2], Paul Brennan[1], Mattias Johansson[1]
[1]*International Agency for Research on Cancer (IARC/WHO), Genomic Epidemiology Branch, Lyon, France;* [2]*MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom*

Obesity is a risk factor for several cancers; however, it is unclear whether body weight in early life affects cancer risk independently of body weight later in life.

To estimate the independent associations of early life body weight and adult body weight for six obesity-related cancers, we performed i) two-sample Mendelian randomization (MR) analyses using large cancer genome-wide association studies and ii) a prospective cohort analysis of 185,361 participants from the European Prospective Investigation into Cancer and Nutrition (EPIC).

Both the MR and longitudinal analyses indicated that larger early life body size was associated with higher risk of endometrial (OR_{MR}=1.61, 95% CI=1.23–2.11 for children larger than average compared to average) and kidney cancer (OR_{MR}=1.40, 95% CI=1.09–1.80). These associations were attenuated after accounting for adult body size in both the MR and cohort analyses. Early life body size was not consistently associated with the other investigated cancers.
The lack of clear independent risk associations suggests that early life body mass index influences endometrial and kidney cancer risk mainly through pathways that are common with adult body mass index.

107

Comparing Methods to Adjust for Fine-Scale Population Structure in Rare Variant Analyses

Katie M. Marker[1,2,*], Ruhollah Shemirani[3], Meng Lin[1], Eimear E. Kenny[3], Gillian M. Belbin[3], Christopher R. Gignoux[1,2]
[1]*Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America;* [2]*Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America;* [3]*Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America*

Previous simulations have shown, when there is a difference in mean trait value or disease prevalence between subpopulations, fine-scale genetic differences in rare variants can induce population structure that cannot be corrected with traditional approaches. This is increasingly a problem as very large sample sizes are needed to detect rare variant associations, datasets from multiple studies will be combined or external controls will be used to increase sample size. If fine-scale population structure is not properly adjusted for, fine-scale ancestry differences between cohorts will induce population structure bias that can confound rare variant association studies. We compare four methods for capturing and adjusting for fine-scale population differences; these include identity-by-descent (IBD) clusters, kinship clusters, PCA using only rare variants, and Uniform Manifold Approximation and Projection (UMAP) projections. To accomplish this, we have developed a simulation framework to generate 9 populations from a continental European demographic model with stepping stone migration using msprime. The differences within these populations replicate fine-scale ancestry found within the White British population in the UK Biobank. Using a phenotype simulator, APRICOT, we simulate phenotypes under varying geographic conditions to induce fine-scale population structure bias. Finally, we run rare variant association studies using these simulations to compare adjustment by IBD clusters, kinship clusters, rare variant PCA, and UMAP projections.

108

Using a Population-specific Reference Panel Improves Genotype Imputation Accuracy in Individuals of African Ancestry

Richard Mayanja[1,2,*], Abram B. Kamiza[1,3], Opeyemi Soremekun[1,4], Konstantinos Hatzikotoulas[6], William N. Rayner[6], Eleftheria Zeggini[6], Andrew P. Morris[7], Mia Crampin[3], Tinashe Chikowore[8,9], Segun Fatumo[1,4,5,6]
[1]*The African Computational Genomics (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda;* [2]*College of Health Sciences, Makerere University, Kampala Uganda;*

[3]Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi; [4]Medical Research Council/ Uganda Virus Research Institute/London School of Hygiene and Tropical Medicine (MRC/UVRI/LSHTM) Uganda research unit, Entebbe, Uganda; [5]London School of Hygiene and Tropical Medicine London, United Kingdom; [6]Institute of Translational Genomics, Helmholtz Zentrum München, Germany; [7]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Manchester, UK; [8]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [9]MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Genotype imputation uses densely genotyped haplotypes to predict genotypes of untyped variants in a target dataset. It boosts the power of genome-wide association studies (GWAS) and can be used for meta-analysis, fine mapping, and developing genetic risk scores. Factors such as the reference panel's representation of the target population influence the quality of imputation.

We investigated the best performing imputation reference panel in people of African ancestry by comparing the performance of the five most commonly used imputation reference panels containing African ancestry data (Trans omics for Precision Medicine (TopMed), Haplotype Reference Consortium (HRC), 1000 Genomes Project (1000G), Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA), and African Genome Resource (AGR)). Data from 288 Malawi Epidemiology Intervention Research Unit (MEIRU) cohort members who were genotyped on the H3A genotyping array were used in this study.

The total number of SNPs imputed with an imputation quality > 0.3 and mean imputation quality scores across a range of minor allele frequency (MAF) bins were used to evaluate imputation performance. TopMed and AGR reference panels imputed the most SNPs (73,553,040 and 26,131,641, respectively), while HRC reference panel imputed the fewest (17,998,233). TopMed and AGR had similar mean r2 values and outperformed all other reference panels with MAF > 0.2. However, AGR had a higher mean r2 than all other reference panels for SNPs with MAF >0.2.

In conclusion, when performing genotype imputation on people of African ancestry, AGR and ToPMed should be prioritized to improve imputation coverage and accuracy.

109

Exploring the Effects of the Maternal and Fetal Proteome on Birthweight: A Mendelian Randomization Study

Nancy McBride[1,2]*, Alba Fernández-Sanlés[1,2], Marwa AL Arab[1,2], Robin N. Beaumont[3], Chris (Jie) Zheng[1,2], Tom Gaunt[1,2], Tom Bond[1,2], Rachel M. Freathy[3], Deborah A. Lawlor[1,2], Maria C. Borges[1,2]

[1]MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK; [2]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK; [3]Institute of Biomedical and Clinical Science, College of Medicine and Health, University of Exeter, Exeter, UK

Birthweight is a widely used proxy for fetal growth, a valuable indicator of offspring perinatal health. Human genetics provides evidence of both maternal and fetal contributions to birthweight, but the causal relationships between maternal or fetal protein levels and birthweight are unclear. Here, we used two-sample Mendelian Randomization (MR) to explore the effects of the maternal and fetal plasma proteome on birthweight.

We used summary data from the Early Growth Genetics (EGG) consortium genome-wide association study (GWAS) of birthweight (N= 406,063 with maternal and/or fetal genotype), in which independent maternal and fetal effects were estimated. We aimed to identify effects of protein quantitative trait loci (pQTL) on birthweight pQTLs were selected from previous GWAS (959 cis-pQTL of 751 proteins, 315 trans-pQTL of 244 proteins and 209 cis-and-trans-pQTL for 81 proteins). Maternal levels of six proteins (instrumented by two cis-pQTL, two trans-pQTL and two cis-and-trans-pQTL) were causally related to offspring birthweight. For example, we found evidence that higher levels of PCSK1 potentially cause higher birthweight. Genetically instrumented fetal levels of three proteins (instrumented by one cis-pQTL, and two cis-and-trans-pQTL) were associated with birthweight. For example genetically higher fetal LEPR was associated with lower birthweight. These proteins are examples of biologically plausible candidates.

Future work will undertake analyses to explore pleiotropy, confounding by linkage disequilibrium (LD), enrichment for Mendelian disorders, replication in an independent sample (N = 30,000 mother-child pairs) and validation of the pQTLs, using proteomic data collected in pregnancy (N=4,000).

110

Robust Inference of Gene-Environment Interaction from Heterogeneous Samples of Case-Parent Trios

Pulindu Ratnasekera[1], Brad McNeney[1,*]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada

In a case-parent trio study we collect genotypes on affected children and their parents. Information may also be collected on the child's environmental exposures. The design permits estimation and testing of genetic effects and gene-by-environment interaction. Inference of genetic effects is robust to population structure, but when genotypes are measured at a non-causal test locus, population stratification can create spurious interaction. That is, the exposure can appear to modify the disease risk of genotypes at the test locus without actually modifying the disease risk of genotypes at the causal locus. We review previous methods to reduce bias from population stratification and propose a new method in which we adjust the risk model by principal components computed from a genome-wide panel of markers. The method is illustrated on simulated data and on data from a study of genetic modifiers of exposures known to affect the risk of cleft palate.

111

Estimating the Genetic Relationship Between Psychiatric- and Cardiometabolic Traits Using the Large National Patient Registers of Denmark and Sweden

Joeri Meijsen[1,2]*, Kejia Hu[5], Raquel Nogueira Avelar E Silva[1,2], Yorgos Athanasiadis[1,2], Richard Zetterberg[1,2], John Shorter[1,2], CoMorMent Consortium, Fang Fang[5], Thomas Werge[1,2,3,4], Alfonso Buil[1,2]

[1]Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark; [2]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen, Denmark; [3]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark; [4]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [5]Institute of Environmental Medicine, Karolinska Institute, Stockholm, Sweden

Psychiatric disorders are leading causes of morbidity, mortality, and disability worldwide. Individuals with severe psychiatric disorders such as Affective Disorder (AFF), bipolar disorder (BIP), attention deficit hyperactivity disorder (ADHD), and schizophrenia (SCZ) present substantial higher rates of mortality as compared to the general population, mainly due to somatic comorbidities such as cardiometabolic diseases. The largest epidemiological study to date using Danish national registries (5.9 million persons) observed a substantial comorbidity and absolute lifetime risk for cardio-metabolic diseases in psychiatric disorders. These results do not show how biological and environmental risk factors contribute to this comorbidity.

In this study we leverage the unprecedented data unique to Scandinavia to examine the genetic association between six psychiatric- and 14 cardio-metabolic disorders using: the genealogies and national patient register of millions of individuals in Denmark and Sweden, polygenic scores (PRS), and LD-Score Regression (LDSC).

First, we calculated the cumulative incidence based genetic correlation between participants and, their parents and full siblings using the Danish and Swedish National Registries. We observed strong genetic correlations between specific combinations of cardiometabolic and psychiatric outcomes. Furthermore, we compared these results against iPSYCH PRS and LDSC estimates and observed that any cardiometabolic diseases are significantly associated with ADHD, and to a lesser extend to SCZ, ASD, and AFF. Moreover, many PRS associations were observed to be independent of the individual's psychiatric predisposition.

This study represents the most exhaustive study to date investigating the pleiotropy between psychiatric- and cardiometabolic disorders comparing multiple methods in two population representative cohort.

112

Individual-specific Networks and Representation Learning to Capture Dynamics of Microbiome interactions

Behnam Yousefi[1,2,3*], Federico Melograna[3*], Gianluca Galazzo[4], Niels van Best[5,6], Monique Mommers[6], John Penders[6,7] Benno Schwikowski[1], Kristel Van Steen[3,8]

[1]Systems Biology Group, Department of Computational Biology, Institut Pasteur, Paris, France; [2]Sorbonne Universite, École Doctorale Complexite du vivant, Paris, France; [3]BIO3 – Laboratory for Systems Medicine, KU Leuven, Belgium; [4]School of Nutrition and Translational Research in Metabolism (NUTRIM), Department of Medical Microbiology, Maastricht University, Maastricht, The Netherlands; [5]Institute of Medical Microbiology, RWTH University Hospital Aachen, RWTH University, Aachen, Germany; [6]Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands; [7]Care and Public Health Research Institute (CAPHRI), Department of Medical Microbiology, Maastricht University, Maastricht, The Netherlands; [8]BIO3 – Laboratory for Systems Genetics, GIGA-R Medical Genomics, University of Liège, Liège, Belgium

[*]Equal contribution

Background: Tracking multivariate microbiome profiles over time or across conditions remains a daunting task. Often, available statistical tools and methods fail to accommodate individual-specific dynamics of microbial interactions. Such dynamics may hold relevant information about health-to-disease transitions.

Materials and Methods: We present a novel framework that combines representation learning and individual-specific microbiome co-occurrence networks to capture individual-specific microbial interaction dynamics. In particular, the framework consists of first inferring cross-sectional sparse microbial association networks, and then constructing individual-specific networks (ISNs), following Kuijjer's (2019) strategy. ISNs are defined as networks of nodes and edges unique to each individual. A novel developed shallow encoder-decoder neural network forms the core of our framework, referred to as Multiplex Network Differential Analysis framework (MNDA).

Results: We applied MNDA to a subset of the LucKi birth cohort. Our multiplex-ISN embedding allows quantifying variations over time in local ISN neighborhoods for each microbe. Used in SVM models, at least as good prediction performance could be achieved compared to standard microbial abundance or ISN edges: AUCs of 0.767 and 0.710, when modelling mode of delivery and diet, respectively, compared to AUCs within 0.407-0.570 for mode of delivery with ISN edge weights as cross-sectional predictors. Finally, MNDA-based similarity measures to cluster individuals into homogeneous groups according to similar microbial neighborhood dynamics were significantly different from, and complementary to, more traditional Dirichlet Multinomial Mixtures clustering.

Conclusion: The alternative view of MNDA to extract information from matched microbiome profiles opens new avenues to personalized prediction or stratified medicine with temporal microbiome data.

113

Multi-tissue Transcriptome-wide Association Study Identifies 12 Novel Candidate Genes Associated with the Immune Traits in Cancer

Pooja Middha[1*], Rosalyn W. Sayaman[2], Mohamad Saad[3,4], Vésteinn Thorsson[5], Davide Bedognetti[6,7,8], Elad Ziv[9]
[1]Department of Medicine, University of California, San Francisco, San Francisco, California, United States of America; [2]Department of Laboratory Medicine, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, United States of America; [3]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [4]Neuroscience Research Center, Faculty of Medical Sciences, Lebanese University, Beirut, Lebanon; [5]Institute for Systems Biology, Seattle, Washington, United States of America; [6]Research Branch, Sidra Medicine, Doha, Qatar; [7]College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar; [8]Department of Internal Medicine and Medical Specialties, University of Genoa, Genoa, Italy; [9]Department of Medicine, Institute for Human Genetics, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, United States of America

Immune infiltration in solid tumors is a strong predictor of improved survival in many tumor types. We have recently demonstrated the heritability of immune infiltration components in solid tumors by using germline genetic data from The Cancer Genome Atlas (TCGA). We also identified over 20 individual loci that are associated with over 10 immune signatures including genes that are involved in autoimmune disorders and genes that are proposed targets of immunotherapy. Here, we sought to identify additional genes associated with variation in the immune microenvironment. We performed transcriptome-wide association study (TWAS) to predict immune signatures in the pan-cancer analyses of 30 non-hematological cancers in TCGA.

We integrated the results from the pan-cancer genome-wide association study with large-scale expression quantitative trait loci (eQTLs) from whole blood, spleen, and EBV-transformed lymphocytes tissues in GTEx (version 8). We conducted a TWAS using the Summary-MulTiXcan approach. We used an FDR *P* value of <0.05 to adjust for multiple hypothesis testing. Cis-eQTLs (genes that are also part of the immune signature) were excluded in order to prioritize genes that are most likely functionally driving the immune microenvironment.

After exclusion of cis-eQTLs, we identified 12 novel genes (*ZBTB8OS, RAB43, TNNT1, TSSK3, ZNF134, SAP30B, LRFN3, UPK3BL, EPHB6, GKAP1, NT5C,* and *RP11-290F24.6*) whose genetically predicted expression was associated with different immune signatures in tumors. One of these genes, *EPHB6* was found to be inversely associated with a Th1 enrichment (P=2.2x10^{-6}). EphB6 plays a crucial role in T-cell activation with EphB6-deficient mice displaying reduced activation, phosphorylation and recruitment of the T cell signaling molecules. *RAB43* was found to be inversely associated with activated a signature of natural killer cells (P=3.8x10^{-7}). Rab proteins are involved in signal transduction pathways regulating cell invasion, cell apoptosis and innate immune response, particularly in gliomas, leading to poor clinical outcomes.

This is the first TWAS investigating the relationship between genetically predicted gene expression and immune traits. Of these, *EPHB6,* and *RAB3* are strong candidates for a mechanistic role in modifying the immune response to tumors.

114

Non-parametric de Novo Network Identification of Gene-Environment Interactions Affecting Asthma Symptom Control

Joshua Millstein[1*], Sung Chun[2,3], Ahmad Samiei[2], Benjamin Raby[2]
[1]Division of Biostatistics, Keck School of Medicine of USC, Los Angeles, California, United States of America; [2]Division of Pulmonary Medicine, Boston Children's Hospital, Boston, Massachusetts, United States of America; [3]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States of America

Many approaches have been developed and applied in recent decades to identify statistical interactions between genes and environmental exposures that affect susceptibility to complex diseases. Although some evidence of such effects has been found, the discoveries to date have not lived up to early expectations. This lack of evidence is surprising considering the highly interconnected relationships between genes in both regulation and function. Part of the explanation may involve the notoriously low statistical power of conventional hypothesis tests for interaction. We propose a novel permutation-based approach that could increase statistical power by combining evidence of association across multiple gene-gene or gene-environment pairs. If we think of these pairs as links in a network, then we would hypothesis that connected network components are more likely to occur in observed as compared to permuted data, where the disease response is randomly permuted. A novel statistic will be described in which direct evidence of association is combined with evidence due to connected network component structure complexity. The steps of the approach are as follows, 1) conduct first-order interaction analysis, 2) identify 'interesting' pairs using a lenient discovery threshold, 3) identify connected network components, 4) compute the statistic for each component, 5) repeat steps 1-4 in permuted data, 6) compute false discovery rates (FDR) for each component. An application to the Asthma Biorepository for Integrative Genomic Exploration (Asthma BRIDGE) study to identify bipartite structures between SNP variation and environmental exposures affecting asthma symptom control will be discussed.

115

Epigenetic Aging and Colorectal Cancer Survival: A Mendelian Randomization Study

Fernanda Morales Berstein[1,2*], Daniel L. McCartney[3], Ake T. Lu[4], Carolina Borges[1,2], George Davey Smith[1,2], Steve Horvath[4,5],

Riccardo E. Marioni[3], Tom G. Richardson[1,2,6], Rebecca C. Richmond[1,2], Caroline L. Relton[1,2]
[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, Bristol, United Kingdom; [3]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom; [4]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, United States; [5]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, United States; [6]Novo Nordisk Research Centre, Oxford, United Kingdom

Background: Observational and Mendelian randomization (MR) studies suggest that epigenetic ageing may increase colorectal cancer (CRC) risk. In contrast, the influence of epigenetic ageing on CRC survival remains uncertain. Although observational evidence suggests that some epigenetic clocks may be better predictors of CRC prognosis than others, it is less clear whether they influence CRC progression.

Methods: We used a two-sample MR framework to investigate the effect of genetically predicted epigenetic age acceleration as measured by GrimAge (four SNPs), PhenoAge (11 SNPs), HannumAge (nine SNPs) and Intrinsic HorvathAge (24 SNPs) on CRC survival. Genome-wide association data for biological ageing were obtained from a meta-analysis (N=34,710) and for CRC survival from the International Survival Analysis in Colorectal Cancer Consortium (ISACC) (N deaths=3,586 of 16,168 CRC cases). We conducted the main analyses using the inverse variance weighted (IVW) MR method and applied the MR-Egger, weighted median and weighted mode as sensitivity analyses.

Results: We found no convincing evidence of an effect of genetically predicted epigenetic age acceleration on CRC survival (GrimAge IVW OR= 1.04 per year increase in epigenetic age acceleration, 95% CI 0.92–1.18; PhenoAge OR= 1.01, 95% CI 0.96–1.07; HannumAge OR= 1.12, 95% CI 0.99–1.26; Intrinsic HorvathAge OR= 1.03, 95% CI 0.99–1.08).

Conclusions: Our findings suggest that epigenetic ageing does not play a role in CRC survival. As prior evidence suggests that it may influence CRC risk, this study highlights the importance of intervening on factors that influence epigenetic age acceleration before rather than after CRC diagnosis.

116

Evaluation of Rare Variant Association Methods When Incorporating External Controls

Jessica I. Murphy[1]*, Megan Null[2], Christopher R. Gignoux[3,4], Audrey E. Hendricks[1,4,5]
[1]Department of Biostatistics and Informatics, University of Colorado School of Public Health, Aurora, Colorado, United States of America; [2]Department of Mathematics and Physical Sciences, The College of Idaho, Caldwell, Idaho, United States of America; [3]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [4]Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [5]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, United States of America

Identification of rare variant associations is crucial to understanding the genetic contribution to complex traits and diseases. Although individual studies are often smaller and thus have low power to detect rare variant associations, large publicly available datasets can be leveraged as external controls to increase power. However, these large datasets often differ in characteristics such as sample ascertainment, ancestry, and processing, which can lead to biased results if not accounted for appropriately. Here, we evaluate the performance of rare variant association methods designed for common controls (e.g., iECAT-O, ProxECAT, ProxECAT v2) across a variety of simulation scenarios. We simulate using RAREsim, a simulation framework that emulates the distribution of rare variants, functional annotation, and haplotype structure seen in real data. By identifying the optimal method(s) across the simulation scenarios, we increase the utility of publicly available genetic resources for use as external controls. Our comprehensive simulation study along with best practice guidelines for incorporating external control data will aid in the discovery of new genetic associations.

117

Pathogenic Variants in Breast Cancer Susceptibility Genes and Polygenic Risk among US Latinas and Mexican Women

Jovia L. Nierenberg[1,2]*, Esther M. John[3,4,5], Gabriela Torres-Mejia[6], Christopher A. Haiman[7], Lawrence H. Kushi[8], Stephen Gruber[9], Jeffrey N. Weitzel, Laura Fejerman[10,11], Elad Ziv[2], and Susan L. Neuhausen[9]
[1]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America; [2]Department of Medicine, University of California San Francisco, San Francisco, California, United States of America; [3]Department of Epidemiology & Population Health, Stanford University School of Medicine, Stanford, California, United States of America; [4]Department of Medicine, Stanford University School of Medicine, Stanford, California, United States of America; [5]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, United States of America; [6]Instituto Nacional de Salud Pública, Cuernavaca, Mexico; [7]Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [8]Division of Research, Kaiser Permanente Northern California, Oakland, California, United States of America; [9]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, California, United States of America; [10]Department of Public Health Service, University of California, Davis, Davis, California, United States of America; [11]UC Davis Comprehensive Cancer Center, University of California, Davis, Davis, California, United States of America

Introduction: Few breast cancer studies have examined polygenic risk scores (PRS) and pathogenic variants (PVs)

in susceptibility genes jointly. Here, we report the relation between PVs and PRS in Latinas.

Methods: In a pooled case-control analysis of breast cancer in Latinas from California and Mexico (1,776 cases, 1,589 controls), we examined a 180-SNP PRS from known breast cancer SNPs and presence of a PV in 9 breast cancer risk genes (*ATM*, *BARD1*, *BRCA1*, *BRCA2*, *CHEK2*, *PALB2*, *PTEN*, *RAD51C*, and *TP53*). We used logistic regression to examine associations of PVs and PRS with breast cancer risk. We next evaluated, among cases and controls separately, if PRS was associated with the odds of having a PV. Analyses were adjusted for age, study, and ancestry.

Results: Women with a PV had higher risk of breast cancer (OR=5.9, 95% CI: 3.8-9.6) as did women with a higher PRS (OR per SD=1.5, 95% CI: 1.5-1.7). Among cases with PRS below the median, the odds of having any PV were increased by 1.9-fold (95% CI: 1.4-2.7), a *BRCA2* PV by 2.4-fold (95% CI: 1.2-5.1), and a *PALB2* PV by 3.2-fold (95% CI: 1.3-9.9), with consistent effect directions for 6 of the 7 genes where at least 2 participants have a PV. No associations were found among controls.

Conclusion: PVs were more prevalent among cases with low PRS, likely due to collider bias by different types of genetic risk factors when conditioning on case status. Sequencing cases with low PRS could facilitate identification of novel disease-associated genes.

118

The Relationship between Adiposity and Cognitive Function: A Bidirectional Mendelian Randomization Study in UK Biobank

Tom Norris[1]*, Ghazaleh Fatemifar[2], Spiros Denaxas[2], Chris Finan[3,4,5], Victoria Garfield[3*] & Snehal M. Pinto Pereira[1*]
[1]Institute of Sport, Exercise and Health, Division of Surgery & Interventional Science, University College London, London, United Kingdom; [2]Institute of Health Informatics, University College London, London, United Kingdom; [3]Institute of Cardiovascular Science, University College London, London, United Kingdom; [4]UCL British Heart Foundation Research Accelerator; [5]Department of Cardiology, Division Heart and Lungs, University Medical Centre Utrecht, Netherlands
[*]Joint senior authors

Background: We aimed to determine whether reported bidirectional relationships between cognitive function are causal by undertaking a bidirectional Mendelian Randomization (MR) study.

Methods: 378,877 UK Biobank participants, with three indicators of adiposity (body fat percentage (BF%), body mass index (BMI), waist-hip-ratio (WHR)) and two indicators of cognitive function (reaction time (RT) and visual memory (VM)), were included. Using bidirectional MR, we estimated the strength of the adiposity-cognitive function association using genetic instruments for each adiposity indicator as our exposure and repeated this in the opposite direction using genetic instruments for each cognitive function indicator.

Results: Observational analyses indicated that higher BMI and WHR were associated with better cognitive function. Higher BF% was associated with a worse RT and a better VM. MR analyses were generally consistent with observational findings (albeit non-significant). In the direction cognitive function to adiposity, observational analyses indicated that worse RT was associated with higher BF% and lower WHR and BMI; worse VM was associated with lower BF%, WHR and BMI. MR estimates of RT for all three adiposity measures were imprecise and directionally inconsistent. MR estimates for the effect of VM on adiposity indicated that worse VM was associated with higher adiposity, e.g., a 1-unit worse VM score was associated with a 3.64% (95% CI: 1.87,5.43) higher BMI.

Conclusions: Observational associations of adiposity on cognitive function are likely not to be causal. In the reverse direction, worse VM was associated with greater BF%, BMI and WHR, providing support for a causal link between VM and adiposity.

119

Classification of Variants of Uncertain Significance (VUS) in the *BRCA1* and *BRCA2* Genes Based on Ovarian Tumor Pathology Characteristics

Denise O' Mahony[1,2]*, Susan J. Ramus[3,4], Melissa C. Southey[5,6,7], Andreas Hadjisavvas[2], ENIGMA[8], CIMBA[9], AOCS[10], OTTA[11], Antonis C. Antoniou[12], David E. Goldgar[13], Amanda B. Spurdle[14], Kyriaki Michailidou[1]
[1]Biostatistics Unit, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; [2]Cancer Genetics, Therapeutics & Ultrastructural Pathology Department, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; [3]School of Women's and Children's Health, Faculty of Medicine, University of New South Wales, Sydney, Australia; [4]Adult Cancer Program, Lowy Cancer Research Centre, University of New South Wales, Sydney, Australia [5]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia; [6]Department of Clinical Pathology, University of Melbourne, Melbourne, Victoria, Australia; [7]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Australia; [8]Evidence-based Network consortium for the Interpretation of Germline Mutant Alleles (ENIGMA) ; [9]Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA); [10]Australian Ovarian Cancer Study Group (AOCS); [11]Ovarian Tumor Tissue Analysis (OTTA); [12]Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK; [13]Huntsman Cancer Institute, University of Utah School of Medicine, Department of Dermatology, Salt Lake City, Utah, United States of America; [14]Division of Genetics and Population Health, QIMR Berghofer Medical Research Institute, Brisbane, Australia

Introduction: Ovarian cancer histopathology and other tumor characteristics exhibit distinct distributions in *BRCA1* and *BRCA2* pathogenic variant carriers compared to non-carriers. These characteristics can be used as evidence to aid the interpretation of *BRCA1* and *BRCA2* variants of uncertain significance, identified by genetic testing, which are not currently considered in existing variant interpretation models. We aimed to assess ovarian cancer characteristics as predictors of *BRCA* variant pathogenicity, for use as clinical points of the ACMG/AMP variant classification system.

Methods: Ovarian cancer samples were collected from the CIMBA, OTTA and AOCS consortia and ENIGMA consortium collaborators, including germline *BRCA1* (2,068) and *BRCA2* (786) carriers and non-carriers (4,480). The histopathology subtypes defined included, serous carcinomas of high-grade (HGSC), low-grade (LGSC) and unknown-grade (UGSC), mucinous, endometrioid, clear-cell and 'other' carcinomas. Ovarian tumor histopathology was associated with *BRCA1* and *BRCA2* variant pathogenicity by Likelihood Ratio (LR) calculations and estimates were aligned to ACMG/AMP evidence strengths.

Results: Under the ACMG/AMP Bayesian framework, supporting benign strength was assigned to LGSC histopathology association with *BRCA1* (LR_{BRCA1}=0.40) and mucinous histopathology with *BRCA2* (LR_{BRCA2}=0.37). The association of mucinous tumors with *BRCA1* (LR_{BRCA1}=0.21) and clear-cell tumors with *BRCA1* (LR_{BRCA1}=0.21) and *BRCA2* (LR_{BRCA2}=0.17) reached moderate benign evidence. Supporting pathogenic evidence weighted the association of UGSC with *BRCA2* (LR_{BRCA2}=2.23). Refined LR estimates were also calculated for histopathology in combination with tumor grade, behavior and age diagnosis.

Conclusions: We provide detailed estimates for predicting *BRCA1* and *BRCA2* variant pathogenicity based on ovarian tumor characteristics, to aid variant interpretation in combination with other evidence.

120

Leveraging Healthy Population Data to Assess the Pathogenicity of Rare Variants in WGS: Extension of PSAP Method to the Non-coding Genome

Marie-Sophie C. Ogloblinsky[1]*, Ozvan Bocher[1], Chaker Aloui[2], Elisabeth Tournier-Lasserve[2,3], Donald F. Conrad[4], Emmanuelle Génin[1,5], Gaëlle Marenne[1]

[1]Inserm, Université de Brest, Établissement *Français du Sang, Unité Mixte de Recherche 1078, Génétique, G*énomique fonctionnelle et *Biotechnologies, F-29200 Brest, France; [2]Université de Paris, NeuroDiderot, Inserm Unité Mixte de Recherche 1141, F-75019 Paris, France; [3]Assistance publique-Hôpitaux de Paris, Service de Génétique Moléculaire Neurovasculaire, Hôpital Saint-Louis, F-75010 Paris, France; [4]Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Portland, Oregon, United States of America; [5]Centre Hospitalier Régional Universitaire de Brest, F-29200 Brest, France*

For many rare diseases, different genes and different variants within these genes can be involved resulting in extreme situations where each affected individual could carry a specific pathogenic variant. In such situations of high heterogeneity, it is extremely difficult to identify the molecular causes of disease using sequencing data and traditional methods of analysis. This leads to a very low diagnosis rate. The PSAP (Population Sampling Probability) method was developed to overcome this issue by assessing the probability of observing a given genotype in a healthy population. This framework is based on gene-specific null distributions of CADD pathogenicity scores calibrated using allele frequencies from the GnomAD database.

Here, we propose an extension of the PSAP method to the non-coding genome by using some predefined genomic regions as testing units. These regions, referred to as "CADD regions", span the entire genome and their boundaries reflect the functionality of variants within the region according to their CADD scores. The novelty of our method is to broaden the spectrum of variants detectable by PSAP, especially in introns and splicing regions, but also to improve the performance of PSAP for cases where pathogenic variants are located in a more constrained sub-region of the gene. We demonstrate the validity of our approach by reproducing signals already found in exome data. We also show how the method performs on a novel cohort of patients with genetically unresolved cases of Cerebral Small Vessel Disease (CSVD), a very heterogeneous disease that accounts for 25% of ischaemic strokes.

121

Meta-regression of Cross-sectional GWAS Studies to Estimate Trajectories of Genetic Effects

Panagiota Pagoni[1,2*], Julian P T Higgins[2,3], Deborah A. Lawlor[1,2,3], Evie Stergiakouli[1,2], Nicole M. Warrington[4,5], Kate Tilling[1,2],Tim Morris[1,2]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3] National Institute for Health Research Bristol Biomedical Research Centre, University of Bristol, Bristol, United Kingdom; [4] University of Queensland Diamantina Institute, University of Queensland, Brisbane, QLD Australia; [5] K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

Genome–wide association studies (GWAS) test associations of millions of single nucleotide polymorphisms (SNPs) across the whole genome with one or more phenotypes. Fixed-effect meta-analysis, which assumes a common true underlying genetic effect for all studies, has been extensively used to summarize genetic effects across multiple GWAS. However, fixed-effect meta-analysis ignores heterogeneity of genetic effects between studies. Evidence suggest that genetic effects may not stay constant over time, therefore meta-analyzing GWAS of age-diverse samples with a fixed-effect model, without considering potential heterogeneity of genetic effects due to age, could produce misleading results. Meta-regression extends meta-analysis, allowing adjustment for study specific characteristics and the modelling of heterogeneity between studies. If heterogeneity exists due to differences in participants' ages between studies, meta-regression could be used to recover and estimate age-dependent effects of genetic variants. We compared the performance of meta-regression to meta-analysis in accurately summarizing (i) main genetic effects and (ii) age dependent genetic effects (SNP by age interactions) using multiple cross-sectional GWAS studies, and under a range of scenarios. Fixed-

effect and random-effects meta-analysis accurately estimate genetic effects when these are unrelated to age but produce biased estimates when age-dependent genetic effects exist. Meta-regression produces unbiased estimates of both the main genetic effects and the age-dependent genetic effects, regardless of the dependence of genetic effects on age, when overlap of age ranges between studies are not greater than 75%.

122

DeLIVR: A Deep Learning Approach to Testing for Non-linear Causal Effects in Transcriptome-Wide Association Studies

Ruoyu He, Mingyang Liu, Zhaotong Lin, Zhong Zhuang, Xiaotong Shen, Wei Pan*
University of Minnesota, Minnesota, United States of America
*Email: panxx014@umn.edu

Transcriptome-wide Association Studies (TWAS) have been increasingly applied to identify (putative) causal genes for complex traits and diseases. TWAS can be regarded as a two-sample two-stage least squares (2SLS) method for instrumental variable (IV) regression for causal inference. The standard TWAS (called TWAS-L) only considers a linear relationship between a gene's expression and a trait in stage 2, which may lose statistical power when not true. Recently an extension of TWAS (called TWAS-LQ) considers both a linear and a quadratic effect of a gene on a trait, which however is not flexible due to its parametric nature and may be low-powered for non-quadratic non-linear effects. On the other hand, a deep learning (DL) approach, called DeepIV, has been proposed to non-parametrically model a non-linear effect in IV regression. However it is both slow and unstable due to the ill-posed problem of solving an integral equation with Monte Carlo approximations. Furthermore, in the original DeepIV approach statistical inference such as hypothesis testing was not studied. Here we propose a modification to DeepIV, called DeLIVR, and a hypothesis testing framework. We show through simulations that DeLIVR was both faster and more stable than DeepIV. We applied both parametric and DL approaches to GTEx and UK Biobank data, showcasing that DeLIVR detected additional 11-54 genes associated with high-density lipoprotein cholesterol (HDL-C) or low-density lipoprotein cholesterol (LDL-C) that would be missed by TWAS-L or TWAS-LQ. In particular, DeLIVR detected genes NT5DC2 and BUD13 associated with HDL-C, and HP with LDL-C, all of which were supported by previous studies but would be missed by TWAS-L, TWAS-LQ and DeepIV.

123

SLCO5A1 and Synaptic Assembly Genes Contribute to Impulsivity in Juvenile Myoclonic Epilepsy

Naim Panjwani[1]*, Amy Shakeshaft[2,3], Delnaz Roshandel[1], Annalisa Pastore[2], Manolis Fanto[2], Deb K. Pal[2,3,4] & Lisa J. Strug[1,5,6], and the BIOJUME Consortium
[1]Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada; [2]Department of Basic & Clinical Neurosciences, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom; [3]MRC Centre for Neurodevelopmental Disorders, King's College London, London, United Kingdom; [4]King's College Hospital, London, United Kingdom; [5]Departments of Statistical Sciences and Computer Science and Division of Biostatistics, The University of Toronto, Toronto, Canada; [6]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada

Elevated impulsivity is a key component of attention deficit hyperactivity disorder (ADHD), bipolar disorder and epilepsy. We performed a genome wide association, colocalization and pathway analysis of impulsivity in juvenile myoclonic epilepsy (JME). We identify genome wide associated SNPs at 8q13.3 (P=7.5 × 10^{-9}) and 10p11.21 (P=3.6 × 10^{-8}). The 8q13.3 locus colocalizes with *SLCO5A1* expression quantitative trait loci in cerebral cortex (P=9.5 × 10^{-3}). *SLCO5A1* codes for a membrane bound organic anion transporter and upregulates synapse assembly and organization genes. Pathway analysis also demonstrates 9.3 fold enrichment for synaptic assembly genes (P=0.03) including *NRXN1*, *NLGN1* and *PTPRD*. RNAi knockdown of *Oatp30B*, the *Drosophila* homolog of *SLCO5A1*, causes both overreactive startling behaviour (P=8.7 x 10^{-3}) and increased seizure like events (P=6.8 x 10^{-7}). Polygenic risk score for ADHD correlates with impulsivity scores (P=1.60 x 10^{-3}), demonstrating shared genetic contributions. *SLCO5A1* loss of function represents a novel impulsivity and seizure mechanism. Synaptic assembly genes may inform the aetiology of impulsivity in health and disease.

124

Features of X Chromosomal SNPs Associated with Significant Sex-difference in Allele Frequency in High Coverage Whole Genome Sequence Data

Zhong Wang*[1], Lei Sun[2,3], Andrew D. Paterson[3,4,5]
[1]Department of Statistics and Data Science, Faculty of Science, National University of Singapore, Singapore; [2]Department of Statistic Sciences, Faculty of Arts and Science, University of Toronto, Ontario, Canada; [3]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Ontario, Canada [4]Genetics and Genome Biology, The Hospital for Sick Children, Ontario, Canada; [5]Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Ontario, Canada

Recently, using multiple whole genome sequencing datasets we have shown that ~2% of SNPs on the X chromosome show significant sex differences in allele frequencies (sdMAF, P value <5x10^{-8}). We used the high coverage whole genome sequencing data from the 1000 Genomes project, as well as the non-Finnish Europeans (NFE) and African/African Americans from gnomAD v3.1.2. We found that SNPs near the boundaries of the pseudo-autosomal regions show strong sdMAF, likely due to sex linkage. However, there are hundreds of other SNPs in the non-pseudo-autosomal regions (NPR) that also show sdMAF, which may in part be due to bioinformatic reasons. We sought to identify features of the X chromosomal SNPs associated with sdMAF, focusing on common variants from the NPR. We analyzed 46,352 SNPs with call rate >93% and

sex-combined minor allele frequency (MAF) >5% using the 34,029 NFE subjects from gnomAD v 3.1.2. Initial univariate associations showed that more significant sdMAF was associated with (i) lower call rate, depth of coverage, quality, and HWE p value in females, and (ii) higher female than male MAF, proportion of reference allele=A (as opposed to C,G,T), and Combined Annotation Dependent Depletion (CADD) score. Multivariate analyses revealed that more significant HWE P value in females, lower call rate, high female MAF, and reference allele=A remained associated with more significant sdMAF ($p<10^{-3}$). Sensitivity analyses, linkage disequilibrium adjustment, replication in the African/African American sample from gnomAD v.3.1., and implications of these findings for genetic association analysis will be discussed.

125

Contribution of Rare Coding Variants to Complex Trait Heritability

Nazia Pathan[1, 2*], Mohammad Khan[1,2], Wei Q. Deng[3,4], Michael Chong[1,5,6], Matteo Di Scipio[1,2], Shihong Mao[1], Rob Morton[1,6], Marie Pigeyre[1,2], Ricky Lali[1,2], Guillaume Paré[1,2, 5, 6, 7]
[1]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada; [2]Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada; [3]Peter Boris Centre for Addictions Research, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada; [4]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Canada; [5]Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada; [6]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Canada; [7]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada

It has been postulated that rare variants (RVs; MAF<0.01) contribute to the "missing" heritability of complex traits. We developed a novel framework, the Rare variant heritability estimator (RARity), to estimate rare variant heritability (h^2_{RV}) for complex traits by using large scale multiple regressions, without assuming a particular genetic architecture. To understand the contribution and significance of RV to complex traits, we applied RARity to 31 continuous traits from the UK Biobank (N=167,348). 27 traits had estimated h^2_{RV}>5%, with height having the highest h^2_{RV} estimate at 21.9% (95% CI:19.0-24.8%). On the other hand, gene-level variant aggregation, widely used in gene-burden testing, led to an average loss of 80% (95% CI:68-93%) h^2_{RV} when compared to RARity. Importantly, the total estimated h^2 including both common and rare variants almost fully recovered pedigree-based estimates for height (87.8%) and BMI (39.5%). Gene-level analyses, using unaggregated variants, revealed 152 genes that were significantly enriched for h^2_{RV} in ~7000 biological pathways. 12 of these genes represent novel gene-phenotype relationships not previously reported. Finally, we leveraged RARity to assess whether existing *in silico* pathogenicity prediction (variant-level) and gene-level annotations are enriched for RVs that over-contribute to complex trait variance, but observe only modest supporting evidence. Together, these results confirmed that (1) RVs can account for a significant portion of the complex trait heritability, (2) gene-level rare variant aggregation leads to a substantial loss of information, (3) identification of genes significantly enriched for h^2_{RV} can help with gene discovery, and (4) novel methods are needed to predict variant-level functionality.

126

The Relationship Between Major Depressive Disorder and the Circadian System: A Mendelian Randomization Study in the UK Biobank

Valentina Paz[1, 2*], Dylan Williams[1], Marcus Richards[1], Bettina Tassino[3], Ana Silva[4] & Victoria Garfield[1]
[1]MRC Unit for Lifelong Health and Ageing, Institute of Cardiovascular Science, University College London, London, United Kingdom; [2]Instituto de Psicología Clínica, Facultad de Psicología, Universidad de la República, Montevideo, Uruguay; [3]Sección Etología, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay; [4]Laboratorio de Neurociencias, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay.

The sleep-wake cycle is the most conspicuous human circadian rhythm, and it is characterized by multiple dimensions such as sleep duration, quality, and timing or chronotype. Inadequate duration, poor quality and eveningness have been associated with various conditions in observational studies, including major depressive disorder. However, despite advances in this area, it is still uncertain whether these associations are causal or not. We will present results from bidirectional Mendelian randomization (MR) analyses of the relationship between major depressive disorder and multiple sleep dimensions: sleep duration (and short/long durations), insomnia (frequent and any symptoms) and chronotype (morning/evening person). Data will be from the UK Biobank, and Inverse-variance weighted MR will be implemented with sensitivity analyses, including MR-Egger and the Weighted Median Estimator for horizontal pleiotropy. Our findings will provide insights into the association between major depressive disorder and the circadian system, allowing a better understanding of some of the etiopathogenic processes that characterize this disorder and contribute to developing more effective prevention, diagnosis, and treatment strategies.

127

Investigating Genetic Effects on Clinical Heterogeneity in Major Depression: Symptoms, Subtypes, and Cardiometabolic Traits

Roseann E. Peterson[1,2*], Tim B. Bigdeli[2], Eva E. Lancaster[1], Bradley T. Webb[3], Kenneth S. Kendler[1]
[1]Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, School of Medicine, Virginia Commonwealth University, Richmond, Virginia, United States of America; [2]Department of Psychiatry and Behavioral Sciences, State University of New York, Downstate Health Sciences University, Brooklyn, New York, United States of America;

[3]GenOmics, Bioinformatics, and Translational Research Center, Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, North Carolina, United States of America.

Background: A diagnosis of Major Depression (MD) requires that at least 5 of 9 DSM accessory symptoms be present, although patients vary with respect to the combination of symptoms endorsed. Vegetative and reversed-vegetative symptoms, reflecting depression-related changes in weight, appetite, and sleep, seem to implicate energy balance and metabolism.

Methods: Using detailed clinical information from the CONVERGE study of MD in Han Chinese women (n=10,640), we consider the evidence in support of widespread pleiotropy between MD and a range of anthropometric traits. Subsequent genome-wide association studies (GWAS) employ a 'case-only' approach to identify associations between SNPs and symptom dimensions.

Results: Adverse metabolic outcomes such as obesity, CAD, and T2D showed negative genetic correlations with MD, as did *C*-reactive-protein levels. For sleep-related traits, we observed a positive genetic correlation between MD and insomnia. Within-case GWAS identified symptom specific signals for weight gain/loss at intergenic SNPs downstream of *SGK1*(P=2.37×10^{-9}). *SGK1* is a compelling candidate given an established role in stress response via glucocorticoid signaling and hippocampal functioning. An additional association was seen between a SNP in *SORCS2* and increased/decreased appetite(P=1.87×10^{-8}), lending additional support for hippocampal function and stress response as the encoded protein has been shown to facilitate BDNF-dependent synaptic plasticity. For reversed/vegetative symptoms, we observed associations over an extended region of MHC(P=6.1×10^{-9}).

Discussion: Both specific genetic factors and aggregate genetic effects influence clinical heterogeneity in MD. The functional relevance of associated loci and robustness of polygenic effects highlight the importance of studies focusing on genetic risk factors for MD subtypes.

128

An Evaluation of Race-specific Polygenic Risk Scores (PRS) for Uterine Fibroids Across Populations

Jacqueline A. Piekos[1]*, Jacklyn N. Hellwege[1,2], Ozan Dikilitas[3], Iftikhar J. Kullo[3], Daniel J. Schaid[4], David R. Crosslin[3], Dan Roden[1,5], Todd L. Edwards[1,6], Digna R. Velez Edwards[1,7,8]
[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of American; [2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America;[3]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota, United States of America; [4]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America; [5]Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Center, Nashville, Tennessee, United States of America; [6]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [7]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [8]Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Uterine fibroids (UF) are the most common benign pelvic tumor for women and are known to be heritable, complex traits with racial disparities in risk. Using two previous race-specific (European Ancestry [EA] and African Ancestry [AA]) genome-wide association studies' summary statistics and the polygenic risk score (PRS) software PRS-CSx, we generated separate PRSs for EA and AA populations and a meta-analyzed cross ancestry (CA) PRS. Each PRS was optimized in BioVU by adding a P value threshold (P_t) step, determined by the model with the largest pseudoR^2 from logistic regression. The optimal PRSs had the following P_t: EA P_t:0.1 (pseudoR^2=1.14x10^{-3}), AA P_t:1.0 (pseudoR^2=1.63x10^{-3}), and CA P_t:0.5 (pseudoR^2=8.02x10^{-3}). Each PRS was validated using 10-fold cross-validation in the Electronic Medical Records and Genomics (eMERGE) Network. Matching on race, we assessed PRS predictive utility using statistical methods of area under receiver operator curve (AUROC) and Delong's test for ROC curves. The 10-fold cross-validation produced the following AUROC values for EA: 0.70(95%CI:0.65-0.74), AA: 0.73(95%CI:0.61-0.84), and CA: 0.69(95%CI:0.64-0.74). Delong's test showed that the model UF~BMI+Age was significantly different from the model UF~BMI+Age+10PC+PRS in all racial groups (EA P=8.4x10^{-5}, AA P=3.9x10^{-3}, CA P=1.9x10^{-7}) but the predictive information of the model UF~PRS+10PCs is the same as that of the model UF~BMI+Age+10PC+PRS in all racial groups (EA P=0.70, AA P=0.78, CA P=0.32). The AA PRS explains the most variance within its racial group (pseusdoR2: AA: 0.16, EA: 0.032, and CA: 0.052). We can conclude that the predictive utility of our model is being driven by our PRS.

129

Identification of Brain Cell-Types Underlying Genetic Risk for Reading and Correlated Traits

Kaitlyn M. Price*[1,2,3], Karen G. Wigg[1], Kirsten Blokland[2], Margaret Wilkinson[2], Elizabeth N. Kerr[4,5], Sharon L. Guger[4], Maureen W. Lovett[2,5], Lisa J. Strug[6,7], Cathy L. Barr[1,2,3,8]
[1]Division of Experimental and Translational Neuroscience, Krembil Research Institute, University Health Network, Toronto, Ontario, Canada; [2]Program in Neuroscience and Mental Health, Hospital for Sick Children, Toronto, Ontario, Canada; [3]Department of Physiology, University of Toronto, Toronto, Ontario, Canada; [4]Department of Psychology, Hospital for Sick Children, Toronto, Ontario, Canada; [5]Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada; [6]Genetics and Genome Biology, Hospital for Sick Children, Toronto, Ontario, Canada; [7]Departments of Statistical Sciences and Computer Science, Faculty of Arts and Science and Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [8]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada

Multiple regions of the cortex are implicated in reading ability/disability from neuroimaging studies. However,

the brain regions, developmental period, and the neural cell types integral to the reading process are unknown. To identify relevant brain regions/cell types, we used linkage disequilibrium score regression (LDSC) to integrate genetic results from genome-wide association studies for word reading, a strong predictor of reading ability, with gene expression data from fetal and adult human brain. Gene expression data was from the Kriegstein lab, Allen Brain Bank, and PsychENCODE. We identified preliminary enrichment ($p<0.05$) in adult excitatory neurons, identifying these as critical cell type and supporting evidence from neuroimaging of excitatory-inhibitory imbalances in the etiology. We also tested correlated genetic traits (attention deficit/hyperactivity disorder (ADHD), educational attainment, cognitive ability). For ADHD, we identified preliminary enrichment in fetal replicating neuronal progenitor cells. For educational attainment and cognitive ability, we confirmed previous studies identifying adult cortical excitatory and inhibitory neurons and identified enrichment in multiple fetal cell types including cortical excitatory and inhibitory neurons, quiescent, astrocytes, neuroepithelial, fetal oligodendrocytes and intermediate progenitor cells. These findings further the understanding of the neurobiological basis of reading and correlated genetic traits indicating key cell types for stem cell models that may facilitate development of targeted interventions in the future.

130

Predicting Developmental Stuttering Cases in DNA Biobank-linked Electronic Health Records Using Comorbidity Driven Machine Learning

Dillon G. Pruett[1]*, Douglas M. Shaw[2], Hannah G. Polikowksy[2], Hung-Hsin Chen[2], Lauren E. Petty[2], Shelly Jo Kraft[3], Jennifer E. Below[2], and Robin M. Jones[1]

[1]Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Communication Sciences and Disorders, Wayne State University, Detroit, Michigan, United States of America

Background: Developmental stuttering is a speech condition characterized by syllable repetitions, prolongations, and involuntarily pauses, and has a prevalence of 1%. Despite being common, stuttering is underreported in medical center electronic health records (EHRs). The absence of documentation has thus far been a barrier to EHR based studies of stuttering, including EHR linked genome wide association studies (GWAS). The purpose of this study is to present a method for predicting under documented cases of developmental stuttering within a large medical center EHR using machine learning. The process utilizes comorbidities as variables to predict high likelihood stuttering cases without the need for text documentation, enabling use of large-scale DNA biobank resources for well-powered GWAS and epidemiological studies of stuttering.

Method: We used a keyword search, text mining algorithm, and manual review to filter 3.4 million records and identified 1,143 developmental stuttering cases. Following initial case identification, we compared the frequency of phecodes (hierarchical diagnostic groupings for EHR data derived from ICD 9 codes) within the stuttering cohort to a matched control cohort. Significantly enriched phecodes were used as predictor variables in a gini-index based classification and regression tree (CART) classification model, with a binary determination of developmental stuttering *cases* or *controls* as the outcome. The classification model was tested in an independent dataset and found to have a positive predictive value of 83%. Ultimately, the classification model was used to select cases for a stuttering GWAS, demonstrating the utility of the approach for case acquisition of poorly documented conditions.

131

Multi-phenotype GWAS Uncovers Shared Genetic Loci between Type 2 Diabetes, BMI, Colorectal, Pancreatic, Breast and Prostate Cancers

Igor Pupko[1]*, Ayse Demirkan[1], Liudmila Zudina[1], Zhanna Balkhiyarova[1,3], Anna Ulrich[1], Vincent Pascat[2], Jared Maina[2], Philippe Froguel[2,3], Marika Kaakinen[1,3], Inga Prokopenko[1,2]

[1]University of Surrey, Guildford, United Kingdom; [2]Institut Pasteur de Lille, National Center for Scientific Research, University of Lille, Lille, France; [3]Imperial College London, London, UK

Introduction: There are established relationships between type 2 diabetes (T2D) and cancer, and higher body-mass index (BMI) is a risk factor for both diseases. We aimed to elucidate genetic risk loci shared between T2D, BMI and four cancers through multi-phenotype genome-wide association studies (MPGWAS).

Methods: We analyzed five million quality-controlled SNPs from 36,173 individuals from European Prospective Investigation into Cancer (EPIC) cohort, including 10,855 T2D, 4,126 postmenopausal breast, 2,111 colorectal, 473 pancreatic and 419 prostate cancer cases with pooled controls. We performed MPGWAS for five diseases (T2D and four cancers) as well as BMI and four cancers using the reverse regression approach implemented in the SCOPA software. We then evaluated the phenotype combinations driving the top associations using Bayesian Information Criterion (BIC). We also compared the discovery power between MPGWAS and single-phenotype (SP) GWAS from the same dataset.

Results: MPGWAS identified 193 association signals ($P<5x10^{-8}$) either for T2D-cancers or BMI-cancers models. Out of the 24 signals which were previously established in SPGWAS for at least one of the included phenotypes, we classified *DND1P1* (for breast cancer and T2D) and *PTHLH* (five disease/four disease and BMI models) as multi-phenotype loci, contributing to disease comorbidity. Of the remainder, we re-classified 17 established loci, with primary effects either via SP or MP model. Overall, MP models were better powered compared to SP association.

Conclusions: Improved power of MPGWAS enables identification of novel associations and dissection of complex relationships between phenotypes at already established DNA loci.

Funding: WCRF-2017/1641; LONGITOOLS, H2020-SC1-2019-874739; PreciDIAB, ANR-18-IBHU-0001.

132

GENLIB: New Function to Simulate Haplotype Transmission in Large Complex Genealogies

Mohan Rakesh[1*], Hélène Vézina[2,3,4], Catherine Laprise[4,5], Ellen E Freeman[1,6], Kelly M Burkett[7], Marie-Hélène Roy-Gagnon[1]
[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada; [2]Projet BALSAC, Université du Québec à Chicoutimi, Chicoutimi, Canada; [3]Département des sciences humaines et sociales, Université du Québec à Chicoutimi, Chicoutimi, Canada; [4]Centre intersectoriel en santé durable, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada; [5]Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada; [6]Centre de Recherche, Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada; [7]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

Some founder populations have extensive parish and civil registers allowing for reconstruction of deep and large genealogies from present-day probands. These populations are well suited for investigating rare variants contributing to diseases. Characteristics of pedigree structures, for example the number of founders and their relatedness, can influence potential distributions of haplotypes and identity-by-descent (IBD) sharing. Here, we present a new function added to the genealogical analysis R package "GENLIB", which can simulate the transmission of haplotypes from founders to probands along a user-specified genealogy. This simulation tool allows for multiple models of meiosis (Poisson, zero-truncated Poisson, gamma renewal process), sex-specific physical/genetic maps, and is fast even for very large and deep pedigrees. Additionally, the history of transmissions is stored so it is possible to trace segments in a haplotype back along the transmission path through all internal ancestors to the originating founder. We present two example applications of our function: comparing the relationship between IBD segment lengths and the distance to internal ancestors from whom the IBD segments are inherited, and exploring the effect of pedigree structure and different models of meiosis on the amount of IBD sharing between pairs of probands. We performed our simulations in genealogical data from two studies in the French-Canadian founder population: a hospital-based study of individuals from Montreal and a family study from Saguenay–Lac-Saint-Jean. Genealogies were reconstructed using the BALSAC database.

133

Characterizing Common and Rare Variations in Non-traditional Glycemic Biomarkers Using Multi-trait Approaches on Multi-ancestry ARIC Study

Debashree Ray[1,2,*], Stephanie J. Loomis[1], Sowmya Venkataraghavan[1], Adrienne Tin[3], Josef Coresh[1,4], Eric Boerwinkle[5], Bing Yu[6], Nilanjan Chatterjee[2,7], Elizabeth Selvin[1,4], Priya Duggal[1]
[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; [3]School of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, United States of America; [4]Welch Center for Prevention, Epidemiology, & Clinical Research, Johns Hopkins University, Baltimore, Maryland, United States of America; [5]The UTHealth School of Public Health, Houston, Texas, United States of America; [6]Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth School of Public Health, Houston, Texas, United States of America; [7]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America

Type 2 diabetes is a complex disorder whose biological mechanisms are not fully understood. Because HbA1c, fasting glucose, glycated albumin, and fructosamine are biomarkers that reflect different aspects of the glycemic process, genetic studies of these glycemic biomarkers can improve our understanding of type 2 diabetes pathophysiology. While several GWAS of HbA1c and fasting glucose have been conducted, very few GWAS have focused on glycated albumin or fructosamine. Heritability estimates of glycated albumin and fructosamine suggest there are yet-unidentified associated genetic variants, and their co-heritability estimates indicate shared genetic mechanisms. We performed a multi-trait GWAS of glycated albumin and fructosamine from 9,411 White and Black participants in the Atherosclerosis Risk in Communities (ARIC) study on the common variants from genotyped/imputed data. The resultant genome-wide significant findings mapped to either known type 2 diabetes genes (e.g., *ARAP1*, $p=2.8x10^{-8}$) or novel genes using expression quantitative trait loci and 3D-chromatin mapping strategies in diabetes-relevant tissues. We identified additional loci that were ancestry-specific (e.g., *FCGRT* locus in White only, $p=5.2x10^{-8}$) and sex-specific (e.g., *ARAP1* locus in females only, $p=7.4x10^{-8}$), and these loci were not all identified when each biomarker was analyzed separately. Further, we implemented multi-trait gene-burden tests on whole-exome sequence data from 8,899 White and Black ARIC participants. Rare protein-truncating variants in *QSER1* ($p=3.8x10^{-8}$), alongwith genes from other rare-variant aggregation strategies, were exome-wide significant only in ancestry-combined analysis. We illustrate improved locus discovery and potential effector gene discovery by leveraging joint patterns of related biomarkers across entire allele frequency spectrum in multi-ancestry analyses.

134

Evaluating the Potential Benefits and Pitfalls of Combining Protein and Expression Quantitative Trait Loci in Evidencing Drug Targets

Jamie W. Robinson[1*], Thomas Battram[1], Denis A. Baird[2], Philip C. Haycock[1], Jie Zheng[3], Gibran Hemani[1], Chia-Yen Chen[2], Tom R. Gaunt[1]
[1]MRC Integrative Epidemiology Unit, Bristol Medical School,

University of Bristol, Bristol, United Kingdom; [2]Translational Biology, Research & Development, Biogen Inc., Cambridge, Massachusetts, United States of America; [3]Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Molecular quantitative trait loci (molQTL), which can provide functional evidence on the mechanisms underlying phenotype-genotype associations, are increasingly used in drug target validation and safety assessment. In particular, protein abundance QTLs (pQTLs) and gene expression QTLs (eQTLs) are the most commonly used for this purpose. However, questions remain on how to best consolidate results from pQTLs and eQTLs for target validation.

In this study, we combined blood cell-derived eQTLs and plasma-derived pQTLs to form QTL pairs representing each gene and its product. We performed a series of enrichment analyses to identify features of QTL pairs that provide consistent evidence for drug targets based on the concordance of the direction of effect of the pQTL and eQTL. We repeated these analyses using eQTLs derived in 49 tissues. We found that 25-30% of blood-cell derived QTL pairs have discordant effects. The difference in tissues of origin for molecular markers contributes to, but is not likely a major source of, this observed discordance. Finally, druggable genes were as likely to have discordant QTL pairs as concordant.

Our analyses suggest combining and consolidating evidence from pQTLs and eQTLs for drug target validation is crucial and should be done whenever possible, as many potential drug targets show discordance between the two molecular phenotypes that could be misleading if only one is considered. We also encourage investigating QTL tissue-specificity in target validation applications to help identify reasons for discordance and emphasize that concordance and discordance of QTL pairs across tissues are both informative in target validation.

135

The Contribution of Rare Variants to the Heritability of Coronary Artery Disease Based on 38,544 Whole Genome Sequences from the NHLBI TOPMed Program

Ghislain Rocheleau[1,2*], Shoa L. Clarke[3,4], Natalie R. Hasbani[5], Patricia A. Peyser[6], Ramachandran S. Vasan[7,8,9], Jerome I. Rotter[10], Danish Saleheen[11,12,13], Themistocles L. Assimes[3,4,14], Paul S. de Vries[5], Ron Do[1,2], on behalf of the National Heart, Lung and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Atherosclerosis Working Group

[1]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [2] Department of Genetics and Genomic Sciences, New York, New York, United States of America; [3]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, California, United States of America; [4]VA Palo Alto Health Care System, Palo Alto, California, United States of America; [5]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [6]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America; [7]Framingham Heart Study, Boston University and National Heart, Lung, and Blood Institute, Framingham, Massachusetts, United States of America; [8]Department of Medicine, Cardiology and Preventive Medicine Sections, Boston University School of Medicine, Boston, Massachusetts, United States of America; [9]Epidemiology Department, Boston University School of Public Health, Boston, Massachusetts, United States of America; [10] The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, California, United States of America; [11]Center for Non-Communicable Diseases, Karachi, Pakistan; [12]Department of Medicine, Columbia University Irving Medical Center, New York, New York, United States of America; [13]Department of Cardiology, Columbia University Irving Medical Center, New York, New York, United States of America; [14]Cardiovascular Institute, Stanford University School of Medicine, Stanford, California, United States of America

Introduction: Current heritability estimates of coronary artery disease (CAD) are based on genotyped and imputed variants, possibly missing a substantial contribution from rare variants. Whole genome sequence (WGS) data enables discovery of rare variants and can contribute a portion of missing heritability of CAD that is not explained by common variants.

Methods: After extensive quality control of the Trans-Omics for Precision Medicine (TOPMed) WGS data freeze 9, we obtained 9,978 cases and 28,566 controls from 14 studies representing five different genetic ancestry groups, including European, African, Hispanic, East Asian and South Asian ancestry. A principal component analysis was used to infer genetic ancestry for each sample. Then, we used the GREML-LDMS method to estimate SNP-heritability in each ancestry group. Variants were binned according to their minor allele frequencies (MAF) and linkage disequilibrium scores. All heritability estimates were adjusted for age, sex, study and the first 10 principal components.

Results: Precise estimates were obtained with observed heritability 0.38 (SE 0.21) in African, and 0.34 (SE 0.14) in European ancestries, respectively. Ultra-rare (MAF < 0.1%) and rare (0.1% < MAF < 1%) variants accounted for 54% and 36% of the total CAD heritability in those of European ancestry, and 23% and 29% of CAD heritability in those of African ancestry. Less precise or biased heritability estimates were observed in the other ancestry groups presumably due to low sample size (East Asian), inbreeding (South Asian) or admixture (Hispanic).

Conclusions: Ultra-rare and rare variants contribute a substantial fraction to the total heritability of CAD.

136

Genotype Imputation Quality Prediction using Machine Learning

Khalid Kunji[1*], Mohamad Saad[1]
[1]*Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar*
[*]Presenting author

Genotype imputation is a common practice in genome-wide association studies. The quality of imputed genotypes is typically assessed via a metric that can be considered as a proxy between imputed and ground truth genotypes (e.g., *Rsqr* for Minimac and *INFO* Impute). Here, we propose a new approach to evaluate imputation quality with machine learning using many features/variables of the pre- and post-imputation data. Data from the 1000 Genomes Project was used to evaluate the performance of our approach. We constructed training and testing datasets (70% vs 30%) and we used 30%, 40%, and 50% of each dataset as a reference for imputation. We used various proportion of genotyped SNPs, i.e., 5%, 10%, 15%, and 20%. The data was imputed with Minimac v3. Machine learning model Xgboost was applied on many features with the goal to predict R^2, which is the correlation between imputed and true genotypes. Features included minor allele frequencies (MAF), variance and entropy of imputed dosages, linkage disequilibrium patterns, and *Rsqr*. The correlation between R^2 and Minimac Rsqr was 0.67. The correlation between R^2 and the Xgboost prediction increased to 0.77. Performance varied between rare (MAF<0.01) and common (MAF≥0.01) SNPs. For rare variants, correlation between R^2 and *Rsqr* was 0.59, and it increased to 0.69 using Xgboost. For common variants, Xgboost improved from 0.85 to 0.89. These results suggest that predicting imputation quality can be improved using machine learning and features extracted from data. This allows better filtering of good vs. bad imputed SNPs.

137

Population Genetic Diversity and Anthroponymic Variation in Brittany

Aude Saint Pierre[1*], Daniel Le Bris[2], Anthony Herzig[1], Mael Jézéquel[2], FranceGenRef consortium, FREX consortium, Claude Férec[1], Christian Dina[4], Pierre Darlu[3], Emmanuelle Génin[1]
[1]*Univ Brest, Inserm, Etablissement Français du Sang, Centre Hospitalier Universitaire Brest, UMR 1078, GGB, Brest, France* ; [2]*Univ Brest, Centre de Recherche Bretonne et Celtique, EA 4451, Brest, France* ; [3]*UMR 7206 Eco-anthropologie et ethnobiologie, MNHN-CNRS-Université Denis Diderot, Paris, France ;* [4]*Inserm UMR 1087 / CNRS UMR 6291 IRS-UN, Nantes, France ;* [5]*Centre National de Recherche en Génomique Humaine (CNRGH), Direction de la recherche fondamentale, CEA, Institut de biologie François Jacob, Université Paris Saclay, Evry, France.*

The analysis of genetic diversity within European countries has provided important information on the demographic history of populations. This is the case in particular for France where we found fine-scale population stratification that corresponds closely to geographic, historical and linguistic divisions of France. At the finer scale of the Brittany region, we highlighted patterns of stratification which could reflect the observed anthroponymics and geolinguistics specificities of the region. In this work, we propose to focus on western Brittany and to study how to combine the information on surnames distribution and the information on genetic stratification.

The genetic data comes from two sequencing projects where exomes (FrEx project) and whole genomes (FranceGenRef project) were obtained on French individuals for whom the birthplaces of the four grandparents were known and located within a distance of 30 kilometers. Fine-scale population stratification was studied using FineSTRUCTURE software. We used INSEE files that provide a list of surnames by city of birth and year. Two periods of 25 years were considered (1891-1915) and (1916-1940).

Our first results show a genetic stratification between the north-west and the south-east of Brittany consistent with the territorial bipartition highlighted by the anthroponymic and dialectic maps. By combining the information provided by genetics, geolinguistics and anthroponymy, we seek to better understand the history of the settlement of the Armorican peninsula and its repercussions on the prevalence of genetic diseases more frequent in Brittany.

138

Assessment of Intergenic Polymorphisms Functional Impact on Late Onset Alzheimer's Disease Risk

Rebecca L. Sale[1], Michael J. Betti[1], Garrett Kaas[1,2], Eric R. Gamazon[1,3]
[1]*Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee;* [2]*Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee;* [3]*Clare Hall, Cambridge University, Cambridge, UK*

Genome Wide Association Studies (GWAS) have identified several independent single SNPs associated with Late Onset Alzheimer's Disease (LOAD). The majority of these SNPs are located in intergenic or intronic regions and are thought to regulate expression of genes in various cell types. The ability to map these associated variants to promoter and enhancer regions can elucidate which genes and cell types may contribute to disease risk. To assess contributions of genetic variants in promoter and enhancer regions to the risk of developing LOAD, we developed a methodology implementing Polygenic Risk Scores (PRS) and variant annotations to determine which variants and cell types contribute the greatest risk for AD. To generate PRS, we utilized GWAS summary statistics published in Jansen et al 2019, which is the second largest GWAS of AD to date encompassing 79,000 cases (including proxy cases where individuals report family history of AD) and 376,000 controls. Genetic instruments are determined using clumping and thresholding, and PRS are then generated and tested in an independent data set composed of cases and controls derived from the BioVU database. Then, to generate PRS using genetic instruments linked to functional regions of the genome, we first annotated LOAD GWAS variants leveraging ChIP-sequencing data from the ENCODE and Roadmap databases

to determine in a cell type specific manner, which variants are likely to be in promoter or enhancer regions. We then built PRS from using genetic variants in promoter and enhancer regions for each cell type represented in the ENCODE and Roadmap data sets. Finally, we tested each cell type specific PRS in our independent BioVU cohort. We report the prediction performance of each cell type specific PRS against the standard PRS methodology. We can utilize these results to prioritize highly predictive variants and cell types in LOAD to further understand their potential disease contributions to LOAD.

139

Epigenomic Signatures of Insulin Resistance Associated with Alzheimer's Disease and Related Traits

Chloé Sarnowski[1*], Marie-France Hivert[2,3], Chunyu Liu[4], Honghuang Lin[5], Alexa Beiser[4,6,7], Charles S. DeCarli[8], Josée Dupuis[4], Philip L. De Jager[9], Alanna C. Morrison[1], Sudha Seshadri[6,7,10]

[1]Department of Epidemiology, Human Genetics, and Environmental Sciences, University of Texas Health Science Center, School of Public Health, Houston, Texas, United States of America; [2]Division of Chronic Disease Research Across the Life Course, Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, Massachusetts, United States of America; [3]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [4]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [5]Department of Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America; [6]Department of Neurology, Boston University School of Medicine, Boston, Massachusetts, United States of America; [7]Boston University and the National Heart Lung and Blood Institute's Framingham Heart Study, Boston, Massachusetts, United States of America; [8]Center for Neuroscience, University of California at Davis, Sacramento, California, United States of America; [9]Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia ; University Medical Center, New York, New York, United States of America; [10]Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, University of Texas Health Science Center, San Antonio, Texas, United States of America

The mechanism by which insulin resistance (IR) predisposes to Alzheimer's disease (AD) is unknown. Epigenomic studies may help identify molecular signatures of IR associated with AD or related traits, thus contributing to an improved understanding of the biological and regulatory mechanisms of IR in AD. We conducted an epigenome wide association analysis of IR adjusted for body mass index, using the homeostasis model assessment, in 3,167 Framingham Heart Study (FHS) participants without type 2 diabetes (T2D). We evaluated the association of four DNA methylation markers (outcome) associated with IR (P<0.05/450,000=1.1×10^{-7}) with neurological traits in participants from the FHS (N=3,040) and the Religious Orders Study/Memory and Aging Project (ROSMAP, N=708). DNA methylation profile was measured in blood (FHS) or prefrontal cortex (ROSMAP) using the Illumina HumanMethylation450 BeadChip. Linear regressions or mixed-effects models, accounting for familial relatedness, were adjusted for age, sex, study, ancestry, and batch. Brain volume analyses were adjusted for intracranial volume. Significant associations were defined using P_{FHS}<0.05/$N_{markers}$/N_{traits}=0.05/4/7=0.002 and P_{ROSMAP}<0.05/$N_{markers}$/N_{traits}=0.05/4/4=0.003. DNA methylation at cg17058475 (*CPT1A*) was significantly associated with total brain volume (TBV: P_{FHS}=7.0×10^{-4}) and one AD biomarker measure (CERAD: P_{ROSMAP}=4.8×*10*$^{-4}$). DNA methylation at cg00574958 (*CPT1A*) was significantly associated with clinical diagnosis of cognitive status (P_{ROSMAP}=9.1×10^{-6}) and two AD biomarker measures (BRAAK: P_{ROSMAP}=2.7×10^{-4}; CERAD: P_{ROSMAP}=2.4×*10*$^{-5}$), and suggestively associated with TBV (P_{FHS}=0.02). The carnitine palmitoyl transferase (CPT) system, crucial for the mitochondrial beta-oxidation of long-chain fatty acids, is involved in T2D, cardiovascular and neurological diseases, including AD. Our analysis sheds light on *CPT1A* as potentially implicated in both IR and AD.

140

Leveraging Electronic Health Record Studies to Identify Shared Genetic Architecture of Eye Diseases

Alexandra Scalici[1,2,*], Tyne W. Miller-Fleming[1,2], Jibril Hirbo[1,2], Ela W. Knapik[1,2], Nancy J. Cox[1,2]

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

Proper vascularization and innervation of the eye is critical for normal eye development and vision. Eye diseases such as myopia and glaucoma are associated with optic neuropathy, however, genetic correlations between these two vary greatly. Methods used to calculate genetic correlations rely on GWAS summary statistics for SNP associations that are often not causal. Leveraging electronic health record (EHR) based studies that integrate predicted gene expression and eye disease comorbidities have the potential to shed greater insight into disease mechanisms. To examine shared genetic architecture across eye diseases, we broadly defined cases as subjects with at least one eye disease phecode in their EHR. We conducted a transcriptome-wide association study (TWAS) and identified two genes (*GPX7 & AC016590.3*) with altered predicted gene expression associated with eye disease status in BioVU (N=70,493). To identify comorbidities associated with eye disease, we conducted a phenome-wide association study (PheWAS) in non-genotyped subjects with at least three visits to VUMC in five years (N= 663,228). Using the beta estimates from the significantly associated comorbidities, we constructed a phenotypic risk score (PheRS) representing a weighted sum of a subject's comorbidities. This PheRS is predictive of disease status and associated with the altered predicted expression of six genes (*POU1F1, PAK1, AC091100.1, TDRKH, RPL41, NEU2*) in an independent population (BioVU). Taken together, both approaches identified known and novel

genes related to eye disease. Further functional analyses of these genes can expand the known pathways involved in eye disease pathogenesis and give greater insight into shared disease mechanisms.

141

Statistical Challenges in Multi-omics Integration

Paola Sebastiani[1]*, Anastasia Leshchk[2], Zeyuan Song[3], Tanya T. Karagiannis[1], Anastasia Gurinovich[1], Harold Bae[4], Mengze Li[2], Stefano Monti[2,3,5]

[1]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, United States of America; [2]Bioinformatics Program, Boston University, Boston, MA, United States of America; [3]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States of America; [4]Biostatistics Program, College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon, United States of America; [5]Department of Medicine, Geriatrics Section, Boston University School of Medicine, Boston, MA, United States of America

Several observational studies are in the process of augmenting genome-wide genotype data and phenotypic data with multi-omics data that often include gene expression, serum metabolomics and proteomics, DNA methylation, and microbiome data. The promise of such massive data generation effort is to be able to identify molecular signatures of the genetic fingerprints of many traits that can be used as therapeutic targets. Our group has focused attention on the use of probabilistic networks for integration of genetic, genomic, and complex phenotypic data. Probabilistic networks are multivariable models that factorize the joint probability distribution of the variables using marginal and conditional independences that are represented by a Markov graph. A strength of probabilistic networks is that they can be used to predict how perturbations of some nodes in the network affect other variables using probabilistic reasoning algorithms and therefore can be used for many purposes. There are well-established methods to generate probabilistic networks from random samples that satisfy some restrictive assumptions. However, generating and using probabilistic networks for multi-omics integration present some unique challenges that are related to restriction of the likelihood function due to study design and the need for interpretable data reduction summaries for scalability of the learning process. I will describe these challenges and our proposed solutions in the context of building a probabilistic network that links genetic and genomic data to exceptional longevity traits in centenarians, their offspring and unrelated controls from the New England Centenarian Study. I will also discuss many open questions.

142

Multi-Trait Genome-Wide Association Study of Neuropathology-Based Endophenotypes Identified Novel Risk Loci for Tau Pathology

Lincoln M.P. Shade[*,1,2], Yuriko Katsumata[1,2], Peter T. Nelson[2,3], and David W. Fardo[1,2]

[1]Department of Biostatistics, University of Kentucky, Lexington, Kentucky, United States of America; [2]Sanders-Brown Center on Aging, University of Kentucky, Lexington, Kentucky, United States of America; [3]Department of Pathology, University of Kentucky, Lexington, Kentucky, United States of America

Genome-wide association studies (GWAS) have to date identified over 70 genetic risk loci for Alzheimer's Disease (AD). Most AD GWAS have focused on clinical AD; however, clinical AD with AD neuropathology. We hypothesized that investigating the genetic architecture of neuropathology-based endophenotypes (NPE) would partition AD genetic risk between NPE and identify novel genetic risk for NPE.

We performed multi-trait GWAS of NPE in a harmonized data set of four aged, genotyped autopsy cohorts: National Alzheimer's Coordinating Center (NACC), Religious Orders Study and the Memory and Aging Project (ROSMAP), Adult Changes in Thought (ACT), and the AD Neuroimaging Initiative (ADNI). Conditional, mediation, and colocalization analyses were performed to investigate the shared genetic risk between NPE.

In total, 12 NPE were harmonized for analysis, and analyzed sample sizes ranged from 922 to 7285. We identified two novel risk loci for primary age-related tauopathy (PART), mapped to *PDZRN3 (P value = $4.7x10^{-10}$) and MAP3K20 (P value = $3.1x10^{-8}$),* and one for Braak stage *PIK3R5 (P value = $4.8x10^{-8}$)*. We also replicated multiple known AD and NPE risk loci. *Mediation analysis results were consistent with the association between BIN1 and AD being entirely mediated by tau pathology*.

In conclusion, we identify multiple novel risk loci for NPE and contextualize genetic risk loci for clinical AD. Novel risk loci were identified for tau endophenotypes, including PART. We also provide additional evidence that *BIN1 is associated with AD through tau pathology, consistent with previous gene expression and imaging studies.*

143

Genetic Association Analysis of Epilepsy Prognosis Using Whole Exome Sequencing

Ravi G. Shankar[1*], Andrew P. Morris[1,2], Graeme Sills[3], Tony Marson[4], Andrea Jorgensen[1]

[1]Department of Health Data Science, University of Liverpool, Liverpool, UK; [2]Centre for Musculoskeletal Research, University of Manchester, Manchester, UK; [3]School of Life Sciences, University of Glasgow, Scotland; [4]Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK

Understanding of the genetic basis of epilepsy has expanded over the years due to advancement in sequencing technology. Pharmacogenetics, the study of how genetic factors influence drug-efficacy and toxicity, plays a pivotal role in determining genetic variants associated with heterogeneity in response to anti-epileptic drugs (AEDs) commonly observed among epilepsy patients. Although time-to-event outcomes are often most important when studying genetics of treatment response, there exists a lack of analysis tools aimed at such outcomes in both genome wide association studies (GWAS) and whole-exome sequencing (WES) settings. To address this analytical bottleneck, our group focuses on developing

methodologies and software capable of GWAS and WES analysis with time-to-event outcomes.

In this study, using WES data from 421 epilepsy affected individuals from Standard and New Antiepileptic Drugs (SANAD) studies, we evaluate evidence of both common and rare variant associations with time-to-event outcomes representing AED-response. There are 297 (71.3%) subjects achieving remission of at least 365days, with a median time to remission of 389 days. After standard sample and variant quality control, survival endpoints are analyzed using *SurvivalGWAS_SV* and *rareSurvival* software, developed by our group. The results are compared to those obtained by dichotomizing time-to-event outcomes to binary variables, using standard software. The associations were stronger for analyses incorporating time-to-event outcomes than binary responses. Using appropriate methods and software to analyze GWAS and WES data with censored time-to-event outcome without having to rely on converting to binary outcomes for use with standard software helps identifying potential biomarkers for treatment response in epilepsy and benefit personalized medicine approaches.

144

Model Complexity and Explainability in Prediction for Coronary Artery Disease in the UK Biobank

Natasha Sharapova[1*], Jessye M. Maxwell[1], Saskia P. Hagenaars[1], Richard A. Russell[2], Zina M. Ibrahim[3], and Cathryn M. Lewis[1,4]
[1]Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UK; [2]Global Research and Data Analytics, RGA UK Services Limited, London UK; [3]Department of Biostatistics and Health Informatics, King's College London, London, UK, and [4]Department of Medical and Molecular Genetics, King's College London, London, UK.

Cox Proportional Hazards (CPH) regression is used widely for survival analysis due to its ease of use, computation speed and explainable output. In recent years, machine learning techniques have delivered more accurate predictions relating to disease onset. However, this has often been at the cost of explainability, meaning it is unclear how the models reach their predictions. In this study, we aimed to explore the balance between increasing predictive accuracy through model complexity and retaining explainability in prediction models for coronary artery disease (CAD). We used the UK Biobank, a health study of 500 000 participants aged 40-70 at recruitment with a follow-up period of up to 14 years. CAD cases were defined using self-reported health data, hospital episode statistics and the death register. Demographic, lifestyle, environmental, biomarkers and genetic factors known to contribute to CAD onset were used as risk factors for the disease. We first built classical statistical CPH models before applying machine learning algorithms and evaluated changes to predictive accuracy and explainability. Our CPH model gave a C-index of 0.761, which ridge regression reduced to 0.758 and our LASSO and elastic net analyses showed improvements to 0.801 and 0.823. For our XGBoost model, a C-index of 0.930 was achieved, likely due the model's ability to deal with complex interactions and nonlinearities, and SHapley Additive exPlanations (SHAP) were used to explain its predictions. We concluded that machine learning techniques could be effectively applied to survival analyses to gain predictive accuracy and maintain explainability.

145

Linkage Analysis Identifies Novel Genetic Modifiers of Microbiome Traits in Families with Inflammatory Bowel Disease

Arunabh Sharma[1*], Silke Szymczak[2], Malte Rühlemann[3], Sandra Freitag-Wolf[1], Carolin Knecht[1], Janna Enderle[4], Stefan Schreiber[3,5], Andre Franke[3], Wolfgang Lieb[4], Michael Krawczak[1] and Astrid Dempfle[1]
[1]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [2]Institute of Medical Biometry and Statistics, University of Lübeck, University Hospital Schleswig-Holstein Campus Lübeck, Lübeck, Germany; [3]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany; [4]Institute of Epidemiology, Kiel University, Kiel, Germany; [5]Department of Internal Medicine I, University Hospital Schleswig-Holstein, Kiel, Germany

Dysbiosis of the gut microbiome is a hallmark of inflammatory bowel disease (IBD) and both, IBD risk and microbiome composition, have been found to be associated with genetic variation. Using data from families of IBD patients, we examined the association between genetic and microbiome similarity in a specific IBD context, followed by a genome-wide quantitative trait locus (QTL) linkage analysis of microbiome traits. SNP, microbiome and phenotype data were obtained from the Kiel IBD family cohort, an ongoing prospective study in Germany currently comprising 256 families (455 IBD patients, 575 first- and second-degree relatives). Considering known IBD risk loci, we noted a statistically significant (FDR<0.05) association between genetic similarity at SNP rs11741861 and overall microbiome dissimilarity among IBD discordant relative pairs. In our linkage analysis, 12 chromosomal regions were found to be linked to the abundance of one of seven microbial genera, namely *Barnesiella* (chromosome 4, region spanning 10.34 cM), *Clostridium_XlVa* (chr4, 3.86; chr14, two regions spanning 7.05 and 13.02 cM), *Pseudoflavonifractor* (chr7, 12.80 cM) *Parasutterella* (chr14, 8.26 cM), *Ruminococcus* (chr16, two overlapping regions spanning 8.01 and 16.87 cM), *Roseburia* (chr19, 7.99 cM) and *Odoribacter* (chr22, three regions spanning 0.89, 5.57 and 1.71 cM), as well as α diversity (chr3, 1.47 cM). Our study shows that, in families of IBD patients, pairwise genetic similarity for at least one IBD risk locus is associated with overall microbiome dissimilarity among discordant relative pairs, and that hitherto unknown genetic modifiers of microbiome traits are located in at least 12 human genomic regions.

146

Credible Set Determination for Multi-Ancestry esFine-Mapping

Jiayi Shen[1], e[2]Anqi Wang[1], Fei Chen[1], Christopher A. Haiman[3,4], David V. Conti[1, 3, 4]

[1]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, 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

Trans-ethnic genetic studies may increase power to detect novel risk variants and improve fine-mapping resolution. Here, we expand upon our previous approach for single and multi-population fine-mapping through Joint Analysis of Marginal SNP Effects (mJAM) to develop an approach for determining credible set SNPs. The approach utilizes mJAM to use population-specific summary statistics to fit a single conditional model that incorporates ancestry-specific reference linkage disequilibrium (LD). The mJAM framework can be used to first select index SNPs using any feature selection approach, such as forward selection, Bayesian selection, or regularized regression. Then, given a set of index SNPs within each region, the posterior inclusion probability (PIP) of a SNP is defined as a combination of two probabilities: one models the marginal association between the candidate SNP and the outcome; the other models the mediation effect of the index SNP on the candidate SNP, borrowing from a mediation framework. These PIPs are then used to construct credible sets. We first illustrate mJAM for selecting index SNPs through two implementations for selection: mJAM-SuSiE (a Bayesian approach) and mJAM-forward selection. We then compare these approaches to fixed-effect meta-analysis, COJO stepwise selection, and MsCAVIAR. When available, we also compare credible set performance. Through simulation studies, we demonstrate that mJAM performs better than other existing multi-ethnic methods for identifying index SNPs and corresponding credible sets that include the underlying causal variants. In a real data application, we apply this approach to the most recent summary statistics from a trans-ethnic prostate cancer GWAS.

147

Assessing the Effectiveness and Concordance of Different Bioinformatics Tools to Detect ASD Candidate Variants in Whole Exome Sequencing (WES) Data

Apurba Shil[1,2,3 *], Noa Sadigurschi[2,3,4], Hadeel Abu-Kaf[2,4], Gal Meiri[2,5], Analya Michaelovski[2,6], Yair Tsadaka[2,7], Adi Aran[8], Ilan Dinstein[2,3,9], Hava Golan[2,3,4], Idan Menashe[1,2,3]

[1]Department of Epidemiology, Biostatistics, and Community Health Sciences, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [2]Azrieli National Centre for Autism and Neurodevelopment Research, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [3]Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [4]Department of Physiology and Cell Biology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [5]Preschool Psychiatric Unit, Soroka University Medical Center, Beer-Sheva, Israel; [6]Child Development Center, Soroka University Medical Center, Beer-Sheva, Israel; [7]Child Development Center, Ministry of Health, Beer-Sheva, Israel; [8]Psychology Neuropediatric Unit, Shaare Zedek Medical Center, Jerusalem, Israel; [9]Psychology Department, Ben-Gurion University of the Negev, Beer-Sheva, Israel

There are multiple tools for detecting disease candidate single-nucleotide variants (SNVs) in whole-exome sequencing (WES) data, but none of them is tailored towards autism spectrum disorder (ASD).

We examined the ability of different bioinformatics tools to detect candidate ASD SNVs in WES data from 250 ASD probands registered at the National Autism Database of Israel. SNVs were detected from the WES data using the Genome Analysis Toolkit (GATK) and summarized in a unified variant calling format (vcf) file for all subjects in the study. InterVar and TAPES tools were used to identify pathogenic or likely-pathogenic variants according to the American College of Medical Genetics (ACMG) guidelines in these data. In addition, we developed an algorithm (Psi-Variant) to identify likely gene-disrupting SNVs based on seven in-silico prediction tools. The ability of all three approaches to detect ASD candidate SNVs was assessed by the odds ratio (OR) of detecting SNVs in high-confidence ASD genes (according to SFARI Gene database) in these data.

Overall, 605 SNVs in 499 genes distributed in 193 probands were detected by these tools. The overlap between the tools was 64.1%, 17.0% and 21.6% for InterVar-TAPES, InterVar-Psi-Variant, and TAPES-Psi-Variant respectively. The combination of InterVar and Psi-Variant was the most effective approach in detecting ASD genes (OR=10.15; 95%CI=4.68-19.40). They detected 102 SNVs in 99 genes among 80 probands (36.5% yield) overall.

The integration of InterVar and Psi-Variant can effectively detect ASD SNVs in 1/3 probands. Inclusion of additional criteria may further improve the effectiveness and diagnostic yield of this approach.

148

Evaluating the Biomedical Relevance of Identity by Descent Genetic Communities in The Biobank at The Colorado Center for Personalized Medicine

Jonathan A. Shortt[1*], Meng Lin[1], Nick M. Rafaels[1], Christa Caggiano[2], Ruhollah Shemirani[3], Gillian M. Belbin[3], Noah A. Zaitlen[2], Eimear E. Kenny[3], Kristy R. Crooks[1], Christopher R. Gignoux[1] on behalf of CCPM

[1]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [2]Department of Neurology, University of California, Los Angeles, Los Angeles, California, United States of America; [3]Institute for Genomic Health, Icahn School of Medicine at Mt. Sinai, New York, New York, United States of America

Human demographic history and selection have shaped the genetic variation found within human populations for

tens of thousands of years, and thus play a profound role in the risk of disease during an individual's life. However, understanding an individual's risk of disease is complicated by interacting environmental exposures that vary by population. Here, we use a data-driven approach to derive high-resolution population substructure information and use it to characterize the risk of complex disease. We identified identity by descent (IBD) segments indicative of recent shared ancestry among nearly 80,000 unrelated individuals genotyped in The Biobank at the Colorado Center for Personalized Medicine (CCPM). Individuals were clustered into 40 different genetic communities using pairwise shared IBD as input to iterative Louvain clustering. Communities have both overlapping and distinct global ancestry profiles indicative of population substructures within CCPM. We estimated the prevalence of over 1,200 phecodes derived from linked electronic health record data within each genetic community and used logistic regression models to identify community-specific enrichment or depletion across the phenome. At FDR=5%, we identify 65 associations between 14 unique communities and 48 phecodes reflecting health disparities between communities, including two European ancestry communities with contrasting risk for skin cancer (OR[95% CI]=0.73[0.62-0.86] and 1.6[1.37-1.88]). While genetics is used to identify the communities, differences in risk between communities may largely reflect sociocultural, socioeconomic, and other environmental differences between communities. Communities may therefore be useful in accounting for these factors.

149

Depression and Coronary Artery Disease Share Genetic Risk Factors Enriched in Inflammation and Cardiomyopathy-Associated Pathways

Kritika Singh[1,2*], Tyne M. Flemming[1,2], Lea K. Davis[1,2]
[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Coronary artery disease (CAD) and depression are highly prevalent (7% and 8% respectively), comorbid diseases. Approximately 40% of adults who are initially diagnosed with either CAD or depression will also develop the other condition. Our previous work demonstrated that polygenic risk scores for depression were significantly associated with coronary atherosclerosis in a large hospital population even in patients without any psychiatric diagnoses and after adjusting for known CAD risk factors. This suggests that some proportion of genes that increase susceptibility to depression act pleiotropically to increase CAD risk. Therefore, to identify the genes increasing risk for both CAD and depression we performed a cross-tissue S-MultiXcan analysis on the summary statistics for CAD and depression. We identified 185 genes significantly associated with both depression and CAD, representing a three-fold enrichment of shared associations (*P* value <1.718e-43). These genes were enriched for inflammation and cardiomyopathy-associated pathways. We then mined electronic health records at Vanderbilt, to investigate the prevalence of cardiomyopathy in patients with comorbid depression and CAD. We found that the depression-CAD cohort had a significantly increased rate of incident cardiomyopathy diagnoses (*P* value < 2.2e-16) compared to patients with CAD alone. We also observed that the increased predicted expression of the top genes from our S-MultiXcan analysis (sample size = 13) were significantly associated with altered levels of blood and immune markers. These results highlight cardiomyopathy and inflammation as an important biological link in the comorbid manifestation of CAD and depression.

150

Distinct Metabolic Features of Genetic Liability to Type 2 Diabetes and Coronary Artery Disease: A Reverse Mendelian Randomization Study

Madeleine L. Smith[1,2*], Caroline J. Bull[1,2,3], Michael V. Holmes[1,2,4,5,6], George Davey Smith[1,2], Emma L. Anderson[1,2], Joshua A. Bell[1,2]
[1]Medical Research Council (MRC) Integrative Epidemiology Unit, University of Bristol, UK; [2]Bristol Medical School, Population Health Sciences, University of Bristol, United Kingdom; [3]School of Translational Health Sciences, University of Bristol, Bristol, United Kingdom; [4]Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom; [5]MRC Population Health Research Unit at the University of Oxford, Oxford, United Kingdom; [6]National Institute for Health Research, Oxford Biomedical Research Centre, Oxford University Hospital, Oxford, United Kingdom

Type 2 diabetes (T2D) and coronary artery disease (CAD) both have roots in disordered metabolism but the mechanisms through which their associated genetic variants lead to disease onset remain poorly understood. We performed two-sample reverse Mendelian randomization (MR) to estimate effects of genetic liability to T2D and CAD on 249 circulating metabolites from targeted nuclear magnetic resonance spectroscopy in the UK Biobank (N=118,466). We examined the potential for medication use to distort effect estimates by examining effects of disease liability on metformin and statin use and conducting age-stratified metabolite analyses. Using inverse variance weighted models, higher genetic liability to T2D was estimated to increase triglycerides and branched chain amino acids (BCAAs). Estimates for CAD liability suggested an effect on reducing high-density lipoprotein cholesterol (HDL-C) as well as raising levels of very-low-density lipoprotein cholesterol, low-density lipoprotein cholesterol (LDL-C), LDL triglycerides, and apolipoprotein-B. In pleiotropy-robust sensitivity models, T2D liability was still estimated to increase BCAAs, but several effect estimates for higher CAD liability were reversed and estimated to decrease LDL-C and apolipoprotein-B. Higher T2D liability increased the odds of using metformin, whereas higher CAD liability increased the odds of using statins. Estimated effects of CAD liability differed substantially by age tertile for non-HDL-C traits, e.g., pleiotropy-robust models

suggested that higher CAD liability lowers LDL-C only at older ages when use of statins is common. Our results support largely distinct metabolic features of genetic liability to T2D and to CAD, and substantial effect modification by medication use in adulthood for CAD in particular.

151

The Blood Proteome of Cancer Risk

Karl Smith-Byrne[1], Åsa Hedman[2,3], Trishna Desai[1], Ruth C. Travis[1], James McKay, Paul Brennan, Mattias Johansson[4*], Anders Mälarstig[2,3*]

[1]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, England; [2]Department of Medicine, Karolinska Institute, Stockholm, Sweden; [3]Emerging Science & Innovation, Pfizer Worldwide Research & Development, Cambridge, United States of America; [4]Genomic Epidemiology branch, International Agency for Research on Cancer, Lyon, France

**Authors Contributed Equally*

Introduction: Cancer is the leading cause of death worldwide and accounts for up to ten million deaths annually. Cancer may be preventable by altering established modifiable risk factors or by chemoprevention currently limited to high-risk groups by targeting known causal proteins markers. Identifying further proteins with a credible role in cancer aetiology may help identify novel targets for potential chemoprevention targets.

Methods: We conducted a two-sample Mendelian randomization (MR) study for the association of up to 2,329 proteins using 6,476 *cis* SNPs with risk for head and neck, pancreas, lung, skin, breast, endometrial, ovarian, cervical, renal, and bladder cancer risk in over 250,000 cancer cases. We considered proteins with MR Wald ratios significant at a cancer-specific Bonferroni correction (0.05/N proteins) and strong evidence of colocalization (PP4 > 0.7) robust aetiological cancer biomarkers.

Results: We found significant evidence for associations with cancer risk for 83 *cis*-predicted protein concentrations, of which 52 had strong evidence of a shared causal signal from colocalization. For three proteins, there was evidence of an association with multiple cancers. Additionally, four of the 52 proteins are encoded by genes targeted by existing drugs.

Discussion: We present an expansive investigation into the association of the blood proteome with risk of common cancers, identifying novel aetiological markers. If further replicated, the identified proteins represent a significant advance in our understanding of pathways for cancer risk and may highlight novel drug targets for chemoprevention.

152

Using Imputed Genotype Data in the Joint Score Tests for Genetic Association and Gene–environment Interactions in Case-control Studies

Minsun Song[*]

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

Genome-wide association studies (GWAS) are now routinely imputed for untyped single nucleotide polymorphisms (SNPs) based on various powerful statistical algorithms for imputation trained on reference datasets. The use of predicted allele counts for imputed SNPs as the dosage variable is known to produce valid score test for genetic association. In this paper, we investigate how to best handle imputed SNPs in various modern complex tests for genetic associations incorporating gene–environment interactions. We focus on case-control association studies where inference for an underlying logistic regression model can be performed using alternative methods that rely on varying degree on an assumption of gene–environment independence in the underlying population. As increasingly large-scale GWAS are being performed through consortia effort where it is preferable to share only summary-level information across studies, we also describe simple mechanisms for implementing score tests based on standard meta-analysis of "one-step" maximum-likelihood estimates across studies. Applications of the methods in simulation studies and a dataset from GWAS of lung cancer illustrate ability of the proposed methods to maintain type-I error rates for the underlying testing procedures. For analysis of imputed SNPs, similar to typed SNPs, the retrospective methods can lead to considerable efficiency gain for modeling of gene–environment interactions under the assumption of gene–environment independence. Methods are made available for public use through CGEN R software package.

153

Fast-REML Can Analyze Biobank-sized Datasets and Provide a More Detailed Understanding of Complex Traits

Doug Speed[1*], Xuan Zhou[1]

[1]Centre for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark

We have developed Fast-REML, a highly-efficient algorithm for performing restricted maximum likelihood. Fast-REML is orders of magnitude quicker than existing algorithms, which means that it can be routinely applied to biobank-sized datasets. For example, when analyzing 100,000 individuals, Fast-REML takes four CPU hours and requires 40 GB memory, which is about ten times less time / memory than required by the softwares GCTA and MTG2. We apply Fast-REML to 39 quantitative traits and 229 ICD-10 diseases using data from up to 250,000 UK Biobank individuals. For all traits, we use Fast-REML to estimate family and SNP heritabilities, the enrichments of conserved regions and genic variants, and the relationship between heritability and allele frequency (the latter can be interpreted as a measure of selection). Then we use Fast-REML to identify significant interactions between genetic factors and environment for the 39 quantitative traits, and show how these findings enable improved power to detect associated loci and more accurate prediction models. Fast-REML is freely available within our software LDAK (www.ldak.org).

154

Polygenic Risk Scores Based on Penalized Generalized Linear Mixed Models

Julien St-Pierre[1]
[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada*

In genome-wide association studies (GWAS), generalized linear mixed models

(GLMMs) are now widely used to model population structure and/or cryptic relatedness by including a polygenic random effect with variance-covariance structure proportional to the genetic similarity matrix used to infer the Principal Components (PCs). Following a GWAS, a polygenic risk score (PRS) can be constructed by summing the risk alleles in an individual to obtain a single overall genetic risk. Based on the work of GMMAT [Chen et al., 2016] for univariate genetic association tests in GLMMs, we propose to use a penalized quasi-likelihood loss function with a LASSO regularization to select important genetic predictors and estimate their effects, while controlling for population structure and/or cryptic relatedness by including a polygenic random effect, to derive a multivariate PRS for binary traits. We perform simulation studies to evaluate the performance of our proposed method in a variety of scenarios.

155

Estimating Local Ancestry Proportions from Genetic Summary Data

Hayley R. Stoneman[1,2,3*], Audrey E. Hendricks[1,2,3]
[1]*Mathematical and Statistical Sciences, University of Colorado Denver, Colorado, United States of America;* [2]*Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus, Aurora, Colorado, United States of America;* [3]*Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America*

The development of high-thr0ughput sequencing has increased the number of large, publicly available genetic data sets. The data is often made available at the summary level to protect individual privacy, but this aggregation can mask population structure. We have developed Summix, a method that accurately and precisely estimates global ancestry proportions from summary data using only allele frequencies. There is potential for Summix to be used for estimation of local ancestry proportions. Local ancestry can be useful for studies such as mapping traits or diseases to regions of the genome and detecting regions of selection. Existing local ancestry deconvolution methods require individual level data, which is not always accessible.

Here, we evaluate the ability of Summix to estimate local ancestry proportions from genetic summary data. In a variety of simulation scenarios, we show that Summix can produce accurate and precise local ancestry proportion estimates within one percent. To identify regions of selection, we develop a test statistic to determine local ancestry blocks that are significantly different from the global ancestry. Using simulations, we found that Summix maintains at or below the expected type I error rate of 5% and has >80% power to detect a two percent difference in local ancestry proportions. This work expands the utility of publicly available genetic data for novel uses such as identification of regions of selection and correction for local ancestry proportions in summary level data, especially in populations of admixed ancestry.

156

PIn Test: A Novel Method for Detecting Pairwise Pure Interaction Effect in Single Nucleotide Polymorphism (SNP) Data

Rui Sun[1,2*], Xiaoxuan Xia[2,3], Jinqiu Yuan[1,2], Zilong Zheng[4], Tianshun Gao[1,2], Shuo Fang[5,6], Benny Chung-Ying Zee[2,7], Yihang Pan[1,2*], Maggie Haitian Wang[2,7*]
[1]*Scientific Research Center, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen, P. R. China;* [2]*The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China;* [3]*Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China;* [4]*Department of Information Management, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, P. R. China;* [5]*The department of clinical oncology, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, P. R. China;* [6]*The department of clinical oncology, The University of Hong Kong;* [7]*CUHK Shenzhen Research Institute, Shenzhen, China*

Epistasis plays a crucial role in the etiology of complex human diseases. Biologically well-known genes that play important roles in disease progression are often found to be insignificant by main effect evaluation in the genome-wide association studies (GWASs), and they may contribute to disease through epistasis. Although a large number of methods are available to detect the interaction effect in categorical genetic data, few could capture it in the absence of strong single marker effect. In this study, we propose the Pure Interaction (PIn) test to evaluate the epistatic relation of two loci unaffected by their main effect. The method compares the linkage disequilibrium pattern of two loci between the cases and controls conditioning on their marginal distributions. Simulation studies showed that the PIn test is highly effective in capturing the pure epistatic effect of linear and nonlinear genetic models. In the application on real GWAS data of the Alzheimer's disease (AD), the PIn test identified the biologically meaningful genes, *EML1*, *EYA4*, and *MARCHF1*, in significant epistatic relations while their main effects were modest. These genes were also found to be involved in significant pure interactions in the independent sequencing data of AD in the UK Biobank. The proposed method provides a novel approach to uncover the important genes contributing to complex diseases through epistasis.

157

Soluble CD14-associated DNA Methylation Sites predict Mortality among Men with Human Immunodeficiency Virus Infection

Boghuma K. Titanji,[1] Zeyuan Wang,[2] Junyu Chen,[2] Qin Hui,[2] Kaku So-Armah,[3] Matthew Freiberg,[4] Amy C. Justice,[5,6] Ke Xu,[5,6] Vincent C. Marconi,[1,7,8,9] Yan V. Sun[2,7*]

[1]Division of Infectious Diseases, Emory University School of Medicine, Atlanta, Georgia, United States of America; [2]Department of Epidemiology, Rollins School of Public Health, Emory University Atlanta, Georgia, United States of America; [3]Boston University Medical School, Massachusetts, United States of America; [4]Cardiovascular Medicine Division and Tennessee Valley Healthcare System, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Yale University School of Medicine, New Haven, Connecticut, United States of America; [6]Connecticut Veteran Health System, West Haven, Connecticut, United States of America; [7]Atlanta Veterans Affairs Health Care System, Decatur, Georgia, United States of America; [8]Hubert Department of Global Health, Rollins School of Public Health, Atlanta, Georgia, United States of America; [9]Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, United States of America

Objectives: Elevated plasma levels of soluble Cluster of Differentiation 14 (sCD14) predict all-cause mortality in people with human immunodeficiency virus (PWH). Epigenetic regulation plays a key role in infection and inflammation. To reveal the epigenetic relationships between sCD14, immune function and disease progression among PWH, we conducted an epigenome-wide association study (EWAS) of sCD14 and investigated the relationship with mortality.

Design and Methods: DNA methylation (DNAm) levels of peripheral blood samples from PWH in the Veterans Aging Cohort Study were measured using the Illumina Infinium Methylation 450K (n=549) and EPIC 850K BeadChip (n=526). Adjusted for covariates and multiple testing, we conducted an epigenome-wide discovery, replication, and meta-analysis to identify significant associations with sCD14. We then examined and replicated the relationship between the DNAm sites and survival using Cox regression models.

Results: We identified 118 DNAm sites significantly associated with sCD14 in the meta-analysis of 1,075 PWH. The principal associated DNAm sites mapped to genes (e.g., *STAT1*, *PARP9*, *IFITM*1, *MX1,* and *IFIT1*) related to inflammation and antiviral response, particularly interferon-stimulated pathways. Adjusting for multiple testing, 10 of 118 sCD14-associated DNAm sites significantly predicted survival time conditional on sCD14 levels.

Conclusions: The identification of DNAm sites independently predicting survival may improve our understanding of prognosis and potential therapeutic targets among PWH.

158

Trait Selection Strategy in Multi-trait GWAS: Boosting SNPs Discoverability

Yuka Suzuki[1*], Hervé Ménager[2], Cyril Nerin[1], Rachel Torchet[2], Pierre Lechat[2], Lucie Troubat[1], Hugues Aschard[1], Hanna Julienne[1,2]

[1]Institut Pasteur, Université de Paris, Department of Computational Biology, Paris, France, [2]Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France

With a profusion of *Genome-Wide Association Study* (GWAS) summary statistics becoming publicly available, investigators have developed a variety of multi-trait GWAS methods, including joint association tests to increase statistical power and detect new variants missed by univariate analyses. Yet, the joint analysis of GWASs carries many challenges, including data harmonization, computational cost, and missing values. More critically, although combinatorial possibilities are overwhelming, there is no established strategy to astutely select sets of traits to be analyzed jointly. To solve these challenges, we implemented a python package (*Joint Analysis of Summary Statistics* (JASS)) with specific attention to computational efficiency (on 62 traits and ~7 million SNPs, JASS runs in 17 minutes). We conducted multi-trait GWASs using JASS on ~2k sets of two to 62 traits randomly selected out of 122 traits, retrieved from a database of curated summary statistics (https://jass.pasteur.fr/). Furthermore, we investigated the impact of genetic architecture on the capacity of the joint test to discover new associations (i.e. associations missed by the univariate test). Mean effect size and mean genetic correlation both impact this capacity (*P* value < 5e-6, <2e-20, respectively). When included in the multivariate regression model, interaction terms between these variables almost doubled the explained variance (R-squared from 0.18 to 0.35). Further investigation of the influence of genetic architecture on joint association tests should allow investigators to effectively uncover associations with low discoverability in univariate GWAS and ensure the novelty of their findings.

159

Phenome-wide Association Analysis Identifies Novel Traits Associated with XXY and XYY Syndromes

Craig C. Teerlink[1,*], Shanlee Davis[2], Julie Lynch[1], Judith L. Ross[3,4], Meghana Pagadala[5,6,7], Guneet K. Jasupa[8,9], Cristina Perez[1], Nai-Chung Chang[1], Matthew Panizzon[10,11], Richard L. Hauger[10,11]

[1]Veterans Administration Salt Lake City Healthcare System, Salt Lake City, Utah, United States of America; [2]Department of Pediatrics, University of Colorado, Aurora Colorado, United States of America; [3]Nemours Children's Hospital, Wilmington, Delaware, United States of America; [4]Thomas Jefferson University, Philadelphia, Pennsylvania, United States of America; [5]Biomedical Science Program, University of California San Diego, La Jolla, California, United States of America; [6]Medical Scientist Training Program, University of California San Diego, La Jolla, California, United States of America; [7]Veterans Administration San Diego Healthcare System, San Diego, California, United

States of America; [8]Section of General Internal Medicine, Boston University School of Medicine, Boston, Massachusetts, United States of America; [9]Center for Healthcare Organization and Implementation Research, VA Bedford Healthcare System, Bedford, Massachusetts, United States of America; [10]Department of Psychiatry, Center for Behavior Genetics of Aging, University of California San Diego, La Jolla, California, United States of America; [11]Center of Excellence for Stress and Mental Health, Veterans Administration San Diego Healthcare System, San Diego, California, United States of America

Background: The phenotypic expression of the karyotypes 47, XXY and XYY, are not completely specified and there is limited data on how such individuals age. Hence, we pursued a phenome-wide association analysis (phewas) of XXY and XYY individuals in the Veteran's Administration's Million Veterans Program dataset that is combined with a rich, uniform, nation-wide medical record system (VISTA).

Methods: Individuals with XXY and XYY karyotypes were identified via DNA analysis of blood samples. 1,876 phecode categories were specified from ICD-9 or ICD-10 billing codes. XXY and XYY subjects were matched on age, sex, and race to 5x matched controls. Phewas analysis was conducted with the PHEWAS R package.

Results: 862 XXY males (1/691.0) and 747 XYY males (1/797.3) were identified. 306 (35%) of males with XXY had a clinical diagnosis of Klinefelter syndrome, whereas only 2 (0.2%) of males with XYY carried a diagnosis, indicating many subjects were unaware they had a condition. For XXY, the leading phewas categories were testicular hypofunction (*P* value=$1.3e^{-104}$), deep vein thrombosis (*P* value=$4.1e^{-44}$), and chronic venous insufficiency (*P* value=$2.0e^{-43}$). For XYY, the leading phecode categories were chronic venous insufficiency (*P* value=$3.6e^{-44}$), deep vein thrombosis (*P* value = $2.1e^{-34}$), and peripheral vascular disease (*P* value=$3.1e^{-30}$).

Conclusions: These results emphasize the importance of the vascular phenotype of XXY and XYY which is understudied, and present novel phenotypic expressions of XXY and XYY syndromes that may be relevant for clinical care, risk analysis, and anticipatory guidance. Replication in external datasets is also a current focus of investigation.

160

Sex-biased Genetic Regulation of Methylation and Transcript Levels in Placenta

Fasil Tekola-Ayele[1*], Rich Biedrzycki[2]

[1]*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*

Sex differences are present in many complex phenotypes in early- and later-life. Sex-dependent genetic regulation of methylation and transcript levels in tissues such as the placenta can inform mechanisms in early origins of sex differences in phenotypes; however, these data remain scarce. We performed methylation and gene expression quantitative trait locus analyses on placental methylation (mQTL) and transcript levels (eQTL) to identify sex-interacting QTLs, followed by integration of these QTL loci with genome-wide association study (GWAS) summary statistics to determine overlaps with complex diseases/phenotypes and identify multi-trait colocalization. Analyses were performed using genome-wide fetal genotype, placental methylation and RNA-seq data from the NICHD Fetal Growth Studies cohort (maximum n=291) using linear regression models that included fetal sex and genotype as an interaction term and adjusted for covariates. We identified 1839 sex-interacting *cis*-mQTLs, and 13 sex-interacting *cis*-eQTLs (P_{FDR} <0.05). 71.2% of the mQTL associations and 92.3% of the eQTL associations exhibited single-sex effect in sex-stratified analysis. The eQTLs identified are distinct from previously reported sex-biased eQTLs across 48 tissues in the Genotype Tissue Expression (GTEx) portal. Only 0.4% sex-interacting mQTL target methylation sites showed sex-biased placental levels, as opposed to nearly two-thirds of the sex-interacting eQTL target genes. Several QTLs have previously been reported to be GWAS loci in complex phenotypes, including blood cell count and adiposity measures known to vary between males and females. In conclusion, we identified novel genetic loci exhibiting sex-biased effect on placental molecular traits, and some of these loci are implicated in complex diseases and phenotypes.

161

Inferring Sweep Parameters for Recent Adaptive Selection Using Identity-by-descent Segments

Seth D Temple[1*], Sharon R Browning[2]

[1]*Department of Statistics, University of Washington, Seattle, Washington, United States of America; [2]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America*

Recent positive selection can result in an excess of long identity-by-descent (IBD) segments. We present a novel statistical framework that relates selective sweep models to the length distribution of such IBD segments. Our aim is to infer sweep parameters such as the selection coefficient and the variant age as these parameters influence the trajectory of the population allele frequency over time. This task is challenging because the length distribution arises as the output of a noisy stochastic process, whose variation we decompose in terms of a random walk for variant counts and dependence among IBD segments conditional on a coalescent tree. Using our framework, we propose an estimation procedure based on a chi-squared goodness-of-fit statistic and report confidence intervals (CIs) derived from parametric bootstraps of a fast coalescent process. In simulation studies, our method accurately infers the selection coefficient with tight CIs. We use our method to estimate selective sweep parameters for genetic loci showing signals of positive selection in the UK Biobank (Browning and Browning, 2020). Our method can be applied to both SNP array and whole

genome sequence data to study recent positive selection in other contemporary populations.

162

Cox-MKF: A Knockoff Filter for High-Dimensional Mediation Analysis with a Survival Outcome in Epigenetic Studies

Peixin Tian[1], Minhao Yao[1], Tao Huang[2,3,4], Zhonghua Liu[1]
[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong SAR, China; [2]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, China; [3]Center for Intelligent Public Health, Academy for Artificial Intelligence, Peking University, Beijing, China; [4]Key Laboratory of Molecular Cardiovascular Sciences (Peking University), Ministry of Education, Beijing, China.

Motivation: It is of scientific interest to identify DNA methylation CpG sites that might mediate the effect of an environmental exposure on a survival outcome in high-dimensional mediation analysis. However, there is a lack of powerful statistical methods that can provide a guarantee of false discovery rate (FDR) control in finite sample settings.

Results: In this article, we propose a novel method called Cox-MKF, which applies aggregation of multiple knockoffs to a Cox proportional hazards model for a survival outcome with high-dimensional mediators. The proposed Cox-MKF can achieve FDR control even in finite-sample settings, which is particularly advantageous when the sample size is not large. Moreover, our proposed Cox-MKF can overcome the randomness of the unstable model-X knockoffs. Our simulation results show that Cox-MKF controls FDR well in finite samples. We further apply Cox-MKF to a lung cancer data set from the The Cancer Genome Atlas (TCGA) project with 754 subjects and 365,306 DNA methylation CpG sites, and identify four DNA methylation CpG sites that might mediate the effect of smoking on the overall survival among lung cancer patients.

163

Multiethnic Polygenic Risk Prediction in Diverse Populations through Transfer Learning

Peixin Tian[1], Tsai H. Chan[1], Yong-Fei Wang[2], Wanling Yang[2], Guosheng Yin[1], Yan D. Zhang[1,4]
[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Hong Kong SAR, China; [2]Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Pokfulam Hong Kong SAR, China; [4]Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

Polygenic risk scores (PRS) leverage the genetic contribution of an individual's genotype by estimating disease risk. Traditional PRS prediction methods are predominantly for European population. The accuracy of PRS prediction in non-European populations is diminished due to much smaller sample size of genome-wide association studies (GWAS). In this article, we introduced a novel method to construct PRS for non-European populations, abbreviated as TL-Multi, by conducting transfer learning framework to learn useful knowledge from European population to correct the bias for non-European populations. We considered non-European GWAS data as target data and European GWAS data as informative auxiliary data. TL-Multi borrowed useful information from auxiliary data to improve the learning accuracy of the target data while preserving the efficiency and accuracy. To demonstrate the practical applicability of the proposed method, we applied TL-Multi to predict systemic lupus erythematosus (SLE) risk in Hong Kong population by borrowing information from European population. TL-Multi achieved better prediction accuracy than alternative methods including Lassosum, meta-analysis and linkage disequilibrium (LD)-informed pruning and *P* values thresholding for multiethnic PRS (PT-Multi), and substantially improved the prediction performance with moderate cross-population genetic correlation in both simulations and SLE application.

164

Epigenetics of Parkinson's Disease: A Case and Control Study on Brain and Blood Samples Derived from the Japanese Male Kuakini Honolulu Heart Cohort

R.C. Go[1,2,3*], G.W. Ross[2,4,5], H. Petrovich[2,4,5], K.H. Masaki[1,,4], Q. He[1,4], M.J. Corley[6], A.K. Maunakea[7], M. Tiirikainen[8]
[1]Kuakini Health Systems, Honolulu, Hawaii, United States of America; [2] Pacific Health and Education Institute, Honolulu, Hawaii, United States of America; [3]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, United States of America; [4]Department of Geriatric Medicine, John A. Burns School of Medicine, University of Hawaii Manoa, Honolulu, Hawaii, United States of America; [5]Neurology Veterans Affairs Pacific Islands Health Care Systems, Honolulu, Hawaii, United States of America; [6]Cornell Center for Immunology, Cornell University, New York, New York, United States of America; [7]Department of Native Hawaiian Studies, John A. Burns School of Medicine, University of Hawaii Manoa, Honolulu, Hawaii, United States of America; [8]University of Hawaii Cancer Center, University of Hawaii Manoa, Honolulu, Hawaii, United States of America

Recent research has increasingly linked microglia to the pathogenesis of Parkinson's Disease (PD); a common neurodegenerative disease. Activation of microglia and reactive astroglia lead to microsomal dysfunction, accumulation of extracellular toxins and death of dopaminergic neurons in the substantia nigra. PD cases (n=22) and controls (n=7) were drawn from a prospective population of 8,006 Japanese males with 48 years of follow-up. Utilizing DNA extracted from brain temporal lobe autopsy samples of glial origin and their matching blood samples, we performed a genome-scale DNA methylation study on the Illumina HumanMethylation450K BeadChip. Analysis of Covariance (ANCOVA) was used to detect Differentially Methylated Loci (DML) between PD cases and controls. Preliminary results found 240 DML with P values <0.001, and 22 with <0.0001 in brain, and 355 DML and 36 respectively in blood. Nine DML were concordant between the tissues. Top two DML from ANCOVA analyses on brain were EPHA7 and PRKG1 with unadjusted P values of 7.04E-06 and 1.25E-05, respectively. Top two DML for peripheral blood were

WHSC1 and POGZ with unadjusted P values of 1.56E-07 and 8.8E-07, respectively. 180 unique genes with DML in brain were subjected to hierarchical clustering and showed clear separation of cases and controls also by Principal Component Analysis. Pathway analyses revealed interaction networks related to cell signaling, cellular movement, and nervous system development and function. 298 blood DML genes formed networks related to cell death, and nervous system development and function. Further results will be reported on the neuropathologic DML connected to glial pathways.

165

A Study of the Genetic Etiology of Familial Pulmonary Fibrosis in Patients from the Canary Islands (Spain)

Eva Tosco-Herrera[1*], Aitana Alonso-Gonzalez[1,4], Ibrahim Véliz-Flores[5], Felipe Rodríguez de-Castro[5], Rosario Perona[6], Beatriz Fernández-Vara[6], Luis A. Rubio-Rodríguez[2], Jose M. Lorenzo-Salazar[2], Rafaela González-Montelongo[2], Carlos Flores[1,2,3]
*[1]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [2]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [3]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [4]Universidad de Santiago de Compostela, Santiago, Spain; [5]Servicio de Neumología Hospital Universitario de Gran Canaria *Dr Negrín, Spain; [6]Instituto de Investigaciones Biomédicas CSIC-UAM, Spain.*

Pulmonary Fibrosis (PF) is a rare progressive scarring lung disease with poor prognosis. Genetic variation is underlying its etiology, although genetic diagnosis is not typically assessed in the clinical practice. We relied on whole-exome sequencing (WES) to screen the genetic causes of familial PF (FPF) in Canary Islanders. WES was obtained from 54 subjects from 11 families using a HiSeq4000 and small germline variant identified with BWA-GATK v3.8 against the GRCh37/hg19 reference. Relative telomere length (TL) was measured by quantitative PCR and severe reduction denoted when length was <10th percentile against age-matched controls. Rare (MAF<0.1%) non-synonymous exonic and splicing variants from PF or interstitial lung disease genes were considered for manual review. Initial analysis excluded shared genetic variants among affected individuals. A total of 22 rare variants from 15 genes were classified following ACMG recommendations. Pathogenicity was supported for two variants: in *NAF1* gene (NM_138386.3:c.1104T>G) found in one family and the TL of the index case was in the 1st percentile, and in *PARN* (NM_002582.4:c:1271 T>A) found in two individuals from the same family where the TL of both asymptomatic carriers was <1st percentile. A diagnostic yield of 18% (95% CI: 5-48%) resulted from genetic testing of 11 FPF families. Our results evidence the value of genetic diagnosis considering family history and patient stratification by clinical features.

Funding: Ministerio de Ciencia e Innovación (RTC-2017-6471-1; AEI/FEDER, UE); ITER agreement (OA17/008); Gobierno de Canarias & Fondo Social Europeo Canarias Avanza con Europa" (TESIS2021010046), Ministerio de Universidades (modality Margarita Salas).

166

Multi trait GWAS for Diverse Ancestry: Mapping the Knowledge Gap

Lucie Troubat[1*], Hugues Aschard[1], Hanna Julienne[1,2]
[1]Institut Pasteur, Université de Paris, Department of Computational Biology, Paris, France, [2]Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France

~95% of samples analyzed in univariate genome-wide association study (GWAS) are from European ancestry (EUR) individuals. As illustrated on Polygenic Risk Scores, existing data in non-European populations, which often present modest sample sizes, can benefit from innovative methodological developments. Here we conducted a multi trait GWAS using JASS (Joint Analysis of Summary Statistics) on blood traits across five super populations. We detected 393 new associations in non-European populations (seven in South Asian (SAS), 15 in American (AMR), 55 in African (AFR), 316 in East Asian (EAS)). For the SAS and AMR population, 25% and 37% of associations were detected with the joint test only. Six of the seven new associations in SAS mapped to genes annotated for blood trait related pathways including Hemoglobin Subunit Beta (HBB) involved in Beta Thalassemia and Sickle Cell Anemia diseases, Tubulin Beta 1 Class VI (TUBB1) expressed in platelets, and MYB Proto-Oncogene playing a key role in the regulation of hematopoiesis. These seven associations were found at least in one other ancestry, suggesting the robustness of the new loci detected by JASS. For other non-European populations, we detected new ancestry specific loci : two for AMR, four for AFR, eight for EAS. Several of these specific loci were mapped to relevant genes. For instance, a specific association for AFR (rs1149189) mapped to CD53, that might be involved in growth regulation in hematopoietic cells. We argue that multitrait GWAS methods can be a valuable tool to narrow the genetic knowledge gap between European and non-European populations.

167

Calcium Channel Blockers: Clinical Outcome Associations with Reported Pharmacogenetics Variants in 32,000 Patients

Deniz Türkmen[1], Jane A.H. Masoli[1,2], João Delgado[1], Chia-Ling Kuo[3,4], Jack Bowden[5], David Melzer[1], Luke C. Pilling[1]
[1]Epidemiology and Public Health Group, College of Medicine and Health, University of Exeter, Exeter, UK; [2]Department of Healthcare for Older People, Royal Devon and Exeter Hospital, Barrack Road, Exeter, UK; [3]UConn Center on Aging, University of Connecticut, Farmington, Connecticut, United States of America; [4]Connecticut Convergence Institute for Translation in Regenerative Engineering, University of Connecticut, Connecticut, United States of America; [5]Exeter Diabetes Group (ExCEED), College of Medicine and Health, University of Exeter, Exeter, UK

Background: Pharmacogenetic variants impact dihydropyridine calcium channel blockers (dCCB) treatment

efficacy, yet evidence on clinical outcomes in routine primary care is limited. We aimed to estimate associations between reported pharmacogenetic variants and incident adverse events in a community-based cohort prescribed dCCB

Methods: We analyzed up to 32,360 UK Biobank participants prescribed dCCB in primary care (from UK General Practices, 1990 to 2017). We investigated 23 genetic variants. Outcomes were incident diagnosis of coronary heart disease (CHD), heart failure (HF), chronic kidney disease (CKD), edema, and switching antihypertensive medication.

Results: Participants were aged 40 to 79 years at first dCCB prescription. Carriers of rs877087 T allele in *RYR3* had increased risk of HF (Hazard Ratio 1.13: 95% Confidence Intervals 1.02 to 1.25, p=0.02). We estimated that if rs877087 T allele could experience the same treatment effect as non-carriers, the incidence of HF in patients prescribed dCCB would reduce by 9.2% (95%CI 3.1 to 15.4). In patients with a history of heart disease prior to dCCB (N=2,296), rs877087 homozygotes had increased risk of new CHD or HF compared to CC variant. rs10898815 in *NUMA1* and rs776746 in *CYP3A5* increased likelihood of switching to an alternate antihypertensive. The remaining variants were not strongly or consistently associated with the outcomes.

Conclusions: Patients with common genetic variants in *NUMA1*, *CYP3A5* and *RYR3* had increased adverse clinical outcomes. Work is needed to establish whether outcomes of dCCB prescribing could be improved by prior knowledge of pharmacogenetics variants supported by clinical evidence of association with adverse events.

168

Leveraging Genetic Variation to Evaluate Risk Factors and Therapeutic Opportunities for Aortic Valve Stenosis: A Mendelian Randomization Analysis

Helena Urquijo[1,2*], Rebecca J Richardson[2], Tom R Gaunt[1], Tom G Richardson[1,3]

[1]MRC Integrative Epidemiology Unit, Bristol Population Health Science Institute, University of Bristol, Bristol, United Kingdom; [2]Faculty of Life Sciences, School of Physiology, Pharmacology and Neuroscience, University of Bristol, Bristol, United Kingdom; [3]Novo Nordisk Research Centre, Headington, Oxford, United Kingdom

Aims: Aortic valve stenosis (AVS) is an increasingly prevalent disease with no pharmacotherapies available. In this study, we used mendelian randomization (MR) to estimate the genetically predicted effects of established and emerging risk factors on AVS risk, as well as evaluate drug repurposing opportunities.

Methods and Results: Univariable MR analyses using genetic data of up to 462,927 UK Biobank participants supported previous MR findings indicating several established risk factors, such as elevated blood pressure and lipoprotein lipids, increase AVS risk. We additionally identified calcium (OR=1.26 per SD increase, 95% CI=1.10-1.43, P=7x10^{-4}) and lifetime smoking (OR=1.92 per SD increase, 95% CI=1.55-2.39, P=5x10^{-9}) as risk factors for AVS. Applying the same analysis on coronary artery disease (CAD) risk found comparable estimates with the notable exception of calcium which provided limited evidence of an effect on CAD (OR=1.06, 95% CI=0.97-1.15, P=0.23). Multivariable MR provided evidence that apolipoprotein B is the predominating lipoprotein trait increasing AVS risk when conditioned on low density lipoprotein cholesterol (LDL-C) and triglycerides. Finally, we found a genetically predicted effect of inhibition of antihypertensive target *ADRB1* on lower AVS risk. LDL-C lowering target *PCSK9*, lipoprotein(a) lowering target *LPA* and triglyceride lowering targets *ANGPTL4* and *LPL* also were genetically predicted to lower AVS risk.

Conclusions: We provide a comprehensive evaluation of modifiable risk factors driving the increased healthcare burden of AVS, which should be considered in risk stratification and prevention strategies. Furthermore, our findings provide valuable insight into the prioritization of the pharmacotherapies which may yield clinical benefit towards treating AVS.

169

Integrating Mendelian Randomization and Literature-Mined Evidence for Breast Cancer Risk Factors

Marina Vabistsevits[1,2*], Tim Robinson[1,2], Yi Liu[1,2], Tom Gaunt[1,2]

[1]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom

Identifying evidence from multiple sources when studying disease risk factors is one of the main challenges in population health research. Biomedical data integration platforms such as EpiGraphDB (epigraphdb.org) can facilitate evidence triangulation from different sources, improving confidence in the causal relationships of interest. In this study, we aimed to integrate Mendelian randomization (MR) and literature-mined evidence from EpiGraphDB to build a comprehensive overview of breast cancer risk factors.

We queried MR-EvE (MR "Everything-vs-Everything") data in EpiGraphDB to generate a list of causal risk factors for breast cancer and extracted literature-mined relationships for these traits to dissect how they may be linked to breast cancer. Integrating these two sources of evidence allowed us to identify potential mediators of the risk factors' effect on breast cancer. We used multivariable MR to separate the direct effects of the traits as a validation step.

We identified 175 lifestyle and molecular traits with evidence of an effect on breast cancer, including established and novel risk factors. In the mediators analysis, we determined that the negative effect of cardiotrophin-1 on ER+ breast cancer (OR:0.97[0.95:0.99]) may be mediated via leukemia inhibitory factor. IGF-1 effect (OR:1.07[1.01:1.13]) may be linked to breast cancer via KIT ligand. None of the detected mediator candidates of childhood obesity protective effect (OR:0.64[0.56:0.73]) explained it.

Our work demonstrates that using MR-EvE to identify disease risk factors is an efficient hypothesis-generating approach. Moreover, we show that integrating MR evidence

with literature-mined data may identify causal effect intermediates or uncover mechanisms behind observed effects.

170

Calcium Intake and Risk of Colorectal Cancer: A Genome-Wide Interaction Study

Virginia Díez-Obrero[1,2,3,4], Fränzel J. van Duijnhoven[5,*], Andre E. Kim[6], Ulrike Peters[7], W. James Gauderman[6], Victor Moreno[1,2,3,4] on behalf of the CCFR, CORECT, and GECCO

[1]*Oncology Data Analytics Program, Catalan Institute of Oncology (ICO). L'Hospitalet de Llobregat, Barcelona, Spain*; [2] *Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL). L'Hospitalet de Llobregat, Barcelona, Spain*; [3] *Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain*; [4] *Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain;* [5] *Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, the Netherlands;* [6] *Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California, United States of America;* [7] *Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America*

Calcium intake has been associated with a decreased risk of colorectal cancer (CRC). However, the molecular mechanisms underlying this association have not been fully elucidated. Previous gene-environment interaction studies have identified several SNPs, but with limited statistical power. In this study, we investigated SNP interactions with calcium intake in relation to CRC risk in the largest dataset to date. We used genotype and self-reported total (dietary and supplemental) calcium intake information from 33,731 CRC cases and 43,824 controls of European descent. Genome-wide interaction models were applied with the GxEScanR package, adjusting for sex, age, study, total energy intake and genetic ancestry. In addition to standard interaction tests we applied joint and two-step tests.

Higher total calcium intake was associated with a reduced CRC risk (OR per quartile increase, 0.87; 95% CI, 0.85–0.89). Similar results were observed in stratified analyses by sex and tumor anatomic location. Two-step testing identified a statistically significant interaction signal at the lactose intolerance-related 2q21.3 locus, which has been associated with lactose metabolism and gut microbiome composition. Each quartile increase in total calcium consumption was associated with a reduced CRC risk of 17% (OR 0.83, 95%CI (0.79-0.87) in people with the CC genotype, 14% (OR 0.86, 95%CI (0.84-0.88) in those with the CA genotype, and 10% (OR 0.90 95%CI (0.88-0.92) in those with the AA genotype. We demonstrate that well-powered studies which incorporate established environmental risk factors can provide discoveries of interactions between calcium intake and genetic variants in relation to CRC risk.

171

Development of Novel Epigenetic Clock in Targeted Methylation Sequencing Data

Denitsa I. Vasileva[1*], Ming Wan[1], Allan B. Becker[2], Edmond S. Chan[3], Yuka Asai[4], Ann Clarke[5], Catherine Laprise[6,7], Andrew J. Sandford[1], Celia M. T. Greenwood[8], Denise Daley[1]

[1]*Center for Heart Lung Innovation, Faculty of Medicine, University of British Columbia, Vancouver, Canada;* [2]*Department of Pediatrics and Child Health, University of Manitoba, Manitoba, Canada;* [3]*BC Children's Hospital Research Institute, Faculty of Medicine, Vancouver, Canada;* [4]*Department of Medicine, Queen's University, Ontario, Canada;* [5]*Cummings School of Medicine, University of Calgary, Alberta, Canada;* [6]*Centre intersectoriel en sante durable (CISD) de l'Universite du Québec á Chicoutimi, Saguenay, Canada,* [7]*Centre intégre universitaire de sante et de services sociaux (CIUSSS) du Saguenay–Lac-Saint-Jean, Saguenay, Canada;* [8]*Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada*

Background: The epigenetic clock is a biomarker which relies on changes in DNA methylation at age-informative Cytosine-phosphate-Guanine (CpG) dinucleotides. Most clocks have been developed using adult samples and methylation arrays. Targeted methylation sequencing—an increasingly popular novel approach which measures methylation at millions more CpGs than the arrays—is likely to identify novel age informative CpGs.

Objectives: 1. Identify novel age informative CpG sites in targeted methylation sequencing data from both pediatric and adult samples; 2. Develop a novel sequencing-based epigenetic clock

Methods: Targeted methylation sequencing using the Illumina TruSeq Methyl Capture library (San Diego, California) was conducted on 932 samples from three Canadian studies-Canadian Asthma Primary Prevention Study (CAPPS, n=632, age range: birth – 43years); Saguenay-Lac-Saint-Jean study (SLSJ, n=180, age range:5 – 93years); Canadian Peanut Allergy Registry (CanPAR, n=120, age range:5 – 38years). Age informative CpGs were identified in the CAPPS and SLSJ datasets using linear and mixed effects modeling. Novel epigenetic clocks were developed using elastic net regression and gradient-boosting trees. Training studies included CAPPS and SLSJ, with CanPAR as the validation set. Accuracy of the novel clocks was calculated using absolute error (AE= |predicted-reported age|).

Results: Thousands of novel CpGs associated with age were identified in the linear regression (582,187 CpG with *P* values: $<1x10^{-8}$ – $1x\ 10^{-260}$) and mixed effects modeling (559,870 with *P* values: $<1x10^{-8}$ – $1x10^{-250}$). Two novel age predictors with lower AE than current clocks (e.g., Horvath and PedBE clocks) will be presented.

Conclusions: Accuracy of epigenetic clocks is achieved when using CpGs from targeted methylation sequencing.

172

How Reliable is the Salivary Microbiome Information Obtained From the Whole Genome Sequencing of Human Saliva Samples?

Lourdes Velo Suarez[1,2]*, Anthony Herzig[1], Gaëlle Le Folgoc[1], Stephanie Gouriou[1], Marie Zins[3], Marcel Goldberg[3], Geneviève Hery-Arnaud[1,2], Emmanuelle Genin[1]

[1]Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200, Brest, France [2]Centre Brestois d'Analyse du Microbiote (CBAM), CHU Brest, Brest, France; [3]Inserm-Paris Saclay University, Université de Paris, Villejuif, France

Topic -- OMICs: methods and applications

The past two decades have seen tremendous advances in understanding human genetic variation and its implication in disease. Similarly, the microbiome was also shown to play a significant role in human health and disease. However, the relationship between host genetic variation and microbiome composition is largely unknown, mainly because of the elevated cost of processing both types of samples with high throughput sequencing. This study explores how well an individual salivary microbiome can be characterized by using whole-genome sequencing (WGS) data obtained from DNA saliva kits designed to study human genome variation. WGS was performed on saliva samples from 35 healthy individuals who received saliva kits at home. The relative abundance distributions obtained from the analysis of non-human reads were compared against those obtained by specific 16S rRNA gene resequencing of the same DNA samples. The results showed that 16S sequencing detects only part of the salivary microbiome revealed by WGS analysis. Low abundant taxa were identified on the WGS data that could not be captured by 16S sequencing. Interestingly, some of these taxa could be linked to oral diseases such as periodontitis.

Our microbiome communities were very similar to those described in the Human Microbiome Project (HMP) and core salivary microbiome studies, showing that accurate microbiome profiles can be obtained from read data generated for human whole-genome sequencing.

173

A Signature of Platelet Reactivity in Complete Blood Count Scattergrams identifies genes with Thrombotic disease Associations

Hippolyte Verdier [1,2,*], Kate Downes[3,4], William Astle[1,5], Ernest Turro[3,6]

[1]MRC Biostatistics Unit, University of Cambridge, United Kingdom [2]Decision and Bayesian Computation, USR 3756 (C3BI/DBC) & Neuroscience department CNRS UMR 3751, Institut Pasteur, CNRS, Paris, France; [3]Department of Haematology, University of Cambridge, United Kingdom; [4]East Genomic Laboratory Hub, Cambridge University Hospitals National Health Service Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom; [5]National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom; [6]Genetics and Genomic Sciences Department, Icahn School of Medicine at Mount Sinai, United States

Genetic studies of platelet reactivity (PR) offer a hypothesis-free approach to antiplatelet drug target discovery. However, due to the technical challenges of measuring PR on a large scale, genetic studies of PR have been hampered by small (N<4,000) sample sizes, identifying, in aggregate, only 15 distinct PR-associated loci. We have trained a predictor of flow cytometry-measured PR from complete blood count (CBC) scattergrams from hematology analyzers widespread clinical use. A GWAS of this predicted phenotype in 29,806 blood donors of the INTERVAL study identified 21 distinctly associated loci, six of which were previously known to be associated with agonist-specific PR phenotypes. We built a genetic score of PR from the effect sizes of the identified variants determined using flow cytometry-measured phenotypes on an independent cohort of 1,373 donors. We then showed that this score was associated with thrombosis-related health outcomes in UK Biobank, linking a genetic score of PR to thrombosis for the first time. Our approach provides a blueprint for studying the determinants of hard-to-measure but biologically important traits by phenotype imputation.

174

Construction of Artificial Most Representative Trees by Minimizing Tree-based Distance Measures

Björn-Hergen von Holt [1]*, Ana Westenberger[2], Inke R. König [1]

[1]Institute of Medical Biometry and Statistics, University of Lübeck, Germany; [2]Institute of Neurogenetics, University of Lübeck, Germany

Selecting most representative trees from a random forest is one way to improve the interpretability of the otherwise very complex ensemble. For this, pair-wise tree distances are estimated, and the tree with the minimal mean distance is selected to represent the forest. To decorrelate the trees within the random forest, at each node in a tree only a small subset of independent variables is used to select the optimal split point. This leads to the problem that each split is only a local and not necessarily a global optimum. Thus, the interpretation of the selected most representative trees is still challenging.

To overcome this issue, we developed an algorithm which generates an artificial most representative tree using a greedy approach selecting the split point at each node that minimizes the distance to all trees of the original random forest. It iteratively adds new splits to the artificial representative tree until no new split is found, which decreases the mean distance to the forest.

In an extensive simulation study, we compared artificial most representative trees with those directly selected from the forest regarding their prediction performance, ability to condense the information of the ensemble and coverage of the meaningful predictors. Additionally, we applied both methods to a genetic data set of X-linked dystonian-parkinsonism (XDP) and evaluated the resulting most representative trees with regard to recent results on genetic modifiers of age at onset in XDP.

Finally, we added the new method to our existing R package timbR (https://github.com/imbs-hl/timbR).

175

Using Structural Equation Modelling to Partition Genetic Effects on Birth Weight at Human Leukocyte Antigen Loci into Maternal and Fetal Components

Geng Wang[1, 2*], Nicole M. Warrington[1, 2, 3, 4], David M. Evans[1, 2, 3]
[1]*The University of Queensland Diamantina Institute, The University of Queensland, Brisbane, Australia;* [2]*Institute for Molecular Biosciences, The University of Queensland, Brisbane, Australia;* [3]*Medical Research Council Integrative Epidemiology Unit, University of Bristol, United Kingdom;* [4]*K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway*

Birth weight (BW), as a proxy for intrauterine growth, is influenced by both fetal and maternal genetic factors. SNPs in the human leukocyte antigen (HLA) region in both maternal and fetal genomes have been robustly associated with BW in previous studies. However, no study to date has specifically partitioned the association between BW and classical HLA alleles into maternal and fetal components. We used structural equation modelling (SEM) to estimate the indirect maternal (i.e. via the intrauterine environment) and direct fetal effect of classical HLA alleles on BW. Our SEM leverages the data structure of the UK Biobank (UKB), which includes participants' own BW and/or the BW of their firstborn children (in the case of UKB females). We show via simulation that our model yields asymptotically unbiased estimates of maternal and fetal effects on BW and appropriate type I error rates, in contrast to simple regression models. Asymptotic power calculations show that we have sufficient power to detect moderate sized maternal or fetal effects of common HLA alleles on BW in the UKB. Applying our SEM to imputed classical HLA alleles and own and offspring BW of ~250,000 participants from the UKB replicated previous reported association at the *HLA-C* locus (*C*04:01*, maternal effect, $P = 3.67\times10^{-3}$) and revealed strong evidence for maternal (*HLA-A*03:01*, *B*35:01*, *B*38:01*; $P < 0.001$) and fetal effects (*HLA-DRB1*11:04*, $P < 0.001$) of non-*HLA-C* alleles on BW. These novel allelic associations with BW provide insight into the immunogenetics of fetal growth *in utero*.

176

Integration of Genetically Regulated Expression and Neurophysiological Traits Implicates 93 Neuro Imaging-derived Phenotypes in Alzheimer's Disease

Ting-Chen Wang [1]*, Xavier Bledsoe[1], Douglas Shaw[1], Hung-Hsin Chen[1], International Genomics of Alzheimer's Project, Adam C. Naj[2], William Bush[3], Eric R. Gamazon[1], 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]*Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America;* [3]*Department of Population & Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America*

Late-onset Alzheimer's disease (LOAD) is a highly polygenic neurological disorder, and genome-wide association studies (GWAS) have identified dozens of associated genetic loci. However, the functional role of many genetic loci identified through SNP-based analysis of LOAD risk and the impact of LOAD-associated genetic architecture on brain physiology and structure remain largely unknown. Genetically regulated gene expression analysis (GReX) provides a more powerful methodology to investigate the effect of genetically determined expression of genes and gene networks on complex diseases. Here we report the integration of these genetic and expression quantitative trait loci-based associations in models built from brain feature data in the UK Biobank. By modeling the relationship of GReX with 3,079 neurophysiological traits, our approach leverages a broad collection of brain endophenotypes, i.e., neuro imaging-derived phenotypes (NIDPs), representing brain structure and function. The NIDPs enable physiological insights into the LOAD-associated gene expression effects and the corresponding altered brain regions.

GReX analysis on 39 LOAD datasets composed of 58,713 cases and controls from the International Genomics of Alzheimer's Project across 51 tissues, 32 significant genes were found cross all tissues and cross brain tissues. We adopted a methodology that identifies NIDPs associated with these genes whose genetically determined expression level is associated with brain structure and function. Ninety-three NIDPs were significantly associated (Bonferroni-adjusted *P* value < 0.05) with the LOAD-associated genes. Future studies on these genes and NIDPs can help uncover expression-mediated changes to the brain physiology and structure in individuals with elevated risks of Alzheimer's disease.

177

Improved Risk Prediction using Functionally Calibrated Polygenic Risk Scores

Xuexia Wang[1*], Jianjun Zhang[1], Samantha Gonzales[1]
[1]*Department of Mathematics, University of North Texas, Denton, Texas, United States of America*

Correctly predicting disease risk for an individual plays an important role in disease prevention and early treatment selection. Even though genome wide association studies (GWAS) have identified thousands of genetic variants underlaying complex traits, prediction accuracy with polygenic risk scores (PRSs) remain moderate for most diseases. This is largely due to the challenges in accurately estimating the effect sizes of genetic variants and whether these variants are functionally relevant. In this study, we propose a Bayesian method that more accurately estimates the effect sizes of functionally relevant genetic variants by incorporating multiple GWAS and functional annotations into a prior distribution. PRS employing these more accurate effect sizes improves prediction accuracy for disease status. Simulation studies demonstrate that the proposed method outperforms the existing comparison methods regarding the area under receiver operating characteristic curve (ROC) for binary traits and the correlation coefficient be-

tween the predicted and observed trait values for quantitative traits. Application to schizophrenia data from the Psychiatric Genomics Consortium, Type 2 diabetes (T2D) data from the UK Biobank, and the 41 traits in the electronic Medical Records and Genomics (eMERGE) Network data reveals that the proposed method performs more accurately in risk prediction than other comparable methods in most of the diseases/traits.

178

An Optimally Weighted Combination Method to Detect Novel Trait–gene Association using GWAS Summary Data

Xuexia Wang[1]*, Samantha Gonzales[1], Callum Doyle[1]
[1]Department of Mathematics, University of North Texas, Denton, Texas, United States of America

Genome-wide association studies (GWAS) have identified over 30,000 disease-related genetic variants. However, these variants explain only a small proportion of heritability for most traits. Missing heritability may be due to small effect sizes, which need large sample sizes and/or powerful methods to investigate. Increasing evidence suggests that certain traits are associated with multiple variants shared between traits, rendering standard 1-trait 1-variant association test which is less powerful for not considering between-trait and between-variant correlation. A gene-based multiple traits method can not only increase the power of detecting trait-variant associations, but also provide insight into intertwined genetic mechanisms underlying complex disease. We propose a statistical method-testing an optimally weighted combination of multiple traits using GWAS Summary data (TOMS) to test the association between multiple traits and multiple variants in a genomic region. TOMS is based on the score test under a linear model. The weighted combination of traits is tested by the maximized score statistic over weights. Simulation studies demonstrate that TOMS is valid and powerful: it is the most powerful test or comparable to the most powerful test in varying scenarios. We applied TOMS and comparison methods to two European GWAS summary datasets: 1) Global Lipids Consortium meta-analysis for three plasma lipids (100, 000 individuals); 2) international MAGIC consortium meta-analysis for two glycemic traits (46,186 individuals). TOMS identified multiple novel genes missed by the other methods. Functional analyses indicate that these genes are biologically relevant to plasma lipids or glycemic traits that we investigated.

179

Kernel Machine Regression Pathway Analysis with Longitudinal Phenotypes

Bernadette Wendel[1]*, Markus Heidenreich[1], Monika Budde[2], Maria Heilbronner[2], Mojtaba Oraki Kohshur[2], Sergi Papiol[2,3], Peter Falkai[3], Thomas G. Schulze[2,4,5], Urs Heilbronner[2], Heike Bickeböller[1]
[1]Department of Genetic Epidemiology, University Medical Center, Georg-August-University, Göttingen, Germany; [2]Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, Ludwig Maximilian University, Munich, Germany; [3]Department of Psychiatry and Psychotherapy, University Hospital, Ludwig Maximillian University, Munich, Germany; [4]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, New York, United States of America; [5]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

We developed and implemented a kernel machine regression (KMR) approach for pathway association analysis with longitudinal phenotypes using original genotype array data. It allows interaction testing between the considered pathway and the longitudinal course and the integration of (part of) the network structure of the pathway.

For longitudinal data we used additional random effect terms in a linear mixed model. For the corresponding altered variance component pathway test based on a genetic similarity kernel, we enabled the classical linear as well as a network kernel allowing for topology information. In addition, our approach makes the investigation of interaction between longitudinal course and pathway computationally feasible. We implemented the approach in form of an R package "kalpra".

For our approach we studied type I error and power for main, interaction and joint pathway effects using both kernels via simulations with samples of 1000 individuals and 100,000 replications. We tested pathway topologies of equal size (19 genes) but of different densities 0.2, 0.5 and 0.8, selecting nine causal SNPs allocated in three causal genes with equal effect sizes. The type I error was generally maintained. The network kernel was slightly conservative, the linear kernel was more conservative. For the different scenarios of main and interaction effect, the network kernel demonstrated an increase of power with decreasing density. The network kernel with density 0.2 yielded the highest power in general when simulating only main or interaction effect. We, further, plan to apply the method on PsyCourse data, a longitudinal multicenter study of psychosis.

180

Accurate Detection of Genetic Sharing Between Rare and Common Diseases Enables more Powerful Association Discovery in the Rare Disease Context

Thomas W. Willis[1]*, Chris Wallace[1,2] on behalf of the INTREPID Consortium
[1]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [2]Cambridge Institute of Therapeutic Immunology and Infectious Disease, University of Cambridge, Cambridge, United Kingdom

Study of the genetic relationships between rare and more common diseases can help elucidate their aetiology and aid variant discovery, but this is made difficult by small case numbers. The genetic correlation is one means of quantifying such relationships, but its use is infeasible in a small-sample context: the popular Linkage Disequilibrium Score Regression method requires a minimum of around 5,000 cases for plausibly precise estimation of the genetic correlation. As an alternative we explored a nonparametric test statistic, the genome-wide pairwise-association signal sharing (GPS) statistic proposed by Li et al. We show the statistic's null distribution can be modelled by an extreme value distribution.

We applied the test to simulated and UK Biobank genome-wide association studies (GWAS) of varying sample sizes. The GPS test was more powerful than both the genetic correlation and a standard test of bivariate independence in the small-sample context (1,000 - 5,000 cases), whilst maintaining type 1 error rate control. We used the GPS to search for diseases genetically related to rare antibody-deficient primary immunodeficiencies (AD-PID) using an AD-PID GWAS of 733 cases, identifying several related immune-mediated diseases with large GWAS, such as type 1 diabetes (T1D). One approach to increasing the power to detect variant-disease associations is the use of cross-phenotype, pleiotropy-informed methods, which leverage known associations from related traits. Applying a conditional false discovery rate procedure to leverage T1D, we identified three credible new loci for AD-PID, doubling the number of associations reported in the original study.

181

Differences in Genetic Effects on Glomerular Filtration Rate Between Individuals with and without Diabetes

Thomas W. Winkler[1*], Humaira Rasheed[2,3,4], Alexander Teumer[5,6,7], Mathias Gorski[1,8], Bryce X. Rowan[9,10], Adriana M. Hung[10], Florian Kronenberg[11], Anna Köttgen[12], Cristian Pattaro[13], Iris M. Heid[1] on behalf of the CKDGen Consortium.
[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]K. G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology, Trondheim, Norway; [3]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [4]Division of Medicine and Laboratory Sciences, University of Oslo, Oslo, Norway; [5]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany; [6]DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany; [7]Department of Population Medicine and Lifestyle Diseases Prevention, Medical University of Bialystok, Bialystok, Poland; [8]Department of Nephrology, University Hospital Regensburg, Regensburg, Germany; [9]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [10]Department of Veteran's Affairs, Tennessee Valley Healthcare System (626)/Vanderbilt University, Nashville, Tennessee, United States of America; [11]Department of Genetics and Pharmacology, Institute of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria; [12]Institute of Genetic Epidemiology, Department of Data Driven Medicine, Faculty of Medicine and Medical Center–University of Freiburg, Freiburg, Germany; [13]Eurac Research, Institute for Biomedicine (affiliated with the University of Lübeck), Bolzano, Italy

Estimated glomerular filtration rate based on serum creatinine (eGFRcrea) is an important biomarker for kidney function. Reduced eGFRcrea can lead to kidney failure. Diabetes mellitus (DM) is a major risk factor for reduced eGFRcrea. However, little is known about its interaction with the genetic determinants of eGFRcrea. Here, we conducted genome-wide association study (GWAS) meta-analyses of eGFRcrea separately in individuals with and without DM (n_{DM}~180,000; n_{NoDM}~1.2M). Analyses incorporated 72 studies from the CKDGen Consortium and four large biobanks including UK Biobank. We identified seven loci: six showing genome-wide significant differences between DM and noDM ($P_{Diff}<5x10^{-8}$) and one additional from a candidate screen of 634 known variants ($P<5x10^{-8}$ and $P_{Diff}<0.05/634$). Five of these exhibited more pronounced effects in DM (near *UMOD/PDILT*, *TPPP*, *DCDC5*, *NRIP1*, and *SLC22A2*), one was DM-specific (near *CSRNP1;* no effect in noDM) and one was noDM-specific (near *MED1/NEUROD2*). By accounting for differences using a joint 2-degrees-of-freedom test taking main and interaction effects into account, we identified 32 novel genome-wide significant eGFRcrea loci ($P_{Joint}<5x10^{-8}$). These included four additional loci with suggestive differences ($P_{Diff}<0.05/32$, near *SH3BP4*, *ALPL*, *LOXL4* and *PIK3CG*) and *CUBN*, which is a prominent GWAS region for kidney damage markers but previously undetected genome-wide for eGFRcrea. Gene prioritization analysis at the highlighted loci yielded 18 genes that may inform drug development. We highlight the existence of DM-specific effects, which can inform the target group for medication, but our results also suggest that most interventions aimed at altering eGFRcrea should be equally effective among persons with and without diabetes.

182

Transferability of Polygenic Risk Scores to Admixed Groups Reveals Trait-specific Interaction of Genetic Ancestry and Non-Genetic Factors

Genevieve L. Wojcik[1]*, Mariaelisa Graff[2], Charles Kooperberg[3], Ulrike Peters[3], Steven Buyske[4], Tara C. Matise[5], Christopher Haiman[6], Kari E. North[2], Christopher R. Gignoux[7], Eimear E. Kenny[8]
[1]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America; [2]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States of America; [3]Public Health Sciences Division, Fred Hutch Cancer Center, Seattle, Washington, United States of America; [4]Department of Statistics, Rutgers University, Piscataway, New Jersey, United States of America; [5]Department of Genetics, Rutgers University, Piscataway, New Jersey, United States of America; [6]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [7]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States of America; [8]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America

Polygenic risk scores (PRS) have been shown to have diminished performance in populations that differ from the training data, most notably in terms of genetic ancestry, but also for demographic and environmental factors. However, there has been limited investigation into the role of heterogeneity in ancestry and non-genetic factors within populations, specifically within recently admixed groups.

Using the Population Architecture using Genomics and Epidemiology (PAGE) Study, we examined in African American (N=15,993) and Hispanic/Latino (N=18,950) participants the performance of PRS for height (PRS_{height}) and body mass index (BMI; PRS_{BMI}) previously trained on European ancestry data. Using ancestry proportions estimated with ADMIXTURE, we find positive correlations for non-European ancestry with PRS_{BMI} but negative correlations with PRS_{height}, indicating a trait-specific role of ancestries on PRS estimation. Model fit improvements after adjustment for ancestry depended upon both the trait and specific ancestry, with increased performance after adjustment for Indigenous to the Americas (AME) ancestry for PRS_{BMI} and African (AFR) ancestry for PRS_{height}, but decreased performance for other pairwise relationships (PRS_{BMI} and AFR, PRS_{height} and AME). Proportion of PRS variance explained by genetic ancestry and epidemiologically-relevant variables, such as smoking, was heterogeneous by background within Hispanic/Latino participants (i.e. Puerto Rican, Mexican, etc.) for both genetic and non-genetic factors. Our results highlight the need for PRS to be developed and fine-tuned in populations that are diverse both in terms of genetic ancestry and epidemiologically-relevant non-genetic variables to ensure equity in the translation of genomic findings to improve human health.

183

A Rule-based Approach for Identifying Genetic Heterogeneity in Survival Data

Alexa A. Woodward[1*], Ryan J. Urbanowicz[2], Jason H. Moore[2]
[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, California, United States of America

Genetic heterogeneity is a prevalent mechanism underlying many complex diseases. Most current survival analysis methods such as Cox proportional hazards models and random survival forests struggle to detect and characterize genetic heterogeneity and other complex patterns of association due to insufficient power, violated assumptions, overfitting, or other challenges. Learning classifier systems (LCS) are a type of rule-based machine learning algorithms that are uniquely suited to heterogeneous problem domains, but to date, have not been adapted for survival analysis. Most traditional statistical and machine learning models seek to develop a single model that represents all the data; LCSs evolve a generalizable and interpretable population of rules that more flexibly models underlying heterogeneity. We propose a novel "survival-LCS" that fully accounts for right-censored observations and makes no assumptions about baseline hazard or survival distributions. As proof of concept, we evaluated the survival-LCS on simulated genetic survival datasets of increasing complexity and compared the results and performance against traditional survival analysis methods. The four genetic models included main effect, epistatic, additive, and heterogeneous models, simulated across a range of heritability values, minor allele frequencies, and number of features. From these preliminary analyses, we demonstrated the ability of a survival-LCS to identify complex patterns of association including epistasis and genetic heterogeneity in survival data. We show that survival-LCS can also reliably predict survival times and survival distributions for individual subjects, an element that may prove useful for clinical applications such as informing self-controls in single-arm clinical trials.

184

A Varying Coefficient Model to Jointly Test Genetic and Gene–environment Interaction Effects

Chao Xing[1]
[1]McDermott Center of Human Growth and Development, Department of Bioinformatics, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

Most human traits are influenced by the interplay between genetic and environmental factors. Many statistical methods have been proposed to screen for gene-environment interaction (GxE) in the post genome-wide association study era. However, most of the existing methods assume a linear interaction between genetic and environmental factors toward phenotypic variations, which diminishes statistical power in the case of nonlinear GxE. In this paper, we present a flexible statistical procedure to detect GxE regardless of whether the underlying relationship is linear or not. By modeling the joint genetic and GxE effects as a varying-coefficient function of the environmental factor, the proposed model is able to capture dynamic trajectories of GxE. We propose a likelihood ratio test with a fast Monte Carlo algorithm for hypothesis testing. Simulations were conducted to evaluate validity and power of the proposed model in various settings. Real data analysis was performed to illustrate its power, in particular, in the case of nonlinear GxE.

185

Bayes Factor with Conjugate Prior for Region-Based Rare Variant Analysis

Jingxiong Xu[1], Nanwei Wang[2], Laurent Briollais[1,3]
[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [2]Department of Mathematics and Statistics, University of New Brunswick, Fredericton, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Next Generation Sequencing technology provides opportunities to discover rare variants (RVs) associated with complex human diseases. We previously introduced a region-based test for case-control designs using a Bayes Factor (BF) statistic (Xu et al., Biometrics, 2020) where the association between a set of RVs in the same region (e.g. a gene) and a disease was assessed. In particular, we showed that the new BF statistic outperforms standard methods (SKAT, SKAT-O, Burden test) in case-control studies with moderate sample sizes and is equivalent to them under large sample size scenarios. Here we extend this approach using the generalized linear model framework and its conjugate prior (Chen et al., Statistica Sinica, 2003), which can handle outcomes of different

types (binary, continuous, count), informative functional annotation and unbalanced designs. We also demonstrate that implementing a variable selection step (i.e. using the birth-death algorithm) of RVs that contribute the most to the region-based association can both improve the power of the region-based statistic and also help identify functional RVs associated with the outcome. We also warn against using functional annotation when the functional information is mis-specified. Simulation studies and application to whole-exome sequencing data from UK Biobank and cancer outcomes are conducted to assess the finite sample properties of our method.

186

A Novel Penalized Inverse-Variance Weighted Estimator for Mendelian Randomization with Applications to COVID-19 Outcomes

Siqi Xu[1]*, Peng Wang[2], Wing Kam Fung[1], Zhonghua Liu[1]
[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China; [2]Department of Epidemiology and Biostatistics, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Mendelian randomization (MR) utilizes genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure variable on an outcome of interest even in the presence of unmeasured confounders. However, the popular inverse-variance weighted (IVW) estimator could be biased in the presence of weak IVs, a common challenge in MR studies. In this article, we develop a novel penalized inverse-variance weighted (pIVW) estimator, which adjusts the original IVW estimator to account for the weak IV issue by using a penalization approach to prevent the denominator of the pIVW estimator from being close to zero. Moreover, we adjust the variance estimation of the pIVW estimator to account for the presence of horizontal pleiotropy. We show that the recently proposed debiased IVW (dIVW) estimator is a special case of our proposed pIVW estimator. We further prove that the pIVW estimator has smaller bias and variance than the dIVW estimator under some regularity conditions. We also conduct extensive simulation studies to demonstrate the performance of the proposed pIVW estimator. Furthermore, we apply the pIVW estimator to estimate the causal effects of five obesity-related exposures on three coronavirus disease 2019 (COVID-19) outcomes. Notably, we find that hypertensive disease is associated with an increased risk of hospitalized COVID-19, and peripheral vascular disease and higher body mass index are associated with increased risks of COVID-19 infection, hospitalized COVID-19 and critically ill COVID-19.

187

Using Human Genetics to Evaluate the Causal Role of Circulating Inflammatory Markers in Risk of Adult Cancer

James Yarmolinsky[1,2]*, Jamie Robinson[1,2], Kostas K. Tsilidis[3,4], Abbas Dehghan[4], Mattias Johansson[5], Daniela Mariosa[5], Marc J. Gunter[6], Lambertus A. Kiemeney[7], George Davey Smith[1,2], Richard M. Martin[1,2,8]
[1]*MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; [4]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St Mary's Campus, London, United Kingdom; [5]Genomics Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France; [6]Nutrition and Metabolism Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France; [7]Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands; [8]University Hospitals Bristol, NHS Foundation Trust, National Institute for Health Research Bristol Biomedical Research Centre, University of Bristol, Bristol, United Kingdom*

Tumor-promoting inflammation is a "hallmark" of cancer and laboratory and epidemiological studies have reported links between various inflammatory markers and cancer risk. The causal nature of these relationships and, thus, the suitability of these markers as intervention targets for cancer prevention is unclear. We meta-analyzed 6 genome-wide association studies of circulating inflammatory markers comprising 59,969 participants of European ancestry. We then used combined *cis*-Mendelian randomization and colocalization to evaluate the causal role of 75 circulating inflammatory markers in risk of 32 adult cancers in 347,300 cancer cases and up to 1,015,204 controls. Genetic instruments for inflammatory markers were constructed using genome-wide significant ($P<5.0\times10^{-8}$) *cis*-acting SNPs (i.e. ±250KB from the gene encoding the relevant protein) in weak linkage disequilibrium (LD, $r^2<0.10$). Effect estimates were generated using inverse-variance weighted random-effects models and standard errors were inflated to account for weak LD between variants with reference to the 1000G Phase 3 European panel. We find strong evidence for effects of genetically-proxied circulating adrenomedullin on breast cancer risk (OR:1.19, 95%CI:1.10-1.29, $P=2.02\times10^{-5}$, H_4=84.3%), interleukin-23 receptor on pancreatic cancer risk (OR:1.42, 95%CI:1.20-1.69, $P=6.72\times10^{-5}$, H_4=73.9%), antithrombin on triple-negative breast cancer risk (OR:3.62, 95%CI:1.70-7.70, $P=8.30\times10^{-4}$, H_4=72.7%), and macrophage migration inhibitory factor on bladder cancer risk (OR:1.14, 95%CI:1.05-1.23, $P=1.43\times10^{-3}$, H_4=76.1%), among other findings. Our comprehensive analyses represent the largest human genetics evaluation of circulating inflammatory markers in risk of adult cancers to date. Our findings highlight various novel inflammatory markers implicated in cancer development and suggest pharmacological targeting of these markers as a potential strategy for primary cancer prevention.

*Corresponding author: James Yarmolinsky, PhD, MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK, james.yarmolinsky@bristol.ac.uk

188

Genome-wide Cross-trait Analysis and Bi-directional Mendelian Randomization Study of COVID-19 with Venous Thromboembolism

Xin Huang[1], Minhao Yao[2*], Peixin Tian[2], Zhonghua Liu[2], Jie V. Zhao[1]

[1]Division of Epidemiology and Biostatistics, School of Public Health, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China; [2]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

Venous thromboembolism (VTE) and COVID-19 may share a common genetic architecture in etiology, which has not been comprehensively investigated. In this study, we leveraged summary-level data from the latest COVID-19 host genetics consortium and UK Biobank to study the genetic commonality between VTE and COVID-19. The cross-trait analysis identified 5, 5 and 4 shared genetic loci for severe COVID-19, COVID-19 hospitalization and SARS-COV2 infection with VTE, respectively, including five novel mapped genes (*NME7, ADAMTS13, TSPAN15, F5* and *SLC39A8*), which were enriched for expression in the liver, whole blood and spleen. Five genetic loci were found to share causal variants between COVID-19 and VTE, which are localized in the *ABO* gene and the nearby *ADAMTS13* gene regions. Mendelian randomization analysis showed that genetically predicted VTE was associated with an increased risk of severe COVID-19 (OR=1.08, *P* value=3.8×10^{-2}), COVID-19 hospitalization (OR=1.07, *P* value=2.8×10^{-2}) and SARS-COV2 infection (OR=1.04, *P* value=4.9×10^{-2}), and the reverse associations were not significant. Our novel findings advance the understanding of COVID-19 and VTE at the molecular and functional levels.

189

Country-specific Calibration of Polygenic Risk Scores for Breast Cancer in European Ancestry Populations

Kristia Yiangou[1,2*], Nasim Mavaddat[3], Joe Dennis[3], Andreas Hadjisavvas[1], Maria A. Loizidou[1], Jacques Simard[4], Antonis C. Antoniou[3], Douglas F. Easton[3,5], Kyriaki Michailidou[2] on behalf of the Breast Cancer Association Consortium.

[1]Department of Cancer Genetics, Therapeutics and Ultrastructural Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; [2]Biostatistics Unit, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; [3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [4]Genomics Center, Centre Hospitalier Universitaire de Québec—Université Laval Research Center, Québec City, Canada; [5]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom

A 313 SNP polygenic risk score (PRS_{313}) has been recently developed for the prediction of breast cancer risk in women of European ancestry. Here, we aimed to specifically explore the mean PRS_{313} in different countries in Europe and explain what might be causing variation across the different populations. For the analysis, 111,814 breast cancer cases and 94,718 controls were used from women of European ancestry participating in the Breast Cancer Association Consortium (BCAC), from 17 countries in Europe, as well as Australia, Canada, Israel, and the United States of America.

Mean and SD of the PRS_{313} were calculated for each country, separately in cases and controls. Additionally, we examined the distribution of the PRS when excluding variants with large frequency differences across countries, with coefficient of variation greater than 0.3.

Mean PRS_{313} differs significantly across countries in Europe, being highest in Greece and lowest in Ireland. When the risk categories in each country were calculated based on the distribution of the pooled control dataset, we observed misclassification of individuals, up to 3.6% in the highest decile. Excluding the 17 SNPs with the most variable frequency across countries, did not explain the entire observed variability.

These results suggest that country specific calibration of the PRS is necessary for accurate risk estimation and classification. We are going to further explore the extent to which the variability in the mean PRS_{313} can be explained by principal components and whether more accurate estimations can be achieved using empirical Bayes approaches.

190

A Simulation Comparison of vQTL Mapping Approaches

Xiaopu Zhang [1,*], Jordana T. Bell [1]

[1]Department of Twin Research and Genetic Epidemiology, School of Life Course Science, King's College London, London, United Kingdom

Gene-environment (GxE) and gene-gene (GxG) interactions lead to changes in phenotypic variability. Identification of quantitative trait loci for phenotypic variance (vQTLs) is an efficient approach to identify genetic factors that may form GxG and GxE interactions. Genome-wide vQTL mapping has been conducted for multiple human traits, including BMI and gene expression. Multiple methods have been developed to detect vQTLs, but a systematic comparison is lacking.

We carried out a simulation study to evaluate the performance of 9 previously proposed vQTL detection methods for a series of underlying genetic models. The genetic models assumed that a phenotype is affected by environmental factors (E), genetic factors (G), and joint GxE effects in up to 2,000 independent samples. Detection of genetic effects was assessed using standard QTL analysis (mean-QTL) only, variance QTL analysis (vQTL) only, both, and neither. After 1000 simulations, the Brown-Forsythe test and deviation regression model (DRM) showed consistently greatest power and lowest type I error across a range of genetic effect sizes, minor allele frequencies, sample sizes, and phenotype distributions. The performance of two methods was also stable across varying mean effect and variance, suggesting that these approaches minimize phantom vQTLs resulting from the mean-variance association.

In conclusion, our study considers vQTL as an efficient

approach to reveal GxG and GxE interactions. Our simulation results suggest that the Brown-Forsythe and DRM models are optimal choices to map vQTLs in independent samples. Detection of such non-additive genetic effects can affect many phenotypes and uncover the mechanism underlying human complex traits.

191

PV-CS: A Method for Individualized Disease Prediction by Dissecting Angular-based Relationship to Subpopulations

Yexian Zhang[1,2*], Xiaoxuan Xia[2,3], Qi Li[1,2], Rui Sun[4,5], Marc Ka Chun Chong[1,2], Benny Chung-Ying Zee[1,2], William Ka Kei Wu[1,6,7], Matthew Tak Vai Chan[6], Maggie Haitian Wang[1,2]

[1]The Chinese University of Hong Kong Shenzhen Research Institute, Shenzhen, China; [2]JC School of Public Health and Primary Care, the Chinese University of Hong Kong, Hong Kong SAR, China; [3]Department of Statistics, the Chinese University of Hong Kong, Hong Kong SAR, China; [4]Department of Information Management, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, China; [5]The Department of Clinical Oncology, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, China; [6]Department of Anaesthesia and Intensive Care and Peter Hung Pain Research Institute, The Chinese University of Hong Kong, Hong Kong SAR; [7]State Key Laboratory of Digestive Disease and LKS Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR

Correspondence: maggiew@cuhk.edu.hk

Keywords: prism vote, Alzheimer's disease prediction, genome sequencing, subpopulations, angular distance, cosine similarity

Background: Disease prediction using human genome sequencing data offers great opportunity in precision medicine. Accruing evidence showed that complex diseases have heterogeneous genetic characteristics influenced by the interactions of gene, environment, and population structure, which caused the genetic biomarkers estimated in different population clusters to exhibit prevailing variations in effect size.

Method: In this paper, to better utilize the genetic architecture informed by subpopulations, we propose to capture individual's relationship to subpopulation clusters using the angular distance by cosine similarity (CS) in the principal component space. The directional mapping resulted more accurate decomposition of individual's probability of identity to subpopulations, in which stratum-specific disease risk can be estimated and integrated (the prism vote, PV). The proposed PV-CS method is applied with alternative effect size estimation approaches including the summary statistics, SBayseR, Lassosum, LDpred2-Inf, and Bayesian regression with cosine shrinkage (BRCS); and the stratum-wised prediction is derived using the polygenic risk score (PRS).

Results and conclusion: Application of the PV-CS on real genome-wide sequencing data of the Alzheimer's disease in the UK Biobank data and GenADA data showed that the method improved prediction accuracy in various scenarios, with the best 5-fold cross-validation AUC of 72.12% and 81.03%, respectively. Without adding external information, the proposed method provides a practical and robust way to enhance complex trait prediction by leveraging genetic heterogeneity and architecture of the data. We have also developed an R-package named *pv* that enables convenient application.

192

Prediction Method of Association Between Microorganism and Disease in Bidirectional Heterogeneous Selection Network

Jian Guan[1], ZhaoGong Zhang1

[1]School of Computer Science and Technology, University of Heilongjiang, Harbin, China

Microorganisms in the human body have a great impact on health. Therefore, mastering the potential relationship between microorganisms and diseases is helpful to understand the pathogenesis of diseases, which is of great significance for disease prevention, diagnosis and treatment. In order to predict the potential microbial disease relationship, a novel computational model is proposed. Firstly, a bi-directional heterogeneous network of microorganism disease is constructed by integrating a variety of similarities, including Gaussian kernel similarity, microbial function similarity, disease semantic similarity and disease symptom similarity. Secondly, learning network neighbor information through asymmetric random walk; Finally, the selection algorithm is used for information aggregation, and the microbial disease node pair is analyzed. In the left one cross validation and five times cross validation, the new calculation method is better than the existing methods. Moreover, in case studies of different diseases, this method is effective.

193

SDPRX: A Statistical Method for Cross-population Prediction of Complex Traits

Geyu Zhou[1], Tianqi Chen[2], Hongyu Zhao[1,2]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America; [2]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America

Polygenic risk score (PRS) has demonstrated its great utility in biomedical research through identifying high risk individuals for different diseases from their genotypes. However, the broader application of PRS to the general population is hindered by the limited transferability of PRS developed in Europeans to non-European populations. To improve PRS prediction accuracy in non-European populations, we have developed a statistical method called SDPRX that can effectively integrate genome wide association study summary statistics from different populations. SDPRX characterizes the joint distribution of the effect sizes of a variant in two populations to be both null, population specific or shared with correlation. It automatically adjusts for linkage disequilibrium differences between populations. Through simulations and applications to seven traits, we compared the prediction performance of SDPRX with three other methods

PRS-CSx, LDpred2 and XPASS. LDpred2 is a single population method that takes non-EUR GWAS summary statistics as input, while SDPRX, PRS-CSx and XPASS are multi-discovery methods that jointly integrate GWAS summary statistics from multiple populations. We show that SDPRX outperforms other cross population prediction methods in the prediction accuracy in non-European populations, with an average 22% better than PRS-CSx, 33% better than LDpred2, 39% better than XPASS for quantitative traits and binary traits considered.

194

LUCID: An Integrative Clustering Model for Multi Omics Data

Yinqi Zhao[1*], David V. Conti[1]
[1]*Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, California, United States of America*

Technological advances in high-throughput biochemical data generation have made it possible to obtain multiple omics data on the same individuals for genetic epidemiology studies. Analyzing each omics data individually might not reveal the biological complexity of disease phenotypes. Integrative analysis of multi omics data has emerged as an approach to facilitate full understanding of the biological interplay between genetic or environmental risk factors and complex traits.

As an alternative to high-dimensional pairwise mediation or clustering approaches, Latent Unknown Clustering Integrating multi omics Data (LUCID), is an integrative statistical model that jointly estimates latent clusters characterizing each omic profile and genetic or environmental factors, while simultaneously associating these clusters to the phenotype. LUCID incorporates biological process into its conceptual model and is tailored for prospective cohort studies. Simulation studies indicate the proposed approach can obtain consistent estimates reflective of the true underlying simulated values and accurate cluster assignments. We demonstrate the application of the integrated model in Human Early Life Exposome (HELIX) Study. We apply LUCID to better characterize the association between maternal exposure to organochlorines and childhood body mass index (BMI) by integrating serum metabolomics, urine metabolomics and proteomics measurements in 1192 children in Europe. Our aim is to estimate metabolomics profiles for latent clusters with high BMI and to determine how those clusters may be associated with maternal exposure to organochlorines.

195

Approaches to Estimate Bidirectional Causal Effects Using Mendelian Randomization

Jinhao Zou[1*], Rajesh Talluri[2], Sanjay Shete[1, 3]
[1]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America;* [2]*Department of Data Science, The University of Mississippi Medical Center, Jackson, Mississippi, United States of America;* [3]*Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*

Mendelian Randomization (MR) is an epidemiological framework of using genetic variants as instrumental variables (IVs) to examine the causal effect of exposure on medical outcomes in observational data. Unidirectional MR are widely used in current observational studies for causal estimation. In bidirectional MR (BMR) model, bidirectional causal effects between exposure and outcome leads to a feedback loop between exposure and outcome, which biases the estimation of causal effects in real data applications. We considered BMR in light of the underlying feedback loops and demonstrated the properties of these feedback loops under various scenarios. We propose two novel MR methods for BMR model: BiRatio and BiLIML extended from Ratio and limited information maximum likelihood (LIML) methods, respectively. We evaluated the new BMR methods by comparing them with the Ratio and LIML under three different casual relationships: unidirectional causation, bidirectional causation with finite feedback cycles, and bidirectional causation with infinite feedback cycles. Our simulations show that BiRatio and BiLIML methods provide more accurate estimation of causal effects when the underlying causal mechanism is either unidirectional or bidirectional. We applied these bidirectional causal models to understand relationship between obesity and diabetes using the Multi-Ethnic Study of Atherosclerosis cohort. Our results revealed the bidirectional causal relationship between body mass index (BMI) and fasting glucose (FG). One kg/m^2 increase in BMI increased the FG by 0.70 mg/dL (P=8.43*10^{-5}) and also 1 mg/dL increase in the FG increased the BMI by 0.10 kg/m^2 (P=6.80*10^{-4}).

196

Power and Sample Size Computation for Association Analysis of a Binary Trait: Accounting for Covariate Effects

Ziang Zhang[1], Lei Sun[1,2*]
[1]*Department of Statistical Science, University of Toronto, Toronto, Ontario, Canada;* [2]*Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada*

Accurate power and sample size estimation are crucial to the design and analysis of genome-wide association studies (GWAS). It is well known that replication studies with underestimated sample sizes can result in false negatives, missing truly associated SNPs.

In GWAS of a binary trait via logistic regression, important covariates such as age and sex are typically included in the model. However, their effects are rarely properly considered, for example, in power or sample size computation. Unlike with continuous traits, powers of association testing of genetic variants with binary traits also depend on covariate effects, even under the assumption of gene-environment independence. Earlier work recognizes this phenomenon but implemented methods are not flexible.

We thus propose and implement a generalized method of estimating power and sample size for genetic association studies of binary traits that a) accommodates different types

of non-genetic covariate E , b) deals with different types of GE relationships, and c) is computationally efficient.

Extensive simulation studies show that the proposed method is accurate and computationally efficient for both prospective and retrospective sampling designs with various covariate structures. A proof of principle application to the UK Biobank data, focusing on the understudied African sample, shows that ignoring covariate age effect leads to overestimated power (hence underestimated replication sample size) when analyzing the binary hypertension trait, while the computation for the continuous blood pressure trait is invariant.

The R package *SPCompute* that implements the proposed method can be found at https://cran.r-project.org/web/packages/SPCompute.