Abstracts

INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

The 2021 Annual Meeting of the International Genetic Epidemiology Society

1

Obesity Partially Mediates Sex Differences in Cardiovascular Profiles Associated with Polycystic Ovary Syndrome Genetic Risk

Ky'Era V. Actkins[1,2]*, Lea K. Davis[2]
[1]*Department of Microbiology, Immunology, and Physiology, Meharry Medical College, Nashville, Tennessee, 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 heritable endocrine disorder in premenopausal females characterized by ovarian dysfunction, hyperandrogenism, and numerous metabolic comorbidities like type 2 diabetes (T2D). Despite the high prevalence of metabolic dysfunction, the genetic etiology between these conditions still remains poorly understood. Using polygenic risk scores (PRS), we showed that PCOS genetic risk drives sex differentiated cardiovascular signatures and is influenced by routinely collected electronic health record (EHR) body mass index (BMI) measurements. Therefore, we aimed to understand the mediating causal effect of BMI between PCOS and cardiometabolic diseases and to determine whether these effects were modified by sex. To do this, we performed a mediation analysis on 72,824 European descent individuals with $PCOS_{PRS}$ as the exposure variable and dichotomized clinical diagnosis as the outcome. When we examined the mediating role of BMI extracted from EHRs, which captures both genetic and environmental variance, we found that BMI was a strong mediator for cardiometabolic outcomes in both sexes ($T2D_{Females}$=29%, $T2D_{Males}$=23%, $Hypertension_{Males}$=17%, *P*-value=<2e-16; $Hypertension_{Females}$=41%, *P*-value=0.002). However, once we partitioned out the genetically regulated BMI variance, our findings revealed that the residual environmental BMI was not mediating the pathway from $PCOS_{PRS}$ to T2D (*P*-value=0.78) or hypertension (*P*-value=0.82) in males. Overall, our results implicate genetically regulated BMI as an important risk factor in the early development of cardiovascular diseases in males with high genetic predisposition for PCOS. Therefore, implementation of intervention programs and monitoring procedures are warranted for both sexes with family history of PCOS.

2

An Efficient Score Test Procedure for Association Analysis of Genomic Sequences

Abdulrahman Alshammari*, M. Fazil Baksh
Department of Mathematics and Statistics, University of Reading, Reading, United Kingdom

Standard association methods do not take into account possible calling errors for somatic mutations and are therefore limited in their suitability for investigating functional consequence of such mutations. A recent somatic mutation association test with measurement errors (SAME) that addresses this issue via the likelihood ratio test has shown that taking account of uncertainty in somatic mutation calling improves power for detecting an association. In the spirit of SAME, this talk develops and evaluates a score procedure that models actual somatic mutation as an unobservable variable and uses read-depth to increase the mutation calls. The score test is computationally efficient as only optimisation under the null model is required for each genetic variant. Additionally, risk of non-convergence of optimisation routines is reduced. These computational advantages are particularly beneficial in genomewide settings.

We implement our proposed score test in the R software and appraise it by comparing its performance, in terms of type I error and power, with SAME and a generalised linear model (GLM) that does not consider somatic mutation calling error. The results of simulation studies for a wide range of scenarios reveal that while all tests control type I error, the proposed score test procedure is more efficient than SAME and the GLM. For

example, for gene-level mutation analysis and gene-based somatic mutation frequency of 0.1, the score test has power of 0.815, for effect size 1.6 and sample size of 400. In contrast, SAME and the GLM have 0.604 and 0.596 power, respectively.

3

HaploGC - Constructing Haplotypes of Exceptional Quality in Families

Jason A. Anema[1]*, Laura Escobar [2], E. W. Daw[1], Michael A. Province[1]

[1]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri, United States of America; [2]Department of Mathematics and Statistics, Washington University in St. Louis, St. Louis, Missouri, United States of America

Ideally, we would have accurate full molecular sequence haplotypes. However, currently whole genome sequence (WGS) short-read technology allows individual genotype calls, which may then be constructed mathematically into haplotypes. Haplotype construction in family data with existing methods has been hindered by exponentiality in time complexity, and is often sensitive to genotyping errors producing excessive recombination. Access to accurate haplotypes will greatly enhance the ability of researchers to discover complex causal variants.

Using graph theory, we developed a novel efficient family-based haplotyping algorithm, Haplotyping with Graph Colorings (HaploGC), which is deterministic and produces haplotypes of exceptional quality very quickly (computation time is polynomial in family size, linear in number of markers). No limitation is present for variant type, number of alleles, inbreeding status, or pedigree size. It is robust to genotype errors and also detects and flags them. This method also provides a framework to localize recombination events with greater accuracy than previous algorithms. Additionally, no excess recombination results from genotyping error, bypassing a major drawback of the Lander-Green method.

Input to HaploGC includes the identity-by-descent structure on a genomic interval of interest. Using extensions of Elston-Stewart, Lander-Green, or other methods on a small set of high-quality markers, one can construct this identity-by-descent structure with high confidence.

We implemented HaploGC in our Long Life Family Study (LLFS), and constructed haplotypes in large multigenerational families with runtimes of a few seconds on ~100,000 markers. We expect this approach will be instrumental in resolving variants and haplotypes driving our linkage peaks in LLFS.

4

A Multivariate Genome-Wide Association Study of Psycho-Cardiometabolic Multimorbidity

Vilte Baltramonaityte[1]*, Priyanka Parmar[2], Charlotte A.M. Cecil[3,4,5], Janine Felix[4,6], Marjo-Riitta Järvelin[2,7], Jean-Baptiste Pingault[8], Yuri Milaneschi[9], Sylvain Sebert[2], Esther Walton[1]

[1]*Department of Psychology, University of Bath, Bath, United Kingdom; [2]Center for Life Course Health Research, University of Oulu, Oulu, Finland; [3]Department of Child and Adolescent Psychiatry/Psychology, Erasmus University Medical Center Rotterdam, Netherlands; [4]Department of Epidemiology, Erasmus Medical Center, Rotterdam, Netherlands; [5]Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands; [6]The Generation R Study Group, Erasmus Medical Centre, Rotterdam, Netherlands; [7]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [8]Department of Clinical, Educational, and Health Psychology, at University College London, London, United Kingdom; [9]Department of Psychiatry, Amsterdam Public Health and Amsterdam Neuroscience, Amsterdam UMC, Vrije Universiteit/GGZ inGeest, Amsterdam, Netherlands*

Coronary artery disease (CAD), type 2 diabetes (T2D) and depression are among the leading causes of morbidity and mortality worldwide. Epidemiological studies indicate a substantial degree of mutimorbidity, which may be explained by shared genetic influences. However, research exploring the presence of pleiotropic variants and genes that simultaneously influence CAD, T2D and depression is lacking. The present study aimed to identify genetic variants with effects on cross-trait liability to psycho-cardiometabolic (PCM) diseases. We used genomic structural equation modelling to perform the first multivariate genome-wide association study of PCM multimorbidity ($N_{effective}$=91,000), using as input summary statistics from univariate genome-wide association studies for CAD, T2D and depression. First, we assessed genetic correlations among each pair of traits and modelled their shared genetic architecture with a latent multimorbidity factor. Subsequently, we identified genetic variants associated with multimorbidity and performed functional gene mapping. We observed weak-to-moderate genetic correlations between depression and CAD (r_g = 0.20), and CAD and T2D (r_g = 0.39). The genetic overlap between depression and T2D was not significant (r_g = 0.13). The latent multimorbidity factor explained the largest proportion of variance in CAD (59%), followed by T2D (26%) and depression (6%). We identified 13 independent SNPs associated with multimorbidity

across 8 genomic risk loci. The strongest evidence was present for chromosome 9p21 locus. We observed enrichment in immune and inflammatory pathways, elucidating putative biological mechanisms underlying PCM multimorbidity. These findings reveal the genetic architecture of multimorbidity and advance our understanding of the shared genetic aetiology of CAD, T2D and depression.

5

Multi-Trait Analysis Of Multiple Related Cardiovascular Traits Identifies Novel Loci for Fibromuscular Dysplasia

Takiy Berrandou[1,2], Adrien Georges[1], Nabila Bouatia-Naji[1]
[1]Paris Cardiovascular Research Center U970 HEGP Research Center, INSERM, Paris, France; [2]Quantitative Genetics and Genomics (QGG), Aarhus University, Denmark.

Fibromuscular dysplasia (FMD) is a neglected arterial disease that shares several genetic determinants with more common cardiovascular diseases (CVDs). Complex imaging-based diagnosis prevent access to large cohorts to conduct large scale GWAS. With only four known susceptibility loci, we aimed to leverage genetic correlation between FMD and related cardiovascular and neurovascular diseases and traits to increase the power for loci discovery.

Using LD score regression, we demonstrated substantial positive genetic correlations between FMD and systolic blood pressure (r_g=0.44, P=5x10^{-10}), migraine (r_g=0.32, P=1x10^{-4}), intracranial aneurysm (r_g=0.34, P=7x10^{-6}), and cervical artery dissection, a rare cause of stroke (r_g=0.78, P=1×10^{-2}). Interestingly, FMD was negatively correlated with low-density lipoprotein (r_g=-0.19, P=8 ×10^{-3}), and coronary artery disease when adjusting for blood pressure genetics using mtCOJO (r_g= -0.31, P=5×10^{-5}). Multi-trait analysis of GWAS (MTAG), using summary statistics of these CVDs and traits, identified 99 genome-wide significant loci for FMD, including 95 new ones. We detected substantial local contribution to FMD heritability at 19 MTAG loci ($P_{Bonferroni}$=0.0005), which we considered as less subject to be MTAG artefacts due to important differences in sample size among traits analysed. Significant enrichment of FMD MTAG loci among open chromatin regions generated by ATAC-Seq in key arterial tissues (aorta, tibial and coronary arteries) supported further the genetic and biological relevance of the new FMD MTAG loci.

In summary, leveraging genetic correlation between carefully selected CVDs and traits is an effective strategy to considerably increase gene discovery pace for neglected arterial diseases like FMD, where only small samples can be collected.

6

Genotype-Based Microrna Expression and Gallbladder Cancer Risk

Alice Blandino[1*], Dominique Scherer[1], Justo Lorenzo Bermejo[1]
[1]*Statistical Genetics Research Group, Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany*

Circulating microRNAs are good candidates for cancer risk prediction, but the association between individual genotypes and expression-levels of circulating microRNAs is largely under-exploited compared to other molecular endophenotypes such as proteins and metabolites. We explored the potential and limitations of combining independent datasets with information on microRNA-expression alone, genotype alone, and both to identify circulating microRNAs associated with the risk of gallbladder cancer (GBC).

In a first dataset we identified circulating microRNAs differentially expressed in GBC cases. In a second independent dataset, we selected single nucleotide polymorphisms (SNPs) in close proximity and associated with circulating levels of the differentially expressed microRNAs (cis-miR-QTLs). In a third independent dataset with genotype information only, we predicted the circulating microRNA expression levels relying on the previously identified cis-miR-QTLs, and tested the association between genotype-based microRNA-expression and GBC risk. We applied robust logistic and linear regression models, and chose the best set of cis-miR-QTLs for microRNA expression prediction using a robust version of Akaike's information criterion.

We identified two microRNAs overexpressed in GBC cases in the first dataset, which were associated with eight cis-miR-QTLs in the second dataset. The risk of GBC increased with genotype-based microRNA expression in the third dataset, but the risk increase did not reach statistical significance (P-value > 0.05). The limitations of this strategy, in particular regarding the transferability of SNP-microRNA and microRNA-disease association statistics between independent datasets, and potential improvements will be discussed at the meeting.

7

Polymethylation Scores for Prenatal Maternal Smoke Exposure Persist Until Age 15 and Are Detected in Saliva

Freida A. Blostein[1*], Jonah Fisher[2], John Dou[1], Lisa Schenper[3], Daniel A. Notterman[3], Colter Mitchell[2], Kelly M. Bakulski[1]

[1]*Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America;* [2]*Institute for Social Research, University of Michigan, Ann Arbor, Michigan, United States of America;* [3]*Department of Molecular Biology, Princeton University, Princeton, New Jersey, United States of America*

Background: Prenatal exposure to smoking is associated with DNA methylation in children's cord blood, particularly described in European ancestry populations. The goal of this study is to assess the persistence of this DNA methylation signature into adolescence, its applicability to saliva DNA, and generalizability to other ancestry groups.

Methods: In the Fragile Families and Child Wellbeing study, a longitudinal birth cohort, at birth mothers self-reported prenatal smoking. Saliva DNA from children ages nine and fifteen were processed on the Illumina HumanMethylation 450K array. Polymethylation scores for prenatal smoke exposure were calculated using regression weights from a published meta-analysis of prenatal maternal smoking and child blood methylation. Cross-sectionally at each time point, we tested the association of prenatal smoke exposure with these polymethylation scores using multivariable linear regression, adjusting for sociodemographic, behavioral, and technical covariates.

Results: Among 779 children, 158 were exposed to prenatal smoke, 63% were of African ancestry, and 21% were of Hispanic ancestry. Polymethylation scores for prenatal smoke exposure were correlated (Pearson r=0.89, *P*-value<0.001) between ages nine and fifteen. Prenatal maternal smoking was associated with 0.1 unit higher polymethylation scores at age nine (95% CI: 0.05, 0.14) and with 0.08 unit higher polymethylation scores at age fifteen (95% CI: 0.03, 0.13). Results were consistent across ancestry groups. Global DNA methylation and DNA methylation clocks did not associate with prenatal maternal smoking.

Conclusion: DNA methylation signatures for prenatal smoke exposure developed in blood samples can be applied to saliva samples and are consistent across ages and ancestry groups.

8

Comparison of Power to Detect Epistatic Interactions of Causal Variants Between Recurrent Weighted Replanting and Other Machine Learning Approaches

Joan E. Bailey-Wilson[1*], James D. Malley[1], and Anthony M. Musolf[1]

[1]*Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America*

Much effort is being expended to detect the causes of "missing heritability" of complex traits, including effects of rare, moderate to high penetrance risk variants and epistatic interactions. Many machine learning (ML) methods can produce strongly predictive models when the number of predictive variables is very large compared to sample size and in the presence of complex interactions between the predictors. However, identifying which of the predictor variables are important in the prediction and therefore biologically important in genetic studies, is not possible for many ML methods. We have developed r2VIM, based on Random Forests (RF), and have now extended it to Recurrent Weighted Replanting (RWR) to improve power to detect causal variants.

RWR is a multi-step procedure that runs r2VIM iteratively, using r2VIM importance scores from prior steps to select subsets of features (SNPs here) to include in subsequent runs and novel weights to adjust the probability that each feature is selected as available for splitting (mtry parameter) in any given tree of each RF. At the final stage, r2VIM importance scores are calculated and features that pass a threshold based on importance score variance are selected as important. RWR is powerful and can detect variants involved in epistatic interactions with no marginal effects on the trait (power over 80% across a wide variety of simulated models with 5000 cases, 5000 controls, 100,000 SNPs). We compare these results with observed power from other machine learning methods including RF, r2VIM, Boruta and Vita; RWR has equal or better power.

9

Investigating the Causal Role of Inflammation on Parkinson's Disease by a Bi-Directional Mendelian Randomization Approach

Daniele Bottigliengo[1*], Luisa Foco[1], Philip Seibler[2], Christine Kein[2,3], Inke R. König[4], Fabiola Del Greco M.[1]

[1]*Institute for Biomedicine, Eurac Research Bozen/Bolzano, Italy;* [2]*Institute of Neurogenetics, University of Lübeck, Germany;* [3]*Department of Psychiatry and Psychotherapy, University of Lübeck, Germany;* [4]*Institut für Medizinische Biometrie und Statistik, University of Lübeck, Lübeck, Germany*

In the last decades, several observational studies suggested that inflammatory processes may influence the pathogenesis of Parkinson's disease (PD). However, it remains unclear whether inflammation is a factor that affects the onset of PD or is triggered by the neurodegenerative nature of the disease. In this study, we aim at evaluating the causal relationship between inflammation and PD by Mendelian randomization (MR) design.

Genetic instruments were identified using summary-level data from genome-wide association studies (GWAS) on inflammatory biomarkers of European ancestry participants (from 27,185 to 204,402). Genetic association data on PD and age at PD onset were obtained from the International Parkinson's Disease Genomics Consortium (IPDGC) (1,456,306 participants). Causal associations were evaluated using a bi-directional two-sample MR approach on sets of strong genetic instruments (p-value<$5x10^{-8}$; F-statistic>10) selected with a conservative r^2 threshold (<0.001) to handle linkage disequilibrium. Sensitivity analyses with robust MR methods were performed accounting for pleiotropic effects. We repeated the analysis on several GWAS data of inflammatory biomarkers to check the findings' consistency.

Considerable statistical evidence in support of higher risk and reduced age at onset of PD was observed and associated with higher interleukin-6 (IL-6) blood levels. There was limited statistical evidence of an association with PD and C-reactive protein (CRP), interleukin 1 receptor-antagonist (IL-1ra), and tumor necrosis factor (TNF). No evidence of reverse causation was observed. Results were consistent across the inflammatory biomarkers GWAS datasets.

These findings could provide new insights in the context of anti-inflammatory therapeutic strategies for disease prevention.

10

Simulated Data Provides Insight on Optimal Control Method for Confounders

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

Case/control studies are an incredibly prevalent method. These studies can offer insight on how an exposure affects an outcome. However, population stratification, environmental variables, and other confounders often obscure results. However, different control methods can correct for these problems. One popular approach is regression, where each confounder is accounted for by weighting its impact on the exposure and outcome. Another popular approach is matching, where controls and cases are matched on a potential confounding variable. There are a lot of subtypes within these two approaches, thus there are many different control methods. However, there is little guidance for choosing an appropriate control method for a given data set. Here we hypothesize that the choice of a control method has a significant impact on results, and that knowing which one to choose will lead to more accurate, reproducible science. To test this, directed acyclic graphs (DAGs) modeled complex relationships between the exposure, the outcome, and confounders. Next, an effect strength was inputted between each variable. From there, Bayesian Networks (BN) generated a data set of a desired sample size, and control methods were ranked based on how close the output was to the inputted effect strength. We compared regression-based control methods to matched control methods to determine which approach yielded the most accurate estimate of the true effect size.

11

Evaluation of A Region-Based Approach for Localization of Causal Variants (73/150 Char)

Myriam Brossard[1*], Kexin Luo[1], Delnaz Roshandel[2], Fatemeh Yavartanoo[4], Andrew D. Paterson[2,3], Yun J. Yoo[4], Shelley B. Bull[1,3]
[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, 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

To bridge the gap between region discovery by genome-wide regional association testing and fine-mapping within identified regions, we investigate an analytic approach that partitions the genome for region definition, followed by regional association testing. Genome partitioning is based on a haplotype block detection method (BigLD, Kim 2019) which identifies non-overlapping quasi-independent linkage disequilibrium (LD) blocks. Within each such region, a multi-SNP regression is used to construct a constrained test statistic based on multiple linear SNP combinations (MLC, Yoo 2017) of SNPs. MLC (multiple linear combination) thus provides a region-level statistic as well as multiple cluster-level statistics corresponding to within-region correlated SNP clusters. In this study, we investigate the ability of MLC statistics to localize signals at different resolutions in the presence of haplotype heterogeneity and long-range LD. Based on the example of *MC1R* harbouring five causal variants with independent effects on melanoma, we simulate haplotypes for 107292 SNPs on the 16q arm in 40000

cases and controls using 1000G European-ancestry haplotypes. The partitioning of 16q identifies 2394 regions, out of which MLC detects 35 regions with $P_{Bonferroni} \leq 2.1x10^{-5}$. The most significant region includes all the *MC1R* variants and exhibits an improved regional *P*-value compared to single-SNP *P*-values; other associated regions show signal attenuation with distance to *MC1R*. Localization within the top region successfully identifies the most significant SNP cluster including the causal variant with the strongest effect size, however localization of two lower effect-size *MC1R* variant clusters is challenged by other cluster associations with varying degrees of LD with the causal variant(s).

12

Metabolic Features of Colorectal Cancer Liability: Life Course Study Integrating Genetic Risk with Repeated Metabolomics

Caroline J. Bull[1,2,3]*, Joshua A. Bell[1,2], Neil Murphy[4], Jeroen Huyghe[5], Marc J. Gunter[4], Nicholas J. Timpson[1,2], Emma E. Vincent[1,2,3]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, United Kingdom; [3]School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom; [4]Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France; [5]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

Recognizing the early signs of cancer development is vital for informing early detection, primary prevention, and improving survival. Using data from the Avon Longitudinal Study of Parents and Children (ALSPAC) and UK Biobank (UKBB), we examined the relationship between increased genetic susceptibility to adult colorectal cancer and metabolic traits in childhood and adulthood, measured by NMR spectroscopy. We constructed a genetic risk score comprised of 72 single nucleotide polymorphisms strongly and independently associated (*P*<5.0E-08 and LD r2<0.001) with colorectal cancer case status in a large genome-wide association 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). This score was generated within 4,760 ALSPAC participants and analysed using two-sample Mendelian randomization (MR) in UKBB. Linear regression models were applied to examine the relationship between the colorectal cancer genetic risk score and metabolites measured in childhood (age 8y), adolescence (age 15y), and young adulthood (age 18y and 25y). We observed associations between the colorectal cancer genetic risk score and up to 35% of the circulating metabolic traits (Benjamini-Hochberg adjusted *P*-value ≤ 0.05) at a single time point, in particular fatty acids, VLDL, LDL, and IDL subclass lipids at age 18y. Two-sample MR estimates among adults in UKBB indicated broadly persistent patterns of disease liability across metabolic traits. This analysis reveals subtle changes in metabolism over time which precede the onset of clinically detectable cancer by several decades.

13

Gene-based Association Tests Using GWAS Summary Statistics and Incorporating eQTL

Xuewei Cao[1,*], Xuexia Wang[2], Shuanglin Zhang[1], Qiuying Sha[1]

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America; [2]Department of Mathematics, University of North Texas, Denton, Texas, United States of America

Although genome-wide association studies (GWAS) have been successfully applied to a variety of complex diseases and identified many genetic variants underlying complex diseases, there is still a considerable heritability of complex diseases that could not be explained by the identified genetic variants. One alternative approach to overcome the missing heritability is gene-based analyses, which consider aggregate effects of multiple genetic variants in a gene. Another alternative approach is transcriptome-wide association studies (TWAS). TWAS aggregate genetic information into functionally relevant testing units that map to genes and their expression in a trait-relevant tissue. TWAS is not only powerful, but can also increase the interpretability in biological mechanisms of identified trait associated genes. In this study, we propose two powerful and computationally efficient gene-based association tests, Overall and Copula, based on GWAS Summary Statistics and Incorporating eQTL. These two tests aggregate information from three traditional types of gene-based association tests and also incorporate expression quantitative trait locus (eQTL) data from multiple trait-relevant tissues into GWAS using GWAS summary statistics. Overall utilizes the extended Simes procedure and Copula utilizes the Gaussian copula approximation-based method. Simulation studies show that these two tests can control type I error rate very well and have higher power than the tests we compared with. We also apply these two methods to two schizophrenia GWAS summary datasets and two lipids GWAS summary datasets. The results show that these two methods can identify more significant genes than other methods we compared with.

14

Deep Learning-based Feature Extraction in Neuroimaging Genetics for Alzheimer's Disease

Dipnil Chakraborty*, Zhong Zhuang, Haoran Xue, for the Alzheimer's Disease Neuroimaging Initiative*, Wei Pan
Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

The prognosis and treatment of the patients suffering from Alzheimer's disease (AD) have been one of the most important and challenging problems over the last few decades. To better understand the mechanism of AD, it is of great interest to identify genetic variants associated with brain atrophy. Commonly in these analyses, neuroimaging features are extracted based on one of many possible brain atlases with FreeSurf and other popular software, which however may lose important information due to our incomplete knowledge about brain function embedded in these suboptimal atlases. To address the issue, we propose convolutional neural network (CNN) models applied to three-dimensional whole-brain structure MRI data to perform automatic feature extraction. These image-derived features are then used as endophenotypes in Genome-Wide Association Studies (GWAS) to identify associated genetic variants. When applied to the ADNI data, we identified several associated SNPs which have been previously shown to be related to several disorders such as depression, schizophrenia and dementia.

15

Functional Response Regression Model on Correlated Longitudinal Microbiome Sequencing Data

Bo Chen[1], Wei Xu[1,2]
[1] Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada; [2] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Functional regression has been widely used on longitudinal data, but it is not clear how to apply functional regression to microbiome sequencing data. We propose a novel functional response regression model analyzing correlated longitudinal microbiome sequencing data, which extends the classic functional response regression model only working for independent functional responses. We derive the theory of generalized least squares estimators for predictors' effects when functional responses are correlated and develop a data transformation technique to solve the computational challenge for analyzing correlated functional response data using existing functional regression method. We show by extensive simulations that our proposed method provides unbiased estimations for predictors' effect, and our model has accurate type I error and power performance for correlated functional response data, compared with classic functional response regression model. Finally, we implement our method to a real infant gut microbiome study to evaluate the relationship of clinical factors to predominant taxa along time.

16

A Genealogical Estimate of Genetic Relationships to Improve Detection of Population Structure Over Time

Caoqi Fan[1,2], Nicholas Mancuso[1,2,3], Charleston W. K. Chiang[1,2]*
[1] Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [2] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, United States of America; [3] Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America

The application of genetic relationships among individuals, characterized by a genetic relationship matrix (GRM), has far-reaching effects in genetic epidemiology. However, the current standard to calculate the GRM does not take advantage of linkage information and does not reflect the underlying genealogical history of the study sample. Here, we propose a coalescent-informed framework to infer the expected relatedness between pairs of individuals given an ancestral recombination graph (ARG) of the sample. This expected GRM (eGRM) is an unbiased and highly correlated estimate of the latent pairwise genome-wide relatedness and maintains the mathematical properties of canonical GRMs. Through extensive simulations we show that the eGRM is robust when using genealogies inferred from incomplete genetic data, and can reveal the time-varying nature of population structure in a sample. When applied to genotyping data from a population sample from Northern and Eastern Finland (N=2,644), we found that clustering analysis using the eGRM more accurately delineates population structure

than would be possible using the standard GRM. Taken together, our proposed estimator drastically shifts the notion of genetic relatedness from a variant-centric to a tree-centric world view, and will be widely applicable to genetic studies in understudied human or ecological samples where whole genome sequencing data or references might not be readily available.

17

Genetic Analyses of Common Infections in the Avon Longitudinal Study of Parents and Children Cohort

Amanda HW. Chong[1*], Ruth E. Mitchell[1], George Davey Smith[1], Rebecca C. Richmond[1†], Lavinia Paternoster[1†]
[1] MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom
[†]These authors contributed equally to this work.

The individual and public health burden of infections can be profound. Observational studies have shown a relationship between infections and the pathogenesis of noncommunicable diseases such as cancer, autoimmune disease, and cardiovascular disease; however, a greater knowledge of the role of host genetics is essential. We investigated antibodies against 14 infections measured in plasma from children in the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort at age seven: Alpha-casein protein, beta-casein protein, cytomegalovirus, Epstein-Barr virus, feline herpes virus, *Helicobacter pylori*, herpes simplex virus 1, influenza virus subtype H1N1, influenza virus subtype H3N2, measles virus, *Saccharomyces cerevisiae*, Theiler's virus, *Toxoplasma gondii*, and SAG1 protein domain, a surface antigen of *Toxoplasma gondii* measured for greater precision. We performed genome-wide association analyses of the 14 antibodies (N = 357 – 5010) and identified three genome-wide signals ($P < 5\times10^{-8}$), with two associated with antibodies against Measles virus and one associated with *Toxoplasma gondii* antibodies. Furthermore, we performed association analysis focused on the human leukocyte antigen (HLA) region where we identified 15 HLA alleles at a two-digit resolution and 23 HLA alleles at a four-digit resolution associated with five antibodies. Eight HLA alleles were associated with Epstein-Barr virus antibodies and showed strong evidence of replication in the independent cohort, UK Biobank. Our study has highlighted the potential identification of host genetic risk factors for several common infections and contributed to the endeavour to uncover the genetic and biological mechanisms of infection susceptibility.

18

Investigating a Causal Role for Neutrophil Count on P. Falciparum Severe Malaria: A Mendelian Randomization Study

Andrei Constantinescu[1*], Ruth Mitchell[1], David Hughes[1], Siddhartha Kar[1], Nicholas Timpson[1], Caroline Bull[1,2], Borko Amulic[2], Emma Vincent[1,2]
[1]Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [2]Cellular and Molecular Medicine, Faculty of Life Sciences, University of Bristol, Bristol, United Kingdom

P. falciparum malaria leads to great loss of life in Africa, and individuals living in this region often have reduced neutrophil counts in circulation due to a heritable phenomenon called 'benign ethnic neutropenia' (BEN). Neutrophils defend against bacterial infections but have been shown to be detrimental in malaria mouse models, suggesting that neutropenia may be protective against severe *P. falciparum* malaria. We tested this hypothesis by performing a genome-wide association study (GWAS) of circulating neutrophil count, and a Mendelian randomization (MR) analysis of neutrophil counts on severe malaria in individuals of African ancestry. We used ADMIXTURE on UK Biobank (UKBB) participants to identify those with >=80% African ancestry (N=6,653). A principal component analysis (PCA) using EIGENSOFT was done to find related individuals (N=544) and remove outliers (N=197). A K-means cluster analysis was done, and we estimated F_{ST} values between these clusters (N=9). A GWAS (N=6,086) of neutrophil count was done using BOLT/LMM, with SNPTEST used to run a GWAS on each K-cluster, the results of which were then meta-analysed with META. Finally, we performed a bi-directional two-sample MR analysis using summary statistics for neutrophil count (UKBB, N=6,086) and severe malaria (MalariaGEN, N=17,056). Our main GWAS identified 81 loci associated with neutrophil count. The MR results provided some evidence for an effect of severe malaria on neutrophil count. However, a small sample-size reduced the power to identify those variants with low allele frequencies and/or effects sizes in our GWAS. This only highlights the importance of conducting large-scale biobank studies in Africa.

19

Application of Polygenic Risk Scores to Admixed Hispanic Samples

Brandon J. Coombes[1*], Euijung Ryu[1], Anthony Batzler[1], Gregory Jenkins[1], Susan L. McElroy[2], Alfredo Cuellar-Barbosa[3], Miguel Prieto[4,5], Mark A. Frye[6], Joanna M. Biernacka[1,6]
[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America; [2]Lindner

Center of HOPE/University of Cincinnati, Cincinnati, Ohio, United States of America; [3]Department of Psychiatry, Universidad Autonoma de Nuevo Leon, Monterrey, Mexico; [4]Department of Psychiatry, Universidad de los Andes, Santiago, Chile; [5]Mental Health Service, Clinica Universidad de los Andes, Santiago, Chile; [6]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, United States of America

Almost 80% of participants included in genome-wide association studies (GWAS) to date are of European ancestry, thus, for most traits, polygenic risk scores (PRS) performs worse in non-European ancestries due to differences in minor allele frequency and linkage disequilibrium across the genome. Here, we use two Hispanic studies to compare methods to estimate the PRS when either a small GWAS of a Hispanic sample exists or doesn't. In the first study, we estimate the PRS for bipolar disorder (BD), for which no Hispanic GWAS exists, in a Hispanic sample from the Mayo Clinic Bipolar Disorder Biobank (N = 372). In the second study, we estimate the PRS for type II diabetes (T2D), for which a multi-ethnic GWAS exists, in the Sangre por Salud Biobank, a Hispanic biobank (N = 3820) which excluded participants with current diagnosis of T2D and use the PRS to predict hemoglobin A1c, a biomarker that is elevated among individuals with T2D. In both examples, we compare the pruning+thresholding PRS approach to a Bayesian shrinkage approach (LDpred2). Because a small GWAS of T2D in a Hispanic sample exists in the second study, we also compare these approaches to cross-ancestry PRS approaches which leverage the smaller Hispanic GWAS to improve prediction performance. While these cross-ancestry PRS methods have shown that they can improve PRS prediction, they are much more computationally intensive and harder to implement than traditional PRS. Here, we propose a computationally fast and simple alternative to combine the PRS across ancestries using a principal component approach.

20

Prevalence of Individuals with *DHCR7* Variants Consistent with Smith-Lemli-Opitz Syndrome in Subjects with Autism

Jennifer B. Cordero[1]*, Roberto Y. Cordero[1], Elaine Tierney[2], Forbes D. Porter[3], Christopher Wassif[3], Claire L. Simpson[1,4]

[1]Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America; [2]Kennedy Krieger Institute, Baltimore, Maryland, United States of America; [3] Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, United States of America; [4]Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America

Introduction: *DHCR7* encodes the enzyme that catalyzes the conversion of 7-dehydrocholesterol (7-DHC) to cholesterol. Defective cholesterol synthesis results in low cholesterol and increased concentration of 7-DHC. Mutations in *DHCR7* cause Smith–Lemli–Opitz syndrome (SLOS). Most SLOS patients are compound heterozygotes and present with congenital abnormalities, intellectual disability, and autism. Over 140 different mutations in *DHCR7* have been reported. In this study, we identified the prevalence of individuals with known *DHCR7* variants in MSSNG, an autism genomic database.

Methods: We examined whole genome sequencing (WGS) *DHCR7* variant data from MSSNG subjects with autism and identified individuals with ≥ 2 SLOS *DHCR7* known pathogenic variants. We analyzed lipid and behavioral phenotypic data in subsets of subjects in relation to the WGS analysis results. The frequency of identified individuals with the SLOS *DHCR7* pathogenic variants was then compared with control unaffected family members.

Results: A total of 7187 MSSNG subjects (3425 affected; 3762 unaffected) from 2756 families were included in the study, majority (80%) of which belonged to family trios or quads. Preliminary results from a subset of MSSNG database identified Autism-Genetic-Resource-Exchange (AGRE) subjects as having potentially deleterious *DHCR7* mutations. Analysis for the remaining MSSNG subjects is ongoing.

Conclusion: Early recognition and management of autism patients with SLOS-related *DHCR7* variants can help improve the course of the disorder. Further, the study of autism patients with concomitant *DHCR7* mutations affecting cholesterol synthesis may help elucidate pathways and relationships of autism and lipids.

21

Leveraging Transcriptome Imputation to Identify Risk Genes for Crohn's Disease in African Americans

Roberto Y. Cordero,[1,2*] Subra Kugathasan,[3] Dermot P.B. McGovern,[4] Steven R. Brant,[5] Claire L. Simpson[1,]

[1] Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee United States of America; [2] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore,

Maryland United States of America; [3] Department of Pediatrics and Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, United States of America; [4] F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars Sinai Medical Center, Los Angeles, California, United States of America ; [5]Rutgers Crohns and Colitis Center of New Jersey, Department of Medicine, Robert Wood Johnson Medical School and Department of Genetics, School of Arts and Sciences, Rutgers University, New Brunswick and Piscataway, New Jersey, United States of America

Multiple genetic risk loci associated with inflammatory bowel disease (IBD) have been identified through genome-wide association studies (GWAS). However, very large sample sizes are essential in GWAS to increase statistical power to detect disease-related loci. The power to detect IBD loci in African Americans (AA) and ultimately identify risk genes has been limited by modest sample sizes. To overcome this challenge, gene-based association methods such as PrediXcan were developed to detect the relationship between genes and traits to reduce the multiple testing burden. We utilized PrediXcan to integrate expression quantitative trait loci (eQTL) from the Genotype-Tissue Expression (GTEx) study and IBD GWAS summary statistics using independent case-control datasets, totaling 843 Crohn's disease (CD) cases and 1678 controls from unrelated, self-identified AA individuals. An elastic net model in whole blood, transverse colon, small intestine, adipose tissue were used as part of our transcriptome imputation. Our initial results reveal 48 significant associations with CD, with the strongest association coming from *SPDYE6* (p, 4.63x10-8). Other top associations include *OIT3, BIVM, AC245041.1, HROB, CSNK1G1, ZNF488, RASA4, FAHD2B, LOC102724488,* and *TMEM106A*. We found significant gene associations (*C15orf61, FBXW8, FAM219B, CYP1A1, PARG, ASAH2, SCAMP2, MPI, RPP25, WARS2, KNOP1,* and *PPCDC*) in regions in our AA CD published GWAS (Brant, Simpson, Okou et al, 2017). Four other loci (*DHX58, MAP3K8, IFI35,* and *RPS6KL1*) were within a megabase of established European loci. Here, we demonstrate how gene-based methods, informed by other omics data, can improve our ability to detect known and novel genes associated with CD in AA.

22

Phenotypic Manifestations of Genetic Liability to Neuroticism Across Childhood: A UK Prospective Birth Cohort Study

Ilaria Costantini*[1,2], Daphne-Zacharenia Kounali[1,2], Hannah Sallis[2,3], Kate Tilling[2,3], Daniel Smith[2,3], Rebecca Pearson[1,2,3]

[1]Centre for Academic Mental Health at the University of Bristol, Oakfield House, Bristol, United Kingdom; [2]Department of Population Health Sciences, University of Bristol Medical School, Oakfield House, Oakfield Grove, Bristol, United Kingdom; [3]Medical Research Council (MRC) Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom

Neuroticism is the tendency to experience negative emotions. Longitudinal studies suggest that neuroticism increases risk for a variety of psychological and physical problems (e.g. anxiety, depression, and comorbidity between disorders). It therefore represents an important economical and societal burden.

In this study, we investigated how a polygenic risk score (PRS) to neuroticism is expressed across childhood (using 2 to 7 measurements taken from 6 months to 11 years of age) on various psychological outcomes and trajectories of emotional and behavioural difficulties in 7,240 children in the Avon Longitudinal Study of Parent and Children. The PRS was constructed from a GWAS of neuroticism in 329,821 adults.

Using a PRS at a *P*-value threshold of <0.05, we found strong evidence that genetic liability to neuroticism was associated with higher scores on various temperamental (as early as six months of age) and psychological outcomes, including diagnostic classification of anxiety and depression. Genetic liability to neuroticism was also associated with an increase in the trajectories of emotional and behavioural difficulties across childhood. Our results remained robust to various sensitivity analyses employed to examine potential differential misclassification of outcomes reported by the mother due to maternal PRS to neuroticism, attrition, potential genetic confounding via maternal genotype and when using PRS at different *P*-value thresholds.

Our findings suggest that it is possible to detect manifestations of a genetic liability to adult mental health problems as early as infancy and childhood. This approach could be used for identification of early behavioural predictors of later psychopathology in high-risk groups.

23

Effects of Rare, Functional Variants on Risk of Common Phenotypes in 200,000 Exome-sequenced UK Biobank Participants

David Curtis[1, 2]

[1]UCL Genetics Institute, UCL, Darwin Building, Gower Street, London, United Kingdom; [2]Centre for Psychiatry, Queen Mary University of London, Charterhouse Square, London, United Kingdom

Introduction: Genome wide association studies using common variants have been applied to a wide

range of phenotypes but typically only provide an approximate localisation and biological interpretation can be difficult. Large exome-sequenced cohorts provide the opportunity to identify effects of rare coding variants in specific genes.

Materials and Methods: Weighted burden analysis was applied to 200,000 exome-sequenced participants, with variants being weighted more highly if they were predicted to have a more severe functional effect and if they were extremely rare. Phenotypes were derived from available data which might be directly measured, as for BMI, derived from questionnaires, as for heavy and problem drinking, or from a mixture of reported medication, self-reported diagnosis and health records, as for hypertension, type 2 diabetes (T2D) and hyperlipidaemia.

Results: A small number of novel genes were exome-wide significant, such as *DNMT3A* for hypertension or *GIGYF1* for T2D. Some previously implicated genes were recapitulated and the contributions of different categories of variant better characterised, such as the relative contributions to hyperlipidaemia risk from loss of function versus deleterious variants in *LDLR* and *PCSK9*. The effect of rs1229984 in *ADH1B* on heavy and problem drinking was confirmed, with multivariate analyses showing that no other variants in the *ADH* gene cluster had important effects. Variants having discernible effects on phenotypes had cumulatively low frequencies.

Conclusions: These analyses provide insights into biological processes impacting important common phenotypes. This research has been conducted using the UK Biobank Resource.

24

Genome-wide Polygenic Risk Score of Prostate Cancer in African and European Ancestry Men

Burcu F. Darst[1*], Ravi K. Madduri[2], Alexis A. Rodriguez[2], Xin Sheng[1], Rosalind A. Eeles[3], Zsofia Kote-Jarai[3], J. Michael Gaziano[4,5,6], Amy C. Justice[7,8], David V. Conti[1], Christopher A. Haiman[1], on behalf of the Million Veterans Program and the PRACTICAL Consortium

[1]Center for Genetic Epidemiology, Department of Preventive Medicine, 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]The Institute of Cancer Research, London, United Kingdom; [4]VA Boston Healthcare System, Boston, Massachusetts, United States of America; [5]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [6]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [7]VA Connecticut Healthcare System, West Haven, Connecticut, United States of America; [8]Yale School of Medicine, New Haven, Connecticut, United States of America

Genome-wide polygenic risk scores (PRS) are reported to have higher performance than standard genome-wide significant PRS across numerous traits. We evaluated the ability of genome-wide PRS to evaluate prostate cancer risk compared to our recently developed PRS of 269 established prostate cancer risk variants and multi-ancestry weights. Genome-wide PRS approaches included LDpred2, PRS-CSx, and EB-PRS. Models were trained using the largest and most diverse prostate cancer GWAS to date of 107,247 cases and 127,006 controls. Resulting models were tested in independent samples of 1,586 cases and 1,047 controls of African ancestry from the California Uganda Study and 8,045 cases and 191,835 controls of European ancestry from the UK Biobank. Among the genome-wide PRS approaches, LDpred2 had the best performance, with AUCs of 0.649 (95% CI=0.627-0.670) in African and 0.819 (95% CI=0.815-0.823) in European ancestry men. African and European ancestry men in the top PRS decile relative to men in the median 40-60% PRS category had odds of prostate cancer of 3.29 (95% CI=2.47-4.40) and 2.99 (95% CI=2.78-3.23), respectively. However, the PRS constructed using 269 variants had significantly larger AUCs in both African (0.679, 95% CI=0.659-0.700) and European ancestry men (0.845, 95% CI=0.841-0.849), with African and European ancestry men in the top PRS decile having larger odds of prostate cancer (3.53, 95% CI=2.66-4.69 and 4.20, 95% CI=3.89-4.53, respectively). Findings will be further validated in diverse men from Million Veteran's Program. This investigation suggests that genome-wide PRS may not improve the ability to distinguish prostate cancer compared to a genome-wide significant PRS.

25

Individuals of African Ancestry Share HLA Alleles Protective Against Tuberculosis and Sarcoidosis

Bryan A. Dawkins[1*], Lori Garman[1], Nicholas Cejda[1], Nathan Pezant[1], Astrid Rasmussen[1], Benjamin A. Rybicki[2], Albert M. Levin[2,3], Penelope Benchek[4], Thomas R. Hawn[5], Chetan Seshadri[5], Michael C. Iannuzzi[2], Catherine M. Stein[4,6], and Courtney G. Montgomery[1]

[1]Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, United States of America; [2]Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan; [3]Center for Bioinformatics, Henry Ford Health System, Detroit, Michigan United States of America; [4]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, United States of America; [5]Department of Medicine, University of Washington,

Seattle, WA, United States of America; [6]Division of Infectious Diseases and HIV Medicine, Department of Medicine, Case Western Reserve University, Cleveland, OH, United States of America

Tuberculosis (TB) and sarcoidosis are distinct granulomatous disorders with numerous genetic associations. Human leukocyte antigen (HLA) is important in susceptibility and progression in both diseases. Although TB is caused by *Mycobacterium tuberculosis*, HLA associations in case control TB studies are highly variable, potentially due to phenotype heterogeneity, limited allelic resolution, and narrow analytical methods that exclude more complex associations common to biological data. Using four digit HLA alleles and applying more inclusive feature selection methodology, we provide the first HLA association analysis in TB that compares latent and active TB to individuals who display no evidence of infection and maintain negative diagnostic tests over a long term period of exposure to individuals with active TB. To detect a more comprehensive array of statistical effects, we introduce a novel application of nearest neighbor feature selection that uses a consensus approach across three input neighborhood algorithms to define allelic importance for classifying outcomes. This nearest neighbor approach is generally applicable in the context of binary classification and regression, with either categorical or continuous predictors. We compare our findings to sarcoidosis, both persistent and resolving. We provide the first comparison between TB and sarcoidosis resistance phenotypes, showing that HLA-DRB1 alleles *01:02, *03:02, *12:01, and *13:02 are associated with both resolving sarcoidosis in African Americans and long term resistance to latent or active TB in Ugandans.

26

Assisted Reproductive Technologies Reduce Fetal Growth and Alter Maternal and Fetal DNA Methylation

William R.P. Denault[1,2,3], Kristine L. Haftorn[1,2], Christian M. Page[1,4], Maria Magnus[1,5,6], Siri Håberg[1], Jon Bohlin[1,7] and Astanand Jugessur[1,2,3]

[1]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway; [2]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway; [3]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway; [4]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway; [5]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [6]Population Health Sciences, Bristol Medical School, Bristol, United Kingdom; [7]Department of Method Development and Analytics, Norwegian Institute of Public Health, Oslo, Norway

Numerous observational studies have reported associations between assisted reproductive technologies (ART) and perinatal outcomes. However, the causal nature of these associations remains unclear.
We performed a Mendelian randomization study using a Norwegian cohort of 25,105 genotyped mother-father-newborn trios (562 were ART-conceived children) and investigated the causal effects of ART on birth weight, birth length, maternal and fetal DNA methylation (DNAm). Among the 25,105 trios, 1218 had DNAm data from the Illumina EPIC 850K array. We estimated the effect of ART using a new unbiased one-sample MR method we have developed called "Cross-fitting for Mendelian Randomization" (CFMR). We assessed the causality of ART using a powerful pleiotropy-free polygenic risk score as an instrument for ART (area under the curve=0.77 in the test set) using paternally non-transmitted genotypes. We performed an epigenome-wide MR of ART on maternal and fetal DNAm.

89 CpGs and 158 CpGs were causally associated with ART in the children and the mothers, respectively (FDR < 0.05). Among the 158 CpGs detected in the mothers, 60 were also detected in the children, with similar effect of ART on DNAm level. ART reduced the global methylation level in children but increased it in mothers. Further, CFMR showed that ART reduces birth weight by 166 g (95% CI: [-296;-37]) and birth length by -1.04 cm (95% CI:[-2.02;-0.07]). Finally, we show that even when controlling for known confounders, the estimations of the effect of ART differ largely from our CFMR estimates, suggesting that previous epigenome-wide association analyses of ART are strongly confounded.

27

Less is More: An Unbiased and Versatile Estimator of Genetic Variance Using Summary Statistics

Wei Q. Deng[1*], Guillaume Pare[2,3], Radu Craiu[1], Lei Sun[1,4]

[1]Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, Toronto, Ontario, Canada; [2]Population Health Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Ontario, Canada; [3]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada; [4]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Decomposing phenotypic variance to genetic and environmental components is fundamental to quantitative genetics. Embracing the statistical challenge of a large number of genetic markers as compared to the available sample size, many high-dimensional

variance estimators have been proposed to quantify the genetic contribution to continuous complex traits. Recent interests have shifted from an overall summary of variance to more specific hypothesis of whether groups of genetic variants, such as functional units, are enriched for an excess of genetic variance. These trends call for an estimator that is versatile to allow arbitrary partitions of the genome at the gene or regional level yet remains unbiased regardless of data dimensions as a result of such partitioning. From a univariate perspective, we proposed a multi-collinearity measure that enables a moment estimator of genetic variance by aggregating over individual genetic markers. This estimator is easy-to-implement for large-scale analyses, requiring only the univariate regression coefficients and the singular value decomposition (SVD) or truncated SVD of the genotype matrix as a starting point, which is particularly attractive when only summary data are available. Since genetic variance provides an upper bound of prediction variance, as an application, we examined whether additional genome-wide markers (as supposed to genome-wide significant ones) would improve prediction using consortia summary statistics for height, BMI, HDL and LDL. Results suggest the improvement depends on the underlying polygenicity: we observed a nearly 10-fold increase in genetic variance when relaxing the *P*-value threshold to 0.1 for LDL, but little gain for HDL.

28

Leveraging Extreme Phenotype at Multi-Omics to Identify Biomarkers for Early-stage Lung Cancer Survival

Mulong Du[1,2,*], David C. Christiani[1,3]
[1]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China; [3]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital / Harvard Medical School, Boston, Massachusetts, United States of America

Approximately 30% of lung cancer patients in stage I/II die within five years due to the progression and recurrence. Thus, there is a great need to characterize the molecular architecture for early-stage lung cancer patients at short survival.

A total of 39 lung cancer patients in stage I with short survival (median = 19.9 mos) and 38 with long survival (median = 159.93 mos) were recruited from Boston Lung Cancer Study cohort. Multi-omics profiles included genome (somatic mutation), proteome (in serum), and epigenome (DNA methylation in both blood and tumor). Additional in-house (tissue microarray) and publicly available omics datasets (TCGA and CPTAC) were applied for validation.

There was a distinct somatic mutation spectrum between two survival sets, including variant classification and mutated genes, especially *DDR2* mutated specifically in short survival tumors. In serum proteome, two proteins, GAPDH and SPHK1, having similar expression patterns in serum and tissue, served as early-stage specific predictors for lung cancer survival. In DNA methylation, the tumor and blood performed distinct epigenomic patterns, including more significant CpGs with larger fold change identified in short survival tumors, a certain extent of opposite signals for survival sets, and the disrupted correlation between DNA methylation age and diagnosis age by tumor. Importantly, the CpG site cg05697274 was confirmed with the mediation effect for smoking on early-stage lung cancer survival.

This is the first study to apply an extreme phenotype strategy at omics for early-stage lung cancer survival. It is worth extending for further biomarker studies.

29

An Analysis of Methods for Phenotype Prediction from Genetic Data

Megan Duff[1*], Stephanie A. Santorico[1,2,3]
[1]Department of Mathematical and Statistical Sciences,[2]Human Medical Genetics and Genomics Program, and [3]Department of Biostatistics & Informatics, University of Colorado, Denver, Colorado, United States of America

Predicting an individual's phenotypic value from their genetic data is a goal and current research area for the field of genetics. This would not only serve as a public heath tool but could provide researchers with an opportunity to increase power for their analyses by increasing sample size. The primary difficulty in creating such a model lies in the number of loci that contribute to a complex trait compared to the sample sizes used in training the model. Several penalized regressions, Bayesian regression, and non-linear prediction methods have been developed to account for such limitations, but there is no robust method that performs best in every scenario. Here, the most commonly used methods for phenotype prediction will be presented. A comparison between the methods among a variety of genetic architecture assumptions will be presented, as well as discussing limitations of such methods and possible future areas of research.

30

Sparse Canonical Correlation Analysis to Detect Trans-Regulated Genes and Proteins Related to Traits

Diptavo Dutta[1*], Yuan He[2], Ashis Saha[3], Marios Arvanitis[2,4], Alexis Battle[2,3] & Nilanjan Chatterjee[1,5]
[1]Department of Biostatistics; [2]Department of Biomedical Engineering, [3]Department of Computer Science, [4]Department of Cardiology, [5]Department of Oncology, Johns Hopkins University, Baltimore, Maryland, United States of America

Background: Large-scale genetic association studies have identified many trait-related variants but understanding their role in regulating intermediate molecular phenotypes like gene-expressions or protein-levels can uncover important underlying biological mechanisms. While detecting *cis*-regulation has received much attention, detecting distal (*trans*)association remains challenging due to weaker effect sizes and large multiple comparison burden.

Methods: We propose Aggregative *tRans* assoCiation to detect pHenotype speclfic gEne-sets (ARCHIE), which employs sparse canonical correlation analysis to identify gene-sets trans-regulated by a subset of trait-related variants by aggregating multiple *trans*-associations. Further, we adopt a resampling-based approach to test whether these gene-sets reflect significant trait-specific patterns of trans-associations.

Results: We applied ARCHIE to *trans*-eQTL summary statistics from eQTLGen consortium for variants associated to 29 traits and found majority (50.7%) of the detected genes had no significant *trans*-associations. For example, we found 75 genes to be trans-regulated by Schizophrenia-related variants, of which 59 had only weaker trans-associations, including genes like *CXCR4* and *CAV1*, previously implicated in neurological traits. Further, applying ARCHIE to plasma protein-level data in Atherosclerosis Risk in Communities (ARIC) study, we identify 129 and 98 proteins trans-regulated by variants related to urate-levels in European Americans and African Americans respectively. Through a series of downstream analyses including pathway enrichment, transcription-factor targets and expression-imputation, we show that the identified gene (protein)-sets can be biologically associated to the trait.

Conclusion: By aggregating multiple associations, ARCHIE is a powerful tool for identifying target gene-sets through which the effects of trait-related variants might be mediated and can highlight potential causal mechanisms.

31

Stochastic Functional Linear Models for Gene-based Association Analysis of Complex Traits in Longitudinal Studies

Ruzong Fan
Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington DC, United States of America

Longitudinally measured phenotypes are important for exploring genetic and environmental factors that affect complex traits over time. Genetic analysis of multiple measures in longitudinal studies provides a valuable opportunity to understand genetic architecture and biological variations of complex diseases. Many genetic studies have been conducted in cohorts in which repeated measures on the trait of interest are collected on each participant over a period of time and sequence data are available. Such studies not only provide a more accurate assessment of disease condition but enable us to investigate genes influencing on the trajectory of a trait and disease progression, which are likely to help reduce the remaining missing heritability of these traits. Although they are important, there is a paucity of statistical methods to analyze sequence data in longitudinal studies. In this paper, stochastic functional linear models are developed for temporal association analysis at gene levels to analyze sequence data and longitudinally measured quantitative traits. Functional data analysis techniques are utilized to reduce high dimensionality of sequence data and draw useful information. A variance-covariance structure is constructed to model the measurement variation and correlations of the traits based on the theory of stochastic processes. Spline models are used to estimate the time-dependent trajectory mean function. By intensive simulation studies, it is shown that the proposed stochastic models control type I errors well, and have higher power levels than those of the perturbation tests. We test and refine the models and related software using real data sets of Framingham Heart Study.

32

A New Powerful Unsupervised Random Forests Proximity Measure

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

Random Forests (RF) are fast and perform well in high dimensional classification problems. In precision medicine, another use of large scale data is to stratify

individuals into homogeneous subgroups. For this unsupervised setting, unsupervised random forests (URF) have been proposed to compute dissimilarities between individuals, which can then be used as input for dimensionality reduction methods or clustering algorithms. The main idea of URF is to synthetize an artificial data set by resampling the original observations which is combined with the original data set. The standard RF algorithm can then be used to classify observations as original or artificial. Dissimilarities between each pair of individuals are obtained by counting how often they end up in the same terminal nodes across the forest.

We introduce a novel approach of computing dissimilarities of observations, bases on the distances between terminal nodes they belong to. We compare the two URF approaches in terms of their ability to capture population structure with principal component analysis (PCA) as the reference method. We use genotype data from selected populations available in the 1000 genomes project. Our results show that URF based approaches deal better with outliers than PCA, and that our new approach performs better on overlapping populations.

33

Current Methods Integrating Variant Functional Annotation Scores Have Limited Capacity to Improve Power of GWAS

Jianhui Gao[1*], Osvaldo Espin-Garcia[1,2], Andrew D. Paterson[1,3], Lei Sun[1,4]

[1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Department of Biostatistics, Princess Margaret Cancer Center, University Health Network, Toronto, Canada; [3]Program in Genetics & Genomic Biology, The Hospital for Sick Children Research Institute, Toronto, Canada; [4]Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, Toronto, Canada

Functional annotations have the potential to increase the power of genome-wide association study (GWAS) by prioritizing variants according to their biological relevance. Recently, the method, functionally informed novel discovery of risk loci (FINDOR), has been proposed to leverage a set of 75 coding and conserved annotations into GWAS, which detected more loci than GWAS alone for 27 selected complex traits with high heritability. Here we consider the more convenient meta-functional scores such as CADD and Eigen, and we broadly examine the GWAS summary statistics of 1,132 phenotypes from the UK Biobank.

We study four data-integration methods, including meta-analysis, Fisher's method, the weighted p-value approach, and the stratified false discovery control (sFDR), focusing on methods' robustness to the possibility of the functional scores used are uninformative.

We conclude that, while meta-analysis and Fisher's methods are often used to integrate multiple GWASs together, they are unsuitable for integrating functional annotations with GWAS summary statistics. Averaged across the 1,132 phenotypes, sFDR was the most robust method while the weighted p-value method identified slightly more variants. While the earlier application of FINDOR to the 27 selected traits detected more loci than GWAS alone, the median [Q1,Q3] new discovery was 0 [0,2] by FINDOR, 0 [0,1] by weighted p-value and 0 [0,0] by sFDR across all the 1,132 traits examined. Overall, our study suggests more informative functional meta-scores or new data integration methods are needed to further improve the power of GWAS by leveraging functional annotations.

34

Identifying Common Genetic Susceptibility Underlying Comorbid Phenotypes Using Binomial Regression

Prasun Panja, Saurabh Ghosh*

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

The models suggested in Multiphen (O'Reilly et al., 2012) and BAMP (Majumdar et al, 2015) provide an alternative to study population-based genetic association with multivariate phenotypes by exploring the dependence of genotype on phenotype instead of the naturally arising dependence of phenotype on genotype. However, the underlying tests for association are based on the null hypothesis of no association with of the constituent traits versus the alternative hypothesis of association with at least one of the constituent traits of the multivariate phenotype vector. Thus, such tests do not provide evidence of pleiotropy or common genetic factors underlying all the traits constituting the multivariate phenotype. With respect to a pair of comorbid phenotypes (both binary, a combination of binary and quantitative or both quantitative), we aim to modify the proposed BAMP (Binomial regression-based Association of Multivariate Phenotypes) approach to test the null hypothesis of no association with at least one of the phenotypes versus the alternative hypothesis with both the phenotypes. Since the likelihood ratio test requires a constrained maximization (over the two coordinate axes) under the null hypothesis, it is analytically difficult to obtain the asymptotic distribution of the log-likelihood test statistic under the null

hypothesis and permutation strategies need to be employed. We carry out extensive simulations under different genetic models and correlation structures of the bivariate phenotype to evaluate the type 1 error rates and power of the proposed test. We show that an extension of the method to more than two comorbid phenotypes is theoretically straight forward.

35

Tissue-specific Regulation of mtDNA Encoded Genes

Xenofon Giannoulis[1,2*], Na Cai[1]

[1]*Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany;* [2]*Institute of Translational Genomics, Helmholtz Zentrum München, Neuherberg, Germany*

Mitochondrial DNA (mtDNA) is a 16KB circular molecule that is present in multiple copies per tissue, encoding 13 protein-coding genes that form subunits of respiratory complexes (I, III-V), 22 tRNAs, and 2 rRNAs. Recent studies have identified the effects of mtDNA variants on a variety of physiological traits and disease conditions, but few studies have investigated the regulation of the expression of mtDNA encoded genes. Knowing these can help us understand mtDNA associations with phenotypes, and how mtDNA encoded genes may affect those. In this study, we examine the effects of mtDNA and nucDNA effects on mtDNA encoded genes in 48 tissues using data from the GTEx Consortium. We identified 85 mtDNA cis-eQTLs, and 618 trans-eQTLs between nucDNA variants and mtDNA encoded genes. Leveraging the covariance between mtDNA encoded genes that may be due to causal relationships between them, we identified 118 further cis eQTLs for mtDNA genes using the covariates for multi phenotype algorithm (CMS). We further asked if these effects are driven by specific cell types within a tissue using Decon-eQTL, and found interaction effects between cell types and both mtDNA cis-eQTL and trans eQTLs between nucDNA variants and mtDNA encoded genes. Seven of those interaction effects are present in neurons, suggesting the importance of mitochondria function in neuronal activity. Our results represent the largest eQTL study on mtDNA gene regulatory effects to date and form the basis for further investigations into the mtDNA function and mito-nuclear interactions.

37

Combining Mendelian Randomization and Randomized Control Trial Study Designs to Determine Effects of Adiposity on the Plasma Proteome

Lucy J. Goudswaard*[1, 2, 3, 4], Laura J. Corbin[1, 2], David A. Hughes[1, 2], Michael V. Holmes[5], Naveed Sattar[6], Ingeborg Hers[3, 4], Nicholas J. Timpson[1, 2]

[1]*Medical Research Council (MRC) Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom;* [2]*Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom;* [3]*School of Physiology, Pharmacology and Neuroscience, University of Bristol, United Kingdom;* [4]*Bristol Heart Institute, Bristol, United Kingdom;* [5]*Medical Research Council Population Health Research Unit, University of Oxford, Oxford, United Kingdom;* [6]*Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, United Kingdom*

Mendelian randomization (MR) studies using a UK blood donor cohort (INTERVAL) have revealed that higher BMI has a substantial effect on the plasma proteome. We aimed to determine the effect of adiposity on the circulating proteome by combining MR and a randomized control trial (RCT). We used SomaLogic proteomic data from the Diabetes Remission Clinical Trial (DiRECT). Participants with obesity and type 2 diabetes were allocated either guideline-based care (control) or total diet replacement (TDR) (intervention) treatment. Serum samples were taken at baseline and after 12 months. Participants in the TDR group had mean BMI change of -3.6kg/m^2 vs -0.46kg/m^2 in the control group. We performed linear regression to explore the association between BMI change and protein change. To reduce confounding and utilize RCT study design, treatment group was used as an instrument for BMI change in a two stage least squares analysis. INTERVAL analyses suggested a causal effect of BMI on proteins including fumarylacetoacetase (0.51 SDs higher per SD higher BMI, 95% CI 0.30-0.72) and growth hormone receptor (GHR) (0.43 SDs, 95% CI 0.23-0.63). The DiRECT study provided evidence that proteins that are positively associated with BMI can be reduced by lowering BMI: 1 SD reduction in BMI reduces fumarylacetoacetase levels by 0.73 SDs (95% CI 0.55-0.90) and the GHR by 0.57 SDs (95% CI 0.42-0.73). Concordance across study designs, each with different potential sources of bias, gives increased confidence in the estimated causal effect of BMI on selected proteins, pointing to their potential involvement in downstream health outcomes.

38

Adjusting for Principal Components Can Induce Spurious Associations in Genome-Wide Association Studies in Admixed Populations

Kelsey E. Grinde[1*], Brian L. Browning[2], Sharon R. Browning[3]

[1]*Department of Mathematics, Statistics, and Computer Science, Macalester College, Saint Paul, Minnesota, United States of America;* [2]*Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, United States of America;* [3]*Department*

of Biostatistics, University of Washington, Seattle, Washington, United States of America

Principal component analysis (PCA) is widely used to control for population structure in genome-wide association studies (GWAS). It has been shown that the top principal components (PCs) typically reflect population structure, but deciding exactly how many PCs to include in GWAS regression models can be challenging. Often researchers will err on the side of including more PCs than may be actually necessary in order to ensure that population structure is fully captured. However, through both theoretical results and application to TOPMed whole genome sequence data for 1,888 and 2,676 unrelated African American individuals from the Jackson Heart Study (JHS) and Chronic Obstructive Pulmonary Disease Genetic Epidemiology Study (COPDGene), respectively, we show that adjusting for extraneous PCs can actually induce spurious associations. In particular, spurious associations arise when PCs capture local genomic features, such as regions of the genome with atypical linkage disequilibrium (LD) patterns, rather than genome-wide ancestry. In JHS and COPDGene, we show that careful LD pruning prior to running PCA, using stricter thresholds and wider windows than is often suggested in the literature, can resolve these issues, whereas excluding lists of high LD regions identified in previous studies does not. We also show that the rate of spurious associations can be appropriately controlled in these data when we simply adjust for either the first PC or a model-based estimate of admixture proportions. Our work demonstrates that great care must be taken when using principal components to control for population structure in genome-wide association studies in admixed populations.

39

Computationally Efficient, Exact, Multimarker Omnibus Tests by Leveraging Individual Marker Summary Statistics from Large Biobanks

Angela M. Zigarelli[1*], Hanna Venera[2], Brody Receveur[3], Jack M. Wolf[4], Jason Westra[5*], Nathan Tintle[5*]
[1]University of Massachusetts Amherst, Massachusetts, United States of America; [2]Pacific Lutheran University, Tacoma, Washington, United States of America; [3]George Mason University, Fairfax, Virginia, United States of America; [4]University of Minnesota, Minneapolis, Minnesota, United States of America; [5]Dordt University, Sioux City, Iowa, United States of America

As biobanks become increasingly popular, genotypic and phenotypic data has become increasingly accessible in the form of summary statistics. The publishing of these summary statistics alleviates many issues, including that of data privacy and confidentiality, as well as high computational costs. However, questions remain about how useful these summary statistics can be when used for downstream genotype-phenotype association analyses with complex phenotypes. Here we present a novel method for evaluating the association between complex, researcher defined phenotypes (linear combination of existing phenotypes), and an arbitrary large number of Single Nucleotide Variants (SNVs), and researcher defined covariates. We present a method for calculating various omnibus tests including the F test statistic and the Unified Score test statistic for rare variants using only summary statistics as inputs. We validate exact formulas for our method through simulation, and provide an illustrative real data application using fatty acid and genotypic data from the Framingham Heart Study.

40

When DNA Methylation (5-Methylcytosine at CpG) Meets SNP

Jing Qi Hao[1,2*], Andrew D. Paterson[1,2]
[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Program in Genetics and Genome Biology, Hospital for Sick Children Research Institute, Toronto, Canada

Microarray artifacts, such as within-probe SNPs, could interfere with associations between CpGs and traits. The current approach of excluding CpGs with within-probe SNPs prior to downstream analyses could result in reduced power.

In a simulation study, a range of SNP minor allele frequencies (MAFs) (0.1, 0.3, 0.5), SNP effects on CpG, CpG effects on quantitative trait (QT), SNP effects on QT, as well as CpG-SNP interaction on QT (all 0 to 2) were examined. As expected, when QT was modelled by CpG only, the CpG test had inflated type 1 error in the presence of both SNP effects on CpG and QT, as well as CpG-SNP interaction on QT. Adjusting for the SNP, the CpG test also had inflated type 1 error when there existed CpG-SNP interaction on QT.

Modelling QT by CpG, SNP and CpG-SNP interaction, type 1 error was well-controlled in the CpG and CpG-SNP interaction tests. Type 1 error was also well-controlled in a joint test of CpG and CpG-SNP interaction. Without CpG-SNP interaction on QT, the joint test was more powerful than the CpG test. In addition, without CpG effect on QT, the joint test was more powerful than the CpG-SNP interaction test. As such, for CpGs with within-probe SNPs (e.g., ≈10% of the Illumina MethylationEPIC CpGs on chromosome 19 with MAF > 0.05), the joint test

should be employed to control false discoveries, as well as to improve power to detect CpG and SNP-specific CpG effects on QT.

41

Assessing the Causal Impact of Adiposity Variation on Rates of Hospital Admission: Application of Mendelian Randomization

Audinga-Dea Hazewinkel[1,2]*, Rebecca C. Richmond[1,2], Kaitlin H. Wade[1,2], Padraig Dixon[2,3]
[1]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [2]MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Body mass index (BMI) and waist-hip-ratio (WHR) are measures of adiposity, the former being a good marker for overall total body fat, the latter describing regional adiposity. Higher adiposity has been associated with the increased prevalence of many chronic diseases. We analyze how BMI and WHR causally influence rates of hospital admission. Conventional analyses of this relationship are susceptible to omitted variable bias from variables that jointly influence both hospital admission and adipose status. We implement a novel quasi-Poisson instrumental variable models in a Mendelian Randomization framework, identifying causal effects from random perturbations to germline genetic variation. We estimate the individual and joint effects of BMI, WHR, and WHR adjusted for BMI. We also implement multivariable instrumental variable methods in which the causal effect of one exposure is estimated conditionally on the causal effect of another exposure. Data on 310,471 participants and over 550,000 inpatient admissions in the UK Biobank were used to perform one-sample and two-sample Mendelian Randomization analyses. The results supported a causal role of adiposity on hospital admissions, with consistency across all estimates and sensitivity analyses. Point estimates were generally larger than estimates from comparable observational specifications. We observe an attenuation of the BMI effect when adjusting for WHR in the multivariable Mendelian Randomization analyses, suggesting that an adverse fat distribution, rather than a higher BMI itself, may drive the relationship between adiposity and risk of hospital admission.

42

Multiclass Regularized Regression Integrating Prior Information

Jingxuan He[1]*, Chubing Zeng[1], Juan Pablo Lewinger[1], David V. Conti[1]
[1]Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America

Regularized regression with sparsity-inducing penalties is a common approach to select features in high-dimensional contexts such as genomics. To enhance model prediction and interpretation, we introduce a prior-informed penalized regression for predicting multi-categorical phenotypes in high-dimensional data contexts. Specifically, the regression coefficients are regularized by feature-specific Elastic-Net penalty parameters which are modeled as a log-linear function of prior covariates such as gene functional annotations and genes information from previous studies. Penalty parameters are estimated by an empirical Bayes method instead of cross-validation using a partial quadratic approximation and the resulting marginal likelihood is optimized by a an iterative reweighted algorithm. Through simulation studies and an applied example for selecting genomic data with annotation, we demonstrate our method's improved prediction accuracy, feature selection, and effect estimation compared with regular penalized models. We discuss the relationship to relevant vector machine and present extensions for grouped and ungrouped penalty vectors across multiple classes.

43

Comparing Gene Expression Across Paired Human Airway Models for Cystic Fibrosis Precision Medicine

Gengming He[1, 2*], Naim Panjwani[2], Julie Avolio[3], Hong Ouyang[3], Shaf Keshavjee[4,5], Johanna M. Rommens[2,6], Tanja Gonska[3,7], Theo J. Moraes[3,8], Lisa J. Strug[2,9,10]
[1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [2]Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada;[3]Translational Medicine, Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada;[4]Division of Thoracic Surgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada.Toronto Lung Transplant Program, University Health Network, Toronto, Ontario, Canada; [5]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada; [6]Department of Paediatrics, Division of Gastroenterology, Hepatology and Nutrition, The Hospital for Sick Children, Toronto, Ontario, Canada; [7]Division of Respiratory Medicine, Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada; [8]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada; [9]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada

Cystic fibrosis (CF) is caused by loss-of-function variants in the *CF transmembrane conductance regulator (CFTR)*, with modifier genes impacting lung disease

severity and therapeutic efficacy. Cultured human nasal epithelia (HNE) are becoming an important surrogate airway model for the gold standard cultured human bronchial epithelia (HBE) to assess the efficacy of CF therapies, because HNE are more easily accessible from patients. However, it remains unknown whether the HNE and HBE genome-wide transcriptome are similar, which we investigate here.

RNA-sequencing of paired HNE and HBE samples, cultured and fresh (n=71), that were collected from 21 individuals with CF was carried out. We implemented an equivalence testing procedure based on the two one-sided t-test (TOST) to assess the statistical evidence for similarity in transcriptome between HNE and HBE. A comparison of cultured and fresh airway tissues showed that the culturing process had little impact on the expression level of CF lung disease modifier genes identified in genome-wide association studies (FDR<0.1 for equivalence testing). Across cultured HNE and HBE, more than 90% of genes exhibited equivalent expression levels (FDR<0.1). The co-expression relationships of CF lung disease modifier genes estimated using a Gaussian graphical model also overlapped between cultured HNE and HBE, reflecting common biological processes at play in the two tissues.

In conclusion, we demonstrated the similarity of the transcriptome between cultured HNE and HBE, which supports the use of HNE as a surrogate airway model to investigate the efficacy of CF therapeutics and modifier genes in the context of CF precision medicine.

44

Association of Classic HLA Alleles with 28-day Sepsis Survival in GEN-SEP[$]

Tamara Hernandez-Beeftink[1,2]*, Megan L. Paynton[3], Beatriz Guillen-Guio[1], Jose M. Lorenzo-Salazar[4], Almudena Corrales[1,5], Eva Suarez-Pajes[1], M. Isabel García-Laorden[2,5], Louise V. Wain[3,6], Jesús Villar[2,5], Carlos Flores[1,4,5,7]

[1]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [2]Research Unit, Hospital Universitario Dr. Negrin, Las Palmas de Gran Canaria, Spain; [3]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [4]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [5]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [6]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom; [7]Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna, Santa Cruz de Tenerife, Spain

Introduction: Sepsis is a severe inflammatory response to infections and shows a high mortality rate. Given the importance of the major histocompatibility complex in inflammatory and immunological diseases, we assessed the association of the human leukocyte antigen (HLA) locus with 28-day survival in patients with sepsis from the GEN-SEP study.

Methods: Genotypes for 685 patients were phased using SHAPEIT and Impute2 based on SNP2HLA references to impute classic HLA alleles from eight genes, amino acids, and SNPs. Survival analyses were performed on those patients (181 died) with Cox regressions adjusting for gender, age, and the main two principal components derived from about 100,000 SNPs. Significance thresholds were established at $P<2.49e-4$ for HLA alleles, $P<4.83e-5$ for amino acids, and $P<1.50e-5$ for SNPs, after Bonferroni correction.

Results and Conclusions: We analyzed a total of 207 HLA alleles, 1,034 amino acids, and 10,968 SNPs. None of the HLA alleles (lowest P-value=0.0169), amino acids (lowest P-value=0.0169), or SNPs (lowest P-value=6.17e-3) were significantly associated with survival. Given the heterogeneity of sepsis, these results suggest that the association of this region with 28-day sepsis survival may have modest effect sizes or might relate to rarer variants.

[$]The GEN-SEP Network: Miryam Prieto-González, Aurelio Rodríguez-Pérez, Demetrio Carriedo, Jesús Blanco, Alfonso Ambrós, Elena González-Higueras, Elena Espinosa, Arturo Muriel, David Domínguez, Abelardo García de Lorenzo, José M. Añón, & Javier Belda.

Funding: Instituto de Salud Carlos III (CB06/06/1088, FI17/00177, MV19/00017, PI17/00610, PI16/00049, PI19/00141, PI20/00876), Ministerio de Ciencia e Innovación (RTC-2017-6471-1; AEI/FEDER, UE), and agreement OA17/008 with ITER; ECIT CGIEU0000219140.

45

Regional Variation of Imputation Accuracy in France

Anthony F. Herzig[1]*, Lourdes Velo-Suárez[1] Christian Dina[2], Richard Redon[2], Jean-François Deleuze[3], Emmanuelle Génin[1], Frex Consortium, FranceGenRef Consortium

[1]Génétique, génomique fonctionnelle et biotechnologies (UMR 1078) EFS, Université de Brest Faculté de Médecine - IBRBS 22 avenue Camille Desmoulins F-29238 BREST Cedex 3, France; [2]unité de recherche de línstitut du thorax UMR1087 UMR6291 Université de Nantes, Institut National de la Santé et de la Recherche Médicale U1087, Centre National de la Recherche Scientifique : UMR62918 quai Moncousu - BP 70721 - 44007 Nantes Cedex 1, France; [3]Centre National de Recherche en Génomique Humaine *CEA-DRF-IBFJ CNRGH, GENMED, Fondation Jean Dausset*

CEPH, Institut de Biologie François Jacob, CEA, Université Paris-Saclay, F-91057, Evry, France

Funding from the FROGH project (ANR-16-CE12-0033); Funding from Laboratory of Excellence GENMED (ANR-10-LABX-0013); Funding from the French Ministry of Higher Education, Research and Innovation for the POPGEN France Genomic Medicine pilot project; Funding from Inserm Cross-Cutting Project GOLD

Funding from France Génomique (ANR-10-INBS-09-08) for the FREX project.

Imputation of missing genotypes is widely performed to enrich datasets of genotyped individuals. For this practice, advances in software capabilities have been rapid, enormous haplotype reference panels have been assembled, and dedicated computation servers have been created.

France has a population with extensive internal fine-structure; and while public reference panels contain an abundance of European genomes, few are French. Therefore, using a 'study specific panel' (SSP) for France would likely be beneficial. To investigate, we imputed 550 French individuals with array and whole-exome sequencing data from six different regions in France, using either the University of Michigan imputation server with the Haplotype Reference Consortium panel, or in-house using a panel of 850 whole-genome sequenced French individuals.

With approximate geo-localization of both our target and SSP individuals we are able to pinpoint the relevance of the proposed SSP for different groups of target individuals. This helped to illustrate different scenarios where SSP-based imputation would be preferred over server-based; or vice-versa. We also show to a high degree of resolution how the proximity of the reference panel to each target individual determined the accuracy of both haplotype phasing and genotype imputation.

Previous studies have shown the benefits of combining public reference panels with SSPs. Getting the best out of both resources is unfortunately impractical. This is because the largest public panels are only accessible through external servers; limiting possibilities. We put forward a pragmatic solution where server-based and SSP-based imputation outcomes can be combined based on comparing posterior genotype probabilities.

46

Control for Population Stratification in Genetic Association Studies based on GWAS Summary Data

Shijian Yan[1*], Qiuying Sha[1], Shuanglin Zhang[1]

[1]*Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America*

Over the past 16 years, genome-wide association studies (GWAS) have generated a wealth of new information. Summary data from many GWAS are now publicly available, promoting the development of many statistical methods for association studies based on GWAS summary statistics, which avoids the increasing challenges associated with individual-level genotype and phenotype data sharing. However, for population-based association studies such as GWAS, it has been long recognized that population stratification can seriously confound association results. For large GWAS, it is very likely that population stratification and cryptic relatedness exists that will result in inflated type I error in association testing. Although many methods have been developed to control for population stratification, only two of these approaches can be used to control population stratification without individual-level data: one is based on linkage disequilibrium score regression (LDSC) and the other one is based on genomic control (GC). However, the performances of these two approaches are currently unknown. In this research, we use extensive simulation studies including population with subpopulations, spatially structured populations, and population with cryptic relatedness to compare the performance of these two approaches to control population stratification using only GWAS summary data without individual-level data. The results from this research will provide very useful information for researchers using GWAS summary statistics when trying to control for population stratification.

47

Machine Learning-Driven Radiogenomic Analysis Framework With Mediation Analysis for Identifying Prognostic Radiogenomic Biomarkers in Breast Cancer

Qian Liu[1,2,3*], Pingzhao Hu[1,2]

[1] *Department of Biochemistry and Medical Genetics;* [2] *Department of Computer Science;* [3] *Department of Statistics University of Manitoba, Winnipeg, Manitoba, Canada*

Background: Radiogenomics is a field where medical images and genomic profiles are jointly analyzed to answer critical clinical questions. We proposed a novel framework to identify breast cancer (BC) prognostic radiogenomic biomarkers from multi-modal magnetic resonance imaging (MRI) and genomic data, which may serve as a substitute for genetic testing.

Methods: Bayesian tensor factorization was used to extract integrated multi-omics features from gene expression, DNA methylation, and copy number variation data of 762 BC patients. An explainable deep

learning (DL) model was built to extract multi-modal MRI features for 61 of the BC patients with MRI data. Regularized regression models were trained to impute the MRI features. Survival analyses were performed to estimate the prognostic significance of each MRI feature. Statistical mediation analyses were performed to explore underlying biological mechanisms of the identified biomarkers. Traditional semi-auto radiomic features and previously established gene expression features were used for comparison.

Results: Three DL-based multi-modal MRI radiogenomic biomarkers were successfully identified, which were confirmed to have significant differences in overall survival (log-rank test, Bonferroni corrected *P*-value<0.05). The most significant one was associated with 10 BC risk genes (such as *APITD1*, *HNF4*) and several metabolism related pathways (Purine metabolism pathway and Tryptophan metabolism pathway), which has a significant mediation effect on the relationship between the function of natural killer cells and the BC survival time (adjusted *P*-value<0.002).

Conclusion: The results may promote MRI as a non-invasive examination of probing BC prognosis and multi-level molecular status, and ultimately increase precision in BC prognosis.

48

Relationship Between Major Depressive Disorder (Mdd) Symptoms and Mdd Heterogeneity

Lianyun Huang[1]*, Sonja Tang[2], Morten Dybdahl Krebs[3,4], Andrew J. Schork[3,4,5], Thomas M. Werge[3,4,6], Andy Dahl[7], Verena Zuber[2], Na Cai[1]

[1]*Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany;* [2]*Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom;* [3]*Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Roskilde, Denmark;* [4]*The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Copenhagen, Denmark;* [5]*Neurogenomics Division, The Translational Genomics Research Institute (TGEN), Phoenix, Arizona, United States of America;* [6]*Department of Clinical Medicine, Copenhagen University, Copenhagen, Denmark;* [7]*Section of Genetic Medicine, University of Chicago, Chicago, Illinois, United States of America*

Major Depressive Disorder (MDD) is a leading cause of disability with high population prevalence and serious impairments to daily functioning. There is mounting evidence that MDD is a highly heterogeneous disease with multiple causal pathways, represented by different symptom profiles and clinical presentations, yet a mechanistic understanding of its underlying causes remains elusive. Here, we aim to investigate the genetic architecture and heterogeneity of MDD through its relationship with its constituent symptoms, using 14 lifetime MDD symptoms from the online follow-up in 337,545 white British participants in UK Biobank (UKB). Our work has found that each MDD symptom has a partly distinct genetic architecture, demonstrated through genome-wide association studies (GWAS) loci, SNP-based heritability estimates, and their pairwise genetic correlations. Using univariate Mendelian Randomization (MR), we rank symptoms by their MR effect estimates for MDD and find that genetic liability to certain symptoms is significantly associated with MDD and that effect strengths differ between symptoms. Moreover, we used bidirectional MR to investigate how symptoms relate with each other. Using Subtest, we found that the genetic effects contributing to symptoms such as anhedonia and fatigue among cases of MDD also contribute directly to MDD, meaning partitioning MDD cases by these symptoms leads to subgroups with different liabilities. This result is replicated in other cohorts, including GWAS from Integrative Psychiatric Research (iPSYCH) (n=34,230), Psychiatric Genomics Consortium (PGC-MDD) (n=142,646) and 23andMe (n=307,354). This is the first robust demonstration supported by genetic evidence that there are different genetically-driven MDD causal pathways represented by distinct symptoms.

49

The Relationships Between Body Mass Index and Metabolite Response To A Standardized Meal Challenge

David A. Hughes[1]*, Ruifang Li-Gao[2], Caroline Bull[1], Renée de Mutsert[2], Frits R. Rosendaal[2], Dennis O. Mook-Kanamori[2], Ko Willems van Dijk[2], Nicholas J. Timpson[1].

[1]*MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom;* [2]*Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands*

The response, or change, of metabolite abundance to a meal is an emergent trait in studies of disease. Body mass index (BMI) is a recognized risk factor for numerous health outcomes and may influence metabolite response to feeding. Here we use the Netherlands Epidemiology of Obesity (NEO) study to examine associations between BMI and metabolic response to standardized liquid meals and extend this by using Mendelian randomization to estimate causal effects.

The NEO study conducted a standardized liquid meal challenge in 5700 participants and collected metabolome profiles using the Nightingale metabolomics platform. Observational and one-sample Mendelian randomization (MR) analysis were conducted, by linear modelling, to estimate the effect of BMI on

metabolites in the fasting, postprandial, and response (or change in abundance) states.

We observed 95 metabolites (53 fasting, 35 postprandial, 7 response) that are associated with BMI in MR analyses at a P-value threshold of 0.05, with all 95 directionally consistent with observational analyses. After correcting for multiple testing four metabolites have evidence of a BMI effect. Two of which are the branch chain amino acids isoleucine (beta=-0.00069, SE=0.00018 mmol/L change per unit increase in BMI (kg/m^2)) and leucine (beta=-0.00086, SE=0.00023) in the response state.

Our results suggest that BMI has an association with fasting and post prandial abundances of specific metabolites. Additionally, our work suggests that the branch chain amino acids isoleucine and leucine are likely to have feeding response abundance differences influenced by BMI that might mark life course risk exposures derived from regular feeding.

50

Polygenic Risk Scores for Prediction of Breast Cancer in Korean Women

Yon Ho Jee[1*], Weang-Kee Ho[2,3], Douglas F. Easton[4,5], Soo-Hwang Teo[3,6], Peter Kraft[1,7]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]School of Mathematical Sciences, Faculty of Science and Engineering, University of Nottingham Malaysia, Semenyih, Selangor, Malaysia; [3]Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia; [4]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [5]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom; [6]Sime Darby Medical Centre, Subang Jaya, Selangor, Malaysia; [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

Polygenic risk scores (PRS) have been shown to predict breast cancer risk. However, it is unclear how well PRS developed in large European GWAS would perform in Asian women relative to PRS developed in smaller Asian studies. Here we evaluate the predictive ability of four PRS developed using Asian or European training samples: two PRS were restricted to genome-wide significant SNPs ($PRS\text{-}11_{ASN}$ and $PRS\text{-}136_{EUR}$) and two were not ($PRS\text{-}42_{ASN}$ and $PRS\text{-}209_{EUR}$). For each PRS, we compared area under the curve (AUC) and expected-to-observed ratio (E/O) of three absolute risk models in a cohort of 41,031 women from the Korean Cancer Prevention Study-II (KCPS-II) Biobank: (i) a model using incidence, mortality, and risk factor distributions among U.S. women and European-ancestry RRs; (ii) a recalibrated model, using Korean incidence mortality and risk factor distributions but European-ancestry RRs; and (iii) a fully Korean-based model using Korean incidence mortality and risk factor distributions and RR from the KCPS. Both Asian and European PRS improved risk prediction for breast cancer in Korean women (Qx: AUC=0.65, $Qx+PRS\text{-}42_{ASN}$: AUC=0.68, $Qx+PRS\text{-}209_{EUR}$: AUC=0.69 in Korean-based model for age<50). We found that the U.S.-based absolute risk models overestimated the risks for women age ≥50 years, even after incorporation of PRS ($PRS\text{-}42_{ASN}$: E/O=1.93, $PRS\text{-}209_{EUR}$: E/O = 1.92). Our absolute risk projections suggest that risk-reducing lifestyle changes would lead to larger absolute risk reductions among women at higher PRS. This indicates that PRS may be useful for prioritizing individuals for targeted intervention on their lifestyle such as alcohol intake and obesity.

51

Assessing the Impact of Winner's Curse on Mendelian Randomisation

Tao Jiang[1*], Dipender Gill [2], Adam Butterworth[1], Stephen Burgess[3]

[1]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [3]Medical Research Council Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

In genome-wide association studies, Winner's Curse is a form of selection bias due to the focus on variants achieving a stringent level of significance. Consequently, the estimated effect sizes of the significant variants are likely to be upwards biased in the discovery sample, which may impact downstream analyses such as Mendelian Randomization (MR) if these biased estimates are used. We demonstrate the impact of Winner's Curse in a specific applied setting to quantify the extent to which it has an impact on MR results. We use the UK Biobank dataset and take body mass index and coronary artery disease as our exemplar exposure and outcome respectively. By splitting our dataset randomly into three equally sized subgroups, we perform our discovery analysis in one subgroup and get genetic association estimates for our exposure and outcome from all three. We then repeat this process 100 times. This allows us to consider five different scenarios with different patterns of overlap between the discovery, exposure, and outcome datasets. By considering the scenario where the three sets of results come from distinct subgroups

as our gold standard, we show that Winner's Curse can have a statistically significant impact on the MR results (*P*-value=4.32 x 10^{-7}), a change in OR from 1.086 to 1.099. In conclusion, we have demonstrated that although Winner's Curse is insufficient to invalidate the conclusions in this example, care needs to be taken to avoid biased point estimates through study design or appropriate methodology.

52

Assisted Clustering of Gene Expression Using Regulator Data From Overlapping Samples

Wenqing Jiang[1*], Daniel Levy[2,3], George T O'Connor[4], Josée Dupuis[1]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2]Boston University's and National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America; [3]The Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [4]Pulmonary Center, Department of Medicine, Boston University, Boston, Massachusetts, United States of America

As omics measurements profiled on different platforms are interconnected, integrative approaches that incorporate regulatory effect from multi-level omics data are needed. When the multi-level omics data are from the same individuals, gene expression (GE) clusters can be identified using information from regulators like genetic variants and methylation. When the multi-level omics data are from different individuals, the choice of integration approaches is limited. We developed an approach to improve GE clustering by integrating regulator data from different but overlapping samples. We achieve this through (1) decomposing gene expression into the regulated part and the other part that is not regulated by measured factors, (2) optimizing the clustering goodness-of-fit objective function. We do not require the availability of different omics measurements on all individuals. A certain amount of sample overlap between GE data and the regulator data is adequate for modeling the regulation, thus improving GE clustering.

A simulation study shows that performance of the proposed approach depends on the strength of GE-regulator relationship, degree of missingness, data dimensionality, sample size, and number of clusters. Across the various simulation settings, the proposed method shows competitive performance in terms of accuracy compared to the alternative K-means method, especially when the clustering structure is due mostly to the regulated part, rather than the unregulated part. We further validate the approach by applying to 8,902 Framingham Heart Study participants with data on 2,181 genes and regulation information of methylation and genotype from different but overlapping participants. We identify gene clusters associated with pulmonary function.

53

The Association of Accelerated Epigenetic Age with Time-to-death Mediated by Subclinical and Clinical Vascular Outcomes

Rong Jiang[1*], Elizabeth R. Hauser [2,3], Lydia C. Kwee[2], Svati H. Shah [2,4], William E. Kraus[2,4], Cavin K. Ward-Caviness[5]

[1]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, United States of America; [2]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, North Carolina, United States of America; [3]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, United States of America; [4]Division of Cardiology, Department of Medicine, School of Medicine, Duke University, Durham, North Carolina, United States of America; [5]Center for Public Health and Environmental Assessment, US Environmental Protection Agency, Chapel Hill, North Carolina, United States of America

Epigenetic age is a biomarker of aging and has been correlated with cellular and organ system aging. Age acceleration (AA), the difference between epigenetic age and chronological age, is a strong predictor of lifespan and healthspan in the general population, but has been little evaluated in people with chronic conditions. Additionally, the underlying mechanisms of AA's ability to predict mortality is unclear. Using 562 participants from the CATHGEN cohort and three AA measures (Horvath epigenetic age acceleration - HAA, phenotypic age acceleration - PhenoAA, and Grim age acceleration - GrimAA), we evaluated whether AA was associated with all-cause mortality and examined multiple pathways that may mediate AA-mortality associations. The total effect, direct effect, indirect effect and percentage mediated were estimated using a regression adjustment approach. Accelerated epigenetic aging was associated with greater hazard ratios for mortality ($P \leq 0.01$), with PhenoAA and GrimAA more strongly associated than HAA. PhenoAA-mortality association was mediated by angiopoietin-2 (ANG2, a biomarker of peripheral vascular health, 19.8% mediated, $P = 0.016$) and diabetes (8.15%, $P = 0.043$). The GrimAA-mortality association was mediated by ANG2 (12.3%, $P = 0.014$), and potentially by left ventricular ejection fraction (5.3%, $P = 0.065$).

These results indicate that epigenetic age acceleration is strongly predictive in individuals with underlying cardiovascular disease and associations in these populations may be mediated by metabolic and vascular factors. This abstract does not necessarily represent the views or policies of the US Environmental Protection Agency.

54

Automated Classification of Germline and Somatic Variants

Alexander Joyner
Saphetor, North Palm Beach, Florida, United States of America

VarSome integrates, maintains, and updates over 100 of the most relevant variant and cancer databases into our MolecularDB. The MolecularDB serves as the source for VarSome's rich and comprehensive variant annotations. While annotation is important, using those annotations to actively classify variants and ultimately interpret them is the goal of NGS analysis. VarSome offers automated ACMG, AMP, and CNV classifications. These classifiers are subject to user input, including phenotype information for somatic variants. They are designed for clinical decision support and serve as excellent first pass variant filters for exome and whole-genome analysis. The new CNV classifier is guided by the recent paper from ACMG, Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Users will be able to call CNV from FASTQ files, or input CNV VCF files to classify current and historic CNV data. We are constantly updating our classification processes, with efforts underway to improve publication tagging and increased phenotype integration with our classifiers for higher resolution disease-based classifications.

55

Genetic Determinants of Prostate-Specific Antigen Levels Improve Cancer Screening Utility

Linda Kachuri[1*], Thomas J. Hoffmann[1], Rebecca E. Graff[1], John P. Shelley[2], Kerry Schaffer[2], Jonathan D. Mosley[2], Stephen K. Van Den Eeden[3], John S. Witte[1]
[1]*Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, California, United States of America;* [2]*Departments of Internal Medicine and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America;* [3]*Division of Research, Kaiser Permanente Northern California, Oakland, California, United States of America*

Prostate-specific antigen (PSA) testing has been shown to reduce prostate cancer (PCa) mortality, but remains controversial due to its high sensitivity, but low specificity. Using a polygenic risk score (PRS) to correct for PSA variation that is not due to cancer may improve its screening utility.

We conducted the largest genome-wide association study of PSA in cancer-free men (N=65,962; 63,338 European ancestry) using longitudinal measures in the UK Biobank (UKB: N=26,491), BioVU (N=8078), and Genetic Epidemiology Research on Adult Health and Aging (N=30,088) cohorts.

Our analysis identified 87 significantly associated ($P<5\times10^{-8}$) variants, including 44 SNPs in 37 newly discovered loci that are not in linkage disequilibrium with previously reported associations. PRS_{PSA} comprised of these 87 variants accounted for 12.8% of PSA variation. We observed inverse associations for PRS_{PSA} with PCa mortality (HR=0.82, $P=7.4\times10^{-9}$) and Gleason score (≤6 vs. ≥8 OR=0.83, $P=8.4\times10^{-5}$), suggesting that genetic predisposition to elevated PSA increases the detection for low-grade disease. PRS_{PCa} based on 269 variants was highly correlated ($r=0.287$, $P<10^{-308}$) with PRS_{PSA} in an independent UKB sample (n=164,669). Although this may partly reflect pleiotropy, we found evidence of selection bias due to screening ($P=2.1\times10^{-130}$). Correcting for this bias in PRS_{PCa} SNP weights attenuated its correlation with PRS_{PSA} ($r=0.049$, $P=5.5\times10^{-93}$) and revealed a previously absent association with PCa mortality (HR=1.18, $P=0.035$).

Our work provides preliminary evidence that PSA genetics may improve PCa screening and risk prediction. Larger and more diverse study populations are required to fully characterize the genetic determinants of PSA and optimize their clinical utility.

56

Introduction Sources and Early Spread of SARS-CoV-2 in Senegal

Khadim Kébé[1*], Abou A.M. Diouara[1], Fatou Thiam[1], Fatou Fall[2], Yakhya Dièye[1], Alpha O. Touré[1], Cheikh M. Nguer[1],
[1]*Group GRBA-BE, LE3PI laboratory, Department of Chemical Engineering and Applied Biology, Polytechnic Higher School of Dakar;* [2]*Cheikh Anta Diop University, BP 5005, Dakar-Fann*

Senegal, like the rest of the world, is currently exposed to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first laboratory-confirmed Coronavirus disease 2019 (COVID-19) case has been reported there on March 2nd, 2020 and was related to an imported case from international travel. In Senegal, as everywhere else, human-to-human transmission was very efficient. Molecular epidemiology studies could provide important informations about SARS-CoV-2 evolution and transmission that are necessary for

outbreak response. To infer the genomic diversity and evolutionary epidemiology in Senegal, we used a total of 290 SARS-CoV-2 genomes retrieved in public databases (GISAID database) including 25 from Senegal. Collection period of sequences range between December 26, 2019 to March 31, 2020. DNA polymorphism, maximum likelihood and haplotype network analyses revealed multiple introductions of SARS-CoV-2 into Senegal, mainly from European countries followed by local transmission chains. Our results showed also that four groups of viruses deriving from different positions of the reference virus, which we call "epidemiological strains" circulated in Senegal at the early stages of the outbreak the most prevalent carrying the widly documented variant D614G. We suspected that the presence of this variant is linked to the rapid increase COVID-19 cases. Indeed, it's seem to be an urgent need to have more genomic data based on large sample covering all the affected localities to further map the genome variations in conjunction with clinical symptoms and epidemiological data. It is therefore essential to monitor the possible emergence of mutations that might alter transmissibility and pathogenicity of SARS-CoV-2.

57

Genetic Association Study of COVID-19 Severe Versus Non-severe Cases by RNA-seq and Whole Genome Sequencing in a Hong Kong Cohort

Qi Li[1,2^], Zigui Chen[3^], Yexian Zhang[2], Marc KC Chong[1,2], Benny CY Zee[1,2], Maggie H Wang[1,2*], Paul KS Chan[3*]
[1]JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong SAR, China; [2]CUHK Shenzhen Research Institute, Shenzhen, China; [3]Department of Microbiology, Stanley Ho Centre for Emerging Infectious Diseases, Li Ka Shing Institute of Health Sciences, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
^ Co-first authors
* Corresponding: maggiew@cuhk.edu.hk, paulkschan@cuhk.edu.hk

Previous studies suggest that genetic risk factors might contribute to COVID-19 susceptibility and severity. We sought to identify the associated markers related to severe cases of COVID-19.

We conducted a case-control study of 24 COVID-19 patients, including 8 severe and 16 non-severe cases, who were enrolled from the Prince of Wales Hospital in Hong Kong. Among the 24 individuals, sixteen (6 severe vs. 10 non-severe) underwent whole-genome sequencing (WGS) and 21(6 severe vs. 15 non-severe) underwent RNA sequencing. After quality control, Fisher's exact test was used to identify SNPs associated with disease severity; logistic regression and Kolmogorov-Smirnov test were performed to detect genes with differential expression in the severe subjects compared to the non-severe subjects.

Five variants in genes *BAGE2*, *ROCK1P1*, *LOC105376980*, and *ANKRD36BP2* in the WGS dataset, and 6 differentially expressed genes in RNA-seq data had *P*-value <0.05. Among the shortlisted genes from RNA-seq, 100% of them were known to play a role in the development of nervous system disease, 81.25% were enriched in circulatory system diseases, and 75% have pleiotropy effect in muscular and skeletal system diseases. All of the genes identified in the WGS data were also known to be associated with reproductive system diseases and integumentary system diseases.

This study identified susceptible genetic markers associated with COVID-19 severity in a Chinese cohort in Hong Kong. The identified genes with pleiotropic effects on COVID-19 severity, nervous system disease, circulatory system diseases, and reproductive system diseases might explain the observed comorbidity of COVID-19 infection and these disorders.

58

A Powerful Test of Ancestral Heterogeneity in the Effects of Gene Expression on Complex Traits

Katherine Knutson, Wei Pan
University of Minnesota, Department of Biostatistics, Minneapolis, Minnesota, United States of America

The Transcriptome Wide Association Study (TWAS) is a widely used approach which integrates expression and GWAS data to study the role of cis-regulated gene expression (GEx) in complex traits. TWAS models GEx as a function of cis-eQTL genotypes. However, strong evidence suggests that the genetic architecture of GEx varies across populations. Furthermore, recent findings point to possible ancestral heterogeneity in the effects of GEx on complex traits, heterogeneity which may be amplified in TWAS by modeling GEx as a function of cis-eQTLs. We present a novel extension to TWAS which models heterogeneity in the effects of cis-regulated GEx which are correlated with ancestry. By jointly analyzing samples from multiple populations, our multi-ancestry TWAS framework can improve power to detect genes with shared expression-trait associations across populations through increased sample sizes, as compared to existing stratified TWAS approaches. Under our proposed model, we derive score tests for homogeneous, heterogeneous, and total GEx effects. Our preliminary simulations reveal conserved Type-I error rates and high power across a number of scenarios, holding promise for further simulations

on larger simulated datasets. We apply our test to case-control genotypes from the Alzheimer's Disease Sequencing Project (ADSP) and prediction models from the MESA study. We identify a number of genes with suggestive heterogeneous effects in Alzheimer's Disease. In forthcoming work, we will apply our test to an augmented ADSP sample, and consider application to continuous endophenotypes from the UK Biobank.

59

Identification of Representative Trees in Random Forests Based on a New Tree-Based Distance Measure

Björn-Hergen Laabs[1]*, Ana Westenberger[2], Inke R. König[1]
[1]*Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany;* [2] *Institute of Neurogenetics, University of Lübeck, Lübeck, Germany*

In life sciences random forests are often used to train predictive models. However, gaining any explanatory insight into the mechanics leading to a specific outcome is rather complex, which impedes the implementation of random forests into clinical practice. By simplifying a complex ensemble of decision trees to a set of a few representative trees, it is assumed to be possible to observe common tree structures, the importance of specific features and variable interactions. Thus, representative trees could also help to understand interactions between genetic variants.

Intuitively, representative trees are those with the minimal distance to all other trees, which requires a proper definition of the distance between two trees. Thus, we developed a new tree-based distance measure, which incorporates more of the underlying tree structure, than other metrics. We compared our new method with the existing metrics in an extensive simulation study and applied it to predict the age at onset based on a set of genetic risk factors in a clinical data set.

In our simulation study we were able to show the advantages of or weighted splitting variable approach and observed that removing poorly grown trees could be a further use case of tree-based distance measures. Our real data application revealed that representative trees are not only able to replicate the results from a recent genome-wide association study, but also can give additional explanations of the genetic mechanisms. Finally, we implemented all compared distance measures in R and made them publicly available in the package timbR (https://github.com/imbs-hl/timbR).

60

Genetic Endowments for Education and Social Capital: An Investigation Accounting for Genetic Nurturing Effects

Michael Lebenbaum[1], France Gagnon[2], Claire de Oliveira[1,3], Audrey Laporte[1]
[1]*Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, Ontario, Canada;* [2]*The Dalla Lana School of Public Health (DLSPH), University of Toronto, Toronto, Ontario, Canada;* [3]*Centre for Health Economics and the Hull York Medical School, University of York, Heslington, York, United Kingdom*

Background: The education polygenic score (PGS) is an emerging tool for economic studies. A growing body of literature has examined the association between education PGS and human, financial, and health capital. Few studies have investigated social capital, and no studies examining the education PGS social capital association have accounted for genetic nurturing (i.e., indirect genetic effects). Social science studies accounting for genetic nurturing have not controlled for measures of the environment, such as parental investments in children. The objective of this study was to evaluate the effect of the education PGS on social capital.

Methods: We used European-descent data from the National Longitudinal Study of Adolescent to Adult Health (Add-Health) (N=2,000 observations in sibling sample). The education PGS was constructed using GWAS data that excluded the Add-Health sample. Social capital was measured as volunteering, religious service attendance, team sports participation and the number of friends. We used panel sibling fixed effects (SFE) probit (volunteering) or ordered probit (other dependent variables) models and controlled for birth weight, breastfeeding, parental investments and child abuse to account for genetic nurturing.

Results: Accounting for genetic nurturing, the associations between education PGS and volunteering were reduced in size but still large and significant (β=0.166, p=0.044), while reducing to non-significance the associations between education PGS and religious service attendance and number of friends (p>0.05).

Discussion: These findings demonstrate that genetic endowments for education play an important role in influencing volunteering behavior and further demonstrate the importance of accounting for genetic nurturing.

61

ExPheWas: A Browser for Gene-based PheWAS Associations

Marc-André Legault[1,2,3*], Louis-Philippe Lemieux Perreault[1,2], Marie-Pierre Dubé[1,2,3]

[1]Montreal Heart Institute, Montreal, Quebec, Canada; [2]Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Quebec, Canada; [3]Department of Biochemistry and Molecular Medicine, Université de Montréal, Montreal, Quebec, Canada

The identification and characterization of genetic associations with human phenotypes can improve our understanding of disease etiology and support the discovery of novel therapeutic targets.

The UK Biobank is an excellent resource for such research. It includes >500,000 participants with available genotypes and linkage to health records including cancer, hospitalization and death registries. We used this cohort to conduct a phenome-wide association study (pheWAS) of all protein coding loci. We implemented a gene-based principal component analysis approach based on common genetic variants from 19,114 protein coding regions and tested their association with 1,210 phenotypes including anthropometric measurements, laboratory biomarkers, algorithmically-defined cardiovascular outcomes and health records. The results of this analysis are publicly available in a user-friendly browser https://exphewas.statgen.org.

As a proof of concept, we characterized the 137 genes associated with atrial fibrillation at a false discovery rate of 1%. Using enrichment analysis, the identified genes were strongly enriched for relevant biological processes such as cardiac muscle contraction (P_{adj}=9.5 × 10^{-6}) and antiarrhythmic drug targets in the ChEMBL database. By further investigating possibly novel genes, we prioritized *MYOT* as a likely atrial fibrillation gene. In ExPheWas, this gene is strongly associated with heart rate (P=8.6 × 10^{-31}) and atrial fibrillation (P=4.9 × 10^{-11}). *MYOT* is a component of sarcomeric Z-disks and mendelian mutations of this protein are associated with cardiomyopathy.

ExPheWas is a database of gene to phenotype associations that may be used for follow-up of genetic studies and for enrichment analysis.

62

New Selection Probability Computation for Pleiotropic Variants Associated with Both Quantitative and Qualitative Traits

Kipoong Kim[1]*, Hokeun Sun[1]

[1]Department of Statistics, Pusan National University, Busan, Korea

In recent genetic association studies, statistical methods to identify pleiotropic variants associated with multiple phenotypic traits have been developed, since susceptible variants with small or moderate effects are rarely detected by association methods based on a single trait. However, most of the existing methods to identify pleiotropic variants are designed for only quantitative traits even though pleiotropic variants are often associated with both quantitative and qualitative traits. This is a statistically challenging problem because there does not exist an appropriate multivariate distribution to model both quantitative and qualitative data. There are some meta-analysis methods which basically integrate summary statistics of individual variants associated with either a quantitative or qualitative trait. However, these methods cannot account for correlations between genetic variants. In this article, we propose new selection method to prioritize individual variants associated with both quantitative and qualitative traits. For individual traits, regression coefficients of elastic-net regularization are first estimated and then they are additively combined to compute selection probability of individual variants. In our extensive simulation studies where either homogeneous or heterogeneous variant effects on both quantitative and qualitative traits were considered, we demonstrated that the proposed method outperforms the existing meta-analysis methods in terms of true positive selection. We also applied the proposed method to peanut data with 4 quantitative and 2 qualitative traits, and cowpea data with 2 quantitative and 4 qualitative traits.

Keywords: pleiotropy, quantitative and qualitative traits, elastic-net, selection probability

63

A segregation Analysis of 17,425 Population-based Breast Cancer Families: Implications ror Breast Cancer Genetic Susceptibility and Risk Prediction

Shuai Li[1,2,3*], Andrew Lee[2], Robert J. MacInnis[4], Leila Dorling[2], Sara Carvalho[2], Tu Nguyen-Dumont[3], Melissa C. Southey[3,4,5], Douglas F. Easton[2], John L. Hopper[1#], Antonis C. Antoniou[2#]

[1]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria, Australia; [2]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [3]Precision Medicine, School of Clinical Sciences At Monash Health, Monash University, Clayton, Victoria, Australia; [4]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia; [5]Department of

Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia
# Contributed equally.

Breast cancer (BC) risk models used in family cancer clinics rely on accurate modelling of the familial relative risks; however, rare pathogenic variants (PVs) in known BC susceptibility genes, together with known common genetic variants, do not fully explain the familial aggregation of BC. We aimed to investigate plausible genetic models for the residual familial aggregation by using data on 17,425 UK and Australian families ascertained through population-based female BC probands. Eighty-six percent (86%) of probands were screened for PVs in *BRCA1*, *BRCA2*, *PALB2*, *CHEK2*, *ATM*, *TP53* using gene-panel sequencing, and 881 probands carried PVs in these genes. We conducted complex segregation analyses and fitted genetic models in which BC incidence depended on the effects of PVs in the known susceptibility genes, other unidentified major genes and a normally distributed polygenic component. Maximum likelihood estimation was used to estimate the allele frequencies and risk associated with the PVs.

PVs in the six known genes explained 21% of the familial aggregation. After allowing for these PVs, the best fitting model for the residual aggregation involved a recessively inherited allele with frequency of 13% (95%CI:0.6-21%) and penetrance of 69% (95%CI:43-91%) by age 80 for homozygous carriers, explaining 18% of the residual familial aggregation, and a polygenic component with an age-independent variance of 1.27 (95%CI:0.96-1.63). In conclusion, in addition to the known BC susceptibility genes, and polygenes, unidentified major BC susceptibility genes might exist which explain BC familial aggregation. Our findings have implications for attempts to identify new BC susceptibility genes and risk prediction.

64

Applying Recurrent Weighted Replanting to Detect Gene-gene Interaction in Case-parent Trios

Qing Li[1]*, Anthony M. Musolf[1], and Joan E. Bailey-Wilson[1]
[1]*Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America*

To improve the power to detect causal variants in genomic datasets, a novel procedure called Recurrent Weighted Replanting (RWR) is proposed based on the Random Forest (RF) method. In the past, our group developed trio Random Forest (trioRF) to detect gene-gene interactions in case-parent trio data. TrioRF uses cases and random samples of variant calls from the set of matched pseudo controls, and utilizes a proper classification criterion to detect the difference in variant calls between the pseudo controls and cases. The importance score for each feature is calculated based on permutation tests.

Although trioRF can be scaled up and include a large number of features (SNPs here) in one run to fit a classification tree, the chance of including multiple SNPs in interaction within one run is proportional to the total number of SNPs. As a result, for millions of SNPs, we need to increase the number of trees and improve power. Therefore, we propose to implement RWR for trioRF. We run trioRF multiple times. The initial step is to obtain the importance scores for each SNP (feature). Then we use importance scores from the initial step to select subsets of SNPs to include in subsequent runs and novel weights to adjust the probability that each SNP is available for splitting, a procedure denoted as RWR trioRF. At the final stage, the importance scores from the multiple RWR trioRF runs are calculated. We conducted simulation studies to demonstrate the power of RWR trioRF.

65

Subtyping Individuals with Facial and Genomic Data Views in the Presence of Confounders

Zuqi Li[1]*, Kristel Van Steen[1,2], Peter Claes[3], Mark Shriver[4], Seth Weinberg[5], Susan Walsh[6]
[1]*BIO3 – Laboratory for Systems Medicine, Department of Human Genetics, KU Leuven, Leuven, Belgium;* [2]*BIO3 – Laboratory for Systems Genetics, GIGA-R Medical Genomics, University of Liège, Liège, Belgium;* [3]*Laboratory for Imaging Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium;* [4]*Department of Anthropology, Pennsylvania State University, Pennsylvania, United States of America;* [5]*Center for Craniofacial and Dental Genetics, Department of Oral Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America;* [6]*Walsh FDP Lab, Indiana University Purdue University Indianapolis, Indianapolis, Indiana, United States of America*

Single-view data do not exhibit the huge potentials that multiple data views can offer in giving complementary information to relevant problems in precision medicine. These include finding groups of patients with similar treatment benefits or highlighting population subgroups with comparable risks to disease. At the heart of these examples lie unsupervised learning algorithms. Particularly challenging, yet under-investigated, is obtaining un-confounded clusterings. A confounder to a clustering is any variate that unwantedly drives the clustering. In this project, we integrate facial imaging and genomic data views from a US cohort and assess different strategies to derive subgroups of individuals that are unaffected by age (regarded as confounder). To learn a better representation and reduce

dimensionality, facial images are hierarchically grouped into 63 segments in six levels and SNPs are mapped to genes by genomic positions. Subsequently, Principle Component Analysis is performed on each segment and on each gene. The two views are integrated using sparse Canonical Correlation Analysis (CCA) and various clustering algorithms are employed. Age is taken into account at three possible levels: during pre-processing (e.g. Partial Least Squares regressing out), while clustering (e.g. Kernel Conditional Clustering; KCC) or during multi-view integration (e.g. by contrasting age-informed and -uninformed sparse multiple CCA). Even though correcting for confounders at the pre-processing step has the desired effect, it remains tedious and the utility of residuals derived from inappropriate model formulations is questionable. Surprisingly, in our data application, clusters generated by KCC are still highly affected by age. Extensive simulations support these findings.

66

Distribution of *CYP2C9* Variant Genes in the Healthy Thai Population Associated with Medical Cannabis Metabolic Pathway

Bunnalin Liamputhong[1], Atchara Srisodsai, Ph.D.[2], Patompong Satapornpong, Ph.D.[3*]

[1]Regents International School of Bangkok, Bangkok, Thailand; [2]MedCoach Institute, Bangkok, Thailand; [3]The division of general pharmacy practice, department of pharmaceutical care, College of Pharmacy, Rangsit University, Pathum Thani, Thailand

**Corresponding author*

Introduction: Medical cannabis consists of tetrahydrocannabinol (THC) and cannabidiol (CBD). CYP2C9 is a major metabolizer of THC and the frequency of *CYP2C9* genotypes vary between populations worldwide. THC-induced adverse effects (AE) can be explained in terms of *CYP2C9* gene variants in pharmacogenetics. This study aims to investigate the impact of the frequency of *CYP2C9* variants related to THC metabolic pathways in the healthy Thai population.

Materials and Methods: We have recruited a set of unrelated healthy Thai subjects (n=160). Genotyping for *CYP2C9 (*2 and *3)* were subsequently analyzed through real-time PCR.

Results: We found that *CYP2C9*1* alleles is the most common form of the *CYP2C9* gene among the Thai population, comprising a percentage frequency of approximately 95.94%. *CYP2C9*3* alleles were found to occur at only 4.06%. However, *CYP2C9*2* alleles were absent among the subjects. Furthermore, in the aspect of phenotypes and genotypes, we found that the phenotype of extensive metabolizers (EM) (*CYP2C9*1/*1*, wild-type) genes have the highest frequency. Intermediate metabolizers(IM)(*1*3) and poor metabolizers (*3/*3) were also found from the samples, respectively 6.88% and 0.62%. Our results for *CYP2C9*1/*3* and **3/*3* frequency is also similar to previous studies in Asian populations. The allelic variants *CYP2C9*2* and *CYP2C9*3* have been presented to experience decreased enzymatic activity in the THC metabolism pathway.

Conclusion: In conclusion, the distribution of *CYP2C9*3* in Thai populations might be associated with THC-induced serious adverse effects through metabolic pathways.

Keywords: *CYP2C9* gene, Thai population, Cannabis

67

A Weighted Selection Probability to Locate Rare Variants Associated with Highly Correlated Multiple Phenotypes

Xianglong Liang[1,*], Hokeun Sun[1]

[1]Department of Statistics, Pusan National University, Busan, Korea

In the past few decades, many statistical methods have been developed to identify rare variants associated with a complex trait or a disease. Recently, rare variant association studies with multiple phenotypes have drawn a lot of attentions because association signals can be boosted when rare variants truly associated with more than one phenotype. Most of existing statistical methods to identify rare variants associated with multiple phenotypes are based on a group test, where a gene or a genetic region is tested one at a time. However, these methods are not designed to locate individual rare variants within a gene or a genetic region. In this article, we propose a weighted selection probability to locate rare variants associated with highly correlated multiple phenotypes. Selection probability represents selection frequency of nonzero regression coefficient in a regularization model using bootstrap sampling. Because the strength of an association varies in phenotypes for susceptible rare variants, selection probability weights can be computed based on a distribution of selection frequency of variants each phenotype. Simulation study showed that the weighted selection probability method outperforms unweighted selection methods in terms of true positive selection, when phenotype outcomes are highly correlated with each other. We also applied the proposed method to our genomic data set consisting of 10,783 rare variants and 13 correlated phenotypes.

Keywords: rare variants, multiple phenotypes, selection probability, regularization

68

A Genome-wide Association Study Identifies Two Novel Loci for Respiratory Infection with *Pseudomonas aeruginosa* in Cystic Fibrosis

Boxi Lin[1*], Jiafen Gong[2], Naim Panjwani[2], Katherine Keenan[2], Cheng Wang[2], Lei Sun[1,3], Lisa J. Strug[1,2,3]
[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Hospital for Sick Children, Toronto, Toronto, Canada; [3]Department of Statistical Sciences, University of Toronto, Toronto, Canada

Pseudomonas aeruginosa (*Pa*) is a common pathogen that contributes to progressive Cystic Fibrosis (CF) lung disease. Genetic factors contribute approximately 50% to 85% of variation in age of persistent *Pa* infection in CF individuals but remain unknown.

We conducted a genome-wide association study of genetic modifiers for age of first and persistent *Pa* infection in 2,740 Canadians with CF. Our primary analysis on 1,037 individuals identified two novel genome-wide significant loci, rs62369766 (near *FGF10* on Chromosome 5; *P*-value=1.78E-8) and rs927553 (*SPATA13* on Chromosome 13; *P*-value=1.72E-8), for persistent *Pa* infection age. Through a phenome-wide association study in population-based databases of rs62369766 and rs927553 we observed their association with lung function and immunological phenotypes, respectively in non-CF cohorts. We further investigated the genetic overlap between chronic *Pa* infection age and lung function in CF through a polygenic risk score (PRS), defined in the largest GWAS of CF lung disease to date (n=6,365). CF lung function has a moderate phenotypic correlation with chronic *Pa* infection age (Pearson correlation coefficient=0.12, *P*-value=2E-4). The PRS constructed from ~8,000 SNPs associated with CF lung function is significantly associated with chronic *Pa* infections age (*P*-value = 0.006), supporting the possibility that targeting genetic factors of chronic infections will improve lung function outcomes.

Our study identifies novel loci potentially modifying the age of chronic *Pa* infections in CF, and provides new insights into the genetic bases of *Pseudomonas aeruginosa* infections.

69

Combining the Strengths of Inverse-variance Weighting and Egger Regression in Mendelian Randomization Using a Mixture of Regressions Model

Zhaotong Lin[1*], Yangqing Deng[1], Wei Pan[1]
[1]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America

With the increasing availability of large scale GWAS summary data on various traits, Mendelian randomization (MR) has become commonly used to infer causality between a pair of traits, an exposure and an outcome. It depends on using genetic variants, typically SNPs, as instrumental variables (IVs). The inverse-variance weighted (IVW) method (with a fixed-effect meta-analysis model) is most powerful when all IVs are valid; however, when horizontal pleiotropy is present, it may lead to biased inference. On the other hand, Egger regression is one of the most widely used methods robust to pleiotropy, but it suffers from loss of power. We propose a two-component mixture of regressions to combine and thus take advantage of both IVW and Egger regression; it is both more efficient (i.e. higher powered) and more robust to pleiotropy (i.e. controlling type I error) than either IVW or Egger regression alone by accounting for both valid and invalid IVs respectively. We propose a model averaging approach and a novel data perturbation scheme to account for uncertainties in model/IV selection, leading to more robust statistical inference for finite samples. Through extensive simulations and applications to the GWAS summary data of 48 risk factor-disease pairs and 63 genetically uncorrelated trait pairs, we showcase that our proposed methods could often control type I error better while achieving much higher power than IVW and Egger regression. We expect that our proposed method will be a useful addition to the toolbox of Mendelian randomization for causal inference.

70

PRICKLE1 x *FOCAD* Interaction Revealed by Genome-wide vQTL Analysis of Human Facial Traits

Dongjing Liu,[1*] Hyo-Jeong Ban,[2] Ahmed M. El Sergani,[3,4] Myoung Keun Lee,[3] Jacqueline T. Hecht,[5] George L. Wehby,[6] Lina M. Moreno,[7] Eleanor Feingold,[8,9] Mary L. Marazita,[3,4,8,10] Seongwon Cha,[2] Heather Szabo-Rogers,[4,11,12,13] Seth M. Weinberg,[3,4,8] John R. Shaffer,[3,4,8]
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [2]Future Medicine Division, Korea Institute of Oriental Medicine, Daejeon, Daejeon, South Korea; [3]Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [4]Department of Oral and Craniofacial Sciences, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania United States of America; [5]Department of Pediatrics, University of Texas McGovern Medical Center, Houston, Texas United States of America; [6]Department of Health Management and Policy, University of Iowa, Iowa City, Iowa, United States of America; [7]Department of Orthodontics, University of Iowa, Iowa City, Iowa, United States of America; [8]Department of Human Genetics, Graduate School of Public Health,

University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [9]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [10]Department of Psychiatry and Clinical and Translational Sciences, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [11]Department of Developmental Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [12]McGowan Institute of Regenerative Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [13]Center for Craniofacial Regeneration, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

The human face is a highly complex and variable structure resulting from the intricate coordination of numerous genetic and non-genetic factors. Hundreds of genomic loci impacting quantitative facial features have been identified. While these associations have been shown to influence morphology by altering the mean size and shape of facial measures, their effect on trait variance remains unclear. We conducted a genome-wide association analysis for the variance of 20 quantitative facial measurements in 2447 European individuals and identified several suggestive variance quantitative trait loci (vQTLs). These vQTLs guided us to conduct an efficient search for gene-by-gene interactions (G × G), which uncovered an interaction between *PRICKLE1* and *FOCAD* affecting cranial base width. We replicated this G x G interaction signal at the locus level in an additional 5128 Korean individuals. We used the hypomorphic *Prickle1*^{Beetlejuice} (*Prickle1*^{Bj}) mouse line to directly test the function of Prickle1 on the cranial base and observed wider cranial bases in *Prickle1*$^{Bj/Bj}$. Importantly, we observed that the Prickle1 and Focadhesin protein colocalize in murine cranial base chondrocytes and this colocalization is abnormal in the *Prickle1*$^{Bj/Bj}$ mutants. Taken together, our findings uncovered a novel G × G effect in humans with strong support from both epidemiological and molecular studies. These results highlight the potential of studying measures of phenotypic variability in gene mapping studies of facial morphology.

71

GMEPS: A Fast and Efficient Likelihood Approach for Genome-wide Mediation Analysis Under Extreme Phenotype Sequencing

Janaka S. S. Liyanage[1]*, Jeremie Estepp[2], Kumar Srivastava[1], Yun Li[3,4,5], Tomi Mori[1], Guolian Kang[1]
[1]Department of Biostatistics, Saint Jude Children's Research Hospital, Memphis, Tennessee, United States of America; [2]Department of Hematology, Saint Jude Children's Research Hospital, Memphis, Tennessee, United States of America; [3]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [4]Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [5]Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Due to many advantages such as higher power of detecting the association of rare genetic variants in human disorders and cost effectiveness of the study design, extreme phenotype sequencing (EPS) is a rapidly emerging study design in epidemiological and clinical studies investigating how genetic variations associate with underlying disease mechanisms. However, investigation of the mediation effect of genetic variants in underlying disease mechanisms is strictly restrictive under the EPS design because existing methods cannot well accommodate the non-random extreme tails sampling process incurred by the EPS design. In this paper, we propose a likelihood approach for testing the mediation effect of genetic variants through continuous and binary mediators on a continuous phenotype under the EPS design (GMEPS). Besides implementing in EPS design, it also can be utilized as a general mediation analysis procedure. Extensive simulations and two real data applications of a genome-wide association study of benign ethnic neutropenia under EPS design and a candidate-gene study of neurocognitive performance in patients with sickle cell disease under random sampling design demonstrate the superiority of the GMEPS under the EPS design over widely used mediation analysis procedures, while demonstrating compatible capabilities under the general random sampling framework.

72

Disentangling Genetic Feature Selection and Aggregation in Transcriptome-Wide Association Studies

Chen Cao[1], Devin Kwok[2], Qing Li[1], Jingni He[1], Xingyi Guo[3], Qingrun Zhang[1,2,*], Quan Long[1,2,4,5,*,]
[1]Department of Biochemistry & Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada; [2]Department of Mathematics & Statistics, University of Calgary, Calgary, Canada; [3]Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, United States of America; [4]Department of Medical Genetics, University of Calgary, Calgary, Canada; [5]Hotchkiss Brain Institute, O'Brien Institute for Public Health, University of Calgary, Calgary, Canada

The success of transcriptome-wide association studies (TWAS) has led to substantial research towards

improving the predictive accuracy of its core component of Genetically Regulated eXpression (GReX). GReX links expression information with genotype and phenotype by playing two roles simultaneously: it acts as both the outcome of the genotype-based predictive models (for predicting expressions) and the linear combination of genotypes (as the predicted expressions) for association tests. From the perspective of machine learning (considering SNPs as features), these are actually two separable steps - feature selection and feature aggregation - which can be independently conducted. In this work, we show that the single approach of GReX limits the adaptability of TWAS methodology and practice. By conducting simulations and real data analysis, we demonstrate that disentangled protocols adapting straightforward approaches for feature selection (e.g., simple marker test) and aggregation (e.g., kernel machines) outperform the standard TWAS protocols that rely on GReX. Our development provides more powerful novel tools for conducting TWAS. More importantly, our characterization of the exact nature of TWAS suggests that, instead of questionably binding two distinct steps into the same statistical form (GReX), methodological research focusing on optimal combinations of feature selection and aggregation approaches will bring higher power to TWAS protocols.

73

Modelling Hidden Genetic Risk From Family History for Improved Polygenic Risk Prediction

Tianyuan Lu[1,2], Vincenzo Forgetta[1], J. Brent Richards[1,3,4], Celia M. T. Greenwood[1,3,5,6]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2]Quantitative Life Sciences Program, McGill University, Montreal, Canada; [3]Department of Human Genetics, McGill University, Montreal, Canada; [4]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom; [5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada; [6]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada

With many polygenic risk scores demonstrating research and clinical utilities, it has become questionable whether family history, a traditional genetic predictor, still provides valuable information.

We hypothesize that family history of complex traits may be partially determined by rare pathogenic variants and exposure to other environmental factors shared by parents and offspring, in addition to common genetic predisposition. Therefore, leveraging family history, we propose a latent factor model to quantify genetic risk not captured by a common SNP-based polygenic risk score. This model enables calibration of polygenic risk scores with respect to family history without fitting regression models.

We applied our model to predict adult height for 941 children in the Avon Longitudinal Study of Parents and Children, as well as nine complex diseases among >400,000 White British participants in the UK Biobank. Our predictor explained ~55% of the total variance in adult height, close to the estimated heritability of height and substantially higher than ~40% captured by a polygenic risk score or mid-parental height alone. For all complex diseases investigated in the UK Biobank, parental information brought significant improvements. For instance, combined with age and sex, our predictor achieved an area under the receiver operating characteristic curve (AUROC) of 0.734 in identifying individuals with type 2 diabetes, exhibiting significantly stronger discriminative power than the polygenic risk score (AUROC = 0.712) or the parental disease history (AUROC = 0.707).

Our work provides a paradigm and supports the utility of incorporating family history into polygenic risk score-based genetic risk prediction models.

74

Comparison of Region-based and Single SNP Genome-wide Association Testing Methods in the Canadian Longitudinal Study on Aging

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

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, 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

Compared to single SNP analysis, region-based multi-SNP association analysis has the advantage of reducing multiple test burden, and can better capture signals under complex genetic architectures. In previous work, we developed a genome partitioning algorithm (BigLD) to generate genomic regions containing physically contiguous SNPs in strong local correlation, and a multi-SNP regional regression test (MLC) that incorporates local linkage disequilibrium (LD) within each region. The objectives of this study are: 1. Compare MLC results to other regional tests and to single SNP results; 2. Develop multiple testing criteria for regional test statistics.

We applied BigLD and MLC methods to genotyping data from the Canadian Longitudinal Study on Aging (CLSA) which is a large population cohort of individuals aged 45 to 86 years. Following imputation by the Haplotype Reference Consortium and standard quality control, we partitioned the autosomes into 92,327 regions and analysed genome-wide association in 17,779 individuals of European ancestry using 5,288,020 SNPs (minor allele frequency ≥ 0.05). To facilitate validation, we focused analysis on lipid traits with published genome wide associations, including Global Lipids Genetics Consortium results. Although different regional association test statistics were usually well-correlated, compared to other region-based methods, MLC *P*-values were often equivalent or smaller. Using simulation methods to estimate genome-wide significance thresholds for region-based and single SNP tests, and control family-wise error rate at the same level, we observed that MLC detected association in several regions missed by conventional single SNP tests.

75

Information-guided Gene-environment Interaction Analysis

Shuangge Ma
Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America

For the etiology, progression, and response to treatment of complex diseases, gene-environment (G-E) interactions have important implications beyond the main G and E effects. Compared to main-effect-only analysis, G-E interaction analysis is more seriously challenged by high data dimensionality and weak signals. For both marginal and joint analysis, we propose incorporating additional information to improve G-E interaction analysis. In particular, we consider information that can be described using an adjacency matrix, and such information can be extracted from gene expression networks, physical locations of SNPs, KEGG pathways, as well as gene ontology terms. We develop penalization approaches, which are highly coherent for marginal and joint analysis, for effectively incorporating the adjacency matrix in estimation. A major advancement is that the proposed approaches do not require the information to be complete or accurate. Efficient computational algorithms are developed, and extensive simulations show that incorporating the adjacency information can lead to moderate to big improvement in estimation and identification accuracy. In the analysis of GENEVA diabetes data with SNP measurements and TCGA data with gene expression measurements, the information-guided marginal and joint analyses lead to interesting findings with sound biological interpretations and satisfactory statistical properties (stability, prediction performance, etc.). Overall, this study delivers a new venue for efficiently and cost-effectively improving G-E interaction analysis.

76

A GWAS Summary-statistics Based Approach to Examine the Role of the Serotonin Transporter Promoter Tandem Repeat Polymorphism (5-HTTLPR) in Psychiatric Phenotypes

Arunabha Majumdar[1], Preksha Patel[2], Bogdan Pasaniuc[3], Roel A. Ophoff[2]
[1]Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Telangana, India; [2]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, California, United States of America; [3]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America

In genetic studies of psychiatric disorders, the serotonin transporter (5-HTT/SLC6A4) promoter polymorphism, a 43-base pair insertion/deletion polymorphism (5-HTTLPR), is a commonly studied locus. Since, many antidepressant drugs were reported to selectively inhibit the function of the serotonin transporter, the SLC6A4 gene was long considered a strong candidate gene for neurobehavioral phenotypes. Consequently, genetic variants such as 5-HTTLPR were believed to modulate the expression of the serotonin transporter protein. However, the genetic association findings between 5-HTTLPR and psychiatric phenotypes have been found to be inconsistent across studies. Since the 5-HTTLPR VNTR cannot be tested via available SNP arrays, Lu et al. (2012) proposed an efficient algorithm to predict the genotypes of 5-HTTLPR based on nearby SNPs. The predicted genotypes can then be used for association testing. However, this approach requires access to individual-level genotype and phenotype data. To utilize the advantage of publicly available GWAS summary statistics obtained from very large sample sizes, we develop a GWAS summary-statistics based approach for testing 5-HTTLPR associations with various phenotypes.

We first cross-verify the accuracy of the summary-statistics based approach for 62 phenotypes in UK Biobank. Since we observed a strong similarity between the predicted individual-level 5-HTTLPR genotype-based approach and the summary-statistics based approach, we applied our method to the available neurobehavioral GWAS summary statistics data. We found no genome-wide significant evidence for association between 5-HTTLPR and any neurobehavioral trait. Our approach

provides a systematic way to re-assess and examine the role of specific VNTRs and related genetic polymorphisms in genetic susceptibility to diseases/phenotypes.

77

Cis-regulatory Hubs Constitute a Powerful Model to Understand the Impact of 3d Organization in Schizophrenia

Loïc Mangnier[1,2,3,7*], Charles Joly-Beauparlant[3], Arnaud Droit[3,4,7], Steve Bilodeau[5,6,7,8], Alexandre Bureau[1,2,7]
[1]Centre de Recherche CERVO, Québec, Canada; [2]Département de Médecine Sociale et Préventive, Université Laval, Québec, Canada; [3]Centre de Recherche du CHU de Québec - Université Laval, Québec, Québec, Canada; [4]Département de Médecine Moléculaire, Université Laval, Québec, Canada ; [5]Centre de Recherche du CHU de Québec – Université Laval, Axe Oncologie, Québec, Québec, Canada; [6]Centre de Recherche sur le Cancer de l'Université Laval, Québec, Québec, Canada; [7]Centre de Recherche en Données Massives de l'Université Laval, Québec, Québec, Canada; [8]Département de Biologie Moléculaire, Biochimie Médicale et Pathologie, Faculté de Médecine, Université Laval, Québec, Québec, Canada

The cis-regulatory modules (CRMs) are noncoding regulatory regions, playing a crucial role in the regulation of transcription and the emergence of complex phenotypes. Recent single-cell multi-way analyses show that several CRMs and genes locally co-interact through 3D contacts, building hubs. Despite the importance of CRM hubs in gene regulation, their precise implication in complex diseases such as schizophrenia remains unclear. In the present study, we model cis-regulatory hubs (CRHs), using available Hi-C data in neurons derived from induced pluripotent stem cells and the activity-by-contact model linking active enhancers to promoters. Comparing CRHs to either equivalent tissue-specific or non-tissue-specific structures, we showed that they constitute a relevant organization for schizophrenia. Firstly, we defined CRHs as active structures, associated with gene activity. Then, we assessed the relevance of CRHs in schizophrenia using H-Magma. Considering the noncoding SNPs in 3D contact with genes, we found an enrichment in schizophrenia-associated genes within CRHs compared to genes outside (OR=1.81). Next, using the linkage disequilibrium score regression we showed that a larger portion of schizophrenia heritability is explained by CRHs than non-tissue-specific elements, with enrichment of 3 against 0.43 on average. In addition, we also observed up to 11-fold enrichment in schizophrenia heritability compared to equivalent tissue-specific elements. This result is supported by schizophrenia-associated SNP enrichment (OR=1.29) and tissue-relevant GO pathway analysis. Our results demonstrate that CRHs in neurons constitute a useful model for understanding the 3D organization between CRMs and genes involved in the emergence of complex phenotypes such as schizophrenia.

78

NHLBI Biodata Catalyst and the Future of Cloud Computing

Alisa K. Manning[1,2,3] *, Paul Avillach, MD, PhD [4,5], Rebecca R. Boyles [6], Alison E Leaf [7], Jonathan R. Kaltman [8], Stephanie Suber [9], Ingrid Borecki on behalf of the BioData Catalyst Consortium
[1]Data Science Platform, Broad Institute, Cambridge, Massachusetts, United States of America; [2]Massachusetts General Hospital, Boston, Massachusetts, United States of America; [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [4]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, United States of America; [5]Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, United States of America; [6]Research Computing, RTI International, Durham, North Carolina, United States of America; [7]Seven Bridges, Boston, Massachusetts, United States of America; [8]Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America; [9]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

The rate of population health and biomedical data generation is accelerating rapidly. Cloud computing offers an effective way to store and analyze these data, but substantial investment is necessary to make cloud-based research portals broadly United States of Americable by the researcher community. Such systems require substantial infrastructure (scalable and secure environments), analysis resources (analysis environments, harmonized data) and training resources for data sharing, collaboration mechanisms and new statistical methodologies. The National Heart Lung and Blood Institute (NHLBI) initiated the BioData Catalyst effort with a mission to develop and integrate advanced cyberinfrastructure while championing 'findable, accessible, interoperable, and reusable' (FAIR) principles. Now, BioData Catalyst provides access to a highly secure, scalable, reproducible, collaborative, and extensible cloud analysis ecosystem with controlled access to more than 3 Petabytes of heart, lung, blood, and sleep (HLBS) data, including harmonized data from

NHLBI's TOPMed Study, BIOLINCC, and several NIH COVID-19 datasets. "Bring Your Own Data" functionality enables users to securely upload and privately analyze datasets on the platform provided that such United States of Americage is allowed by their existing data use agreements and IRB policies. In this presentation, we will show examples of current research being performed on BioData Catalyst, demonstrate the utility of combining data sets to explore new hypotheses, and illustrate how BioData Catalyst, as a community-driven ecosystem, is democratizing data access and advancing HLBS science.

79

Machine Learning Approaches for Predicting Phenotypes in Pathophysiology of Multiple Sclerosis

João M. Brandet[1*]

[1]Department of Physiotherapy, Centro Universitário Filadélfia, Londrina, Paraná, Brazil

The presence of demyelination plaques in various parts of the central nervous system and formation of glial scars are characteristic of the pathophysiology of multiple sclerosis (MS). Effectively identifying disease-related genes would contribute to developing new therapeutic tools to treat and new ways of understanding the pathophysiological mechanisms of this disease. The objective of this scientific work was to develop new machine learning approaches for predicting phenotypes in pathophysiology of MS and to propose a novel framework for identifying the disease-related genes MS. The model, computational simulations and analyzes of this scientific work were elaborated with the use of software: ACD/ChemSketch, Swiss-PdbViewer, ABCpred, BepiPred-2.0, ElliPro, DEseq, GOseq, FunRich, Cytoscape, BiNGO, PepSurf, AxonDeepSeg, AxonSeg, Computer-assisted Evaluation of Myelin formation (CEM), PyMol, ICM-Browser, Visual Molecular Dynamics (VMD), Cell Illustrator, C-ImmSim, Simmune and ChemDraw. Network Representation Learning-based algorithms were used for the development of this work. Non-linear topological information of the protein-protein-interaction network based on node2vec, DeepWalk, and LINE was analyzed in this work. Docking studies, computed Infrared-active modes, HOMO-LUMO gaps and molecular dynamics methods were applied in the computational analysis. This research demonstrates that certain phenotypic irregularities dependent on changes in the protein-protein-interaction significantly influence the pathogenesis of MS. The future clinical applicability of this work should help guide future research into MS therapeutics, with particular attention to the long-term management of this disease.

80

Evaluation of SNPs Associated with Mammographic Density in European Women with Mammographic Density in Asian Women from South East Asia

Shivaani Mariapun[1,2*], Weang Kee Ho[2], Mei Chee Tai[1], Nur Aishah Mohd Taib[3], Cheng Har Yip[4], Kartini Rahmat[5], Mikael Eriksson[6], Per Hall[6,7], The Breast Cancer Association Consortium, Soo Hwang Teo[1,3]

[1]Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia; [2]Department of Applied Mathematics, Faculty of Engineering, University of Nottingham Malaysia, Semenyih, Selangor, Malaysia; [3]Faculty of Medicine, University Malaya Cancer Research Institute, University of Malaya, Kuala Lumpur, Malaysia; [4]Subang Jaya Medical Centre, Subang Jaya, Malaysia; [5]Biomedical Imaging Department, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [7]Department of Radiology, South General Hospital, Stockholm, Sweden

Mammographic density (MD) is a strong heritable risk factor of breast cancer. Recent GWAS and meta-analyses have identified 46 independent loci associated with MD among women of European ancestry. Their association in Asian women however is largely unknown.
We evaluated the association of 40 SNPs with area-based and volumetric densities in healthy women in the Malaysian Mammography Study (MyMammo), as determined using STRATUS (N=2,450) and Volpara™ (N=2,257), respectively. Association of MD and each SNP was conducted using linear regression, adjusting for age, body mass index and ancestry-informative principal components. In addition, we evaluated the association of these SNPs with breast cancer risk in a population-based case-control cohort of 15,890 women from Malaysia and Singapore.

Of the 40 variants tested, variants in 1q12-q21 (*SV2A, SF3B4*), 2q14.2 (*RALB, INHBB*), 6q25.1 (*CCDC170, ESR1*) and 22q13.1 (*TMEM184B*) showed strongest and significant associations with area-based density, but not volumetric density, at the Bonferroni-corrected threshold of *P*-value $<1x10^{-3}$. Three of the 40 MD SNPs in two regions; 5q23.2 and 6q25.1, were associated with breast cancer risk in Malaysian/Singaporean women at *P*-value $<1x10^{-3}$. Two of these SNPs were also associated with area-based density in this study at *P*-value <0.01.

Our study confirms the association of four loci with MD in a South East Asian cohort. Given that only 10% of MD-associated variants in European women are also associated breast cancer susceptibility in South East Asians, further analyses in Asian populations are required to characterize the genetic determinants of MD in this relatively understudied population.

81

Non-linear Mendelian Randomisation on Partly Summarized Data: Evaluation of a Collaborative Method

Amy Mason[1], Stephen Burgess[1,2]
[1]*Cardiovascular Epidemiology Unit, School of Clinical Medicine, University of Cambridge, United Kingdom;* [2]*Medical Research Council Biostatistics Unit, School of Clinical Medicine, University of Cambridge, United Kingdom*

Mendelian Randomization (MR) is a technique for using observational genetics data to estimate the causal impact of a risk factor on a disease. One of MR strengths is that it can be applied using summarised genetic associations with the exposure and outcome, utilizing the large databases of GWAS results freely available online.

However standard MR methods rely on an assumption of a linear effect of the exposure on the disease. This is a potentially erroneous assumption when we consider factors affecting heart disease, such as BMI. Non-linear methods have been designed to relax this assumption with the use of piecewise linear and fractional polynomial models. However, these non-linear models require individual level data for both the exposure and the outcome, introducing potential difficulties. Many consortiums cannot easily share their complete dataset for reasons of participant privacy.

This simulation project compares a proposed method for fitting non-linear MR models to partly summarised data to the original non-linear methods. This retains the best of both methods: relaxing the linear assumption while working on aggregated data. We demonstrate the method's viability for a range of potential exposure-outcome relationships and show the method in a practical setting in UK Biobank data, showing no evidence for non-linear effects of LDL-cholesterol on Coronary Heart Disease.

82

Fast and Robust Methods to Detect Gene-environment Interactions in Large-scale Biobanks

Joelle Mbatchou[1*], Andrey Ziyatdinov[1] and Jonathan Marchini[1]
[1]*Regeneron Genetics Center, Tarrytown, New York, United States of America*

The past decade has seen an unprecedented rise in the amount of phenotypic data available through the use of electronic health records and self-reported information. Large-scale biobanks have provided unique opportunities for researchers to make novel findings as well as validate existing targets and discover new indications for existing therapies. These biobanks constitute a rich data resource to explore and identify gene-environment (GxE) or gene-gene (GxG) interaction effects, which require environmental exposure information and larger sample sizes for sufficient power. We have previously proposed REGENIE as an efficient computational method to analyze both quantitative traits (QTs) and binary traits (BTs), including highly imbalanced traits, in large-scale biobanks that can highly reduce the computation time while accounting for population structure and relatedness. Although REGENIE was designed to discover marginal genetic effects, we have now extended it to detect GxE and GxG interaction effects for both quantitative and binary traits. We identified situations for both BTs and QTs where existing methods using robust standard errors break down for rare variants or high case-control imbalance and lead to inflated Type I error. We have developed a new approach that combines heteroscedastic linear models, penalized regression and robust standard errors, and prevents miscalibration across a wide range of scenarios. We demonstrate this approach through simulation and real data applications in UK Biobank with up to ~400,000 exome sequenced samples where we analyze body mass index and type 2 diabetes using multiple exposure variables to identify GxE effects across both common and rare variants.

83

Edge and Modular Significance Assessment in Individual Specific Network

Federico Melograna[1*], Fabio Stella[2], Kristel Van Steen[1,3]
[1]*BIO3 - Laboratory for Systems Medicine, KU Leuven, Leuven, Belgium;* [2]*Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy;* [3]*BIO3 – Laboratory for Systems Genetics, GIGA-R Medical Genomics, University of Liège, Liège, Belgium*

Individual-specific network (ISNs), defined as networks with individual-specific node or edge values, or both, are promising in the context of Precision Medicine. They can be used to infer subnetworks or sets of genes, linked to individual-relevant pathways, or to highlight associations between an individual's network properties and extrageneous data. Focusing on ISNs, with individual-unique edge weights (computed as in Kuijjer et al., 2019), evaluation of their statistical significance remains an under investigated problem. Here, we propose and compare several strategies to assess edge and modular significance in an ISN. These include leave-one-out techniques with a linear (LOO-ISN), and non-linear (MultiLOO-ISN) aggregation across ISN edges, based on resampling procedures from a multivariate normal and computing the impact on the corresponding correlation matrix. We also employ

a customized Cook's distance approach by iterative linear modelling of the edge weights in the targeted module. In view of accommodating generic ISNs with flexible edge weight definitions, we empirically evaluate the aforementioned methods against outlier detection techniques, including DBSCAN, kNN and Spoutlier (Sugiyama et al., 2013). We grid-explore different settings; varying sample and module sizes, number of outliers and outlier distributions. Heterogeneity in results increases with module size and proportion of outlying individuals; Cook's distance shows overall excellent to good performance in all scenarios. Overall, our study shows the value of using network structures in ISNs to establish significance of an individual. This is important to determine the added value of ISNs over similar networks across samples, for risk assessment, disease diagnosis or management.

84

Colorblindness Gene Implicated in Myopia in Pennsylvania Amish Pedigrees

Anthony M. Musolf[1]*, Dwight Stambolian[2], and Joan E. Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2]Department of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Family studies offer good power to identify highly penetrant genetic variants that are rare in a population but enriched within a family; such studies have experienced a renaissance with the advent of affordable whole genome sequencing (WGS). Myopia (nearsightedness) has become a major health concern, reaching epidemic proportions in some countries. It is also the second leading cause of blindness worldwide. Although myopia is known to be caused by both environmental and genetic factors, its genetic etiology remains unclear. GWAS studies have identified common variants of low to moderate effect associated with myopia, yet much of the heritability is still missing.

This study uses a family-based approach; we performed WGS on 97 individuals from seven extended Pennsylvania Amish families with prior evidence of linkage to myopia. Founder populations such as the Amish also allow for utilization of exclusive genomic architecture, like unique haplotypes, to better identify potential risk variants. The Amish also have low exposure to some known environmental myopia risk factors.

We performed genetic linkage analysis on these families assuming an autosomal dominant risk allele with 90% penetrance with no phenocopies. We identified 88 genome-wide significant variants across the families, localized to 4q13.1, 5p15.33, 8q21.3, and 9p24.1; 26 of these localized to the *CNGB3* gene at 8q21.3, including the only two exonic variants, which are predicted damaging. *CNGB3* is an excellent candidate as it is expressed in the eye and is causal for both achromatopsia (total colorblindness) and progressive cone dystrophy. Functional analysis of *CNB3* is currently planned.

85

Associations of Circulatory MicroRNAs and Clinical Traits: A Phenome-wide Mendelian Randomization Analysis

Rima Mustafa[1], Michelle M.J. Mens[2,3], Jian Huang[1,4], Gennady Roshchupkin[5,6], André G. Uitterlinden[2,7], M. Arfan Ikram[2], Marina Evangelou[8], Mohsen Ghanbari[2*], Abbas Dehghan[1,9,10*]

*[1]Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom; [2]Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands; [3]Department of Child and Adolescent Psychiatry/ Psychology, Erasmus MC, Rotterdam, The Netherlands; [4]Singapore Institute for Clinical Sciences (SICS), the Agency for Science, Technology and Research (A*STAR), Singapore; [5]Department of Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands; [6]Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands; [7]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands; [8]Department of Mathematics, Imperial College London, London, United Kingdom; [9]Dementia Research Institute, Imperial College London, London, United Kingdom ; [10]MRC Centre for Environment and Health, Imperial College London, London, United Kingdom*

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression. Previous candidate-based studies have linked miRNAs to complex traits in the observational and experimental settings, but their causal relationships remain poorly understood. Here we agnostically investigate associations of >2,000 miRNAs with a wide range of clinical traits and assess their causality.

We measured 2,083 plasma miRNAs in 2,000 participants of the population-based Rotterdam Study cohort (RS) using HTG EdgeSeq miRNA Whole Transcriptome Assay and Illumina NextSeq. Genome-wide association studies (GWAS) were conducted in RS to identify genetic variants that affect the level of miRNAs (N=1,687). We computed weighted genetic risk scores that predicted the plasma level of miRNAs (miRNA-GRS) and performed a phenome-wide

association study (PheWAS) using hospital episode statistics data that covers 905 clinical diagnoses in the UK Biobank (N=423,442). Two-sample Mendelian randomization (MR) was conducted using the inverse-variance weighted method, with weighted median and MR-Egger as sensitivity analyses to rule out horizontal pleiotropy.

We identified 455 associations between miRNA-GRS and clinical traits (FDR<0.05). MR-PheWAS resulted in 235 associations between 121 miRNAs and 175 traits across 16 disease groups. Using two-sample MR, we attempted to replicate 78 associations using data available from previous GWAS. We successfully replicated 17 associations, of which six associations were significant after Bonferroni adjustment, including associations between miR-10a-5p and cardiovascular diseases, miR-10b-5p and LDL-C, and miR-139-5p and type 2 diabetes. Collectively, our study highlights the potentially causal roles of several miRNAs in cardiometabolic health.

86

A Comparison of Association Methods for Fine-mapping Rare Variants in Case-Control Studies

Payman Nickchi[1*], Charith Karunarathna[1], Jinko Graham[1]
[1]Department of Statistics and Actuarial Sciences, Simon Fraser University, Burnaby, British Canada, Canada

We compare different methods to associate genetic variants with an inherited disease in case-control studies. These methods include single-variant testing, aggregation of variants, and methods based on sequence-relatedness. As the true relatedness among sequences is not known, we describe how we may estimate the relatedness and consider this information for fine-mapping genetic variants. We find the concept of sequence-relatedness to be useful for improving the localization of rare causal variants. We conclude with some general recommendations for fine-mapping rare variants in case-control association studies.

Keywords: fine-mapping, rare variants, detection, localization, sequence-relatedness, association methods.

87

Comparison of Mixed Model Based Approaches For Correcting For Population Substructure With Application To Extreme Phenotype Sampling

Maryam Onifade[1,*], Marie-Hélène Roy-Gagnon[2], Marie-Élise Parent[3], Kelly M. Burkett[1]
[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada; [2]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada; [3]Centre Armand-Frappier Santé Biotechnologie, Institut National de la Recherche Scientifique, Université du Québec, Laval, Canada

Mixed models are used to correct for confounding due to population stratification and hidden relatedness in genome-wide association studies. This class of models includes linear mixed models and generalized linear mixed models. Existing mixed model approaches to correct for population substructure have been previously investigated with both continuous and case/control response variables. However, they have not been investigated in the context of `extreme phenotype sampling' (EPS), where genetic covariates are only collected on samples having extreme response variable values. In this work, we compare the performance of existing binary trait mixed model approaches (GMMAT, LEAP and CARAT) on EPS data. Since linear mixed models are commonly used even with binary traits, we also evaluate the performance of a popular linear mixed model implementation (GEMMA). We use simulation to estimate the type 1 error of all approaches under confounding due to population stratification. We also apply all methods to a real dataset from a Quebec, Canada, case-control study that is known to have population substructure. Our simulation results show that for a common candidate variant, both LEAP and GMMAT control the type 1 error rate. We observe similar type 1 error control with the analysis on the Quebec dataset. However, for rare variants the false positive rate remains inflated even after correction with mixed model approaches.

88

The Reliability and Accuracy of Recombination Inferred by Shapeit2 DuoHMM on Whole Genome Sequence

Samir Oubninte[1,2], Ingo Ruczinski[3], Lisa R. Yanek[4], Rasika Mathias[4], and Alexandre Bureau[1,2]
[1]Département de Médecine Sociale et Préventive, Université Laval, Québec City, Québec, Canada; *[2]Centre de recherche CERVO, Québec City, Québec, Canada*; *[3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America*; *[4]Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America*

Few studies assessed the performance of population-based phasing combined with parental genotypes to infer recombination on whole genome sequence (WGS) data. In this study, our objective was to evaluate whether Shapeit2 duoHMM, a Hidden Markov Model using parental genotypes, infers recombination events reliably on WGS and with narrower intervals than SNP arrays. We based our analysis on the overlap between recombination events inferred by Merlin on SNP genotypes and Shapeit2 on WGS and SNP genotypes.

We used a sample of 62 extended families from the GeneSTAR study with TopMED freeze 8 WGS on 580 sequenced subjects (60% of sample). Quality control included removing variants without a PASS status, missingness rate > 5%, Hardy Weinberg equilibrium *P*-value < $1x10^{-6}$ or likely genotyping errors based on Shapeit2 or Merlin error detection. Shapeit2 was run with a window size of 500 kilobases and 200 states on WGS. To mimic a SNP array, we extracted genotypes of 355,112 autosomal markers on the Illumina OmniExpress array. The number of recombination events per meiosis inferred by Shapeit2 on the WGS data (34.3) was aligned with the expected numbers over autosomes (34.0), although Merlin overestimated this number (209). 72% of Shapeit2 recombination events on WGS were detected by Merlin, a proportion rising to 93% when restricting to events also inferred by Shapeit2 on OmniExpress genotypes. Furthermore, Shapeit2 recombination intervals were narrower on WGS than OmniExpress genotypes (median of 4562 bp vs. 49920 bp). The recombination inference with Shapeit2 on WGS thus seems to be reliable and accurate.

89

Tissue-specific Functional Annotations Highlight Association of Liver Polygenic Risk Score with Alzheimer's Disease and Related Biomarkers

Daniel J. Panyard[1*], Yuetiva K. Deming[1,2,3], Burcu F. Darst[4], Carol A. Van Hulle[2,3], Henrik Zetterberg[5,6,7,8], Kaj Blennow[5,6], Gwendlyn Kollmorgen[9], Ivonne Suridjan[10], Cynthia M. Carlsson[2,3,11], Sterling C. Johnson[2,3,11], Sanjay Asthana[2,3,11], Corinne D. Engelman[1†], Qiongshi Lu[12,13†]

[1]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin United States of America; [2]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin-Madison, Madison, Wisconsin United States of America; [3]Department of Medicine, University of Wisconsin-Madison, Madison, Wisconsin United States of America; [4]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [5]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden; [6]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden; [7]Department of Neurodegenerative Disease, UCL Institute of Neurology, London, United Kingdom; [8]UK Dementia Research Institute at UCL, London, United Kingdom; [9]Roche Diagnostics GmbH, Penzberg, Germany; [10]Roche Diagnostics International Ltd., Rotkreuz, Switzerland; [11]William S. Middleton Memorial Veterans Hospital, Madison, Wisconsin United States of America; [12]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin United States of America; [13]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

[*]*presenting author*

[†]*joint senior authorship*

Alzheimer's disease is a neurodegenerative disease whose causal mechanisms are not yet fully known. The use of functional annotation in genetic analyses has been shown to be able to enhance polygenic risk score (PRS) models, improving their power and identifying new insights into disease mechanisms. Here, we leveraged genomic functional annotations from GenoSkyline-PLUS that provide a tissue-specific measure of which genomic regions are expected to be functional. Using these annotations, individual-level genetic data from participants of European ancestry from the Wisconsin Registry for Alzheimer's Prevention (WRAP, n = 1,198) and Wisconsin Alzheimer's Disease Research Center (WADRC, n = 212) cohorts, and genome-wide association study summary statistics from the International Genomics of Alzheimer's Project (IGAP 2019), we built tissue-specific PRS models for 13 tissues and applied the scores to two longitudinal cohort studies of AD that include both cognitive diagnoses and a rich set of cerebrospinal fluid biomarkers for AD, neurodegeneration, and inflammation measured with the Roche NeuroToolKit immunoassays. The model most strongly associated with AD diagnosis (and the only model statistically significantly associated after the strongly associated *APOE* locus was removed) was the liver PRS: n = 1,116; OR = 2.19 (1.70-2.82), P = 1.46 x 10^{-9}. This liver PRS was also statistically significantly associated with two major AD biomarkers: cerebrospinal levels of amyloid (P = 3.53 x 10^{-6}) and phosphorylated tau (P = 1.45 x 10^{-5}). These findings highlight the potential of functional annotation in PRS studies and provide new evidence highlighting the role of the liver-functional genome in AD.

90

Genome-wide Association Study of Predictive Genetic Polymorphisms for Oxaliplatin Treatment Efficacy in Colorectal Cancer

Hanla A. Park[1*], Federico Canzian[2], Tabitha A. Harrison[3], Xinwei Hua[4,5], Qian Shi [6], Richard M. Goldberg[7], Steven R. Alberts[8], Michael Hoffmeister[9], Hermann Brenner[9,10,11], Ulrike Peters[3,12], Andrew T. Chan[13,14,15], Polly A. Newcomb[3,12], and Jenny Chang-Claude[1,16]

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Genomic

Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany; [3]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United State of America; [4]Department of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United State of America; [5]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, Massachusetts, United State of America; [6]Department of Quantitative Science, Mayo Clinic, Rochester, Minnesota, United State of America; [7]West Virginia University Cancer Institute, Morgantown, West Virginia; [8]Medical Oncology, Mayo Clinic, Rochester, Minnesota, United State of America; [9]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; [10]Division of Preventive Oncology, National Center for Tumor Diseases (NCT) and German Cancer Research Center (DKFZ), Heidelberg, Germany; [11]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany; [12]School of Public Health, University of Washington, Seattle, Washington, United State of America; [13]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United State of America; [14]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United State of America; [15]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, United State of America; [16]Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Oxaliplatin is a platinum drug often given in combination with other anticancer drugs to treat colorectal cancer (CRC). Several candidate gene studies identified susceptibility loci that influence outcome after oxaliplatin treatment. However, studies have not provided robust evidence for candidate SNPs as potential predictive markers. Therefore, we conducted genome-wide association analyses to identify novel predictive genetic variants associated with differential prognosis in CRC patients receiving oxaliplatin-based chemotherapy vs. chemotherapy without oxaliplatin.

Included studies were NCCTG N0147, a randomized phase III trial of stage III resected colon cancer patients, NCCTG N9741, a randomized phase III trial of metastatic colorectal cancer (mCRC) patients, and DACHS, a population-based patient cohort study. Multivariable Cox proportional hazards models were conducted with an interaction term between each SNP and type of treatment for overall survival and progression-free survival. The analysis was performed for each study and the results were combined using fixed-effect meta-analyses separately for stage III colon cancer after complete resection (3,098 patients from N0147 and 549 patients from DACHS) and mCRC (505 patients from N9741 and 437 patients from DACHS).

A locus on chr22 (rs11912167) was associated with worse overall survival in stage III colon cancer patients after oxaliplatin chemotherapy compared to chemotherapy without oxaliplatin ($P_{interaction}= 4.6\times10^{-8}$). For mCRC patients, two loci on chr1 (rs1234556) and chr12 (rs11052270) were found to be suggestively associated with differential overall survival at P-value $<5\times10^{-7}$ ($P_{interaction}=2.0\times10^{-7}$ and $P_{interaction}=2.8\times10^{-7}$, respectively).

Identified variants could be potential predictive markers for oxaliplatin treatment efficacy. These findings require confirmation in further independent studies.

91

Major Sex Differences in Allele Frequency for X-chromosome Variants in the 1000 Genomes Phase 3 Data

Zhong Wang [1], Lei Sun [2,3], Andrew D. Paterson [3,4]*
[1]School of Data Science, Fudan University, China; [2]Department of Statistic Sciences, University of Toronto, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Canada; [4]Genetics and Genome Biology, The Hospital for Sick Children, Canada

Low-coverage whole genome sequence (WGS) has been proposed as a cost-effective method to perform large-scale GWAS. We identified that an unexpectedly high proportion of SNPs on the X-chromosome in the 1000 genomes phase 3 dataset have sex differences in allele frequencies (SDAF). This SDAF persists in the recently released high coverage WGS, and it is consistent between the five superpopulations.

Our primary analysis focused on biallelic SNPs that did not overlap with indels with global MAF ≥5%, analyzing 222872, 13244, 634, and 9075 SNPs from the non-pseudo-autosomal region (NPR), pseudo-autosomal region 1 (PAR1), PAR2, and PAR3, respectively. We obtained P-values from testing for SDAF and to be conservative used p<5e-8 to declare significance. We identified NPR=0.83%, PAR1=0.29%, PAR2=13.1%, PAR3=0.85% SNPs from these four regions with significant SDAF. Of the SNPs with significant SDAF we observed bias in the direction, with females generally having higher MAF than males (% of SNPs: NPR=93%, PAR1=31%, PAR2=59%, PAR3=86%). Although SNPs with significant SDAF are located across the X-chromosome, they tend to cluster in specific regions: the boundary between PAR1 and NPR; around 30Mb (build 37); at the q-arm of the centromere; at the centromeric boundary of

PAR3; as well as in PAR2. For comparison, we performed similar analyses for chromosomes 1, 7 and 22 and found only 6, 1 and 0 SNPs with significant SDAF, respectively. These findings have implications for bioinformatic analyses of X-chromosome variants including genotyping and imputation, as well as developing association analyses that are robust to SDAF.

92

Shared Genetic and Modifiable Risk Factors for Psoriasis and Multiple Sclerosis

Matthew T. Patrick[1*], Rajan P. Nair[1], Kevin He[2], Philip E. Stuart[1], Allison C. Billi[1], Johann E. Gudjonsson[1], James T. Elder[1], Jorge R. Oksenberg[3], Lam C. Tsoi[1,2,4]

[1]Department of Dermatology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America; [2]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America; [3]Department of Neurology, University of California, San Francisco, California, United States of America; [4]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America

Psoriasis is a chronic skin disease that affects more than seven million adults in the United States of America, and multiple sclerosis (MS) is a central nervous system disease impacting around one million adults in the United States of America. These immune-mediated diseases have a higher than expected rate of co-occurrence and share immunological pathways (including the IL-23/IL-17 axis). However, until now there has been limited study to discern the effects of different genetic and modifiable risk factors on their comorbidity. We analyzed electronic health records from 1,264,343 Michigan Medicine patients and found vitamin D deficiency to be highly enriched for both diseases (psoriasis-OR=2.81, MS-OR=2.45), whereas obesity (psoriasis-OR=2.41, MS-OR=1.26) and hypertension (psoriasis-OR=2.02, MS-OR=1.33) were more closely associated with psoriasis than MS. We then applied trans-disease meta-analysis (TDMA) to genome-wide association studies from 11,024 psoriasis cases, 47,429 MS cases, and matched controls, identifying 11 genome-wide significant shared loci outside the major histocompatibility complex that were more significant in TDMA than either disease, including signals near *IL12B* (P=7.1x10^{-58}), *TYK2* (P=9.4x10^{-30}), and *STAT3/STAT5* (P=3.3x10^{-19}). We also identified eight loci with signals in opposite direction of effect comparing psoriasis and MS, including one near *REL* (P=1.8x10^{-19}, psoriasis-OR=0.86, MS-OR=1.08), which is involved in NF-κB signaling. Of the signals we identified, only one (a shared locus near *DLEU1*) was associated with a modifiable risk factor (waist-to-hip ratio adjusted for BMI), suggesting they are largely driven by immune mechanisms common to both diseases. These findings can help guide research towards understanding both conditions and identifying effective new treatment strategies for affected individuals.

93

Smoking Dependent and Independent Causal Effects of Educational Ascertainment and Alcohol Use on Lung Cancer

Rowland W. Pettit[1], Jinyoung Byun[1,2], Younghun Han[1,2], Quinn T. Ostrom[2], Rayjean J. Hung[3,4], James D. McKay[5], Christopher I. Amos[1,2,6]

[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]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [4]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Canada; [5]Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France; [6]Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America

To estimate the causal effects of alcohol use and educational ascertainment on lung cancer (LC) we employed univariate (MR) and multivariate two-stage mendelian randomization (MVMR) techniques utilizing genome wide association study summary statistic data from the United Kingdom Biobank (UKBB) and the TRICL-OncoArray LC consortium. We tested several UKBB education and alcohol related traits for causal effects on overall LC and its histological subtypes. We tested for trait – LC effect modulation via confounding or mediation from smoking related traits, including smoking "age of initiation," "cigarettes per day," and "smoking cessation." Using MR and MVMR effect estimation methods, including inverse variance weighting, weighted median, weighted mode, and mr-egger, we found that alcohol and educational ascertainment traits had direct causal effects on lung cancer development independent of smoking related traits. Using the IVW MR method, having no higher education or "Qualifications: None of the above" in the UKBB had 5.94 times increased odds of overall lung cancer (95% CI 3.39, 10.39, p = 4.49x10^{-10}, 62 SNPs), while "Qualifications: College or University degree" had a 0.31 decreased likelihood (95% CI 0.23, 0.42, p = 1.40x10^{-13}, 169 SNPs). Further, the trait "Average weekly beer plus cider intake" had an OR of 3.48 (95% CI 2.15, 5.64, p =

$4.08x10^{-7}$, 19 SNPs) with overall lung cancer risk. "Alcohol usually taken with meals" however had a OR 0.19 (95% CI 0.094, 0.36, $p = 1.06x10^{-6}$, 30 SNPs) with overall lung cancer risk. These trends were modified but remained despite MVMR contingent modeling with smoking related traits.

94

Unique TGF-β Signaling Pathway in African Americans with Fibrotic Sarcoidosis

Nathan Pezant, MS[*,1], Lori Garman, PhD[1], Richard C. Pelikan, PhD[1], Astrid Rasmussen, MD, PhD[1], Chuang Li, PhD[1], Benjamin A. Rybicki, PhD[2], Michael C. Iannuzzi, MD, MBA[3], and Courtney G. Montgomery, PhD[1]

[1]Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma United States of America; [2]Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan, United States of America; [3]Department of Internal Medicine, State University of New York, Upstate Medical University, Syracuse, New York United States of America

Sarcoidosis is a systemic inflammatory disease characterized by the formation of non-caseating granulomas in any number of organs, but with pulmonary manifestation in most patients. Progressive pulmonary disease (PPD), likely including pulmonary fibrosis (PF), leads to worse clinical outcomes and prognosis. Previous genome-wide association studies of PPD in sarcoidosis patients of European ancestry (EA) identified multiple associations with genes in the TGF-β signaling pathway, specifically TGFβ1, TGFβ2, TGFβ3, GREM1, and ANXA11. This is important as TGF-β is widely accepted as a master regulator of fibrosis. The involvement of TGF-β signaling has not previously been investigated in patients of African Ancestry (AA) nor has it been investigated beyond variant association.

In this study we aimed to identify genetic and genomic factors influencing fibrotic disease in a cohort of AA patients with PF compared to non-fibrotic sarcoidosis patients using whole genome sequencing and cell-type specific expression data. We identified suggestive associations between variants in the region of TGFβ3 as well as other TGF-β pathway genes PARS2, LTBP1, PVT1, SLC29A3, OTX2-AS1/EXOC5, and SPOP; all of which are primarily involved in TGF-β signaling and fibrosis or play a regulatory role.

Single-cell RNA sequencing data were obtained and cell-type specific expression quantitative trait loci (eQTL) are being identified to better understand the role risk variants play in the mechanism of fibrotic development. Our findings support both the role of TGF-β signaling in the development of PF as well as differences in the dysregulation of TGF-β signaling by ancestry.

95

Autism Spectrum Disorder Genes in Reading Disabilities: A Hypothesis-Driven Genome Wide Association Study

Kaitlyn M. Price[1,2,3*], Karen G. Wigg[1], Yu Feng[1], Kirsten Blokland[2], Elizabeth N. Kerr[5,6], Sharon L. Guger[5], Maureen W. Lovett[2,6], Lisa J. Strug[4,7], GenLang Consortium[8], Cathy L. Barr[1,2,3]

[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]Genetics and Genome Biology, Hospital for Sick Children, Toronto, Ontario, Canada; [5]Department of Psychology, Hospital for Sick Children, Toronto, Ontario, Canada; [6]Department of Pediatrics, University of Toronto, 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]Max Planck Institute for Psycholinguistics, Nijmegen, Niederlande

Reading Disabilities (RD) is a heritable disorder characterized by difficulties reading print. The leading etiological hypothesis is that disrupted neuronal migration alters connectivity in the developing brain affecting reading. However, the molecular mechanisms are not fully understood, nor are the genes that influence them. To identify genes, we previously conducted genome-wide association studies (GWAS). The analysis did not reach significance; however, we did observe overlap with genes implicated in other neurodevelopmental disorders, including Autism Spectrum Disorder (ASD), and neuronal migration genes. To increase power to identify genes, we now present three Hypothesis-Driven-(HD)-GWAS. HD-GWAS up-weights variants based on prior specified genetic or biological hypotheses. We up-weighted variants implicated in other genetic studies for RD, ASD, and neuronal migration. The analyses were conducted on two samples measured for reading: 1) a family-based RD cohort from Toronto and 2) a large international meta-analysis. For the Toronto sample, no SNPs were associated with reading using the three hypotheses; however, by gene-set analysis, we identified that the joint contribution of ASD genes significantly contributed. For the meta-analysis, the lead SNP, previously identified by conventional GWAS, was significantly associated with reading in each of the hypotheses. This SNP was not up-weighted in any of the analyses and confirms the robust association. These results suggest that in the

clinical sample (Toronto), ASD risk genes are involved in reading. This finding did not replicate in the meta-analysis due to population-based participants. Future studies involving large clinical samples will elucidate the shared underpinnings between RD-ASD.

96

Accounting for Population Structure and Distant Relatedness with Genealogical Data in a French-Canadian Study of Eye Disease and Cognitive Phenotypes

Mohan Rakesh[1*], Kelly Burkett[2], Ellen E. Freeman[1], Marie-Hélène Roy-Gagnon[1]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada; [2]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

Population structure and cryptic relatedness can be accounted for in genetic association studies by using linear mixed models and relatedness estimated from genome-wide genotype data. In founder populations, expected close and distant relatedness can also be estimated from deep genealogical data when available. In this study, we investigated the associations of 33 candidate SNPs with eye disease and cognitive phenotypes in a hospital-based cross-sectional study conducted at the ophthalmology clinics of Maisonneuve-Rosemont Hospital in Montreal, Canada. We retained the 227 participants with deep genealogical data available. Participants were recruited into one of three groups: late-stage age-related macular degeneration (n=67), glaucoma (n=65), or normal vision (n=95). Data on six cognitive tests were obtained. We used the GENLIB R package to analyze the 17-generation genealogical data and obtained estimates of kinship coefficients that we incorporated into generalized linear mixed models implemented in the lme4qtl and GMMAT R packages. For comparison, we used principal components analysis on the proportions of regions of origin of genealogical ancestors to account for structure. Genealogies were 75% complete at the 10th generation. Median kinship among participants was 0.0002, ranging from 0 to 0.016. P-values were significantly changed when using mixed models but not as much when using principal components adjustment. Simulations are needed to compare results using genealogical and genomic sources of relatedness information. When available, deep genealogical data could provide complementary information on relatedness that could be incorporated into analytical models in conjunction with genomic data.

97

Prediction of Coronary Artery Disease using Traditional and Genetic Risk Scores for Cardiovascular Risk Factors

Julia Ramírez[1*], Stefan van Duijvenboden[2], William J. Young[1], Andrew Tinker[1], Pier D. Lambiase[2], Michele Orini[2], Patricia B. Munroe[1]

[1]Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [2]Institute of Cardiovascular Science, University College London, London, United Kingdom

Background: Coronary artery disease (CAD) is a leading cause of death in the general population, but risk stratification remains suboptimal. Genetic risk scores (GRSs) for CAD predict risk independently from traditional risk factors. We assessed if there was improvement in risk stratification when including GRSs for multiple cardiovascular traits.

Methods: We used data from 379,581 European participants in the UK Biobank without known cardiovascular conditions (median follow-up 11.5 years, 2.9% CAD cases). In a training subset (50%), we built four scores using risk factors associated with CAD in Multivariable Cox analyses. Score s1 included sex and age, s2 included s1 and traditional risk factors, s3 included s2 and a GRS for CAD and s4 included s3 and multiple GRSs for cardiovascular traits. In an independent test subset (50%), we evaluated their performance using the area under the curve (AUC), hazard ratios (HRs) and net reclassification index (NRI).

Results: Score s4 had a higher AUC than score s3 (0.753 versus 0.747). The HR (95% confidence interval) for individuals in the top versus bottom 20% of the s4 distribution was 22.1 (18.7 – 26.2), versus 20.2 (17.1 – 23.8) for s3. The overall mean NRI for s4 versus s3 was 1.8%. Score s4 reclassifies 1,757 individuals as ≥10% CAD risk, where 168 would have a CAD event within the follow-up period.

Conclusions: Adding GRSs for multiple cardiovascular traits to a score integrating traditional risk factors and GRS for CAD improves risk stratification, identifying individuals who may benefit the most from early primary prevention measures.

98

Evaluating Viral Etiology of Bladder Cancer Through Analysis of Common Driver Mutations

Nina Rao*[1], Gabriel J. Starrett[2], Michael Dean[1], Nuria Malats[3,4,5], Manolis Kogevinas[6], Debra Silverman[7], Lars Dyrskjøt[8], Christopher B. Buck[2], Stella Koutros[7], Ludmila Prokunina-Olsson[1]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [2]Laboratory of Cellular Oncology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [3]Centro Nacional de Investigaciones Oncológicas, Madrid, Spain; [4]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain; [5]Centro de Investigación Biomédica en Red Cáncer, Madrid, Spain; [6]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain; [7]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [8]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark

FGFR3 and *PIK3CA* are among the most commonly mutated genes in non-muscle invasive bladder cancer (NMIBC), with mutations detected in 65% and 25% of cases, respectively. We hypothesized these highly recurrent mutations could be caused by some common risk factors, such as smoking. We tested this hypothesis in a combined set of ~900 NMIBC patients with FGFR3- or PIK3CA-mutated tumors and available smoking status. *FGFR3*-Y375C, an ERCC2-type mutation, was more common in smokers (*P*-value=0.038), and APOBEC-type driver mutations were enriched in non-smokers: *FGFR3*-S249C (*P*-value=0.019) and *PIK3CA*-E542K/*PIK3CA*-E545K (*P*-value=0.072). To identify risk factors specific for APOBEC-type driver mutations, we analyzed RNA-seq data from ~300 NMIBC tumors with *FGFR3* and *PIK3CA* mutations. We detected the presence of four viruses, most notably a polyomavirus (PyV): 92.9% (13/14) of PyV-positive tumors in this set had APOBEC-type *FGFR3* or *PIK3CA* mutations compared to 67.2% (146/217) of tumors without any detectable viral transcripts. Immunohistochemical (IHC) staining for PyV in another set of NMIBC tumors with *FGFR3* or *PIK3CA* mutations, identified 4.2% (5/119) PyV-positive tumors, all 5 tumors with APOBEC-type mutations in these two genes. In conclusion, our results support two distinct mechanisms in NMIBC, with smoking inducing ERCC2-type mutations and viral pathogens potentially generating APOBEC-type mutations.

99

Re-analysis of a Genome-Wide Gene-By-Environment Interaction Study of Case Parent Trios, Adjusted for Population Stratification

Pulindu Ratnasekera*, Brad McNeney
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

We investigate the impact of confounding on the results of a genome-wide association analysis by Beaty et al., which identified multiple single nucleotide polymorphisms that appeared to modify the effect of maternal smoking, alcohol consumption, or multivitamin supplementation on risk of cleft palate. The study sample of case-parent trios was primarily of European and East Asian ancestry, and the distribution of all three exposures differed by ancestral group. Such differences raise the possibility that confounders, rather than the exposures, are the risk modifiers and hence that the inference of gene-environment (G×E) interaction may be spurious. Our analyses generally confirmed the result of Beaty et al. and suggest the interaction G×E is driven by the European trios, whereas the East Asian trios were less informative. We conclude with a discussion of ongoing work on alternative methods for mitigating the impact of confounding when there are subtle differences in genetic and exposure distributions.

100

Polygenic Risk Scores – Is There a Different Distribution Within Germany and Therefore a Need of More Accurate Determination?

Tanja K. Rausch[1,2]*, Inke R. König[1], Wolfgang Göpel[2], for the German Neonatal Network (GNN)
[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]Klinik für Kinder- und Jugendmedizin, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Polygenic risk scores for complex diseases are widely used in preclinical and clinical research to stratify individuals according to their genetic risk for targeted prevention, therapy, or prognosis. However, they are usually derived and validated within a specific ethnic background, and translation into other ethnicities has been shown to be problematic. Furthermore, even the transfer between populations in the same country can be challenging, as shown, for instance, for Finland and Great Britain.

According to former studies, at least slight genetic differences are present between different parts of Germany. However, the implications for polygenic risk scores have not been evaluated so far. Therefore, this study aims at investigating the impact of geographic regions within Germany on the distribution of polygenic risk scores for common complex diseases.

The German Neonatal Network examines the development of very low birth weight infants with 64 study centers spread across Germany. Umbilical cord tissue frozen after birth is used to genotype the DNA of the infants. Affymetrix Axiom™ Genome-Wide CEU 1

Array Plate 2.0 and Illumina Infinium® Global Screening Array-24 v1.0/v2.0 were used for chip genotyping.

The continuously growing database currently contains genetic data of 10,259 very low birth weight infants. Within this database, we construct polygenic risk scores for common complex diseases, based on the GWAS and PGS Catalog, and compare their distributions between various areas within Germany. Results will provide insight into the transferability of polygenic risk scores between populations but also into the genetic architecture of the investigated traits.

101

Polygenic Risk Score: An Application to the Prediction of Asthma Risk

Jasmin Ricard*[1], Zhonglin Li[1], Sébastien Thériault[1,2], Yohan Bossé[1,3], Aida Eslami[1,4]
[1]Institut universitaire de cardiologie et de pneumologie de Québec, Laval University, Quebec City, Canada; [2]Department of Molecular Biology, Medical Biochemistry and Pathology, Faculty of Medicine, Laval University, Quebec City, Canada; [3]Department of Molecular Medicine, Laval University, Quebec City, Canada; [4]Department of Social and Preventive Medicine, Faculty of Medicine, Laval University, Quebec City, Canada

Context: Asthma is a chronic respiratory disease that affects more than 300 million people worldwide. Genome-wide association studies have confirmed the contribution of several genetic variants as risk factors for asthma and have shown that its hereditary component is further explained by their cumulative effect. Polygenic Risk Scores (PRSs) are defined as a weighted sum of risk alleles and are used to estimate an individual's genetic predisposition for a given trait.

Aim: Compare the association with the presence of asthma of eight sets of PRSs resulting from different methods, across different subsets of variants.

Methodology: Summary statistics reported by the Trans-National Asthma Genetic Consortium meta-analysis (23 948 cases, 118 538 controls) are used to calculate the PRS of Caucasian participants in the UK Biobank cohort (56 176 cases, 352 255 controls). Five *P*-value thresholds set at 1, 5e-2, 5e-4, 5e-5 and 5e-8 are applied to use the methods on different subsets of variants. The eight used methods are Stacked clumping + thresholding (SCT), LDpred2-grid-sparse, LDpred2-grid-nosparse, Lassosum, LDpred2-auto, LDpred2-inf, EBPRS and PRS-CS-auto

Results: Using all variants gives the best results. The SCT method showed the best performance with an AUC of 0.60, which is an improvement of 7.96% over the crude PRS estimation's AUC.

Conclusion: SCT is the optimal method to develop a PRS for asthma in our study. The next step is to use these PRSs and several biomarkers in multivariate methods to classify patients according to asthma severity.

102

DYNAMITE: A Phylogenetic Tool for Identification of Dynamic Transmission Epicenters

Brittany Rife Magalis[1,*], Alberto Riva[2], Simone Marini[3], Marco Salemi[1], and Mattia Prosperi[3]
[1]Emerging Pathogens Institute and Department of Pathology, Immunology, and Laboratory Medicine, University of Florida, Gainesville, Florida, United States of America; [2]Institute for Bioinformatics Research, University of Florida, Gainesville, Florida, United States of America; [3]Department of Epidemiology, University of Florida, Gainesville, Florida, United States of America

Molecular data analysis is invaluable in understanding the overall behavior of a rapidly spreading virus population when epidemiological surveillance is problematic. It is also particularly beneficial in describing subgroups within the population, often identified as clades within a phylogenetic tree, that represent individuals connected via direct transmission or transmission via differing risk factors in viral spread. However, transmission patterns or viral dynamics within these smaller groups should not be expected to exhibit homogeneous behavior over time. As such, standard phylogenetic approaches that identify clusters based on summary statistics (e.g., median genetic distance over the clade) would not be expected to capture dynamic clusters of transmission. For this purpose, we have developed DYNAMITE (DYNAMic Identification of Transmission Epicenters), a cluster identification algorithm based on a branchwise (rather than traditional cladewise) search for cluster criteria, allowing partial clades to be recognized as chains of transmission linked individuals. Using simulated viral outbreaks with varying cluster types and dynamics, as well as a SARS-CoV-2 South African dataset including the variant of concern B.1.351, we show that DYNAMITE is consistently more sensitive than existing tools in detecting both static and dynamic transmission clusters of epidemiological relevance. DYNAMITE has been implemented in R and released as open source at: github.com/ProsperiLab/DYNAMITE.

103

An Efficient and Robust Tool for Genetic Colocalization: Pair-wise Conditional and Colocalization (PWCoCo)

Jamie W Robinson [1]*, Gibran Hemani [1], Tom R Gaunt [1], Jie Zheng[1]

[1]*MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom*

Genetic colocalization has demonstrated value in strengthening causal inference in 'omics analyses and providing evidence for drug target prioritization. However, the single causal variant assumption is a considerable limitation and reduces the accuracy of the method. A novel pipeline is needed for the assumption to hold and to extend the scope for application to high-dimensional 'omics data.

We integrated two robust analyses, conditional analyses (GCTA-COJO) and colocalization analyses ("coloc" package), into a novel pipeline – Pair-Wise Conditional and Colocalization (PWCoCo). PWCoCo performs pair-wise conditional analyses to select each pair of conditionally independent signals for the exposure and outcome which allows for the stringent single-variant assumption to hold for the colocalization analysis.

Comparing PWCoCo with other colocalization analyses (such as coloc and eCAVIAR) in regions with only one causal variant suggested that PWCoCo had a high concordance rate with those methods. In regions with more than one causal variant, false-negative rates increased for those methods but not for PWCoCo. In these analyses, PWCoCo shows greatly increased efficiency (for complex regions with seven signals, analysis takes two minutes – 10x faster compared to other methods).

Preliminary comparisons show that PWCoCo produces robust results in genomic regions with simple or complex linkage disequilibrium structures. Key improvements include: (1) no violation of the single-variant assumption; (2) the pair-wise analyses allow for testing of colocalization between non-primary signals; (3) an easy-to-use and efficient tool to test for colocalization of high-dimensional 'omics data.

104

I'am hiQ – A Novel Pair of Accuracy Indices for Imputed Genotypes

Albert Rosenberger [1 *], Viola Tozzi [1], Marcus Baum [2], Kolja Thormann [2], Rayjean J. Hung [3], Christopher I. Amos [4] and Heike Bickeböller [1] on behalf of the INTEGRAL-ILCCO [5] consortium

[1] *Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Göttingen, Germany;* [2] *Institute for Computer Science, Data Fusion Lab, Georg-August-University of Göttingen, Germany;* [3] *Lunenfeld-Tanenbaum Research Institute, Sinai Health System, University of Toronto, Toronto, Ontario, Canada;* [4] *Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America;* [5] *Integrative Analysis of Lung Cancer Etiology and Risk program of the International Lung Cancer Consortium (INTEGRAL-ILCCO)*

Background: Imputation of untyped markers is a standard tool in genome-wide association studies to close the gap between directly genotyped and other known DNA variants. However, high accuracy with which genotypes are imputed is fundamental. Several accuracy measures have been proposed and some are implemented in imputation software, unfortunately diversely across platforms (e.g. *info* implemented in IMPUTE2*)*. We introduce *I'am* hiQ, an independent pair of accuracy measures that can be applied to dosage files, the output of all imputation software. *I'am* (*imputation accuracy measure*) quantifies the average amount of individual-specific versus population-specific genotype information in a linear manner. *hiQ (heterogeneity in quantities of dosages)* addresses the inter-individual heterogeneity between dosages of a marker across the sample at hand.

Results: Applying both measures to a large case-control sample of the International Lung Cancer Consortium (ILCCO), comprising 27,065 individuals, we found meaningful thresholds for *I'am* and *hiQ* suitable to classify markers of poor accuracy. We demonstrate how Manhattan-like plots and moving averages of *I'am* and *hiQ* can be useful to identify regions enriched with less accurate imputed markers, whereas these regions would be missed when applying the accuracy measure *info*.

Conclusion: We recommend using *I'am hiQ* additionally to other accuracy scores for variant filtering before stepping into the analysis of imputed GWAS data.

Computation: A standalone executable to calculate *I'am* hiQ from several large dosage files is provided.

105

Whole Genome Sequencing of Coronary Heart Disease in a Middle Eastern Cohort Validates Polygenic Risk Scores, Replicates Known Loci, and Suggests New Loci

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

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

Background: Excitement surrounding the potential use of polygenic risk scores (PRSs) in clinical practice is tempered by concern about their portability among diverse populations. To prevent disparities in genomic medicine, there is an urgent need to conduct genome-wide association studies in non-European ancestry cohorts.

Methods: We conducted whole genome sequencing (WGS) with 30x coverage on coronary heart disease (CHD) cases (1,067) and controls (6,170) in a Middle Eastern cohort to compare the performance of available PRSs for CHD (LDpred, metaGRS, lassosum, and P+T) and identify common variants associated with CHD (via generalized linear mixed models).

Results: Besides lassosum, all PRSs performed well. LDpred and metaGRS performed similarly (AUC= ~0.685) and outperformed P+T (AUC=0.667). Based on the OR per 1 SD increase (OR_{1sd}), P+T (OR_{1sd}=1.85 [1.69–2.02], P=3.69x10^{-41}) outperformed all other PRSs (OR_{1sd}=1.61 [1.48–1.74], P=3.02x10^{-31} for LDpred and OR_{1sd}=1.61 [1.49–1.75], P=9.47x10^{-31} for metaGRS). After binning PRSs into 10 deciles, the odds of CHD in the top decile compared to all others was highest for metaGRS (3.87 [3.07–4.88]) and LDpred (3.45 [2.74–4.341]). Thirty-two known GWAS loci (e.g., *ABCG8*, *CELSR2*, and *SLC22A4*) were replicated in our analysis with P<0.05. Seven suggestive new loci/genes (P<10^{-6}) with relevant biological function were identified (e.g., *CORO7*, *RBM47*, and *PDE4D*). The well-established 9p21 locus was not replicated.

Conclusions: Genome-wide PRSs derived from European ancestry cohorts performed well in a Middle Eastern cohort. Further studies are needed to develop and validate an ancestry specific PRS and to confirm the suggestive loci/genes.

106

APOL1 and Biobanking in the West African Terrain – Challenges and Successes

Emmanuella L. Salia[1*], Robert L. Davis[2], Joyce L. Browne[3], Emmanuel K. Srofenyoh[4], Claire L. Simpson[1]
[1]*Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America;* [2]*Centre for Biomedical Informatics, Department of Pediatrics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America;* [3]*Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands;* [4]*Greater Accra Regional Hospital, Accra, Ghana*

Preeclampsia is a pregnancy related condition that affects 4% of pregnancies in the United States of America (USA) and has a higher morbidity and mortality in many low- and middle-income countries (LMICs). As a multi-systemic syndrome, it affects the kidneys and many other organs. Variants in APOL1, which confer resistance to Human Acquired Trypanosomiasis (HAT), have been implicated in kidney disease and more recently, preeclampsia. APOL1 variants are specific to people of African descent and are being extensively investigated as key predictors of preeclampsia. High prevalence of these variants in sub-Saharan African populations demands further investigation, and numerous studies are ongoing in Africa to explore the effects of APOL1 variants. SPOT-BIO is part of the Severe Preeclampsia Adverse Outcome Triage (SPOT) study, an ongoing international research collaboration with multiple sites in Ghana. We collect biological samples from mother/baby dyads to determine effects of APOL1 genotypes on risk of preeclampsia. With only limited research infrastructure existing in LMICs, we will describe challenges in setting up biobanks geared towards improving the quality of maternal and child health in Africa. We will delve into the feasibility of implementing biobanks in a West African country and explore what LMICs in the subregion need to focus on to build robust, sustainable biobanks to develop capacity for cutting-edge genetics research. We will also show that despite these challenges, we have already recruited over 450 women, making it the largest cohort of African women with early preeclampsia in the world.

107

Identification and Characterization of Pleiotropic Loci for Obesity and Inflammation

Yiqian Wenren[1] and Yasmmyn D. Salinas[1*]
[1] *Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, United States of America*

Family- and population-based genetic studies suggest that obesity and inflammation have a shared genetic component. Meanwhile, epidemiologic studies suggest that obesity leads to the development of inflammation. Our study aimed to identify variants that display pleiotropy for obesity and inflammation [respectively measured by body mass index (BMI) and C-reactive protein (CRP)]. Given the causal link between phenotypes, we also aimed to determine whether associations with CRP were mediated through obesity-related pathways. To these ends, we conducted genome-wide association analyses of BMI and CRP within the UK Biobank (n=291,396 unrelated Caucasian subjects) using linear mixed models in fastGWA and searched for cross-phenotype associations (variants with P-value<5x10^{-9} for both phenotypes). We then fine

mapped the associated regions using DAP-G to identify shared putative causal variants (those with posterior inclusion probability > 0.95 for each trait). Lastly, we decomposed the total genetic effects of shared putative causal variants using causal mediation analyses. Mediation models adjusted for multiple confounders of the BMI-CRP relationship. We identified 16 variants (rs79113395, rs58048722, rs199956414, rs12203592, rs2049870, rs2721966, rs10086741, rs179444, rs3808478, rs6265, rs7926362, rs4922793, rs4755725, rs12577464, rs8047395 and rs11075987), residing in eight distinct genetic loci, that display pleiotropy for BMI and CRP. Among these, four variants (rs58048722, rs199956414, rs4755725, rs12577464) affected CRP only through obesity-related pathways. All other SNPs had both direct effects and indirect effects on CRP. The identified variants advance our mechanistic understanding of the pathogenesis of obesity- and inflammation-related diseases and may serve as targets for therapies that simultaneously treat these conditions.

108

Genomic Approaches to Identify Shared Genetic Architecture Among Comorbid Phenomes of Eye Disease

Alexandra Scalici[1,2,*], Tyne W. Miller-Fleming[1,2], Dan Zhou[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

Genome-wide association studies have identified a significant number of SNPs associated with human disease. However, many of these associations provide little insight into the underlying biological mechanisms or pathways involved in disease pathogenesis. To better understand the potential shared underlying mechanisms of eye disease, we used both a gene-based approach to assess shared genetic architecture and a phenome-based approach to identify comorbidities and expand the nascent pathways associated with eye diseases. To assess shared genetic architecture among eye diseases, we used transcriptome-wide association study (TWAS) to test if imputed genetically regulated expression (GReX) in genotyped individuals of European ancestry (N=70,049) within BioVU is associated with eye disease status. We identified two genes (*GPX7 & AC016590.3*) that had significant associations to eye disease status. To identify the comorbid phenomes of eye disease, we conducted a phenome-wide association study (PheWAS) within non-genotyped subjects with at least three visits to VUMC in five years (N=685,820). Using phenomes significantly associated with eye disease case status, we calculated a phenotypic risk score (PheRS) and applied this score to an independent population – genotyped subjects in BioVU. This PheRS based only on non-eye disease comorbid phenome is predictive of eye disease status. The PheRS was significantly associated with the predicted expression of six genes (*BBS5, NEU2, C4B, PAK1, RPL41, AC091100.1*), previously characterized as being involved in eye diseases. Functional analysis of these genes using zebrafish as a model system has the potential to shed light upon some of the common pathways and mechanisms of eye disease.

109

Penalized Mediation Models for Multivariate Data

Daniel J. Schaid, PhD*[1], Ozan Dikilitas, MD[2], Jason P. Sinnwell, MS[1], Iftikhar Kullo, MD[2]
Department of Quantitative Health Sciences[1] and Department of Cardiovascular Medicine[2], Mayo Clinic, Rochester, Minnesota, United States of America

Statistical methods to integrate multiple layers of data, from exposures to intermediate traits to outcome variables, are needed to guide interpretation of complex data sets about which variables are likely contributing in a causal pathway from exposure to outcome. Statistical mediation analysis based on structural equation models provide a general modeling framework, yet they can be difficult to apply to high-dimensional data and they are not automated to select the best fitting model. To overcome these limitations, we developed novel algorithms and software to simultaneously evaluate multiple exposure variables, multiple intermediate traits, and multiple outcome variables. Our penalized mediation models are computationally efficient and simulations demonstrate that they produce reliable results for large data sets. Application of our methods to a study of vascular disease demonstrate their utility to identify novel direct effects of single nucleotide polymorphisms (SNPs) on coronary heart disease and peripheral artery disease, while disentangling the effects of SNPs on the intermediate risk factors including lipids, cigarette smoking, systolic blood pressure, and type2 diabetes.

110

Longitudinal Microbiome and Machine Learning: A CNN-LSTM based Neural Network Model for Disease Prediction

Divya Sharma[1*], Wei Xu[1,2]
[1]Princess Margaret Cancer Center, UHN, Toronto, Ontario; [2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Research shows that human microbiome is highly dynamic along time, changing with diet or medical interventions, enabling discovery of short and long-term trends during disease prediction. We propose a novel deep learning framework combining Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) for feature extraction and analysis of temporal dependency in longitudinal microbiome data, along with the host's environmental factors for disease prediction. Key points of novelty are: (1) two-step approach consisting of firstly, pre-processing the microbiome data to capture correlation between the operational taxonomic units (OTUs) for efficient feature extraction through CNN; secondly, forwarding the extracted features to the LSTM to learn dependency along time, (2) handling unbalanced timepoints and missing data in repeated measures through padding-masking operation in LSTM without requiring explicit data imputation, (3) handling case-control imbalance in the LSTM with weighted loss function, mitigating biased network learning when the ratio of cases to controls is highly-skewed. We have evaluated the model's effectiveness using simulation studies across multiple time points and implemented into two real longitudinal human microbiome studies: (i) DIABIMMUNE three-country cohort with food allergy outcomes (785 samples; 534 OTUs), (ii) DiGiulio study with preterm delivery outcome (3767 samples; 1420 OTUs). Extensive comparison of our proposed model to conventional machine learning methods provided encouraging performance with an AUC of 0.897 (increased by 5%) on simulated studies and AUCs of 0.762 (increased by 19%) and 0.713 (increased by 8%) on the two real longitudinal microbiome studies respectively, in comparison to the next best performing method.

111

Multiethnic Joint Analysis of Marginal Summary Statistics from Genome-wide Association Studies

Jiayi Shen[1*], Lai Jiang[1], Kan Wang[1], Paul J. Newcombe[2], Chris Haiman[3,4], David V. Conti[1, 3, 4]

[1] Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California United States of America; [2] MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [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

Over the last 20 years, Genome-wide Association Studies (GWAS) have been able to identify genetic regions associated with traits and diseases in different populations. As a post-GWAS approach, multiethnic fine-mapping aims to identify underlying causal variants by combining summary statistics and leveraging different linkage disequilibrium (LD) structures across diverse populations. Here, we expand upon our previous approach for single-population fine-mapping through Joint Analysis of Marginal SNP Effects (JAM) to a multiethnic analysis (mJAM). Under the assumption that true causal variants are common across populations, our joint model explicitly incorporates the different LD structures across populations and yields a conditional fixed-effect meta-analysis (FE). To pinpoint causal variants from highly correlated signals efficiently, we incorporate Sum of Single Effects (SuSiE), a Bayesian stepwise selection method, within the mJAM framework. Through simulation studies based on realistic effect sizes and levels of LD, we demonstrate that mJAM performs better than other existing multi-ethnic methods including FE, conditional and joint analysis using summary data (COJO), and multiple study causal variants identification in associated regions (MsCAVIAR). The flexible mJAM framework can be extended to deal with binary disease outcome or missing SNP information in some populations, which is a unique advantage of mJAM over other existing methods. In a real data application, we apply mJAM to recently published summary statistics from a trans-ancestry prostate cancer GWAS.

112

Meta-analysis of dbGaP Data Reveals Population Structure, Admixture, and Known and Cryptic Relatedness across the United States of America

Daniel Shriner, Adebowale Adeyemo, Amy R. Bentley, Charles N. Rotimi

Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Maryland, United States of America

We investigated population structure and relatedness in 23,643 African Americans from eight studies, 13,543 White Americans from three studies, and 5,035 Hispanic Americans from two studies, all genotyped using the same array. First, two-way admixture explained 1.2% of the variance in African Americans and three-way admixture explained 3.5% of the variance in Hispanic Americans. Population structure in White Americans reflected differences among Middle Eastern, Southern European, and Northern European ancestries, explaining 0.2% of the variance. Second, known and cryptic relatedness explained 16.8% of the variance in African Americans, 5.1% in Hispanic Americans, and 7.6% in White Americans. Reflecting these major sources of shared variance, the effective

number of individuals was 73% of the nominal number of individuals for Africans Americans, 86% for Hispanic Americans, and 77% for White Americans. All studies mostly comprised 1st to 6th degree relatives, with 8th degree relatives being effectively independent. The sum of per-study estimates of effective numbers of individuals was systematically larger than the joint estimate, consistent with cryptic relatedness between studies. Using local ancestry data, admixture explained 22.3% and relatedness explained 77.6% of the variance in African Americans. Similarly, admixture explained from 27.6% to 38.1% and relatedness explained from 60.1% to 71.0% of the variance in Hispanic Americans. Consequently, the effective number of individuals was 25.0% of the nominal number of individuals for African Americans and 36.9% for Hispanic Americans. These results lead to concerns of overestimation of significance and power, even more so for admixture mapping than for association testing.

113

A Statistical Framework to Decipher the Genetic Architecture of Combinations of Complex Diseases: Applications to Cardiometabolic Disorders

Liangying Yin[1], Carlos Kwan-long Chau[1], Yu-Ping Lin[1], Shitao Rao[1,9], Yong Xiang[1], Pak-Chung Sham[8], Hon-Cheong So[1-7]*

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong; [2]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China; [3] Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong; [4] CUHK Shenzhen Research Institute, Shenzhen, China; [5]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong; [6]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China; [7]Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong, Hong Kong SAR, China; [8] Department of Psychiatry, University of Hong Kong, Hong Kong; [9] Department of Bioinformatics, Fujian Medical University, Fuzhou, China

**Presenting author*

At present, most genome-wide association studies (GWAS) are studies of a particular single disease diagnosis against controls. However, in practice, an individual is often affected by more than one condition. For example, patients with coronary artery disease (CAD) are often comorbid with diabetes mellitus (DM). Similarly, it is often clinically meaningful to study patients with one disease but without a related comorbidity. For example, obese DM may have different pathophysiology from non-obese DM.

Here we developed a statistical framework to uncover susceptibility variants for comorbid disorders (or a disorder without comorbidity), using GWAS summary statistics only. In essence, we mimicked a case-control GWAS in which the cases are affected with comorbidities or a disease without a relevant comorbidity we may consider the cases as those affected by a specific disease 'subtype', as characterized by the presence or absence of comorbid conditions). We extended our methodology to deal with continuous traits with clinically meaningful categories (e.g., lipids). We also illustrated how the framework may be extended to more than two traits. We verified the feasibility and validity of our method by applying it to simulated scenarios and four cardiometabolic (CM) traits (obesity, DM, CAD and stroke). We also analyzed the genes, pathways, cell-types/tissues involved in CM disease subtypes. Genetic correlation analysis revealed that some subtypes may be biologically distinct from others. Further Mendelian randomization analysis found differential causal effects of different subtypes to relevant complications. The proposed method may open a new avenue to analyzing GWAS data.

114

Testing and Estimation of X-chromosome SNP Effects: Impact of Model Assumptions

Yilin Song[1]*, Joanna M. Biernacka[2], Stacey J. Winham[2]

[1]University of Washington, Seattle, Washington, United States of America; [2]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America

Interest in analyzing X chromosome SNPs is growing and several approaches have been proposed. Prior studies have compared the power of different approaches, but bias and interpretation of coefficients have received less attention.

We performed simulations to demonstrate the impact of X chromosome model assumptions on effect estimates. We investigated the coefficient biases of SNP and sex effects with commonly used models for X chromosome SNPs, including models with and without assumptions of X chromosome inactivation (XCI), and with and without SNP-sex interaction terms. Additional scenarios were assessed assuming non-random XCI (the presence of XCI skewness towards one parental chromosome), equivalent to dominance deviation from the additive genetic model. Sex and SNP coefficient biases were observed when assumptions made about XCI and sex differences in SNP effect in the analysis model were inconsistent with the data-generating

model. However, including a SNP-sex interaction term often eliminated these biases. To illustrate these findings, estimates under different genetic model assumptions are compared and interpreted in a real data example.

Models to analyze X chromosome SNPs make assumptions beyond those made in autosomal variant analysis. Assumptions made about X chromosome SNP effects should be stated clearly when reporting and interpreting X chromosome associations. Fitting models with SNP*sex interaction terms can avoid reliance on assumptions, eliminating coefficient bias even in the absence of sex differences in SNP effect.

115

The Causal Relationships Between Serum Metabolome and Systemic Lupus Erythematosus: A Two-Sample Mendelian Randomization Study

Jun-Seop Song[1,2*]

[1]Division of Cardiology, Yonsei University College of Medicine, Republic of Korea; [2]Geumsan Public Health Center, Republic of Korea

The causal effects of metabolic profiles on systemic lupus erythematosus (SLE) are poorly understood. We aim to examine whether serum metabolite levels are causally associated with the risk of SLE. We performed two-sample Mendelian randomization (MR) analyses using the inverse variance weighted (IVW) and MR-Egger regression methods on publicly available genome-wide association studies (GWAS) summary statistics datasets on 123 unbiased serum metabolites (24,925 individuals of European ancestry) as exposure and SLE (7,219 cases and 15,991 controls of European ancestry) as outcomes. The IVW method yielded that four VLDL derivatives (XXL.VLDL.L, XXL.VLDL.P, XXL.VLDL.TG, and XL.VLDL.P), four fatty acid derivatives (FAw79S, MUFA, FAw3, and otPUFA), and valine causally increased the risk of SLE; however, six HDL derivatives (M.HDL.CE, M.HDL.L, M.HDL.P, M.HDL.C, M.HDL.FC, and M.HDL.PL), glycine, and glucose decreased the SLE risk. The MR-Egger regression method revealed that only M.HDL. CE was causally associated with SLE (OR=0.612, 95% CI 0.432–0.867, P=0.006), which is unlikely to be biased by directional pleiotropy (intercept=0.088, P=0.171). In addition, genes/proteins related to the HDL metabolic pathway are significantly proximal to the SLE disease module on the protein-protein interactome (P<0.001). The results imply the potential therapeutic strategies of SLE by targeting HDL metabolic pathways, which might contribute to restore the abnormal metabolic environment for T cells.

116

Estimating and Visualizing Multivariable Mendelian Randomization Analyses Within a Radial Framework

Wes Spiller[1*], Eleanor Sanderson[1], Jack Bowden[2]

[1] Population Health Sciences, University of Bristol, Bristol, United Kingdom; [2] University of Exeter Medical School, Exeter, United Kingdom

Background: Multivariable Mendelian randomization (MVMR) is a statistical approach using genetic variants to estimate causal associations between multiple exposures and an outcome simultaneously. In univariable MR findings are typically illustrated using scatter or radial plots created using summary data from genome-wide association studies, however, analogous plots for MVMR analyses have so far been unavailable.

Methods: We propose a radial formulation of MVMR, and an adapted Galbraith radial plot, which allow for the direct effects of each exposure within an MVMR analysis to be visualised. Radial MVMR plots facilitate the detection of outlier variants, indicating a violation of one or more assumptions of MVMR. The RMVMR R package is also provided as accompanying software for implementation of the methods described.

Results: We demonstrate the effectiveness of the radial MVMR approach through simulation and applied analyses considering the effect of lipid fractions upon coronary heart disease (CHD). We find evidence of a protective effect of high-density lipoprotein (HDL) and a positive association between low-density lipoprotein (LDL) and CHD, however, the protective effect of HDL appeared to be smaller in magnitude when removing potentially pleiotropic genetic variants. In combination with simulated examples, we highlight how important features of MVMR analyses can be explored using a range of tools incorporated within the RMVMR R package.

Conclusions: Radial MVMR provides a means of effectively visualising causal effect estimates, and can provide valuable diagnostic information with respect to the underlying assumptions of MVMR.

117

Genotype-Phenotype Analysis in African Americans with Inflammatory Bowel Disease

Andrew B. Stiemke[1*], Lisa W. Datta[2], Mark G. Lazarev[2], Inflammatory Bowel Disease Genetics Consortium, Dermot P.B. McGovern[3], Steven R. Brant[2,4], Claire L. Simpson[1]

[1]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America; [2]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine and Department of

Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America; [3]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars Sinai Medical Center, Los Angeles, California, United States of America; [4]Department of Medicine, Department of Genetics and the Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick and Piscataway, New Jersey, United States of America

Inflammatory bowel disease (IBD) is an immune-mediated chronic intestinal disorder that is typically divided into two distinct types; ulcerative colitis (UC) and Crohn's disease (CD). Although each respective disease is classified based on clinical presentation, each disease's presentation is highly heterogenous. Here we aimed to better discern the heterogeneity within disease types by characterizing the many endophenotypes within UC and CD rather than their initial classification.

A total of 1135 African American patients, 772 with CD, 323 with UC, and 41 with inflammatory bowel disease unclassified (IBDU), were included in this study. Genomic DNA data on the HumanOmni2.5 microarray was obtained from all patients and mapped to GRCh37/hg19. Extensive phenotyping was conducted on all patients, which allowed for the creation of 14 direct and derived endophenotypes, including perianal disease, smoking status, and age of onset. Analysis was restricted to the area ±500kb around 15 previously identified candidate single nucleotide polymorphisms previously documented in IBD GWAS. Statistically significant results have been found for a number of stratified endophenotype analyses including disease severity and disease behavior. One such result for this stratification (*P*-value 2.64e-6, OR 0.21) is approximately 380kb proximal to NOD2 in ADCY7. Additional analyses are ongoing.

Utilizing endophenotypes to differentiate various disease presentations will allow for additional genetic markers to be discovered. As additional genetic risk loci are identified, more robust screening will improve patient outcomes and overall quality-of-life. If the disease can be caught in its early stages, treatment options can be taken to ensure the best patient outcome.

118

A Novel Regression-based Method for X-chromosome-inclusive Hardy-Weinberg Equilibrium Test

Lin Zhang[1], Zhong Wang[2], Andrew D. Paterson[3,4], Lei Sun[1,4]*

[1]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada; [2]School of Data Science, Fudan University, Shanghai, China; [3]Program in Genetics & Genome Biology, The Hospital for Sick Children; [4]Dalla School of Public Health, University of Toronto, Toronto Ontario, Canada

How to best perform Hardy-Weinberg equilibrium (HWE) test for an X-chromosomal SNP is not clear, even using a sample of unrelated individuals. One simple strategy is to use female data only and apply the Pearson's Chisq test. Alternatively, earlier work has proposed a 2 *df* test that includes the deviation of male genotype counts from the expected based on pooled allele frequency estimate using both male and female data.

Instead of the Pearson's Chisq test, we propose a new regression-based method that (a) analyzes both autosomal and X-chromosomal SNPs, (b) adjusts for covariate effects if needed, (c) analyzes related individuals, (d) includes the existing tests as special cases, and (e) leads to development of new tests. The proposed method builds from our recent robust allele-based (RA) regression method developed for conducting allelic association analysis.

First, we show that a 2 *df* score test derived from the proposed RA regression includes the existing test as a special case. Second, we show that the existing 2 *df* test can be reformulated as simultaneously testing sex differences in allele frequency, and HWE in female group alone. Thus, we can develop new HWE tests that do not assume that sex differences in allele frequency are due to genotyping error, suitable for analyzing variants subject to sex-specific selection. Finally, the proposed method can analyze samples from multiple populations jointly.

We illustrate the method by application to both phase 3 and high coverage whole genome sequence data from the 1000 genomes project.

119

Association Analysis of Mitochondrial DNA Heteroplasmic Mutations in Deep Sequencing Data

Xianbang Sun[1], Chunyu Liu[1]

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

Maternally inherited mitochondrial DNA (mtDNA) exists in multiple copies per cell, resulting in heteroplasmy, a phenomenon with two or more alleles presented at a mtDNA locus within an individual. Most heteroplasmic mutations are rare in humans and display low alternative allele fractions per individual. Therefore, the methods that analyze rare genetic variants are not readily applicable to association analyses of heteroplasmy. We propose a statistical

framework that flexibly handles a selected threshold to identify heteroplasmic mutations and incorporates a user-specified weight in association analyses of heteroplasmy using a burden and sequence kernel association test (SKAT). We employ an aggregated Cauchy association test (ACAT) and permutation test to combine information from multiple tests. We evaluate the performance of the methods with twelve scenarios using a simulated data. The type I error rate is well controlled at α=0.001 in all models. A SKAT outperforms a burden test when ≤ 25% mutations are causal or around 50% of causal mutations have opposite effects. A burden outperforms a SKAT, otherwise. The ACAT and permutation tests robustly combine individual test results. We applied the proposed framework in association analyses of heteroplasmy with age and sex. The most significant association is found with the *RNR2* gene: one unit increase in heteroplasmy burden in this gene is associated with a 0.75-year older age (P=1.2×10^{-5}). Sex is not associated with the aggregation of heteroplasmy in any mtDNA genes. The proposed statistical framework will facilitate association testing of heteroplasmy with disease traits in large human population.

120

Genome-wide Association Study of Mild Cognitive Impairment in 1,040 Chinese Subjects

Rui Sun[1, 2†], Fan He[3†], Fudong Li[3†], Xue Gu[3], Tao Zhang[3], Yexian Zhang[2], Junfen Lin[3*], Maggie H. Wang[2*]
[1]The Seventh Affiliated Hospital of Sun Yat-Sen University, Shenzhen, China; [2]The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China; [3]Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China
[†]The authors contributed equally to this work.

Background: Mild cognitive impairment (MCI) occurs in 14.71% of individuals above age 60 in China and is considered as a pre-Alzheimer's disease dementia syndrome with approximately 33% of MCI patients progressing to dementia in later life. To gain insight into the genetic factors for MCI in Asian, we conducted a whole-genome wide association study (GWAS) of MCI in a Chinese population.

Subjects and Method: We built a cohort of 12,500 Elderly Chinese and sequenced 1,040 subjects (1,859,965 SNPs) among them. In stage one analysis, a linear mixed model was applied with covariates of age, gender, and education level for GWAS of MCI. In stage two analysis, we performed association testing on combined data of 801,493 SNPs and 1,620 subjects from three sources: in-house MCI data, 1000 Genome project of East Asia subjects, and in-house Alzheimer's disease data.

Results: Five regions were identified (P-value < 1 × 10^{-5}) in the stage one single cohort analysis, including *SNX18P22*, *GPR39*, *PCDH18*, *EDIL3*, and *SETBP1*. In the combined data, *PCDH17* (rs4886083, P-value = 6.40 × 10^{-8}, OR = 0.59) is identified, marginally reaches the genome-wide significance threshold. Interestingly, abundant evidence from biomedical studies suggest that the *PCDH* family is functionally associated with neurological diseases such as schizophrenia, bipolar disorder, and autism spectrum disorder.

Conclusion: We identified novel genome regions in the GWAS of MCI in a Chinese cohort. The susceptible markers are located near *SNX18P22, GPR39, PCDH18, EDIL3, SETBP1,* and *PCDH17*, suggesting a generally differential association profile compared to the European.

121

Population Differences in Genetic Risk of Disease Cannot be Detected at Current Sample Sizes

Iain R. Timmins[1*], Frank Dudbridge[1]
[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom

Many complex diseases have large disparities in incidence rates between populations, although little is known about the contribution of genetic factors to these observed differences. Recently, this question has been addressed by comparing the mean values of polygenic risk score (PRS) distributions across different populations, with significant mean differences taken as evidence of a genetic basis for the difference in incidence. However, while these analyses account for sampling variation in the data for which PRS are compared, they do not account for sampling variation in the training data used to derive the PRS.

We derive analytic expressions for the power to detect a difference in the mean true PRS between populations, accounting for variation in both training and target data, and support these results through empirical and simulation studies across a range of genetic architectures. Assuming that the genetic architectures differ only in the risk allele frequencies, the power depends on the training sample size, the number of causal variants and the F_{st} between the populations. We show that the power to detect differences between continental ancestral groups is barely above the type-1 error rate at current sample sizes, and that samples of several million are required to approach acceptable power.

We conclude that it is currently infeasible to use PRS to infer differences in true genetic risk between populations, and that such analyses will require samples that are orders of magnitude greater than those currently available.

122

Almost Exact Mendelian Randomization

Matthew J. Tudball[1,2,*], Qingyuan Zhao[3]

[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] Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, United Kingdom

Mendelian randomization (MR) is commonly understood as a study design that uses genetic variants as instrumental variables for modifiable exposures. However, it is typically only implicit in MR methodologies that the inferential basis of MR is the random allocation of alleles from parents to offspring via meiosis and mating. As parent-offspring data becomes more widely available, we advocate an approach to MR which is exactly based on this randomization.

Meiosis has been thoroughly studied and modelled in genetics dating back to Haldane (1919). We propose a statistical framework that enables meiosis models to be used as the "reasoned basis for inference" in MR. Specifically, we develop an approach to inference based on exact hypothesis testing, first described in Fisher (1935)'s original proposal for randomized experiments. We therefore make explicit the common analogy between MR and a randomized controlled trial. Furthermore, we develop a rigorous graphical framework for describing within-family MR, which is used to identify sufficient confounder adjustment sets. Our randomization-based inference also has several practical advantages. First, unlike existing within-family MR methods, it sidesteps the need for correctly specifying phenotype models, although a better model will often lead to more powerful tests. We demonstrate via simulation that propensity scores obtained from the underlying meiosis model can form powerful test statistics. Second, our approach is robust to arbitrarily weak instruments. Finally, by using our sufficient adjustment sets, it is provably robust to biases arising from population structure, assortative mating, dynastic effects and several forms of pleiotropy.

123

Risk for Hospitalization and Case-Fatality-Rate for Different Age Groups: The Alpha Variant of SARS-CoV-2: A Turkish Study

Ayse Ulgen[*]

Department of Biostatistics, Faculty of Medicine, Girne American University, Karmi, Cyprus

Knowing the risks for hospitalization and case-fatality-rate for emerging SARS-CoV-2 variants for different age groups is necessary for hospital management and vaccination planning. It is therefore important to collect and analyze health data related to the variants. We have obtained more than 3700 COVID-19 patients (3100 outpatients and 600 inpatients) where about 30% are infected by Alpha variant. Both logistic regression and cause-specific Cox survival analysis of competing-risk is run on inpatients to examine the impact of the Alpha variant on hospitalization and on mortality, conditional on other factors. Descriptive statistics is used to characterize different subgroups. We observed that the Alpha variant is over-represented in inpatients than outpatients so carrying the Alpha variant increases the chance for hospitalization. The impact of the Alpha variant on mortality seems to depend on the patient's age. For age < 70 group, the case-fatality-rate is 0.84% (5.3%) for patients without (with) the Alpha variant (Fisher's test P-value = 2.4 ×10−10). For age ≥ 70 group, the trend is opposite: the case-fatality-rate is 31.5% (13.6%) for patients without (with) Alpha variant (Fisher's test P-value = 0.0016). The two opposite trends would cancel each other, making other analyses such as cause-specific Cox regression and logistic regression non-significant. The Alpha variant increases the risk for hospitalization, increases the case-fatality-rate for lower age group, and decreases the case-fatality-rate for the upper age group. It is therefore imperative to vaccinate young, middle-aged, and early senior population to counter the impact of wave of the Alpha variant.

124

Deciphering How Early Life Adiposity Influences Breast Cancer Risk Using Mendelian Randomization

Marina Vabistsevits[1,2]*, George Davey Smith[1,2], Eleanor Sanderson[1,2], Tom G. Richardson[1,2,3], Bethan Lloyd-Lewis[4], Rebecca C. Richmond[1,2]

[1]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield House, Oakfield Grove, Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Novo Nordisk Research Centre, Headington, Oxford, United Kingdom; [4]School of Cellular and Molecular Medicine, University of Bristol, Biomedical Sciences Building, Bristol, United Kingdom

Studies suggest that adiposity in childhood may reduce the risk of breast cancer in later life. The biological mechanism underlying this effect is unclear but is likely to be independent of adult body size. Using a Mendelian randomization (MR) framework, we investigated 16 hypothesised mediators of the protective effect of childhood adiposity on later-life breast cancer, including hormonal, reproductive, physical, and glycaemic traits.

Using data from publicly available genome-wide association studies, we designed an MR workflow to assess the causal role of potential mediators. In two-step MR, we evaluated the effect of childhood body size on each trait and the effect of those traits on breast cancer risk. Then, we used multivariable MR to assess the independent effect of childhood adiposity on breast cancer accounting for each mediator. Finally, we used mediation analysis to characterise the indirect effect of body size via the mediators.

The results showed that although most of the reviewed mediators were affected by childhood body size, only IGF-1, testosterone, and ages at menarche and menopause influenced breast cancer risk. However, multivariable MR showed that the protective effect of childhood adiposity remained when accounted for those traits, suggesting a lack of evidence for mediation.

Our work presents a framework for the systematic exploration of potential mediators of disease in MR. We explored many plausible links between childhood adiposity and breast cancer, but none accounted for the protective effect observed. It is feasible that several traits collectively contribute to the mediated effect, or mediation occurs via other pathways.

125

Novel Analysis Pipeline for Microbiome Data Via Individual-Specific Networks

Federico Melograna[1], *Diane Duroux* [6], Gianluca Galazzo[2], Niels van Best[2,3], Monique Mommers[4], John Penders[2,5], Kristel Van Steen[1,6]*

[1]BIO3 - Laboratory for Systems Medicine, KU Leuven, Leuven, Belgium ; [2]School of Nutrition and Translational Research in Metabolism (NUTRIM), Department of Medical Microbiology, Maastricht University, Maastricht, The Netherlands ; [3]Institute of Medical Microbiology, RWTH University Hospital Aachen, RWTH University, Aachen, Germany; [4]Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands; [5]Care and Public Health Research Institute (CAPHRI), Department of Medical Microbiology, Maastricht University, Maastricht, The Netherlands; [6]BIO3 – Laboratory for Systems Genetics, GIGA-R Medical Genomics, University of Liège, Liège, Belgium

In individual-specific networks (ISNs), nodes or edges are individual-specific. These edges may be unweighted or weighted, indicating strengths of association or co-occurrence between nodes. Edge-oriented ISNs can be computed in standard ways when repeated measurements for the same individual are available (f.i. neuroscience applications). In the absence of repeated data, such networks can be computed by assessing the influence an individual has on the total sample based (global) network. In this work, we customize the ISN construction method of Kuijjer et al. (2019) to infer individual-specific microbiome networks. In particular, we used data from the Dutch LucKi project, involving microbiome profiling by 16S sequencing on 69 newborns with measurements available at months 6 and 9; milestones in microbiome maturation. For illustration, we used fastSPAR to develop one microbiome co-occurrence global network per timepoint. Subsequently derived ISNs were then used to assess between-individual heterogeneity: by computing network-simuliarity matrices (e.g., edge difference distance) and implementing an unsupervised hierarchical algorithm to identify latent classes of similar ISNs. An adapted distance-based ANOVA was developed, inspired by ecology, to determine the most optimal number of classes. The observed enterotypes complement earlier findings with Dirichlet Multinomial Mixtures clustering across timepoints. We futhermore illustrated how ISNs can be used to compare microbiome profile heterogeneity between timepoints and show the advantage of subnetwork selection to find novel OTUs linked to mode of delivery. Clearly, global models fail to capture population heterogeneity, unlike ISNs. Drawing conclusings for every individual specifically is believed to have direct consequences for Precision Medicine.

125

Epigenetic Age Prediction in Large-Scale Methylation Sequencing Project

Denitsa Vasileva[1*], Ming Wan[1], Allan B. Becker[2], Edmond S. Chan[3], Catherine Laprise[4], Andrew J. Sandford[1], Celia M. T. Greenwood[5], 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]Centre intersectoriel en santé durable (CISD) de l'Université du Québec á Chicoutimi, Saguenay, Canada, Centre intégré universitaire de santé et de services sociaux (CIUSSS) du Saguenay–Lac-Saint-Jean, Saguenay, Canada; [5]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada

The Horvath epigenetic age prediction algorithm consisting of 353 CpG sites, has demonstrated accuracy in array-based studies and adult samples. Accuracy/applicability to targeted sequencing and childhood samples has yet to been established.

The objective of this study was twofold:

1. To assess the accuracy of the Horvath clock in targeted methylation sequencing data of predominantly child samples and, if necessary, improve it.
2. To evaluate its applicability and utility as a quality control (QC) metric in targeted methylation sequencing experiments.

Study population includes 812 samples from The Canadian Asthma Primary Prevention (CAPPS, n=632 samples, 149 cord-blood, 158 year seven, 210 year 15 and 115 maternal samples) and the Saguenay-Lac-Saint-Jean studies (SLSJ, n=180 samples). Longitudinal samples include 89 children at three time-points and 99 at two. SLSJ consists of three-generational triads of French-Canadian descent. Sequencing was completed using Illumina (San Diego, California)'s MethylCapture EPIC library.

Quality control metrics (QC) included Principal Component Analysis (PCA) on ethnicity, age, cell composition, sex and Mendelian errors. Accuracy of the Horvath epigenetic clock was measured using Relative Difference (RD, RD= abs (predicted age (PA)-chronological age (CA, calendar years since birth))/CA). Horvath's epigenetic clock demonstrated utility as a QC metric. Samples with QC flags also had a RD > mean (RD) +2*SD. The average RD () decreased as the CA increased: cord blood (1.831.75), Age seven (0.860.97), Age 15 (0.440.35), CAPPS >18 (0.250.21) and SLSJ (0.320.37).

A novel age prediction algorithm, consisting of 301 CpG sites associated with age (P-value $<10^{133\text{-}253}$) will be presented.

127

Obesity and Risk of Female Reproductive Conditions: A Mendelian Randomisation Study

Samvida S. Venkatesh[1,2*], Teresa Ferreira[1,3], Stefania Benonisdottir[1], Nilufer Rahmioglu[2,3], Christian M. Becker[3], Ingrid Granne[3], Krina T. Zondervan[2,3], Michael V. Holmes[4,5], Cecilia M. Lindgren[1,2,3,4,6], Laura B. L. Wittemans[1,3]

[1]Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, United Kingdom; [2]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom; [3]Nuffield Department of Women's and Reproductive Health, Medical Sciences Division, University of Oxford, Oxford, United Kingdom; [4]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom; [5]Medical Research Council Population Health Research Unit, University of Oxford, Oxford, United Kingdom; [6]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America

Obesity is observationally associated with altered risk of female reproductive conditions. These include polycystic ovary syndrome (PCOS), abnormal menstruation, endometriosis, infertility, and pregnancy-related disorders. However, the roles and mechanisms of obesity in the aetiology of reproductive disorders remain unclear. We estimated observational and genetically predicted causal associations between obesity, metabolic hormones, and female reproductive conditions using logistic regression and Mendelian randomisation (two-sample and multivariable) applied to data from UK Biobank and publicly available genome-wide association studies. Body mass index (BMI), waist-hip ratio (WHR), and WHR adjusted for BMI (WHRadjBMI) were observationally (ORs = 1.02 – 1.87 per 1 SD obesity trait) and causally (ORs = 1.06 – 2.09) associated with uterine fibroids (UF), PCOS, heavy menstrual bleeding (HMB), and pre-eclampsia. Causal effect estimates of WHR and WHRadjBMI, but not BMI, were attenuated compared to their observational counterparts. Increased waist circumference posed a higher causal risk (ORs = 1.16 – 1.93) for the development of HMB, PCOS, pre-eclampsia, and UF than did increased hip circumference (ORs = 1.06 – 1.10). Leptin, fasting insulin, and insulin resistance each mediated between 20%-50% of the total causal effect of obesity on pre-eclampsia. In this first systematic, large-scale, genetics-based analysis of the aetiological relationships between obesity and female reproductive conditions, we found that common indices of overall and central obesity increased risk of reproductive disorders to heterogenous extents, mediated by metabolic hormones. Our results suggest exploring the mechanisms mediating the causal effects of overweight and obesity on gynaecological health to identify targets for disease prevention and treatment.

128

Fine-mapping of Novel Susceptibility Loci Associated with Eosinophil Granule Proteins (ECP and EDN) Reveals Putative Causal Variants And Candidate Genes

Raphaël Vernet[1]*, Régis Matran[2], Farid Zerimech[3], Anne-Marie Madore[4], Patricia Margaritte-Jeannin[1], Marie-Hélène Dizier[1], Florence Demenais[1], Catherine Laprise[4], Rachel Nadif[5], Emmanuelle Bouzigon[1]

[1]*Université de Paris, INSERM UMR 1124, Group of Genomic Epidemiology of Multifactorial Diseases, Paris, France; [2]Université Lille and CHU de Lille, Lille, France; [3]Pôle de Biologie Pathologie Génétique, Laboratoire de Biochimie et Biologie Moléculaire, CHU de Lille, Lille, France; [4]Basic Sciences department, Université du Québec à Chicoutimi, Saguenay, Québec, Canada, Centre intersectoriel en santé durable, Université du Québec à Chicoutimi, Saguenay,*

Québec, Canada; [5]*Université Paris-Saclay, UVSQ, Univ. Paris-Sud, Inserm, Equipe d'Epidémiologie Respiratoire Intégrative, CESP, Villejuif, France*

Background: Eosinophils play a key role in the allergic response in asthma by the release of cytotoxic molecules such as eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) that generate epithelium damages.

Objective: We aimed to identify genetic variants influencing ECP and EDN levels in asthma-ascertained families of European ancestry.

Methods: We conducted univariate and bivariate genome-wide association analyses of these proteins in 1,018 subjects from the EGEA study with follow-up in 153 subjects from SLSJ study and performed meta-analysis to combine evidence from the two datasets. We then conducted Bayesian statistical fine-mapping together with in silico quantitative trait locus and functional annotation analyses to identify credible sets of variants and target candidate genes.

Results: We identified four genome-wide significant loci ($P<5x10^{-8}$) including six distinct signals associated with ECP and/or EDN levels. These six signals were located on 14q11, 7p21, 1p31 and 9q22 chromosomal regions. Four of the six distinct signals were fine-mapped to small credible sets of putative causal variants (95% credible set size ≤10 SNPs). More particularly, the two signals on 7p21 locus each included one SNP with high posterior inclusion probability (PIP>0.7). The most likely candidate genes targeted by these SNPs were: *RNASE2* and *RNASE3* (14q11), *AK4* (1p31), *CTSL* (9q22), and *NDUFA4* (7p21).

Conclusion: This study highlights the interest of joint analysis of biological phenotypes involved in the pathophysiological mechanisms of asthma to increase power to detect new loci and candidate genes.

Funded: AAP Nord-Pas-de-Calais. ANR-GenCAST, IRSC

129

Effect of Selection Bias on Two Sample Summary Data-Based Mendelian Randomization

Kai Wang[1] and Shizhong Han[2]
[1]*Department of Biostatistics, The University of Iowa, Iowa City, Iowa, United States of America;* [2]*Lieber Institute for Brain Development, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America*

Mendelian randomization (MR) is becoming more and more popular for inferring causal relationship between an exposure and a trait. Typically, instrument SNPs are selected from an exposure GWAS based on their summary statistics and the same summary statistics on the selected SNPs are used for subsequent analyses. However, this practice suffers from selection bias and can invalidate MR methods, as showcased via two popular methods: the summary data-based MR (SMR) method and the two-sample MR Steiger method. The SMR method is conservative while the MR Steiger method can be either conservative or liberal. A simple and yet more powerful alternative to SMR is proposed.

130

Analyzing Longitudinal Zero-inflated Oral Microbiome Count Data using Two-stage Mixed Effects Models

Jian Wang[1]*, Cielito C. Reyes-Gibby[2], and 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 Emergency Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*; [3]*Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*

Recent technological advances have provided valuable resources for investigating the role of the microbiome in human health and disease. It has been of great interest to study the longitudinal changes in the microbiome and its association with risk factors and clinical outcomes. The challenges of such analysis include the zero-inflated microbial abundance counts data and the correlation among the longitudinal abundance collected across different time points within the same patient. The current approaches for longitudinal zero-inflated microbiome abundance data focused on testing the covariate-taxon associations (i.e., time-varying abundance as the dependent variable), but ignored the taxon-outcome associations (i.e., non-time-varying clinical outcome as the dependent variable). To address this issue, we proposed a two-stage mixed effects model for analyzing zero-inflated longitudinal data and the clinical outcome. In this model, the longitudinal microbial abundance count data is first modeled as a function of time using the zero-inflated negative binomial mixed effects model, and at the second stage the summaries of the temporal patterns are used in the regression models to assess their associations with the clinical outcome. Simulations showed that the two-stage mixed effects model can provide accurate estimations for the regression coefficients of the association between the longitudinal trend of microbial abundance and the outcome. We applied the approach to the study of longitudinal patterns in oral microbiome and oral mucositis in the patients with squamous cell carcinoma of the head and neck.

Keywords: oral microbiome, two-stage model, zero-inflated negative binomial mixed effects model, longitudinal count data, oral mucositis

131

Challenges with X chromosome analyses and reporting in Genome-Wide Association Studies (GWAS)

Zhong Wang[1]*, Lei Sun[2,3], Andrew D. Paterson[3,4]
[1]*School of Data Science, Fudan University, Shanghai, China;* [2]*Department of Statistic Sciences, University of Toronto, Canada;* [3]*Dalla Lana School of Public Health, University of Toronto, Canada;* [4]*Genetics and Genome Biology, The Hospital for Sick Children, Canada*

Little has improved regarding the analysis and reporting of X-chromosome variants from GWAS in the eight years since the eXclusion of the X-chromosome was brought to the attention of the community in 2013. Using the EBI-NHGRI GWAS catalog (date downloaded 2020-03-08) we identified studies that reported genome-wide significant loci on the X-chromosome, and then extracted details from each. Out of 3869 studies in the catalog (male-only studies excluded), 195 reported a total of 564 loci on the X-chromosome, drastically fewer than 1308 studies reporting 5593 loci on chromosome 7, which has similar size. Limitations of the analyses include that most applied methods for autosomes to X-chromosome – with sex as a covariate and additive coding of genotype, jointly analyzing males and females, presumably with 0/2 coding of males assuming random X-chromosome inactivation. Rarely are sex-specific analyses reported, or sex differences in trait prevalence/trait distribution provided.

One study identified two different loci in the pseudo-autosomal region PAR1 (PMID: 29808027) but did not describe how the analysis was performed and did not report sex-stratified results. Another study reported variants in the controversial PAR3 region to be associated with ANCA-associated vasculitis but did not provide the sex distribution of controls (PMID: 22808956) potentially making it subject to confounding by sex. Only one locus has been identified in PAR2 (rs306890, associated with BMI and lipids; PMID:29507422; 30108127) but sex-specificity of the associations were not reported.

Despite the major success of GWAS, the X-chromosome continues to be ignored or analyzed and reported in a suboptimal fashion.

132

Development of a Platform- and Study-Independent DNA Methylation Signature Predictive of Ovarian Cancer Recurrence

Chen Wang[1]*, Julie M. Cunningham[2], Sebastian M. Armasu[1], Stacey J. Winham[1], and Ellen L. Goode[1]
[1]*Department of Quantitative Health Sciences,* [2]*Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America*

The existence of methylation-based disease subtypes of tubo-ovarian high-grade serous cancer (HGSC) has been demonstrated. Using Illumina Infinium HumanMethylation450 BeadChip data on >450,000 CpG sites, we previously studied methylation of 337 HGSC tumors revealing two subtypes differing by disease recurrence time adjusted for known prognostic factors (hazard ratio (HR) 1.92, 95% Confidence Interval (CI) 1.25-2.94, $=2.9\times10^{-3}$). Less-favorable subtype tumors showed increased methylation of 6p21.3 immune genes. To enable replication, we sought to overcome two inherent methodological obstacles: 1) study- and platform-dependence which may limit generalizability, and 2) binary dichotomization of subtyping which may misrepresent the continuum of disease abnormalities. We developed a quantitative methodology to identify the most representative CpG sites and construct a methylation subtyping signature applicable to a variety of platforms (e.g., Illumina Infinium 27k and MethylationEPIC). Substantial concordance of methylation signatures was achieved across platforms (correlation coefficient =0.85 to 0.98, $p<1\times10^{-16}$). We applied this to 715 additional tumors from multiple independent HGSC methylation studies. The methylation signature indicative of the less-favorable subtype was again associated with disease recurrence time, adjusting for known prognostic factors (HR 1.55, 95% CI 1.08-2.22, p=0.017; all cases N=1,041, HR 1.82, 95% CI 1.29-2.57, $p=6.1\times10^{-4}$). In subset analyses, we integrated additional tumor immune factors, including TAP1 gene expression, lymphocyte infiltration, and transcriptomic subtypes. Results suggest epigenetic mediation of an HGSC tumor immune response. Altogether, we demonstrate utility of a method to derive a multi-platform methylation signature which shows consistent association with disease outcome and highlights important features of the tumor microenvironment.

133

A Novel Method to Estimate Polygenic Risk Scores in Admixed Populations

Xuexia Wang[1]*, Sicong Xie[2], Callum Doyle[1]
[1]*Department of Mathematics, University of North Texas, Denton, Texas, United States of America;* [2]*Beijing National Day School, Beijing, China*

Polygenic Risk Scores (PRS) are summaries of genetic information, which are typically effect size weighted sums of allele counts regarding disease associated genetic variants. PRS do have the potential to predict the risk of a disease for an individual, affect the

selection of treatments, and benefit the development of a medicine. The enormous amount of genome-wide association study (GWAS) summary and/or genomic data now available enables researchers to estimate PRS of an individual for a disease. The majority (>78%) of GWAS rely on European ancestry participants; it is an important challenge to create PRS useful for other admixed populations such as African Americans and Hispanic Americans, the two largest racial minority groups in the US and accounts for ~30.7% of the US population. We propose a novel method to estimate the PRS for individuals in an admixed population with recalibrated effect size of genetic variants using local ancestry proportions. This method is not only useful for an admixed population with two ancestral populations but is also useful for admixed population with more than two ancestral populations. 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 between the predicted and observed trait values for quantitative traits. Application to the 41 traits in the electronic Medical Records and Genomics (eMERGE) Network data reveals that in most of the traits, the proposed method performs more accurately in risk prediction than other comparison methods.

134

Incorporating Family History in Aggregation Unit-based Tests for Family Studies with Unbalanced Case-Control Ratio with Application to the Framingham Heart Study

Yanbing Wang[1], Han Chen[2,3], Gina M. Peloso[1], James B. Meigs[4,5,6], Alexa Beiser[1,7,8], Sudha Seshadri[7,8,9], Anita DeStefano[1], Josée Dupuis[1]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2]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; [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [4]Division of General Internal Medicine, Massachusetts General Hospital, Massachusetts, United States of America; [5]Harvard Medical School, Boston, Massachusetts, United States of America; [6]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America; [7]Framingham Heart Study, Framingham, Massachusetts, United States of America; [8]Department of Neurology, Boston University School of Medicine, Boston, Massachusetts, United States of America; [9]Glenn Biggs Institute for Alzheimer's Disease and Neurodegenerative Diseases, University of Texas Health San Antonio, San Antonio, Texas, United States of America

Introduction: Standard genetic studies are designed to evaluate the association between variants in balanced study designs. Without accounting for family correlation and unbalanced case-control ratio, these analyses could result in inflated type I error. Family history (FH) contains valuable information about disease heritability and can increase statistical power in a cost-effective manner. Here, we develop methods to address the aforementioned type I error issues, while providing optimal power to analyze aggregates of rare variants by incorporating additional information from FH.

Methods: We perform two generalized linear mixed models to adjust for relatedness in phenotyped and genotyped probands, and phenotyped relatives. To incorporate FH from multiple relatives, we use the phenotype and genotype mean among closest probands in relatives' analysis, and we down-weight the genotypes in the relatives' score prior to meta analyzing scores of probands and relatives. We use the saddle point approximation and efficient resampling to calibrate the distribution of score statistic for unbalanced designs. We evaluate our methods in simulations and apply them to data from the Framingham Heart Study (FHS).

Results: In studies with family samples with low disease prevalence, our methods reduce type I error inflation compared to methods that ignore relatedness and unbalanced designs, while providing higher power when incorporating FH. With enhanced power, our methods detect novel genes with Alzheimer's disease, dementia, and type 2 diabetes in the FHS.

Conclusions: The findings enabled by these methods exploiting FH and accounting for relatedness and unbalanced designs will help further characterize underlying mechanism of complex diseases.

135

Identification of Genetic Loci Impacting COVID-19 Severity Via Gene-Environment Interaction Analysis Incorporating Known Risk Factors

Kenneth E. Westerman[1,2,3*], Magdalena Sevilla-González[1,2,3], Joanna Lin[1], Beza Tadess[1], Alisa K. Manning[1,2,3]

[1]Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [2]Metabolism Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America; [3]Department of

Medicine, Harvard Medical School, Boston, Massachusetts, United States of America

Global meta-analyses have described genomic regions associated with severe COVID-19. Meanwhile, epidemiological research has revealed key non-genetic risk factors, including sex, cardiometabolic disease, and social determinants of health. Because of their strong effects, it is critical to understand whether these established non-genetic risk factors modulate the effect of genetic variants on COVID-19 severity. Here, we sought to uncover interactions between genetic variants and risk factors to shed light on COVID-19 biology and uncover new genetic loci. We undertook a series of three genome-wide gene-environment interaction studies in individuals of European ancestry from the UK Biobank (N=378,000), conducting both interaction tests and joint tests of genetic main and interaction effects. The exposures included sex, cardiometabolic conditions (obesity and type 2 diabetes status, tested jointly), and postcode-based multiple deprivation index. The binary outcome was severe COVID-19, defined as hospitalized, laboratory-confirmed SARS-CoV-2 infection or death from COVID-related symptoms, using the rest of the population as controls. We found five significant loci at $p < 5\times10^{-8}$ using the joint test, one of which (rs11115199) was also significant in the interaction test. Of the five loci, three did not have a significant marginal effect, emphasizing the added value of interaction testing. One variant, rs2268616, was found in both the sex and cardiometabolic joint tests and associates with both circulating testosterone and the expression of *EIF2B2*, a gene involved in the viral mRNA translation process. Our results reveal new genetic regions impacting COVID-19 severity while reinforcing the value of interaction testing for locus discovery.

136

Using Summary Statistics to Evaluate Multiplicative Combinations of Initially Analyzed Phenotypes with a Flexible Choice of Covariates

Jack M. Wolf[1]*, Jason M. Westra[2], Nathan L. Tintle[2]
[1]*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America;* [2]*Department of Math, Computer Science, and Statistics, Dordt University, Sioux Center, Iowa, United States of America*

Although the promise of electronic medical record and biobank data is large, major questions remain about patient privacy, computational hurdles, and data access. One promising area of recent development is pre-computing non-individually identifiable summary statistics to be made publicly available for exploration and downstream analysis. We demonstrate how to utilize pre-computed linear association statistics between individual genetic variants and phenotypes to infer genetic relationships between products of phenotypes (e.g., ratios; logical combinations of binary phenotypes using 'and' and 'or') with customized covariate choices. We evaluate our method's accuracy through several simulation studies and an application modeling various fatty acid ratios using data from the Framingham Heart Study. These studies show consistent ability to recapitulate analysis results performed on individual level data including maintenance of the Type I error rate, power, and effect size estimates. An implementation of this proposed method is available in the publicly available R package pcsstools (https://cran.r-project.org/package=pcsstools).

137

The Database of *CYP2C19* Variant Distribution In Healthy Thai Population

Sadhu Wongsaroj[1], Atchara Srisodsai, Ph.D.[2], Patompong Satapornpong, PhD[3*]
[1]*The Regents International School, Bangkok, Thailand;* [2]*MedCoach Institute, Bangkok, Thailand;* [3]*The Division of General Pharmacy Practice, Department of Pharmaceutical Care, College of Pharmacy, Rangsit University, Thailand*
**Corresponding author*

Introduction: CY2C19 is a liver enzyme responsible for metabolizing clinical drugs such as: omeprazole, clopidogrel, phenytoin, proguanil, diazepam, citalopram, imipramine, amitriptyline and clomipramine. In previous studies, the variants of *CYP2C19* can be used to predict the specific reaction a person might have after receiving medicine. The aim of this study was to investigate the variant of *CYP2C19* genes and the allele distribution in the healthy Thai population.

Materials and Methods: *CYP2C19*2* (c.681G>A; rs4244285), *CYP2C19*3* (c.636G>A; rs4986893), *CYP2C19*17* (g.-806C>T; rs12248560) of 160 unrelated healthy Thai individuals were test using real-time PCR.

Results: The results show that the most common allele frequency was *CYP2C19*1* with a percentage of 68.44%. The second most common allele frequency was CYP2C19*2 with a percentage of 23.75%. Lastly, CYP2C19*3 was found in only 4.69% and CYP2C19*17 with 3.13%. CYP2C19 metabolizer in the healthy sample consist of 4 phenotypes: Extensive metabolizers (EM) (*CYP2C19*1/*1* of 45.00% and *CYP2C19*2/*17* of 1.25%), the Intermediate metabolizers (IM) (*CYP2C19*1/*2* of 34.38% and *CYP2C19*1/*3* of 7.50%), the poor metabolizers (5.0% with *CYP2C19*2/*2* and 1.88% with *CYP2C19*2/*3* genotypes), and the Ultra rapid

metabolizers (UM: 5.00% with *CYP2C19 *1/*17* genotype).

Conclusions: The result shows that more than half of the participants have abnormal metabolism with only 46.25% of the participants having normal (extensive) metabolizers. A concerning 41.88% of participants are intermediate metabolizers. Thus, the database of *CYP2C19* variant distribution in the healthy Thai population should be compared with other ethnicity to support precision medicine for screening prior before administration of medication to individuals.

Keywords: Thai population, *CYP2C19* gene variant, Real-time PCR

138

Joint Analysis of Multiple Phenotypes for Extremely Unbalanced Case-Control Association Studies in Biobanks

Hongjing Xie*, Xuewei Cao, Shuanglin Zhang, Qiuying Sha

Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America

In genome-wide association studies (GWAS) for thousands of phenotypes in large biobanks, most binary phenotypes have substantially fewer cases than controls. Many widely used approaches for joint analysis of multiple phenotypes produce inflated type I error rates for such extremely unbalanced case-control phenotypes. In this research, we propose a method to jointly analyze multiple unbalanced case-control phenotypes to circumvent this issue. We first group multiple phenotypes into different clusters based on a perturbation clustering method, then we merge phenotypes in each cluster into a single phenotype. In each cluster, we use the saddlepoint approximation to estimate the *P*-value of an association test between the merged phenotype and a SNP which eliminates the issue of inflated type I error rate of the test for extremely unbalanced case-control phenotypes. Finally, we use the Cauchy combination method to obtain an integrated *P*-value for all clusters to test the association between multiple phenotypes and a SNP. We use extensive simulation studies to evaluate the performance of the proposed approach in order to show that the proposed approach can control type I error rate very well and is more powerful than other available methods. We also apply the proposed approach to phenotypes in category IX (diseases of the circulatory system) in the UK Biobank. We find that the proposed approach can identify more significant SNPs than the other available methods we compared to.

139

Within-sibship GWAS Improve Estimates of Direct Genetic Effects

Laurence J. Howe[1,2,*], Social Science Genetic Association Consortium, Within Family Consortium, Ben Brumpton[1,3,4], Gibran Hemani[1,2,] Neil M Davies[1,2,4]

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, United Kingdom; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Norway; [4]HUNT Research Center, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway

Estimates from genome-wide association studies (GWAS) represent a combination of the effect of inherited genetic variation (direct effects), demography (population stratification, assortative mating) and genetic nurture from relatives (indirect genetic effects). GWAS using family-based designs can control for demography and indirect genetic effects, but large-scale family datasets have been lacking. We combined data on 159,701 siblings from 17 cohorts to generate population (between-family) and within-sibship (within-family) estimates of genome-wide genetic associations for 25 phenotypes. We demonstrate that existing GWAS associations for height, educational attainment, smoking, depressive symptoms, age at first birth and cognitive ability overestimate direct effects. We show that estimates of SNP-heritability, genetic correlations and Mendelian randomization involving these phenotypes substantially differ when calculated using within-sibship estimates. For example, genetic correlations between educational attainment and height largely disappear. In contrast, analyses of most clinical phenotypes (e.g. LDL-cholesterol) were generally consistent between population and within-sibship models. We also report compelling evidence of polygenic adaptation on taller human height using within-sibship data. Large-scale family datasets provide new opportunities to quantify direct effects of genetic variation on human traits and diseases.

140

SIGHR: Side Information Guided High-dimensional Regression

Yuan Yang[1] *, Christopher S. Mcmahan[1], Yu-Bo Wang[1], James W. Baurley[2], Sung-Shim Park[3]

[1]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, United States of America; [2]BioRealm LLC, Walnut, California, United States

of America; [3]Population Sciences in the Pacific Program (Cancer Epidemiology), University of Hawai'i Cancer Center, Honolulu, Hawaii, United States of America

In this presentation, we develop a novel Bayesian regression framework that can be used to complete variable selection in high dimensional settings. Unlike existing techniques, the proposed approach can leverage side information to inform about the sparsity structure of the regression coefficients. This is accomplished by replacing the usual inclusion probability in the spike and slab prior with a binary regression model which assimilates this extra source of information. To facilitate model fitting, a computationally efficient and easy to implement MCMC posterior sampling algorithm is developed via carefully chosen priors and data augmentation steps. The finite sample performance of our methodology is assessed through numerical simulations, and we further illustrate our approach by using it to identify genetic markers associated with the nicotine metabolite ratio; a key biological marker associated with nicotine dependence and smoking cessation treatment.

141

Identifying Clinically-relevant Circulating Protein Biomarkers for Type 1 Diabetes: A Two Sample Mendelian Randomization Study

Nahid Yazdanpanah[1*], Mojgan Yazdanpanah[1], Ye Wang[2], Vince Forgetta[2], Michael Pollak[2,3,4], Constantin Polychronakos[5,6,7], J. Brent Richards[2,3,6,8,9] and Despoina Manousaki[1,10,*]

[1]Research Center of the Sainte-Justine University Hospital, University of Montreal, Montreal, Quebec, Canada; [2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada; [3]Departments of Medicine, McGill University, Montreal, Quebec, Canada; [4]Departments of Oncology, McGill University, Montreal, Quebec, Canada; [5]Department of Pediatrics, McGill University, Montreal, Quebec, Canada; [6]Department of Human Genetics, McGill University, Montreal, Quebec, Canada; [7]Centre of Excellence in Translational Immunology (CETI), Montréal, Quebec, Canada; [8]Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada; [9]Department of Twin Research, King's College London, London, United Kingdom; [10]Departments of Pediatrics, Biochemistry and Molecular Medicine, University of Montreal, Montreal, Canada

**Corresponding Author*

Objective: To identify circulating proteins influencing type 1 diabetes susceptibility using Mendelian randomization (MR).

Research Design And Methods: We employed a large-scale two-sample MR study, using *cis* genetic determinants of up to n=1,611 circulating proteins from five large genome wide association studies to screen for causal associations of these proteins with type 1 diabetes risk in 9,684 cases with type 1 diabetes and 15,743 controls.

Results: We found that a genetically predicted one standard deviation increase in Signal Regulatory Protein Gamma (SIRPG) level was associated with decreased risk of type 1 diabetes risk (MR OR = 1.66, 95% 1.36- 2.03; P = 7.1×10^{-7}). The risk of type 1 diabetes increased almost two fold per genetically predicted standard deviation increase in interleukin-27 Epstein Barr Virus Induced 3 (IL27-EBI3) protein levels (MR OR=1.97, 95% CI = 1.48 – 2.62, P= 3.7×10^{-6}). However, a standard deviation increase in chymotrypsinogen B1 (CTRB1) was associated with decreased risk of type 1 diabetes (MR OR=0.84, 95% CI = 0.77 – 0.90, P= 6.1×10^{-6}).

Conclusions: We identified three novel circulating protein biomarkers associated with type 1 diabetes risk using an unbiased MR approach. Upon validation in type 1 diabetes case-control cohorts, these biomarkers are promising targets for development of drugs and/or of screening tools for early prediction of type 1 diabetes.

142

Shrinkage Parameter Estimation in Penalized Logistic Regression Analysis of Case-Control Data

Ying Yu[1], Siyuan Chen[1,2], Brad McKeney[1]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada; [2]Faculty of Medicine, BC Children's Hospital Research Institute, British Columbia, Canada

In genetic epidemiology, rare variant case-control studies aim to investigate the association between rare genetic variants and human diseases. Rare genetic variants lead to sparse covariates that are predominately zeros and this sparseness leads to estimators of log-OR parameters that are biased away from their null value of zero. Different penalized-likelihood methods have been developed to mitigate this sparse-data bias for case-control studies. In this research article, we study penalized logistic regression using a class of log-F priors indexed by a shrinkage parameter m to shrink the biased MLE towards zero. We propose a maximum marginal likelihood method for estimating m, with the marginal likelihood obtained by integrating the latent log-ORs out of the joint distribution of the parameters and observed data. We consider two approximate approaches to maximizing the marginal likelihood: (i) a Monte Carlo EM algorithm and (ii) a combination of a Laplace approximation and derivative-free optimization of the marginal likelihood. We evaluate the statistical properties of the estimator through simulation studies

and apply the methods to the analysis of genetic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

143

Identifying Major Depressive Disorder Subtypes Using Polygenic Risk Scores

Clement C. Zai*1-5, Arun K. Tiwari1-2, Gwyneth C. Zai1-3, Daniel J. Mueller1-3, Natalie Freeman1, Nicole King1, Sheraz Y. Cheema1, Deanna Herbert1, Heather Emmerson1, James L. Kennedy1-3

[1]Tanenbaum Centre for Pharmacogenetics, Neurogenetics Section, Molecular Brain Science, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada; [2]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada; [3]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada; [4]Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada; [5]Stanley Center for Psychiatric Research, Broad Institute, Cambridge, Massachusetts, United States of America

Major depressive disorder (MDD) is a serious psychiatric disorder and a leading cause of disability worldwide. Its etiopathophysiology is complex and not well understood. While MDD has a genetic component, it is likely highly polygenic. To better understand the complexities surrounding MDD, we carried out a polygenic risk score analysis of 1,171 MDD patients from the Individualized Medicine: Pharmacogenetic Assessment & Clinical Treatment (IMPACT) study. We performed K-means clustering with cluster number ranging from two to five using polygenic risk scores for MDD, bipolar disorder, educational attainment, insomnia, alcohol addiction, maltreatment, impulsivity, loneliness, and post-traumatic stress disorder (JMP v15). We found the two-cluster solution to have the highest Cubic Clustering Criterion. One cluster has higher average polygenic risk scores for impulsivity, alcohol addiction, and loneliness, and lower average scores for MDD, bipolar disorder, educational attainment, and insomnia than the other cluster. We will compare characteristics between these two clusters, including psychotropic medication use and symptom severity.

144

Inverse-covariance Regularized Sparse Multivariate Regression for Identifying Methylation Quantitative Trait Loci with Missing Data

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

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

[*]Presenting author

Introduction: DNA methylation are locally correlated, and often strongly influenced by proximal SNPs. Such associations can be modeled with multivariate (methylation at nearby CpGs) regressions with input variables (SNPs) lying in a high-dimensional space. In the standard formulation of this problem, the outputs are assumed to be fully observed. However, this assumption is violated in many applications. For example, in high-throughput sequencing, it is common for measurements to be missing at some positions in some samples due to stochasticity in the sequencing capture. Although there are many existing techniques for dealing with missing data, they can be challenging to implement because of non-convexity in optimizations. We introduce an efficient method for inverse-covariance regularized sparse multivariate regression when outputs contain missing values, motivated by methylation data.

Methods: Building on error-in-variables concepts, the intermediate variables in optimization procedures are replaced with designed unbiased surrogates. This approach iterates over two convex sub-problems, and achieves an improvement in computational efficiency compared to non-convex techniques.

Results: The method is implemented in Rcpp. In simulations, while varying missing rate and sample size, we evaluate performance measured by whether it recovers the structured sparsity in regression coefficients and in the network structure among the outputs. Preliminary results indicate that our method shows high efficiency and robustness across different scenarios. We apply our method to a corrupted DNA methylation dataset, to identify the subset of SNPs that contribute to region-wise methylation differences, and also the conditional network structures among the CpGs in regions after correcting for genetic effects.

Discussion: Additional types of data corruption such as additive error, may be considered in future work.

145

Genetic Association Analysis of a Binary Trait Detects More Than Just the Genetic Effect: Implications for Pleiotropy and Replication Studies

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

Pleiotropy analyses of multiple phenotypes are ubiquitous, and summary statistics such as the effect size estimate, test statistic are the building blocks for such analyses.

Although it is straightforward to aggregate summary statistics derived from analyzing multiple continuous traits, we show that a critical problem arises when analyzing binary traits. That is, power of testing the association between a genetic variant G and a binary trait Y *also* depends on the effect of covariates such as age and sex, which may vary between the traits of interests.

To demonstrate this, consider a simple logistic regression model, logit $(P(Y=1))=\beta_0+\beta_G G+\beta_E E$. We show analytically that power of testing the null hypothesis, H_0: $\beta_G=0$, *inversely* depends on the magnitude of β_E, in addition to β_G, sample size and the minor allele frequency (MAF) of the variant of interest. This is not the case when analyzing a continuous trait using a Gaussian linear model.

We confirm our theoretical results by simulation. We assume two binary traits measured in the same set of n=1000 individuals, and MAF=0.3 and $\beta_0=0.5$ without loss of generality. We then let $\beta_G=0.3$ for *both* traits, but $\beta_E=0.8$ for trait Y_1 while $\beta_E=0.3$ for trait Y_2. The empirical power of testing H_0: $\beta_G=0$ is 62% for Y_1 while 86% for Y_2.

Our findings have important implications for the current pleiotropy studies of binary traits, and planning of a successful replication study, because the presumed genetic association evidence is relative to other covariate effects, which are likely heterogeneous between phenotypes

146

Leveraging Family History in Genetic Association Analyses of Binary Traits

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

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2]Division of General Internal Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America; [4]Department of Epidemiology, Human Genetics and Environmental Sciences, The University of Texas Health Science Center at Houston, School of Public Health, Houston, Texas, United States of America

Considering relatives' health history in case-control genome-wide association studies (CC-GWAS) may provide new information that increases accuracy and power to detect disease associated genetic variants. We conduct simulations and analyze type 2 diabetes (T2D) data from the Framingham Heart Study (FHS) to compare two methods that incorporate family history into CC-GWAS. The first method, LT-FH, replaces case-control status with posterior mean genetic liabilities. The second method, fam-meta, adopts a meta-analysis framework to combine analyses from case-controls and their relatives. In our simulation scenario of trait with modest T2D heritability ($h^2=0.28$), variant minor allele frequency ranging from 1% to 50%, and 1% of phenotype variance explained by the variant, fam-meta had the highest overall power, while both methods were more powerful than CC-GWAS. Using data from the FHS, we confirmed the well-known association of *TCF7L2* region with T2D at the genome-wide threshold of P-value$<5\times10^{-8}$, and both familial history methods increased the significance of the region compared to CC-GWAS. We additionally identified two loci 5q35 (*ADAMTS2*, plays a role in cardiovascular disease) and 5q23 (*PRR16*), not previously reported for T2D, using CC-GWAS and fam-meta. Using a suggestive threshold of P-value$<10^{-3}$ for at least one method, we identified six additional known T2D risk loci. Overall, LT-FH and fam-meta performed similarly in real-data analyses, though fam-meta was easier to implement given it is an extension of logistic regression. Future work includes the simulation of more complex scenarios, such as different linkage disequilibrium patterns, and extensions of our application to other heritable diseases.

147

A Prism Vote Framework Enhances Prediction Accuracy of the Polygenic Risk Score on the Alzheimer 's Disease Genome Data of the UK Biobank

Yexian Zhang[1,2*] , Xiaoxuan Xia[1,2], Qi Li[1,2], Rui Sun[1,2], Marc Ka Chun Chong[1,2], Benny Chung-Ying Zee[1,2], Maggie Haitian Wang[1,2]

[1]JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong SAR, China; [2]CUHK Shenzhen Research Institute, Shenzhen, China

Correspondence: maggiew@cuhk.edu.hk

Background: Common complex human traits are often result from multiple genetic and environmental causes. Genomic prediction shows promise for personalized medicine in clinical practice. Alongside genomic information, individuals also have latent population structure that can be beneficial for the improvement of genomic prediction accuracy.

Method: We developed the Prism Vote (PV) framework and software based on Bayesian model that uses an individual's prior probabilities of identity for subpopulations to "prismize" individual's disease

probability. Thus, PV enables personalized disease risk estimation. This framework is applied together with a baseline risk prediction method, such as the generalized linear model. Using genetic and phenotypic data collected in UK Biobank, we identified 651 individuals with a diagnosis of Alzheimer's disease and randomly sampling 3255 individuals from the remaining samples as controls according to the age distribution of cases. We performed comparative studies between PV and the baseline risk prediction methods to demonstrate that PV was superior in prediction accuracy.

Results and Conclusion: PV was able to improve the prediction AUC from 69.73% to 69.83%, and 70.78% to 71.40% for logistic regression and polygenic risk score (PRS) model, respectively. We have also developed an R-package named PrismVote that enables convenient application. The PV framework provides a robust and practical way to improve prediction accuracy by leveraging genetic architecture of target datasets.

148

Ultrahigh Dimensional Learning of Polygenic Risk Scores for Mendelian Randomization Studies

Xinyi Zhang*, Dehan Kong, Linbo Wang, Stanislav Volgushev

Department of Statistical Sciences, University of Toronto, Toronto, Canada

Mendelian randomization (MR) is a statistical method by which genetic variants are leveraged as instrumental variables (IV) to examine the causal relationship between modifiable exposures and disease outcomes from observational data. Finding appropriate genetic variants is crucial to make convincing causal conclusions from MR analysis. Current methods work well when candidate variants are of moderate size. However, for the identification of valid IVs from ultrahigh dimensional genetic variants, normal in practice, empirical evidence implies that existing procedures may miss many or even all the valid IVs, due to inclusion of irrelevant variables that have non-negligible correlation with the exposure.

In response to this challenge, we propose a novel approach to first remove irrelevant variants from the candidate set and then apply existing methods to identify valid instruments. To obtain more accurate causal effect estimate, we aggregate the effect of each estimated valid IV by constructing polygenic risk score (PGS), which may explain a considerable proportion of variation in the exposure and produce an adequately powered MR analysis. In addition, we provided theoretical guarantee of the proposed procedure.

To evaluate the performance, we investigate the causal relationship between tau protein and Alzheimer's Disease (AD). The data we use comes from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and we consider 357 subjects in our analysis. The proposed method has identified a new set of genetic variants that were missed by existing approaches.

149

Gene-Based Analysis of Bi-Variate Survival Traits Via Functional Regressions with Applications To Eye Diseases

Bingsong Zhang

Georgetown University, Washington DC, United States of America

Genetic studies of two related survival outcomes of a pleiotropic gene are commonly encountered but statistical models to analyze them are rarely developed. To analyze sequencing data, we propose mixed effect Cox proportional hazard models by functional regressions to perform gene-based joint association analysis of two survival traits motivated by our ongoing real studies. These models extend fixed effect Cox models of univariate survival traits by incorporating variations and correlation of multivariate survival traits into the models. The associations between genetic variants and two survival traits are tested by likelihood ratio test statistics. Extensive simulation studies suggest that type I error rates are well controlled and power performances are stable. The proposed models are applied to analyze bivariate survival traits of left and right eyes in the age-related macular degeneration progression.

150

A Fresh Look at the Role of Hardy-Weinberg Disequilibrium in Association Testing

Lin Zhang[1,*], Lisa Strug[1,2] and Lei Sun[1,3]

[1]Department of Statistical Sciences, University of Toronto; [2]Department of Computer Sciences, University of Toronto, Toronto, Ontario, Canada; [3]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Current guidance for Hardy-Weinberg equilibrium (HWE) screening in case-control genome-wide association studies (GWAS) is incoherent. First, there is no universal agreement on the HWE p-value threshold. Recommended thresholds vary in the literature, ranging from 10E-7 in controls alone, 10E-6 in controls and 10E-8 in cases, and 10E-10 per n=4,000 participants for the UK biobank study. Second, a truly associated SNP could reasonably be out of HWE in both the case and control groups, even if it is in HWE in the whole population. The degree to which there is Hardy-Weinberg disequilibrium

(HWD) due to true association is typically not large but detectable with biobank-sized data. Consequently, HWE-based quality control may mistakenly screen out truly associated SNPs.

We propose a new case-control association test that is robust to genotyping error, leverages HWD attributed to true association to increase power, and is easy-to-implement at the genome-wide level. The proposed robust allele-based (RA) joint test incorporates the difference in HWD between the case and control groups into the traditional association measure. We demonstrate that type 1 error rate of the RA joint test is well-controlled at the genome-wide significance level of 5E-8. Finally, through a GWAS of meconium ileus in 3,161 individuals with cystic fibrosis, we show that the proposed method can (i) robustly analyze SNPs with genotyping error, (ii) replicate previous genome-wide significant loci, and (iii) identify novel genome-wide significant loci that were missed by the traditional GWAS approach.

151

Integrative Clustering Analysis for Omics Data with Missingness

Yinqi Zhao[1*], Burcu Darst[1*], David V. Conti[1]
Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America

The advances in high-throughput biochemical assays have made more omics data accessible for genetic epidemiology studies. Integration of omics data can facilitate further understanding of the biological mechanisms between genetic risk factors and phenotypes. However, the occurrence of missing data is an inevitable issue in integrative omics analysis. Limiting the analysis to only those individuals with complete information can decrease power and may lead to biased estimation.

As an alternative to high-dimensional pairwise mediation or alternative clustering approaches, Latent Unknown Clustering Integrating omics Data (LUCID), is an integrative model that jointly estimates latent clusters characterized by omics profiles and genetic factors, while simultaneously associating the clusters to the phenotype. To further address the issue of missing data, we extend the LUCID model to account for missing data in omics measurements for two scenarios: (1) listwise missing; and (2) general sporadic missingness. Simulation studies show the proposed LUCID model is less biased compared to a multiple imputation and more powerful in comparison to an analysis limited to observations with complete data.

To demonstrate the application of this approach, we apply LUCID to better characterize the mechanism of a polygenic risk score (PRS) for prostate cancer by integrating pre-diagnostic serum metabolomics measures in African ancestry men from the Multiethnic Cohort. Among 4507 African American men with PRS, 691 observations have metabolomics data. Our aim is to estimate metabolomics profiles for latent clusters with high-risk of prostate cancer and to determine how those cluster may be associated with certain SNPs within the PRS.

152

Genome Wide Pleiotropic Analysis to Identify Novel Variants and Improve Genetic Risk Score Construction

Xiaofeng Zhu[1*], Luke Zhu[2], Heming Wang[3], Richard S. Cooper[4], Aravinda Chakravarti[2]
[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America; [2]Center for Human Genetics & Genomics, Department of Medicine, New York University Langone Health, New York, New York, United States of America; [3]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Harvard Medical School, Massachusetts, United States of America; [4]Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, Illinois, United States of America

Systolic and diastolic blood pressure (S/DBP) are highly correlated and modifiable risk factors for cardiovascular disease (CVD). We report here a bidirectional Mendelian Randomization (MR) and GWAS pleiotropy analysis of S/DBP summary statistics from large published UKB-ICBP BP GWAS and construct a composite genetic risk score (GRS) by adding pleiotropic variants. The composite GRS captures more heritability, ranged from 1.11- to 3.26-fold for BP traits, and was associated with an increase in Nagelkerke's R^2 for HTN and CVD, ranging from 1.09- and 2.01-fold of the traditional GRS in European, African and Asian descent in UK Biobank. Using Million Veteran Program (MVP) summary statistics for replication, we confirmed 118 novel BP pleiotropic variants that are not in linkage disequilibrium with known variants, including 18 novel BP loci. Additional 219 novel BP signals and 40 novel loci were identified by meta-analyzing UKB-ICBP and MVP summary statistics but without further independent replication. We observed significant age-modulated genetic effects on BP, hypertension and CVD in both Europeans and Asians. Our study provides further insight into BP regulation and provides a novel way to construct a GRS by including pleiotropic variants. This strategy of

incorporating pleiotropic variants is readily generalized to generate GRS for other complex diseases.

153

Genome-wide Association Analysis of COVID-19 Mortality Risk

Georg Hahn[1*], Chloe M. Wu[2], Sanghun Lee[1,3], Surender Khurana[7], Lindsey R. Baden[8], Adrienne G. Randolph[4,6], Nan M. Laird[1], Scott T. Weiss[4,5], Katharina Ribbeck[2], Christoph Lange[1,4,5]

[1]Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America; [3]Department of Medical Consilience, Graduate School, Dankook University, Yongin, South Korea; [4]Harvard Medical School, Harvard University, Boston, Massachusetts, United States of America; [5]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [6]Department of Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital, Boston, Massachusetts, United States of America; [7]Food and Drug Administration, Silver Spring, Maryland, United States of America; [8]Division of Infectious Diseases, Brigham and Women's Hospital, Boston, Massachusetts, United States of America

In December 2020, using the methodology of genome-wide association studies (GWAS), we looked at the association between whole-genome sequencing data of the virus and COVID-19 mortality as a potential method of early identification of highly pathogenic strains to target for containment. We analyzed 7,548 single stranded SARS-CoV-2 genomes in the GISAID database. Evaluating 29,891 sequenced loci of the viral genome for association with patient mortality using a logistic regression, two loci at 12,053bp (p=4.09e-09) and 25,088bp (p=4.41e-23) achieved genome-wide significance, though only 25,088bp remained singificant in follow-up analyses. The locus at 25,088bp is located in the P.1 strain, which later (April 2021) became one of the distinguishing loci of the Brazilian strain as defined by the Centers for Disease Control. The mutation frequency of 25,088bp in the Brazilian samples on GISAID rapidly increased from 0.37 in December 2020 to 0.77 in March 2021, thus revealing that our GWAS approach may be valid for early identification of more transmissible and/or pathogenic variants, but that early estimates are not stable over time, an implicit assumption of the GWAS approach that is unlikely in a pandemic. This suggests that GWAS methodology can provide suitable analysis tools for the real-time detection of new more transmissible and pathogenic viral strains in databases such as GISAID, though new approaches are needed to accommodate rapidly changing mutation frequencies over time within geographic regions, in the presence of simultaneously changing case/control ratios. Improvements of the associated patient metadata in terms of quality and availability will also be important.

154

Properties of Feedback Loops in Bidirectional 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 provide more accurate estimations when underlying mechanism is either unidirectional or bidirectional provided strong IVs for estimations. When weak IVs are used, BiLIML provides the least biased estimations. 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}).