DOI: 10.1002/gepi.22298

Genetic Epidemiology
OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org
WILEY

ABSTRACTS

The 2020 Annual Meeting of the International Genetic Epidemiology Society

1 | Accounting for cumulative effects of time-varying exposures in the analyses of gene-environment interactions

Michal Abrahamowicz*, Coraline Danieli

Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada

Assessing gene-environment interactions requires careful modeling of how the environmental exposure affects the outcome.[1] For interactions with time-varying exposures, such as air pollution, one needs to account for *cumulative* effects.[1,2] We propose, and validate in simulations, flexible modeling of interactions between genetic factors and time-varying exposures with cumulative effects. Weighed cumulative exposure (WCE) models cumulative effects through *weighed* mean of past exposure intensities.[3] The weight function *w(t)*, that quantifies the relative importance of exposures at different times, is estimated with cubic B-splines. Interactions with genotype or sex, are assessed by comparing fit to data of three alternative WCE models.[3] Model 1 assumes no interactions that is, common *w(t)* for all subgroups. Model 2 assumes the same shapes of subgroup-specific *w(t)*'s but different effect strengths. Model 3 assumes past exposure effects cumulate differently across the subgroups, implying different shapes of *w(t)*. Likelihood ratio tests help identify the model most consistent with the data.[3]

Simulation results validate the proposed models and tests. We illustrate real-life advantages of the WCE modeling by re-assessing interactions between sex and low-dose radiation in cancer.[4] Flexible cumulative effects modeling may yield novel insights regarding interactions between genotype and time-varying exposures.

[1] McAllister et al, *AJE* 2019:753-61

[2] Cruz-Fuentes et al, *Brain and Behavior* 2014: 290-297

[3] Danieli & Abrahamowicz, *SMMR* 2019:248-262

[4] Danieli et al, *AJE* 2019:1552-1562.

2 | Racial differences in methylation pathway-structured predictive models and breast cancer survival

Tomi Akinyemiju[1]*, Abby Zhang[2], April Deveaux[1], Lauren Wilson[1], Stella Aslibekyan[3]

[1]Duke University, Department of Population Health Sciences, Durham, North Carolina, USA; [2]Duke University, College of Arts and Sciences, Durham, North Carolina, USA; [3]University of Alabama at Birmingham, Department of Epidemiology, Birmingham, Alabama, USA

Background: Black women are more likely to develop aggressive breast cancer (BC) subtypes and have poorer prognosis compared with White women. Differential DNA methylation (DNAm) is associated with BC mortality; race-stratified analysis of DNAm-derived gene pathways may help to identify biological mechanisms driving survival disparities.

Methods: Publically available breast cancer clinical and DNA methylation data were downloaded from TCGA (http://cancergenome.nih.gov). We adapted a two-stage approach (Zhang, 2017) to incorporate pathway information into a predictive survival model: calculating a pathway risk score with Bayesian hierarchical Cox modeling using methylation and gene expression data, then combining clinical data and pathway risk scores into an integrated prediction model.

Results: We included 982 (169 AA, 813 White) patients and analyzed 19,775 genes with methylation data. Using the KEGG pathway annotations, we mapped 2,760 genes to 64 pathways. Among the 982 patients, the mortality event rate was 15.6%. The overall survival (OS) predictive model including all patients showed upregulation of AMPK, estrogen, and focal adhesion pathways, and downregulation of platelet activation. MAPK signaling and cell cycle pathway were upregulated in predictive models among whites and among blacks. Pathways that were uniquely associated with OS in race-stratified models were oxytocin and lysosomal signaling in whites, and Rap1, glycolysis, and insulin signaling in blacks. Pathway associations in the progression-free survival (PFS) model mirrored the OS model.

Conclusion: Significant differences in pathway signatures predictive of OS and PFS between whites and blacks may highlight specific biological mechanisms underlying aggressive BC in Black women, and present new targets for intervention.

3 | Effect of population stratification on SNP-by-environment interaction

Jaehoon An[1]*, Christoph Lange[2,3], Sungho Won[1,4,5]

[1]Department of Public Health Sciences, Seoul National University, Seoul, Korea; [2]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts; [3]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts; [4]Institute of Health and Environment, Seoul National University, Seoul, Korea; [5]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea

Proportions of false-positive rates in genome-wide association analysis are affected by population stratification, and if it is not correctly adjusted, the statistical analysis can produce the large false-negative finding. Therefore, various approaches have been proposed to adjust such problems in genome-wide association studies. However, in spite of its importance, a few studies have been conducted in genome-wide single-nucleotide polymorphism (SNP)-by-environment interaction studies. In this report, we illustrate in which scenarios can lead to the false-positive rates in association mapping and approach to maintaining the overall type-1 error rate.

4 | Estimation of non-reference ancestry proportions in genotype frequency data

Ian S. Arriaga-Mackenzie[1]*, Audrey E. Hendricks[1,2]

[1]Mathematical and Statistical Sciences, University of Colorado Denver; [2]Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus

Large, publicly available genotype frequency databases, such as the Genome Aggregation Database (gnomAD), are invaluable resources in the study of health and disease. However, summarizing data into genotype frequencies can mask heterogeneity, such as population structure, within and between samples. This limits the utility of this data, especially for ancestrally diverse populations. We have recently developed a method that can estimate ancestry proportions in summary data using allele frequencies from reference ancestries. The estimated ancestry proportions can be used to update allele frequency estimates within databases, such as gnomAD, to match the global ancestry proportions of a target sample or individual. The resulting estimated ancestry proportions have high accuracy and precision, but are limited by the ancestry groups present in the reference data set. Simply put, if the ancestry group is not within the reference data set, the proportion of that ancestry cannot be estimated.

Here, we update our method by estimating the proportion of reference ancestries and the least-square error while leaving out one reference ancestry at a time. This enables us to estimate a "missing" ancestry's proportion and similarity (i.e., Fst) with the given reference populations. We use the ancestry proportion estimates, including for the hidden ancestry, to update allele frequencies to match a target sample or individual. We evaluate our method in both simulations and real data using 1000 Genomes as reference data, and gnomAD as our target sample. Our method improves the utility and equity of genetic databases of allele frequencies especially for admixed samples and individuals.

5 | Multi-omic strategies for transcriptome-wide prediction and association studies

Arjun Bhattacharya[1], Hudson J. Santos[2], Michael I. Love[1,3]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill; [2]Department of Nursing, University of North Carolina at Chapel Hill; [3]Department of Genetics, University of North Carolina at Chapel Hill

Traditional predictive models for transcriptome-wide association studies (TWAS) consider only single nucleotide polymorphisms (SNPs) local to genes of interest and perform parameter shrinkage entirely with a regularization process. These approaches ignore the effect of distal SNPs or the functionality of the SNP-gene interaction. Here, we outline multi-omic strategies for transcriptome imputation from germline genetics for testing gene-trait associations by prioritizing distal SNPs to the gene of interest. In one extension, we identify mediating biomarkers (CpG sites, microRNAs, and transcription factors) highly associated with gene expression and train predictive models for these mediators using their cis-genotypes. Imputed values for mediators are then incorporated into the eventual model as fixed effects with cis-genotypes to the gene included as regularized effects. In the second extension, we assess trans- eQTLs for their mediation effect through mediators local to these

trans-SNPs. Highly mediated trans-eQTLs are then upweighted in the eventual transcriptomic prediction model. We empirically show the utility of these extensions in TCGA breast cancer data and in ROSMAP brain tissue data, showing considerable gains in percent variance explained of approximately 1–2%. We then use placental omics data from the ELGAN-ECHO study to train expression models in the placenta. Using these models, we then impute placental gene expression and infer gene-trait associations in GWAS cohorts for a variety of traits and disorders. This integrative approach to transcriptome-wide imputation and association studies aids in understanding the complex interactions underlying genetic regulation within a tissue and identifying important risk genes for various traits and disorders.

6 | Characterization of direct and/or indirect genetic associations for multiple traits in longitudinal studies of disease progression

Myriam Brossard[1]*, Andrew D. Paterson[2,3], Osvaldo Espin-Garcia[1,3], Radu V. Craiu[4], Shelley B. Bull[1,3]

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [2]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [4]Statistical Sciences, University of Toronto, Toronto, Canada

Associations of SNPs and longitudinal factors with time-to-event outcomes are often investigated with a Cox survival model (CM) that includes longitudinal risk factors as time dependent covariates. When longitudinal traits are endogenous and/or are measured with random error, joint modelling of risk factors and outcomes can determine direct and/or indirect association of SNPs with time-to-event traits while accounting for dependences with risk factors. Here, we present a joint model (JM) that consists of: a mixed model for multiple longitudinal traits describing the trajectory of each trait as a function of SNP effects and subject random effects; and a frailty CM for multiple time-to-event outcomes that depends on SNPs and longitudinal trajectories. We develop hypothesis testing methods to assess (a) direct/indirect SNP association with each time-to-event and (b) SNP association with all (or a subset) of the traits based on single parameter and generalized Wald statistics. Motivated by the genetic architecture of Type 1 diabetes complications (T1DC), we show by a realistic simulation study that JM of two time-to-complications (retinopathy, nephropathy) with two longitudinal risk factors (HbA1c, blood pressure) improves performance over CM to characterize direct/indirect SNP associations. By application to the Diabetes Control and Complications Trial, we illustrate feasibility and obtain results for multiple T1DCs comparing contemporaneous and cumulative effects of HbA1c according to established HbA1c exposure effects. In conclusion, JM of multiple longitudinal and multiple time-to-event traits can provide insight into etiology of complex traits.

7 | Quality control in genome-wide association studies revisited: A critical evaluation of the standard methods

Hanna Brudermann[1], Tanja K. Rausch[1,2], Inke R. König[1]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany; [2]Department of Pediatrics, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

In recent years, the focus of genome-wide association studies (GWAS) has shifted and, the task is no longer only the discovery of common genetic loci, but the discovery of loci with small effects by (mega-)meta-analyses or the aggregation of genomic information into genetic risk prediction scores. Although even low error frequencies can distort association results, extensive and accurate quality control of the given data is mandatory. However, after extensive discussions about standards for quality control in GWAS in the early years, further work on how to control data quality and adapt data cleaning to new GWAS aims and sizes is rare. The aim of this study therefore was to perform an extensive literature review to evaluate currently applied quality control criteria and their justification. Our results show that in most published GWAS, no scientific reasons for the applied quality steps are given. Cutoffs for the most common quality measures are mostly not explained. For example, principal component analyses and tests for deviation from Hardy-Weinberg equilibrium are frequently used without analysis of the exact existing conditions and corresponding adjustment of the quality control.

Building on the findings from the literature search, a workflow was developed to include scientifically justified quality control steps. This workflow is subsequently illustrated using a real data set. It is pointed out that researchers still have to decide between universal and individual parameters and therefore between optimal comparability to other analyses and optimal conditions within the specific study.

8 | Extensions to rare variant association tests under an affected sibling pair design

Michela Panarella[1,2], Shelley B. Bull[1,2]*

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada

Complex diseases are thought to be caused by both common and rare variants, and there is increasing evidence that common variants may contribute to risk in the presence of rare causal variants. Population-based genome-wide association studies are typically used to identify common disease susceptibility variants, and have led to the development of individual-level polygenetic risk scores (PRS) that are aggregates of putative variants. On the other hand, family studies are often preferred to detect rare inherited variants associated with disease using, for example, designs that ascertain affected sibling pairs (ASPs) and then conduct whole exome or whole genome sequencing. However, available methods for tests of rare variant association in ASPs have given little consideration to the role of background genetic risk from common variants (e.g., Lin and Zollner, 2015, PMID: 259766809). We propose a sib-pair level regression method that compares the presence of rare alleles on shared haplotypes versus non-shared haplotypes, can include PRS values as covariates, and uses likelihood ratio statistics to test for association. Simulations conducted to evaluate validity and power to detect rare susceptibility variants demonstrate reduced power when ASP ascertainment arises from high polygenic risk. This suggests screening of ASPs for high PRS at ascertainment. However if prior screening is not possible, accounting for PRS in the regression can partially recover lost power. Imposing strict ascertainment criteria such as early age at diagnosis in both siblings is also beneficial when rare variant penetrance strongly affects age at disease onset.

9 | Identification of novel susceptibility loci for lung cancer using cross-ancestry genome-wide meta-analyses

Jinyoung Byun[1,2]*, Xiangjun Xiao[1,2], Younghun Han [1,2], Yafang Li[1,2], Xihong Lin[3], James McKay[4], Rayjean Hung[5], Christopher Amos[1,2], INTEGRAL Consortium

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX; [2]Departement of Medicine, Epidemiology and Population Sciences, Baylor College of Medicine, Houston, TX; [3]Department of Biostatistics, Harvard School of Public Health, Boston, MA; [4]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, France; [5]The Lunenfeld-Tanenbaum Research Institute, Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Canada

Genome-wide association studies (GWAS) have revealed genetic risk factors for lung cancer, highlighting the role of smoking, family history, and DNA damage repair genes in disease etiology. Many studies have focused on European populations; however, lung cancer is a leading cause of cancer incidence and mortality around the world.

Previous GWAS analyses have been focusing on a single population-based analyses to exclude the confounding effects such as the presence of systematic allele frequency differences between populations. Another efficient tool for GWAS of complex genetic diseases and traits is meta-analysis providing a practical strategy for detecting genetic variants with modest effect sizes.

We performed a cross-ancestry fixed-effect meta-analyses in up to 70,161 individuals of European (26,683 cases/25,278 controls), African (1,987 cases/3,779 controls), or Asian (7,062 cases/5,372 controls) ancestry using HRC imputed OncoArray lung cancer data. The novel variants in or near DCBLD1 on 6q22.1 (OR = 0.93, $P = 2.11 \times 10^{-10}$), IRF4 on 6p25.3 (OR = 1.11, $P = 3.96 \times 10^{-8}$), PPIL6 on 6q21 (OR = 1.10, $P = 4.41 \times 10^{-9}$) for overall lung cancer, ACTR2 on 2p14 OR = 0.89, $P = 2.96 \times 10^{-9}$), ATM on 11q22.3 (OR = 3.61, $P = 8.88 \times 10^{-10}$), PSMA4 on 15q25.1 (OR = 0.88, $P = 1.07 \times 10^{-12}$) for lung adenocarcinoma, ABCF1 (OR = 1.35, $P = 4.54 \times 10^{-12}$) on 6p21.33, HCG9 on 6p22.1 (OR = 1.30, $P = 2.34 \times 10^{-10}$), IREB2 on 15q25.1 (OR = 1.20, $P = 9.25 \times 10^{-19}$), ZNRF3 (OR = 0.37, $P = 3.55 \times 10^{-10}$) on 22q12.1 for lung squamous cell carcinoma, and ZC3H15 on 2q32.1 (OR = 2.73, $P = 2.62 \times 10^{-8}$), and NECTIN1 on 11q23.3 (OR = 8.96, $P = 3.27 \times 10^{-8}$) for lung small cell carcinoma were identified.

Our large, cross-ancestry GWAS meta-analyses of lung cancer has identified several novel genetic associations. Further work is required to elucidate the biological mechanisms underlying these associations.

10 | Reducing complex dependency structure by graphical models - with an application to Y-chromosomal haplotypes

Amke Caliebe[1,2]*, Mikkel M. Andersen[3,4], James Curran[5]

[1]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [2]University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany; [3]Department of Mathematical Sciences, Aalborg University, Aalborg East, Denmark; [4]Section of Forensic Genetics, Department of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark;

[5]Department of Statistics, University of Auckland, Auckland, New Zealand

High-dimensional multivariate distributions with complex dependencies are present in many fields. Since the dependency structure is usually both unknown and very intricate, for many applications an approximation of the joint distribution is necessary. A natural way to do this, is to approximate the joint distribution through increasing orders of dependency between variables or, in other words, by marginal distributions of an increasing number of variables. We regard here the situation of a high-dimensional discrete probability distribution. A prime example is the joint distribution of genetic markers where dependency is also called linkage disequilibrium (LD). First order dependency corresponds to independence of markers. Second order dependence takes the pairwise dependency structure into account. Here, we approximate the dependency structure by the use of specific graphical models, namely t-cherry junction trees. The significance of t-cherry trees is that they give the optimal approximation for a fixed degree of dependence with respect to the Kullback–Leibler divergence. In this study, we apply the t-cherry tree approximation to the estimation of Y-STR haplotype population frequencies. This is a demanding task because of the complex dependency structure between the involved STR loci. We apply trees of order one, two and three by which dependencies between up to three STR loci can be taken into account. We show that the t-cherry tree method of order three outperforms the well-established discrete Laplace method in estimation accuracy while being computationally easier and quicker.

11 | Functional variant at the 12p13.31 CRC risk locus regulates LTBR expression through a long-range interaction

Yajie Gong†, Jianbo Tian†, Yao Deng, Hao Wan, Ying Zhu, Jiang Chang*, Xiaoping Miao

Key Laboratory for Environment and Health (Ministry of Education), Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

The 12p13.31 locus was identified to be an important susceptibility locus for colorectal cancer (CRC) in both East Asian and European populations through genome-wide association studies (GWAS). However, the identified tag single-nucleotide polymorphism (SNP) rs10849432 is located in an intergenic region with unclear biological function. Through bioinformatics analysis, we found a potential functional variant rs4764551, which in perfect linkage disequilibrium (LD) with the GWAS-identified tag SNP rs10849432, was located in an enhancer region. Biochemistry assays showed that the rs4764551 variant affected the enhancer activity by altering the binding affinity of Yin Yang 1 (YY1), which was previously proved to be a structural regulator of enhancer-promoter loops. We further found that this enhancer region has a long-range interaction with the LTBR promoter. When knocking down LTBR, the proliferation rate of CRC cells was significantly inhibited. In conclusion, we identified a variant modulates LTBR expression and confers CRC susceptibility through a long-range enhancer-promoter interaction. These results suggest that LTBR is an important susceptibility gene for CRC and provide more insights into the prevention and treatment of this disease.

12 | GWAS meta-analysis study for circulating metabolites identifies new loci, and reveals their implications for human health, drug development, and the causal role on cardio-metabolic traits

Pimphen Charoen[1,2]*, Chris Finan[3], Sandesh Chopade[3], the UCLEB Consortium, Aroon Hingorani[3§], Fotios Drenos[4§]

[1]Department of Tropical Hygiene, Faculty of Tropical Medicine, Mahidol University, Thailand; [2]Integrative Computational BioScience (ICBS) Center, Mahidol University, Thailand; [3]Institute of Cardiovascular Science, University College London, UK; [4]Department of Life Sciences, Brunel University London, UK

Genome-wide association study (GWAS) with high-throughput metabolic profiling provides insights into how genetic variation influences metabolism and metabolic complex diseases. Using a targeted nuclear magnetic resonance metabolomics platform, we report results from GWAS meta-analyses of 228 metabolic measures from an extended data set of 45,031 individuals. These metabolic measures include mainly lipids and lipoproteins associated measures, together with a smaller number of fatty acids, amino acids and other markers relevant to cardio-metabolic health. We identify additional metabolic-measures risk variants as well as report all identified associations on their clustering genes, functional relevance, phenotypic relevance, and drug development status. Using updated instrumental variables with increased variance explained, we further look into

the causal contribution of metabolic measures on cardiometabolic traits including coronary heart disease (CHD). Strong evidence of causality between many lipoprotein measures and CHD are shown as expected. High density lipoprotein (HDL) showed very large heterogeneity between different sized particles while medium HDL particles in particular consistently showed a protective effect on CHD regardless of their particle characteristics.

13 | Causal effects of *GCKR, G6PC2* and *SLC30A8* variants on fasting glucose levels

Guanjie Chen, Adebowale Adeyemo, Ayo Doumatey, Amy Bentley, Daniel Shriner, Charles Rotimi

Center for Research on Genomics and Global Health, National Human Genome Research Institute/NIH, Bethesda, Maryland, USA

Impaired glucose tolerance, usually manifesting as elevated fasting glucose, is a major risk factor for type 2 diabetes (T2D) and several cardiometabolic disorders. To identify genetic loci underlying glucose levels, we conducted an analysis of the Atherosclerosis Risk Communities study (ARIC) based on 9,782 participants of European ancestry who were normoglycemic (i.e., did not have T2D) at enrollment. Multivariate mixture linear regression models were used to test for associations between 81.7 million SNPS and first-visit fasting glucose (FG) with adjustment for age, body mass index (BMI), sex, significant principal components of the genotypes (PCs), and genetic relatedness.

Three loci were genome wide significant, with the leading SNPs being rs1260326 (non-synonymous, *GCKR*, T allele, EAF = 0.41, $\beta = -0.011$, and *P*-value = 1.06×10^{-8}), rs560887 (intron, *G6PC2*, T allele, EAF = 0.30, $\beta = -0.013$, and *P*-value = 3.39×10^{-11}), and rs13266634 (non-synonymous, *SLC30A8*, T allele, EAF = 0.32, $\beta = -0.012$, and *P*-value = 4.28×10^{-10}). The additive effect of the three mutated loci (rs1260326; rs560887; and rs13266634) is associated with a significantly lower glucose level ($\beta = -0.012$, and *P*-value = 8.0×10^{-28}). Each of these variants was replicated in independent population samples of European ($n = 2{,}212$), Chinese ($n = 646$) and Hispanic ($n = 1{,}105$) ancestry in the Multi-Ethnic Study of Atherosclerosis (MESA) study, as well as in an African ancestry study ($n = 7{,}713$). Fine mapping, conditional analysis and functional annotation implicated the listed SNPs as the sole causal variant at each locus. The rs1260326 (*GCKR*) variant – a leucine (T) to proline (C) substitution (P446L) – disrupts an exonic splicing site and alters the ability of *GCKR* (glucokinase regulator) to sequester glucokinase (*GCK*) in the nucleus of hepatocytes and pancreatic islet cells. The other two loci are known to influence FG levels through decreasing gluconeogenesis (*G6PC2* rs560887), or increasing peripheral insulin level (*SLC30A8* rs13266634).

In conclusion, we identified three robustly replicated genetic loci associated with lower FG and provided evidence for the lead SNPs at each locus as causal variants.

14 | Leveraging the relatedness in a large-scale biobank to identify novel serum lipid related genes

Hung-Hsin Chen[1]*, Ryan J. Bohlender[2], Lauren E. Petty[1], Quinn S. Wells[1], Chad Huff[2], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA; [2]MD Anderson Cancer Center, University of Texas, Houston, Texas, USA

Large-scale biobanks have become a popular and important resource to study the genetic mechanism of diseases, and relatedness is a common and untapped resource in biobanks. Genomic regions shared with identity by descent (IBD) are inherited from the same common ancestor without recombination, and can be observed in both close and distinct relatives. IBD segments provide an opportunity to discover genes that harbor low frequency, large effect variants that are undetectable in genome-wide association (GWAS) studies due to low power at rare and heterogeneous alleles. Numerous serum lipid-related genes have been identified in GWAS, but only 9%-12% phenotypic variance can be explained, which is significantly lower than previous estimates of heritability (35%–64%). The Vanderbilt biobank (BioVU) comprises over 280,000 DNA samples from participants with linked electronic medical record (EMR). In this study, we extracted 18,337 European dyslipidemia cases, who have either diagnosed dyslipidemia or record of taking lipid-lowering drugs in linked EMR, and 18,337 sex, age, and ancestry-matched healthy controls. The pairwise IBD shared segments were identified by GERMLINE using Illumina Multi-Ethnic Global Array data. To assess enrichment of IBD, we compared local shared IBD rates in case-case pairs and case-control pairs. The greatest enrichment of IBD sharing, based on 1,000,000 permutations, was found on chromosome 5 (chr5:170,043,173-170,047,775, $P = 3.29 \times 10^{-5}$). This region encodes a potassium channel gene, *KCNIP1*, which has never been reported correlated with serum lipid in previous GWAS. Our results demonstrate the potential of genome segments shared due to relatedness to discover novel disease genes in large biobank.

15 | Deep DNA-sequencing reveals genomic differences between esophageal squamous cell carcinoma and precancerous lesions

Yamei Chen[1]*, Wenyi Fan[1], Yongyong Ren[2], Xiaocheng Zhou[2], Meijie Du[3], Xiannian Zhang[4], Chen Wu[1,5,6], Dongxin Lin[1,6,7]

[1]Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences (CAMS) and Peking Union Medical College (PUMC), Beijing, China; [2]SJTU-Yale Joint Center for Biostatistics and Data Science, Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University; [3]School of Life Sciences and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing, China; [4]Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, College of Engineering, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China; [5]Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China; [6]CAMS Oxford Institute (COI), Chinese Academy of Medical Sciences, Beijing, China; [7]Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou, China

Esophageal squamous cell carcinoma (ESCC) is prevalent in Chinese population, accounting for over half of the world's new cases each year. Esophageal epithelia undergo stepwise process to develop ESCC from premalignancy to invasive cancer, however, the underlying genomic pattern is still unclear. Here we characterize the genomic landscapes of 122 patient samples with esophageal precancerous lesions and 191 ESCC patient samples by deep DNA-sequencing of a 1,382-gene panel. We observe more previous validated driver genes variations in ESCC than premalignancy, including known cancer genes *TP53* and *PIK3CA*, affecting cell cycle and PI3K-AKT pathways. In addition, ESCC presents more copy number variations (CNVs) at chromosome arm level compared to precancerous lesions. We further find possible druggable mutations in ESCC patients targeting *TP53*, *ATM* and *BRCA2*. These findings suggest that although cancer genomics is more chaotic than premalignancy, detecting mutations solely is inadequate for early warning of tumorigenesis. It is important to combine multi-omic data to predict cancer development at a precancer stage.

Invited Abstract

16 | Realising the power of big biobanks in diverse populations for stroke medicine

Zhengming Chen, *Professor of Epidemiology*

Nuffield Department of Population Health, University of Oxford, Oxford, UK

Stroke is a leading cause of premature death and permanent disability worldwide. Many important genetic and nongenetic causes of stroke still await discovery. Understanding what causes stroke in diverse populations with different lifestyles, environments and genetic architectures can lead to improved disease prevention and risk prediction, and the development of "precision medicine." Unique opportunities to fulfill these goals are offered by prospective "biobank" studies, with detailed characterization of large numbers of apparently healthy individuals from the general population, using conventional and novel technologies, and with long-term electronic monitoring of their health status.

Several big blood-based prospective studies have been undertaken this century in the West (e.g., UK Biobank) and East (e.g., China Kadoorie Biobank [CKB]). CKB recruited 512,891 adults during 2004-2008 from 10 diverse areas throughout China, with extensive data collected at baseline and periodic resurveys, on lifestyle (e.g., smoking, alcohol drinking, diet, and physical activity), environmental (e.g., ambient temperature and air pollution), and physiological factors (e.g., blood pressure, adiposity, lung function, bone density, and ECG). To date, >1 million disease episodes, including >55,000 well-characterized strokes (>40,000 IS and >10,000 ICH), have been recorded among participants. These exposure and health outcome data are now being complemented, in nested case-control or cohort-wide settings, by blood assays of genetic (e.g., 800 K SNPs), metabolomic (e.g., ~1000 metabolites), proteomic (e.g., ~500 inflammation and other biomarkers), and infective (~20 pathogens) biomarkers in stored biological samples. Major findings are now emerging in CKB about genetic and environmental determinants of stroke (and other diseases), some expected and some intriguingly unexpected but novel, including assessment of any causal protective effects of moderate alcohol drinking on stroke using the East Asian specific "flushing" genes. The big maturing biobanks in the Eastern and Western populations will greatly improve our understanding about aetiology of stroke and many other diseases.

17 | GWAS transethnic meta-analysis of BMI in ~700k individuals reveals novel gene-smoking interaction in African populations

Tinashe Chikowore[1,2]*, Michael Chong[3], Lisa K. Micklesfield[1], Michele Ramsay[2], Paul W. Franks[4,5,6], Guillaume Pare[3], Andrew P. Morris[7]

[1]MRC/Wits Developmental Pathways for Health Research Unit, Department of Pediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [2]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the

Witwatersrand, Johannesburg, South Africa; [3]*Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada*; [4]*Department of Clinical Sciences, Skåne University Hospital, Malmö, Sweden*; [5]*Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden*; [6]*Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, MA, USA*; [7]*School of Biological Sciences, University of Manchester, Manchester, UK*

Sixty two percent of the 1.12 billion obese people globally reside in low-middle income countries, 77% of which are in Africa. There is paucity of data on gene-lifestyle interactions associated with the increasing prevalence of obesity among Africans. We hypothesised that gene-environment interacting (GEI) variants exhibit heterogenous effects on obesity in transethnic meta-analysis of marginal SNP associations as a result of modification by an unknown exposure that varies across populations.

Body mass index (BMI) genome-wide association study (GWAS) summary statistics for 678,671 individuals representative of the major global ancestries were aggregated at 21,338,816 SNPs via fixed-effects meta-analysis. Lead SNPs attaining genome-wide significance ($P < 5 \times 10^{-8}$) were tested for heterogeneity in effects between GWAS. Lead SNPs with significant evidence of heterogeneity after Bonferroni correction were then selected for interaction analysis with selected lifestyle factors in an independent AWI-Gen study of 10,500 African participants. Significant interaction findings were then replicated in 3,177 individuals of African ancestry in the UK Biobank.

Of 881 lead SNPs, five had significant heterogenous effects on BMI ($P < 5.7 \times 10^{-5}$). Rs471094, at the *CDKAL1* locus had significant interaction with smoking status, which reduced the effect of the BMI raising allele in current smokers ($\text{Beta}_{\text{int}} = -0.949\ \text{kg/m}^2$; $P_{\text{int}} = .002$) compared with non-smokers in AWI-Gen. This finding was validated in the UK Biobank ($\text{Beta}_{\text{int}} = -1.471\ \text{kg/m}^2$, $P_{\text{int}} = .020$; meta-analysis $\text{Beta}_{\text{int}} = -1.050\ \text{kg/m}^2$, $P_{\text{int}} = .0002$). Our results highlight the first gene-lifestyle interaction on BMI in Africans and demonstrate the utility of transethnic meta-analysis of GWAS for identifying GEI effects.

18 | PRSet: Pathway-specific polygenic risk score software

Shing Wan Choi*, Hei Man Wu, Paul F. O'Reilly

Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

Genome-wide polygenic risk scores (PRS) could obscure important information about disease risk at the pathway-level. To address this, we developed *PRSet*, a novel tool to compute pathway-specific PRS.

Utilizing the Swedish schizophrenia and UK Biobank data, we evaluated the performance of *PRSet* against *MAGMA* and *LDSC*. First, we collected gene sets from different databases (e.g., Reactome, Gene Ontology, KEGG, and Mouse Genome Informatics), and computed a "disease pathway score" for each gene set as the sum of *Malacards* gene values, which reflect disease relevance for each gene based on literature. Secondly, we performed cell-type and tissue-type specific analyses using cell-type mouse gene expression data and GTEx v8 data, following the approach of Bryois et al. (2019).

Significant rank correlation was observed between the "disease pathway score" (MGI) and competitive *P*-values obtained from *PRSet* ($R^2 = 0.0516$, $p\text{-value} = 5.54 \times 10^{-12}$) and *MAGMA* ($R^2 = 0.0261$, $p\text{-value} = 1.14 \times 10^{-6}$). Results from the tissue-type specific analyses suggest that genes specifically expressed in the brain are those most associated with schizophrenia (*P*-value: $PRSet = 1.38 \times 10^{-5}$; $MAGMA = 1.19 \times 10^{-12}$; $\text{LDSC} = 3.63 \times 10^{-3}$), as expected, while cell-type specific analyses performed in relation to Alzheimer's disease identified Microglia as the top-ranking cell-type according to all three methods. Overall, *PRSet* appears to have similar performance in ranking pathways according to enrichment of GWAS signal as *MAGMA* and *LDSC*. However, a key advantage of *PRSet* over other pathway-enrichment approaches is that *PRSet* generates a pathway-based PRS for each individual, which can be used for downstream analyses such as patient stratification.

19 | Challenge of collider bias in heart disease progression studies

Amanda HW. Chong[1]*, Alastair W. Poole[2], George Davey Smith[1], Rebecca C. Richmond[1]

[1]*Medical Research Council Integrative Epidemiology Unit, University of Bristol, UK;* [2]*Department of Physiology & Pharmacology, University of Bristol, Bristol, UK*

Genome-wide association studies (GWAS) have played a key role in identifying SNPs associated with heart disease, however they have predominantly focused on disease incidence rather than progression. The challenge of progression studies in case only samples is collider bias. By stratifying on case status, factors (including SNPs) that influence disease incidence can become correlated among cases even when these factors are not correlated in the population giving rise to the case sample. This

could violate the assumptions of the Mendelian randomization (MR) approach for evaluating causality in progression studies.

Linear and logistic regressions between seven heart disease risk factors in UK Biobank (UKB) were compared to one sample MR estimates with polygenic risk scores (PRS) calculated using published GWAS, adjusting for age, sex and 40 principal components. These analyses were performed using UKB overall ($n = 488{,}378$) and a subsample of acute myocardial infarction cases ($n = 8{,}654$).

In UKB overall, body mass index (BMI) and low density lipoprotein were positively associated ($\beta = 0.025$; 95% CI = [0.022, 0.028]; P value = <2E-16), whereas they became inversely correlated in the case only group ($\beta = -0.044$; 95% CI = [−0.062, −0.026]; P value = 2E-6). Additionally, a negative correlation was induced between the BMI and smoking initiation PRS in the case only group ($\beta = -0.032$; 95% CI = [0.054, −0.011]; P value = 0.003).

These results illustrate the potential impact of collider bias for both observational and MR in a case only setting. This warrants acknowledgement that this bias can influence the magnitude and direction of causal estimates.

20 | Evaluating machine learning models for building risk prediction models in complex datasets

James P. Cook[1]*, Yannis Goulermas[2], Andrew P. Morris[1,3]

[1]Department of Biostatistics, University of Liverpool, Liverpool, UK; [2]Department of Computer Science, University of Liverpool, Liverpool, UK; [3]Division of Musculoskeletal and Dermatological Sciences, University of Manchester, Manchester, UK

Large-scale population biobanks offer exciting opportunities to develop risk prediction models for complex diseases because of the availability of genetic data with extensive lifestyle and clinical information. Machine learning methods are ideal tools to build these models as they are capable of incorporating features from a wide range of different data sources, as well as interactions between them.

We have performed a simulation study to compare the predictive accuracy of prediction models generated by several machine learning methods (gradient boosting machines, neural networks, random forests and support vector machines) for a complex disease outcome. 100 causal SNPs and five causal continuous normally distributed clinical factors for 20,000 individuals were simulated, with outcomes generated by combining all causal factors, with five SNP-SNP and three SNP-clinical factor interactions also specified in the model. Logistic regression analyses were performed with and without specifying the interaction effects to provide benchmark accuracy, while all machine learning approaches were applied in a grid framework to test multiple parameter combinations.

Initial results show that the machine learning methods are capable of producing more accurate prediction models than the standard logistic regression approach (without interactions), demonstrating that they are able to detect and utilise non-additive effects within the data set. As machine learning methods become more widely adopted in the field of genetic epidemiology, these results have important implications for the development of risk prediction models for complex disease in the coming years.

21 | Genome-wide meta-analysis of primary biliary cholangitis in 10,516 cases and 20,772 controls identifies potential drug candidates for re-purposing

Heather J. Cordell[1]*, James J. Fryett[1], Kazuko Ueno[2], Konstantinos N. Lazaridis[3], Pietro Invernizzi[4], Xiong Ma[5], Katherine A. Siminovitch[6,7], Minoru Nakamura[8], George F. Mells[9], International PBC Consortium

[1]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK; [2]Genome Medical Science Project, National Center for Global Health and Medicine (NCGM), Tokyo, Japan; [3]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota, USA; [4]Division of Gastroenterology and Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy; [5]Shanghai Institute of Digestive Disease, Renji Hospital, Jiao Tong University School of Medicine, Shanghai, China; [6]Mount Sinai Hospital, Lunenfeld-Tanenbaum Research Institute and Toronto General Research Institute, Toronto, Ontario, Canada; [7]Departments of Medicine, Immunology and Medical Sciences, University of Toronto, Toronto, Ontario, Canada; [8]Clinical Research Center, National Hospital Organization, Nagasaki Medical Center, Omura, Japan; [9]Academic Department of Medical Genetics, University of Cambridge, Cambridge, UK

Primary biliary cholangitis (PBC) is a chronic liver disease in which progressive, autoimmune destruction of the small intra-hepatic bile ducts eventually leads to cirrhosis. Many patients have inadequate response to all recognised medications, leaving them at risk of progressive liver disease, motivating the search for novel treatments.

Previous genome-wide association studies (GWAS) and meta-analyses (GWMA) of PBC have identified genome-wide significant associations at both HLA and

42 non-HLA loci. With additional genotyping funded via UK-PBC, a UK-wide initiative to extend knowledge of PBC, we undertook the largest GWMA of PBC to date, combining new and existing data from five European and two Asian (Chinese and Japanese) cohorts. We identified 56 genome-wide significant loci (20 novel) in either European, Asian or combined cohorts. We also identified a separate European locus via conditional analysis, bringing our total findings to 57. We used several approaches – functional annotation of credible causal variants; methylome, transcriptome and proteome-wide association studies; co-localization, and DEPICT – to prioritise candidate genes at genome-wide significant risk loci. These reaffirm that PBC is an archetypal autoimmune condition, enrichment analysis reiterating the importance of TLR, TNF and NFκB/MAPK signalling, and T_H1 and T_H17 cell differentiation. We used network-based in-silico drug efficacy screening, estimating a drug-disease proximity measure quantifying the closeness of the connection between the candidate genes and known drug targets, to identify treatments that might be suitable for re-purposing to PBC. This highlighted several promising treatment options, including immunomodulators already approved for the treatment of diverse autoimmune diseases.

22 | Asthma, gender and the epigenetic clock

Daley Denise[1], Vasileva Denitsa[1], Wan Ming[1], Becker Allan[2], Chan Edmond S.[3], Laprise Catherine[4], Sandford Andrew[1] and Greenwood Celia[5]

[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]Université du Québec á Chicoutimi, Saguenay, Canada; [5]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada

Background: Methylation is an important DNA epigenetic modification. Methylation profiles are impacted by environmental exposures, age and disease. The epigenetic clock evaluates chronological age versus predicted biological age and may provide mechanistic insight into disease/exposure assessment.

Purpose: Use the Horvath age prediction algorithm to examine the epigenetic versus biological age of participants in two Canadian asthma studies.

Study Cohorts: This study was conducted using 812 samples from two Canadian cohorts: The Canadian Asthma Primary Prevention Study (CAPPS, $n = 632$ samples) and the Saguenay-Lac Saint-Jean (SLSJ, $n = 180$ samples) cohorts. The CAPPS study is a prospective, longitudinal birth cohort which has followed 549 children at high-risk for developing asthma from birth to age 15. The SLSJ study is comprised of multigenerational families of French Canadian descent. Childhood asthmatics are predominately male while adult asthmatics are predominately female, this gender switch occurs in adolescence.

Methods: Methylation Sequencing was performed on the 812 samples using Illumina's MethylCapture (San Diego, CA). The Horvath epigenetic algorithm was used to predict age.

Results: Preliminary results show that asthmatic children (all ages) are older on the epigenetic clock compared to children without asthma (P value = .02). Upon further examination of the CAPPS cohort, at age seven (P value = .01) asthmatic children are older on the epigenetic clock versus their biological age than those without asthma. By age 15 (P value = .74) there was no difference between the groups, while asthmatic adults are younger on the epigenetic clock, which mimics with the gender reversal seen in asthma.

23 | Germline sequencing of DNA repair genes in 5,545 men with aggressive and nonaggressive prostate cancer

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

[1]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA; [2]The Institute of Cancer Research, London, UK; [3]Karolinska Institute, Solna, Sweden

Very little is known regarding the etiology of aggressive prostate cancer (PCa), the second leading cause of cancer death in U.S. men. We investigated whether rare pathogenic and deleterious germline variants in DNA repair genes are associated with aggressive PCa risk. Participants were 5,545 men of European ancestry, including 2,775 with nonaggressive and 2,770 with aggressive PCa, the majority of which died due to PCa ($n = 2{,}052$, 74.1%). Rare (MAF < 0.01) pathogenic or deleterious variants were analyzed for 155 DNA repair genes. Gene-based burden tests showed *BRCA2* and *PALB2* as having the most significant association with aggressive disease. *BRCA2* alleles were found in 2.5% of aggressive and 0.8%

of nonaggressive cases (OR = 2.78, $P = 8.3E-06$), while *PALB2* alleles were found in 0.65% of aggressive and 0.11% of nonaggressive cases (OR = 6.19, $P = 7.6E-04$). *BRCA2* conveyed stronger risk for aggressive disease in men diagnosed <60 than ≥60 years (OR = 5.27 vs. OR = 2.20, respectively; $P = .04$). Effects were diminished when assessing aggregate effect of all DNA repair genes (36.4% of aggressive vs. 33.1% of nonaggressive cases carried one or more alleles; $P = .005$) and the aggregate effect of a subset of 24 literature-cured candidate PCa DNA repair genes (14.2% of aggressive vs. 10.6% of nonaggressive cases carried one or more alleles; $P = 6.8E-06$). Each additional allele among these 24 genes was associated with a 0.93-year younger PCa diagnosis age ($P = 1.2E-04$). Our findings suggest that mutation detection in a small number of DNA repair genes could be used to inform risk prediction of aggressive disease and targeted PCa screening efforts.

24 | Genome wide association study in COPD

Ahra Do[1], Sungho Won[2], Woo Jin Kim[3]

[1]Bioinformatics Interdisciplinary Program College of Natural Sciences Seoul National University; [2]Department of Public Health Science, Seoul national university; [3]Kangwon National University, Chuncheon

Chronic Obstructive Pulmonary Disease (COPD) is a chronic lung disease that causes airflow restriction. It is induced by the lung parenchymal destruction caused by long-term stimuli by many environmental factors, smoking, air pollution, and genetic factors. In this study, we performed genomic association studies (GWAS) of COPD with Korean samples. Discovery samples were obtained from Kang-won University Hospital and Seoul National University Gangnam center, and significant SNPs were replicated with KARE data set ($n = 7{,}498$). We found two significant SNPs in this genome-wide association study. One was in the intergenic area of the "PCDH7" gene on chromosome 4 (rs13118325, P value: 2.2×10^{-10}, odds ratio: 2.14) and the other was close to the gene "AGER" on chromosome 6 (rs41268928, P value: 1.1×10^{-8}, odds ratio: 0.48).

25 | Fast detection of unmeasured GxE and GxG interactions using distribution-free assumptions

Claus T. Ekstrøm[1]*, Christian B. Pipper[2]

[1]Biostatistics, Department of Public Health, University of Copenhagen, Denmark; [2]Leo Pharma, Denmark

Environmental factors receive a lot less attention in studies involving genomics for several reasons: they change over time, they can be more expensive to measure on an individual level, and they often contain less structure than genetic data. Consequently, environmental measurements may not be available to include in statistical analyses.

Finite mixtures of regression models provide a flexible modeling framework for many phenomena including gene-environment, gene-gene-interactions and personalized medicine but parametric approaches require full specification of the mixture components. Using moment-based estimation of the regression parameters, we develop unbiased estimators of the regression coefficient and mixture probability with a minimum of assumptions on the mixture components which is particularly useful for large-scale analysis since computation time is minimal. In particular, only the average regression model for one of the components in the mixture model is needed with no requirements on the distributions. The consistency and asymptotic distribution of the estimators is derived and the method is applied to a large-scale omics study to hunt for predictors that were undiscovered using traditional approaches.

26 | Improved mediation analyses in case-control studies

Michael P. Epstein[1]*, Elizabeth J. Leslie[1], Glen A. Satten[2]

[1]Department of Human Genetics, Emory University, Atlanta, Georgia, USA; [2]Centers for Disease Control and Prevention, Atlanta, Georgia, USA

There is substantial interest in assessing how the relationship between known risk variation and complex disease are influenced by intermediate biological, environmental, and phenotypic factors. We can explore these relationships using techniques of mediation analysis, which disentangle associations between exposures and outcomes and assess mechanisms by which such relationships are influenced by intermediate (mediating) variables. Under the popular case-control study design, the most common procedure for mediation analysis uses a counterfactual framework that estimates indirect and direct effects for dichotomous outcomes on the odds ratio scale, allowing for exposure-mediator interactions and nonlinear effects. While this framework has proven valuable, we show it does not fully leverage all relevant data collected by the case-control study and can lead to suboptimal performance. To remedy this, we develop novel likelihood-based approaches for mediation analysis in

case-control studies that fully leverages all available data, thereby leading to more precise estimates and more powerful testing of indirect and direct effects compared to existing frameworks. We demonstrate these improvements using both simulated data as well as GWAS data from a large case-control study of orofacial clefting. For this latter analysis, we refine relationships between risk SNPs and cleft lip/palate and deduce whether these relationships are mediated by factors like maternal tobacco use during pregnancy.

27 | Metabolomics enhances understanding of genomic and metagenomic variation to provide novel insights into human health

Michael J. Evans

Metabolon inc

Metabolic derangements form the basis of any transition from homeostasis to disease, reflecting both genetic and nongenetic (e.g., environment, diet, microbiome) perturbations that drive alterations in health status. Metabolomics, the high-throughput profiling of changes in biochemical composition, is now recognized as a powerful modality to leverage both genetic and nongenetic information for providing unique insights into disease onset, progression, and response to therapies. In large cohort studies, metabolomics is being used to correlate genotype and genomic variants with phenotypes, to identify the function of unknown genes, and to functionally map both common and rare genetic variants that modify the blood metabolome to identify drug targets and signatures of human disease. This approach can be leveraged to provide mechanistic insights into individuals with genetic knockouts that predispose to either harmful or beneficial phenotypes, demonstrating the power of systems biology approaches. High-throughput biochemical profiling also reports on microbiome and environmental changes to provide a comprehensive view of deviation of a homeostatic state. Recent work reveals that microbiome activity, diet, and environment alter biochemical composition that has profound effects on disease onset, severity, and response to therapies that are not influenced by host genetics. Our findings further validate that metabolomics could be an effective tool in precision medicine for disease risk assessment and customized drug therapy in clinics.

28 | Single-cell transcriptomic analysis in a mouse model deciphers cell transition states in the multistep development of esophageal cancer

Jiacheng Yao[1], Qionghua Cui[2], Wenyi Fan[2]*, Yuling Ma[2], Yamei Chen[2], Wen Tan[2], Yanyi Huang[3], Chen Wu[2,4,5], Jianbin Wang[1], and Dongxin Lin[2,5,6]

[1]School of Life Sciences and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing, China; [2]Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences (CAMS) and Peking Union Medical College (PUMC), Beijing, China; [3]Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, College of Engineering, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China; [4]Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China; [5]CAMS Oxford Institute (COI), Chinese Academy of Medical Sciences, Beijing, China; [6]Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou, China

Esophageal squamous cell carcinoma (ESCC) is prevalent in some geographical regions of the world. ESCC development presents a multistep pathogenic process from inflammation to invasive cancer; however, what is critical in these processes and how they evolve is largely unknown, obstructing early diagnosis and effective treatment. Here, we create a mouse model mimicking human ESCC development and construct a single-cell ESCC developmental atlas. We identify a set of key transitional signatures associated with oncogenic evolution of epithelial cells and depict the landmark dynamic tumorigenic trajectories. The transcriptomic alterations in stromal cells indicate that the aberrant immune response of the carcinogen-induced lesion is critical to ESCC formation and progression. An early downregulation of $CD8^+$ response against the initial tissue damage accompanied by the transition of immune response from type 1 to type 3 results in accumulation and activation of macrophages and neutrophils, which may create a chronic inflammatory environment that promote carcinogen-transformed epithelial cell survival and proliferation. These findings illustrate a complete interplay between epithelial cells and their microenvironment at the single cell level and shed new light on how ESCC is initiated and developed. The newly identified molecules in the present study might benefit early detection and targeting therapy of ESCC.

Keywords: tumorigenesis, RNA expression, 4NQO, esophageal cancer, animal model

29 | Cross-cancer cross-tissue Transcriptome-wide Association Study (TWAS) of 11 cancers identifies 56 novel genes

Helian Feng[1]*, Arunabha Majumdar[2], Bogdan Pasaniuc[2], Hongjie Chen[3], Sara Lindstrom[3], BCAC, OCAC, PRACTICAL, Jeroen Huyghe[4], Stephanie L. Schmit[5], Tracy A. O'Mara[6], Deborah J. Thompson[7], Stuart MacGregor[6], Paul Brennan[8], James McKay[8], Richard S. Houlston[9], Beatrice S. Melin[10], Christopher Amos[11], Anne E. Cus[12], Mark M. Iles[13], Siddhartha Kar[14], Paul Pharoah[15], Rayjean J. Hung[16], Peter Kraft[1]

[1]Harvard T. H. Chan School of Public Health, Boston, MA; [2]University of California, Los Angeles, CA; [3]University of Washington, WA; [4]Fred Hutchinson Cancer Research Center, WA; [5]H. Lee Moffitt Cancer Center and Research Institute, FL; [6]QIMR Berghofer Medical Research Institute, Australia; [7]University of Cambridge, UK; [8]World Health Organization, France; [9]The Institute of Cancer Research, UK; [10]Ume. University, Sweden; [11]Baylor College of Medicine, TX; [12]The University of Sydney, Australia; [13]University of Leeds, UK; [14]University of Bristol, UK; [15]Cambridge University Hospitals NHS Foundation Trust, UK; [16]University of Toronto, ON, Canada; [17]Sinai Health System, Chicago, IL, USA

Though heterogeneous, multiple tumor types share hallmark mechanisms. Thus, identifying genes associated with multiple cancer types may shed light on general oncogenic mechanisms and identify genes missed in single-cancer analyses. TWAS have been successful in testing whether genetically-predicted tissue-specific gene expression is associated with cancer risk. Although cross-cancer genome-wide association studies (GWAS) analyses have been performed previously, no cross-cancer TWAS has been conducted to date. Here, we implement a pipeline to perform cross-cancer, cross-tissue TWAS analysis. We use newly-developed multi-trait TWAS test statistics to integrate the TWAS results for association between 11 separated cancers and predicted gene expression in 43 GTEx tissues, including a "sum" test and a "variance components" test, analogous to fixed- and random-effects meta-analyses. We then integrated the results across different tissues using the Aggregated Cauchy Association Test (ACAT) combined test.

A total of 403 genes were significantly associated with at least one cancer type for at least one tissue; 96 additional genes were identified when combining test results across cancers; and 35 additional genes when further combining test results across tissue. Among these significant genes, 70 were not near previously-published GWAS index variants. 14 of the 70 novel genes were identified from the single-cancer single-tissue test; an additional 43 were identified with the cross-cancer test; and another 13 were identified when further combined across tissues. The newly identified genes, including *RBBP8* and *TP53BP*, are involved in chromatin structure, tumorigenesis, apoptosis, transcriptional regulation, DNA repair, immune system, oxidative damage and cell-cycle, proliferation, progression, shape, structure, and migration.

30 | Detecting time-varying genetic association with mixed effects models

Zeny Feng

University of Guelph, Guelph Ontario Canada

Genetic effects can be time-dependent if the change of the phenotype over time is under genetic influence. Genome-wide association studies (GWAs) has been widely used to screen the association between a trait and genetic variants (Single nucleotide polymorphisms, SNPs) throughout the genome. On the other hand, for disease related traits that changes over time, a longitudinal study design is often used to investigate the development and progression of a disease, multiple longitudinal traits are measured repeatedly overtime as well. Detecting longitudinal genetic effect on each trait separately or jointly becomes appealing when genetic information is collected for a longitudinal study. Here, we proposed a 2-step approach to joint analysis genetic association with multiple traits and as well as genetic longitudinal effects on multiple traits. In Step 1, generalized linear mixed effects models in which a random intercept for capturing subject-specific genetic effect and a random slope for capturing subject-specific time-vary genetic effects are used for each trait of interest. The predicted random intercepts and random slopes that captures subject-specific genetic and time-varying genetic effects are treated as responses to be tested for their association with each SNP simultaneously in Step 2. Our method allows the flexibility to detect the genetic pleiotropic effects and genetic longitudinal effects that subjects to the number of traits to be included in the analysis.

31 | Japanese specific imputation reference panel using 7,000 whole genome sequences reveals novel rare variant association with serum uric acid

Jack Flanagan[1,2]*, Nana Matoba[3], Yukihide Momozawa[5], Kaoru Ito[6], Koichi Matsuda[4,7], Yoshinori Murakami[8], Yoichiro Kamatani[3,9], Andrew P. Morris[1,10], Momoko Horikoshi[2], Chikashi Terao[3,11,12]

[1]Department of Biostatistics, University of Liverpool, Liverpool, UK; [2]Laboratory for Endocrinology, Metabolism and Kidney diseases, RIKEN, Center for Integrative Medical Sciences, Yokohama, Japan; [3]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; [4]Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan; [5]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; [6]Laboratory for Cardiovascular Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; [7]Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan; [8]Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan; [9]Kyoto-McGill International Collaborative School in Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; [10]School of Biological Sciences, University of Manchester, Manchester, UK; [11]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan; [12]The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan

We created two Japanese population specific reference panels using whole genome sequence (WGS) data from the BioBank Japan Project to augment all-ancestry haplotypes from the 1000 Genomes Project (1KG): 3000 Japanese WGS (1KG + 3K) at 30× depth; and 7000 Japanese WGS (1KG + 7K) of mixed depth (3000 at 30× and 4000 at 3×).

A cohort of 30,042 Japanese individuals genotyped on the Illumina OmniExpress and HumanExome arrays was imputed up to each of the reference panels and 1KG alone. Comparisons across reference panels demonstrated that the addition of WGS data substantially increased the number of "well imputed" variants and improved their imputation quality, particularly for those with minor allele frequency (MAF) <5%.

We conducted a genome-wide association study of serum uric acid (UA) on a subset of 17,638 individuals after imputation up to the 1KG, 1KG + 3K, and 1KG + 7K panels. We compared association signals at five previously-reported UA loci attaining genome-wide significance across the three panels. At the *NRXN2* locus, we identified a missense variant present in all three panels encoding *SLC22A12* (rs121907892, MAF = 0.025, $p = 8.00 \times 10^{-230}$) that has been implicated in familial renal hypouricemia (FRH). After conditioning on rs121907892, a distinct association driven by a rare missense variant encoding *SLC22A12* was revealed (rs121907896, MAF = 0.0024, $p = 8.65 \times 10^{-24}$), which was not present in the 1KG panel and has also been implicated in FRH. These results highlight the benefits of including population-specific WGS in imputation reference panels, particularly for the identification of rare variant associations with complex human diseases.

32 | Permutation-based variable importance measures for unsupervised random forests

Césaire J. K. Fouodo*, Inke R. König

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Random forests (RF) have been shown to be fast and to produce good results in many high-dimensional applications. In addition to using supervised RF to solve classification problems, RF permutation-based Variable Importance Measures (VIMs) are frequently used for feature selection to differentiate variable signal from noise, and many permutation approaches have been proposed for this purpose. These are expected not only to clearly differentiate signal from noise, but also, for noise variables, to have a distribution of importance estimates centered around zero. In principle, permutation-based VIMs can also be utilized for unsupervised RF (URF). For example, understanding effects of genetic markers on population structure could help to avoid spurious findings. Particularly for URF, the target variable is generated based on the marginal distributions of predictors. Therefore, the distribution of the VIM could be affected by the marginal distribution of predictors. We investigate different proposed VIMs in the case of URF, present their limitations, and propose a new permutation based VIM inspired by cross-validation and the holdout trick as suggested in the literature. Our new approach is expected to produce better results in terms of recognizing signal as well as to have importance estimates of noise variables centered around zero for both RF and URF. The proposed approach will be evaluated on artificial and real data, consisting of three samples drawn from different geographic areas of Germany, with the aim to point out which SNP are relevant for the geographical distribution of our samples.

33 | Methylome-wide association study identifies CpG sites associated with 30 complex traits

James J. Fryett[1]*, Andrew P. Morris[2] and Heather J. Cordell[1]

[1]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK; [2]Division of Musculoskeletal and Dermatological Sciences, University of Manchester, Manchester, UK

Transcriptome-wide association studies (TWAS), in which gene expression is imputed using SNP genotypes and tested for association with a phenotype, are a popular post genome-wide association study (GWAS) approach for elucidating the role of gene expression in complex traits. Like gene expression, CpG methylation is known to play an important role in many traits. Here, we explored how well CpG methylation could be predicted from SNP genotypes, before using CpG methylation prediction models to conduct a methylome-wide association study (MWAS) for 30 complex traits.

We investigated how well three methods - ridge regression, elastic net and LASSO - predicted CpG methylation based on local SNP genotypes. Methylation was poorly predicted at most CpG sites, although for a small subset, it was relatively well predicted. Overall, elastic net and LASSO performed similarly, outperforming ridge regression. For the set of CpG sites where methylation was well predicted, we trained CpG methylation prediction models and applied them to GWAS summary data for 30 complex traits. We identified trait associations with predicted methylation, including at CpGs tagging genes in known GWAS risk loci. To further elucidate the function of these CpGs, we tested association between imputed methylation levels and imputed expression of nearby genes, finding CpG-gene associations and providing insight into the interplay between CpG methylation, gene expression and complex traits. We conclude that this approach represents a powerful method for investigating the role of CpG methylation in complex traits, which may help improve understanding of the biological mechanisms underlying GWAS signals.

34 | Allele-based association mapping of longitudinal phenotypes via binomial regression and Mahalanobis distance

Saurabh Ghosh

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

A complex end-point clinical trait is usually characterized by multiple quantitative precursors and hence, it has been argued that analyses of these correlated traits may be more prudent compared to analyzing the binary end-point trait itself. Moreover, since the values of such traits vary over time, considering phenotype data in a longitudinal framework is likely to lead to increased power in detecting genetic association. We (Majumdar et al., 2015) had developed two allele-level tests of association for analyzing multivariate phenotypes: one based on a Binomial regression model in the framework of inverted regression of genotype on phenotype in the lines of Multiphen (O'Reilly et al., 2012) and the other based on the Mahalanobis distance between the two sample means of vectors of the multivariate phenotype corresponding to the two alleles at a single-nucleotide polymorphism (SNP) in the lines of Lee et al. (2013). In this study, we explore the modification of these procedures to incorporate data on quantitative phenotypes in a longitudinal framework. Using extensive simulations, we evaluate the enhancement in the power of detecting association in comparison with cross-sectional phenotypes. We also explored for possible imputation strategies when phenotype values may be missing at certain time points. The advantage of using longitudinal data is also illustrated via association analyses carried out on triglyceride levels as part of Genetic Analysis Workshop 20.

35 | Whole genome sequencing analysis of the cardiometabolic proteome

Arthur Gilly[1,2]*, Young-Chan Park[2,3], Grace Png[1,2], Thea Bjornland[2,4], Lorraine Southam[1,2,5], Daniel Suveges[2,6], Sonja Neumeyer[1], Iris Fischer[1], Andrei Barysenka[1], N. William Rayner[2,7,8], Emmanouil Tsafantakis[9], Maria Karaleftheri[10], George Dedoussis[11], Eleftheria Zeggini[1,2]

[1]Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK; [3]University of Cambridge, Cambridge, UK; [4]Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway; [5]Wellcome Centre for Human Genetics, Oxford, UK; [6]European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SH, UK; [7]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK; [8]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK; [9]Anogia Medical Centre, Anogia, Greece; [10]Echinos Medical Centre, Echinos, Greece; [11]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Greece

The human proteome is a crucial intermediate between complex diseases and their genetic and environmental components, and an important source of drug development targets and biomarkers. Here, we conduct high-depth (22.5×) whole-genome sequencing (WGS) in 1,328 individuals to fully assess the genetic architecture of 257 circulating protein biomarkers of cardiometabolic relevance. We discover 132 independent sequence variant associations ($P < 7.45 \times 10^{-11}$) across the allele frequency spectrum, including 44 new *cis*-acting and 11 new *trans*-acting loci, all of which replicate in an independent cohort ($n = 1{,}605$, $18.4 \times$ WGS). We identify replicating evidence for rare-variant *cis*-acting protein quantitative trait loci for five genes, involving both coding and non-coding variation. We find causal links between protein biomarkers and cardiovascular, inflammatory and immune-related diseases. We construct and validate polygenic risk scores that explain up to 45% of protein level variation, and find significant correlation between genetically-predicted biomarker levels and cardiovascular disease risk in UK Biobank.

36 | Whole exome sequencing analysis of complex "Time-To-Event" outcomes in epilepsy patients

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

[1]Department of Biostatistics, University of Liverpool, Liverpool, UK; [2]School of Biological Sciences, University of Manchester, Manchester, UK; [3]School of Life Sciences, University of Glasgow, Scotland; [4]Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK

Epilepsy affects millions of individuals worldwide and uncontrolled seizures are a major burden on resources. Whilst seizures are effectively controlled by antiepileptic drugs (AEDs) in a majority (~70%) of individuals, there is significant heterogeneity in AED response. Since, genome-wide association studies (GWAS) have been largely unsuccessful so far in identifying genetic determinants of AED response, whole-exome sequencing (WES) which includes low-frequency and rare variants not typically analysed in GWAS is anticipated to be more successful in identifying variants explaining variability in AED response.

Statistical methods and software to test association with GWAS/WES variants are largely aimed at binary and quantitative traits, being the most common outcomes in disease genetics. However, when studying genetics of treatment response (pharmacogenetics), outcomes such as time to adverse drug reactions and treatment remission are often most important and there exists a lack of analysis tools aimed at such outcomes in both GWAS and WES settings. To address this analytical bottleneck, our group is focussed on developing methodologies and software capable GWAS and WES analysis with survival outcomes.

In this study, we analyse WES and GWAS data with time to event outcomes representing variable AED response from over 1300 individuals with epilepsy, applying software packages *SurvivalGWAS_SV* and *rareSurvival*, recently developed by our group specifically for survival endpoints. Results are compared to those obtained from applying more standard methods and software. The methodology and results for more complex survival outcomes that include competing risks to allow for different endpoints, will also be discussed.

37 | A *cis*-Mendelian randomization approach to evaluate genetic support for lipid-modifying drug targets

Maria Gordillo-Maranon[1]*, Magdalena Zwierzyna[1,3], Aroon D. Hingorani[1,3], Amand F. Schmidt[1,2,4], Chris Finan[1,3]

[1]Institute of Cardiovascular Science, University College London, London, UK; [2]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, the Netherlands; [3]UCL's BHF Research Accelerator centre, University College London, London, UK

Despite considerable interest including several drug development programs, the role of high-density lipoprotein cholesterol (HDL-C) in coronary heart disease (CHD) is still unclear.

In observational studies, circulating low-density lipoprotein cholesterol (LDL-C) exhibits a positive association with CHD risk. Treatment trials of LDL-lowering drugs (statins, ezetimibe and PCSK9 inhibitors), each targeting different mechanisms (HMGCR, NPC1L1 and PCSK9 respectively), show consistent reductions in CHD risk. Naturally occurring genetic variation in the genes encoding these targets, and at other loci identified by the Global Lipid Genetic Consortium, accurately instrument causal effects of LDL-C on CHD using data from the CardiogramPlusC4D Consortium in Mendelian randomisation analysis.

By contrast, circulating HDL-C exhibits a negative association with CHD risk in observational studies, but treatment trials of several HDL-C raising agents, for example, targeting CETP, have been negative, causing

uncertainty about whether this association is causal and if the development of drugs that raise HDL-C is a useful therapeutic strategy for CHD prevention. Mendelian randomisation analyses of HDL-C in CHD have also been equivocal.

Here, we use Mendelian randomisation for drug target validation (*cis*-MR) leveraging genetic variants associated with HDL-C, LDL-C and triglycerides (TG) in-and-around loci encoding drugged and potentially druggable targets. We identify a number drug targets that exhibit similar effects to existing lipid lowering targets. We also demonstrate that prior failure of certain CETP inhibitor drugs are molecule rather than target failures, and that modification of certain drugged/druggable proteins that affect HDL-C and TG metabolism is likely to offer new effective therapeutic mechanisms for CHD prevention.

38 | Trans-ethnic genome-wide association meta-analysis of >195,000 individuals reveal novel loci for kidney function decline

Mathias Gorski[1,2]*, on behalf of the CKDGen Consortium

[1]*Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany;* [2]*Department of Nephrology, University Hospital Regensburg, Regensburg, Germany*

Rapid kidney function decline is a strong risk factor for end-stage kidney disease, cardiovascular events, and early mortality. While a large number of genetic loci has been associated with kidney function cross-sectionally, few studies have investigated the genetic basis of longitudinal kidney function decline.

We conducted meta-analyses of 42 genome-wide association studies from different ancestries within the Chronic Kidney Diseases Genetics (CKDGen) Consortium and the UK Biobank (median follow-up time 4.8 years) to identify genetic loci associated with (a) decline of estimated glomerular filtration rate (eGFR) of ≥3 ml/min/1.73 m^2 per year ("Rapid3"; 34,873 cases, 107,081 controls), and (b) incidence of eGFR <60 ml/min/1.73 m^2 with a ≥25% eGFR drop from baseline ("CKDi25"; 19,621 cases, 175,524 controls).

We identified five genome-wide significant independent variants across four loci: 2 in the *UMOD-PDILT* locus, and 1 each in/near *PRKAG2*, *WDR72*, and *OR2S2*. All variants are in loci previously reported for cross-sectional eGFR, except the *OR2S2*, which is a novel locus. The analysis of 265 variants previously reported in association with cross-sectional eGFR identified two additional variants near *GATM* and *LARP4B* at $P < 0.05/265 = 1.89 \times 10^{-4}$. In the UK Biobank ($n = 487{,}409$) individuals with the highest genetic risk decile of Rapid3/CKDi25 variants had a 1.3/1.4-fold increased odds for chronic renal kidney failure compared with those in the lowest decile ($P = 2.07 \times 10^{-4}/2.02 \times 10^{-6}$).

In summary, we provide a catalog of genetic variants associated with kidney function decline. In-depth analysis of the mechanisms related to the identified variants may pave the way for the identification of underlying molecular mechanisms for renal deterioration.

39 | Effects of body mass index on the human proteome: Mendelian randomization study using individual-level data

Lucy J. Goudswaard[1,2,3]*, Joshua A. Bell[1,2], David A. Hughes[1,2], Klaudia Walter[4], Nicole Soranzo[4,5,6], Adam Butterworth[5,6], Ingeborg Hers[3], Nicholas J. Timpson[1,2]

[1]*Medical Research Council (MRC) Integrative Epidemiology Unit at the University of Bristol, Bristol, UK;* [2]*Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK;* [3]*School of Physiology, Pharmacology and Neuroscience, University of Bristol, UK;* [4]*Wellcome Sanger Institute, Hinxton, UK;* [5]*MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK;* [6]*NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK*

Variation in body mass index (BMI) is associated with cardiometabolic health outcomes such as diabetes and hypertension, but the mechanisms leading from BMI to disease risk are unclear. This study used proteomic data measured by SomaLogic from 2,737 healthy adults from the INTERVAL study to explore the effect of self-reported BMI on 3,622 unique plasma proteins using observational and genetically informed methods. Linear regression models were used, complemented by one-sample Mendelian randomization (MR) analyses. A BMI genetic risk score (GRS) comprised of 654 SNPs from a recent genome-wide association study (GWAS) of adult BMI was used in both observational and MR analysis. Observationally, BMI was associated with 1,576 proteins at $p < 1.4 \times 10^{-5}$ including leptin and sex hormone binding globulin (SHBG). The BMI-GRS was positively associated with BMI ($R^2 = 0.028$) but not with reported confounders. MR analysis indicated a causal association between each standard deviation increase in BMI and eight unique proteins at $p < 1.4 \times 10^{-5}$, including leptin (0.63 SD, 95% CI 0.48-0.79, $p = 1.6 \times 10^{-15}$) and SHBG (−0.45 SD, 95%

CI −0.65 to −0.25, $p = 1.4 \times 10^{-5}$). In addition, there was strong agreement in the direction and magnitude of observational and MR estimates ($R^2 = 0.33$). Finally, there was evidence that the genes which showed associations with BMI were enriched in cardiovascular disease. Altogether, this study provides evidence for a profound impact of higher adiposity on the human proteome and suggests that such protein alterations could be important mechanistic drivers of obesity-related cardiometabolic diseases.

40 | Evaluating the predictive performance of genetic and non-genetic scores in cardiovascular disease

Jasmine E. Gratton[1]*, Chris Finan[1], Reecha Sofat[2], Marta Futema[1], Amand F. Schmidt[1,3], Aroon D. Hingorani[1]

[1]Institute of Cardiovascular Science, University College London, London, UK; [2]Institute of Health Informatics, University College London, London, UK; [3]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands

The practical utility of polygenic scores in the clinic remains unclear. Recent landmark papers have shown that polygenic scores are good predictors of complex diseases such as coronary artery disease (CAD) and some cancers. However, these studies have failed to properly assess the performance of such scores in the clinic using the well-established evaluation metrics that medical tests typically undergo (e.g., disease stratification and disease discrimination).

To assess polygenic score utility in the clinic, we focused on eight disease outcomes: cardiovascular disease, CAD, atrial fibrillation, ischemic and haemorrhagic stroke, type 2 diabetes, moderate to severe chronic kidney disease, and end stage kidney failure. These diseases also have well established non genetic risk prediction tools for the British population from QResearch (QRISK3, QKidney, QStroke and QDiabetes).

We calculated polygenic scores for 341,993 white British participants of the UK Biobank cohort by selecting variants from CARDIoGRAM GWAS summary statistics. We generated weighted and unweighted scores using various *P* value (5e-04, 5e-05, 5e-06, 5e-07, and 5e-08) and linkage disequilibrium (0.8, 0.6, 0.4, 0.2, and 0.01) thresholds for variants, and selected the scores with the lowest Akaike information criterion.

Finally, we evaluated the predictive performance of the genetic, nongenetic and combined risk models for these eight outcomes to establish the usefulness and predictive ability of such scores in the clinic.

41 | Chances and challenges of machine learning based disease classification in genetic association studies illustrated on age-related macular degeneration

Felix Guenther[1,2]*, Caroline Brandl[1,3], Thomas W. Winkler[1], Veronika Wanner[1], Klaus Stark[1], Helmut Kuechenhoff[2], Iris M. Heid[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig Maximilian University of Munich, Munich, Germany; [3]Department of Ophthalmology, University Hospital Regensburg, Regensburg, Germany

Imaging technology and machine learning algorithms for disease classification set the stage for high-throughput phenotyping and promising new avenues for genome-wide association studies (GWAS). Despite emerging algorithms, there has been no successful application in GWAS so far. We establish machine learning based disease classification in genetic association analysis as a misclassification problem. To evaluate chances and challenges, we performed a GWAS based on a neural-network derived classification of age-related macular degeneration (AMD) in UK Biobank (images from 135,500 eyes; 68,400 persons). We quantified misclassification of automatically derived AMD in internal validation data (images from 4,001 eyes; 2,013 persons) and developed a maximum likelihood approach (MLA) to account for it when estimating genetic association. We demonstrate that our MLA guards against bias and artefacts in simulation studies. By combining a GWAS on automatically derived AMD classification and our MLA in UK Biobank data, we were able to dissect true association (*ARMS2*/*HTRA1*, *CFH*) from artefacts (near *HERC2;* OR = 1.26, *P* value = $4.16\text{x}10^{-16}$ in naïve analysis ignoring AMD misclassification and OR = 1.03, *P* value = .76 in MLA accounting for differential misclassification) and to identify eye color as relevant source of misclassification. On this example of AMD, we provide a proof-of-concept that a GWAS using machine learning derived disease classification yields relevant results and that misclassification needs to be considered in the analysis. These findings generalize to other phenotypes and also emphasize the utility of genetic data for understanding misclassification structure of machine learning algorithms.

42 | Measuring population substructure with the robust Jaccard index

Georg Hahn, Christoph Lange

Harvard T.H. Chan School of Public Health, Boston, MA, USA

Genetic association studies are a popular mapping tool; however, they can be vulnerable to confounding due to population substructure. Numerous methods have been proposed to address this issue -- Popular approaches rely on the genetic covariance matrix of the genotype data such as EIGENSTRAT or STRATSCORE, multi-dimensional scaling, or on the genomic relationship matrix. A relatively new approach relies on so-called unweighted and weighted Jaccard indices: The entries of the Jaccard matrix measure the set-theoretic similarity of the genomic data between all pairs of subjects, and they can be computed efficiently using only binary operations. Two recently proposed approaches in the literature for the analysis of rare variant data were shown to provide a higher resolution than the aforementioned popular approaches, however the Jaccard index suffers from the drawback that it is not robust - especially for small datasets, the index is not always defined. In this talk we provide a robust version of the Jaccard similarity index which inherits the computational efficiency and accuracy of the traditional Jaccard similarity measure while being defined for all input data.

43 | A fast and efficient smoothing approach to LASSO regression and an application to a genome-wide association study for COPD

Georg Hahn, Sharon M. Lutz, Nilanjana Laha and Christoph Lange

Harvard T. H. Chan School of Public Health, Boston, MA, USA

We consider solving a high dimensional linear regression problem, using LASSO to account for sparsity. Though the LASSO objective function is convex, it is not differentiable everywhere, making the use of gradient descent methods for minimization not straightforward. To avoid this technical issue, we suggest to use Nesterov smoothing of the LASSO objective function which enables us to compute closed form derivatives for efficient and fast minimization. The contribution of this study is threefold: (a) We propose an algorithm to efficiently compute estimates of the LASSO regression parameters; (b) we prove explicit bounds on the accuracy of the obtained estimates which show that the estimates obtained through the smoothed problem can be made arbitrary close to the ones of the original (unsmoothed) LASSO problem; (c) we propose an iterative procedure to progressively smooth the objective function which facilitates minimization and increases accuracy. A simulation section evaluates accuracy and runtime on simulated data. We illustrate the features of the approach by an application to a genome-wide association study for COPD, COPDGENE.

44 | LD score regression analysis of liver cancer using multi-traits from UK Biobank

Younghun Han[1,2]*, Catherine Zhu[1], Jinyoung Byun[1,2], Donghui Li[3], Manal Hassan[4], Christopher Amos[1,2], Hepatocellular Carcinoma Epidemiology Consortium

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA; [2]Department of Medicine, Section of Epidemiology and Population Sciences, Baylor College of Medicine, Houston, TX, USA; [3]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [4]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Estimating genetic correlations between complex phenotypes can provide valuable insights into etiologic studies. Major challenge for estimating genetic correlation from genome-wide association studies (GWAS) is the insufficient availability of individual-level genotype data and sample overlap among meta-analyses. Linkage disequilibrium (LD) score regression using GWAS summary statistics allow us to quantify the genetic correlation and SNP heritability between pairs of traits. This approach could also provide insights into disease co-development potentially associated with disease risk.

We calculated genetic correlation (rg) between hepatocellular carcinoma (HCC) from Hepatocellular Carcinoma Epidemiology Consortium and other phenotypes from UK Biobank, a large cohort study enrolled over 500,000 adults of age 40–69 years in 2006–2010.

We identified several phenotypes in blood counts, metabolic traits, alcohol consumption, smoking, diabetes, family history, and BMI associated with HCC at the significance level of $P < 5 \times 10^{-2}$. Alcohol intake frequency shows genetic correlation with HCC ($rg = 0.257$, $P = 4.03 \times 10^{-4}$). Both HDL cholesterol ($rg = -0.251$, $P = 3.77 \times 10^{-4}$) and apolipoprotein ($rg = -0.221$, $P = 1.48 \times 10^{-3}$) in the blood were negatively correlated with HCC while both C-reactive protein ($rg = 0.268$, $P = 7.50 \times 10^{-4}$) and haemoglobin concentration

(rg = 0.186, $P = 8.58 \times 10^{-4}$) were positively correlated. Anaemia due to iron deficiency showed the strongest correlation (rg = 0.478, $P = 3.11 \times 10^{-2}$). Diabetes showed the positive association (rg = 0.407, $P = 7.53 \times 10^{-5}$). Family history of diabetes in mother (rg = 0.365, $P = 4.94 \times 10^{-4}$) and in siblings (rg = 0.420, $P = 8.24 \times 10^{-4}$) was positively correlated.

LD score regression analysis provides an improved understanding of the genetic architecture between HCC and biomarkers. Mendelian randomization analyses can validate the potential causal direction between risk of HCC and phenotypes of interest.

45 | Genetics of primary open angle glaucoma differ in individuals of Caucasians and African ancestry

Michael A. Hauser[1,2] and Eyes of Africa Consortium

[1]Department of Medicine, Duke University, Durham, NC, USA, [2]Department of Ophthalmology, Duke University, Durham, NC, USA

Primary open angle glaucoma (POAG) has a complex genetic etiology, and manifests as a health disparity that disproportionately affects individuals of African descent. We have investigated the genetics of POAG in populations of African Ancestry by conducting a genome-wide association study (GWAS) in 8,494 POAG cases and 16,341 controls, including both continental African and African diaspora populations. We find genome-wide significant association ($P = 2 \times 10^{-13}$; OR = 1.19, 95% CI = 1.13-1.24) with a single-nucleotide polymorphism (rs59892895) in *APBB2*, a gene that is involved in the proteolytic processing of amyloid precursor protein (APP). This association is found only in African and African diaspora populations, with no association at this locus in Caucasian or Asian GWAS studies. These findings demonstrate the importance of genetic studies of African populations as the genetic architecture of disease can vary significantly by population. Proteolytic processing of APP produces amyloid beta (Aβ) peptides, which are toxic and which aggregate to form amyloid plaques, one of the neuropathological hallmarks of Alzheimer's disease (AD). Both the retina and primary visual cortex show increased Aβ staining in individuals carrying an *APBB2* risk allele. These findings were the first direct genetic evidence that the same mechanisms of APP processing are involved in neuronal cell death in both AD and glaucoma. Subsequent LD-Score analysis of POAG and AD in European datasets found a genome-wide genetic correlation of 0.14 (95% CI: 0.003-0.28; $P = .049$), suggesting that multiple loci contribute to both diseases. Thus analysis of African datasets can further inform disease associations in other populations.

46 | Evaluation of breast cancer polygenic risk score built on data from women of European decent in predicting breast cancer risk in Asian women

Weang-Kee Ho[1,2]*, Nasim Mavaddat[3], Mei-Chee Tai[2], Breast Cancer Association Consortium, Woon-Puay Koh[4,5], Nur Aishah Mohd Taib[6], Mikael Hartman[7], Douglas F. Easton[3,8], Soo-Hwang Teo[2,6], Antonis C. Antoniou[3]

[1]School of Mathematical Sciences, Faculty of Science and Engineering, University of Nottingham Malaysia, Jalan Broga, Semenyih, 43500 Selangor, Malaysia; [2]Cancer Research Malaysia, 1 Jalan SS12/1 A, Subang Jaya, 47500 Selangor, Malaysia; [3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, CB1 8RN Cambridge, UK; [4]Health Services and Systems Research, Duke-NUS Medical School; [5]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, 12 Science Drive 2, #10-01, 117549 Singapore; [6]Department of Surgery, Faculty of Medicine, University of Malaya, Jalan Universiti, Kuala Lumpur 50630 Kuala Lumpur; [7]Department of Surgery, National University Hospital and NUHS, 1E Kent Ridge Rd, 119228 Singapore; [8]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, 2Worts' Causeway, CB1 8RN, Cambridge, UK

Risk profiles based on combination of low penetrance but common breast cancer susceptibility susceptibility single nucleotide polymorphisms (SNPs), summarised as polygenic risk scores (PRS), have been shown to predict breast cancer risk in European women. Previous efforts in Asian studies have focused on the development of Asian-specific PRS, and have been limited by small sample size. Given the difficulties of defining population-specific PRS, a more practical question is whether the PRS developed using data from women of European ancestry is predictive of risk for women of Asian ancestry. In this study, we independently evaluated the best performing PRSs (313-SNP PRSs) for European-ancestry women using data from 17,262 breast cancer cases and 17,695 controls of Asian ancestry from 13 case-control studies, and 10,255 Chinese women from a prospective cohort (413 incident breast cancers). Compared to women in middle quintile of the risk distribution, women in the highest 1% of PRS distribution had ~2.7-fold risk and women in the lowest 1% of PRS distribution had ~0.4-fold risk of developing breast cancer. The estimated breast cancer odd ratio (OR) per SD of the PRS and the

discriminatory accuracy, measured by area under the receiver operating characteristic curve (AUC), was 1.52 (95% CI = 1.49-1.56) and 0.613, respectively. We showed that European-ancestry based PRS was predictive of breast cancer risk in Asians and can help in developing risk-stratified screening programmes in Asia.

47 | Statistical integration of methylation, transcriptome and proteome data in cell lines

Jeanine J. Houwing-Duistermaat[1,2]*, Said el Bouhaddani[2], Hae Won Uh[2]

[1]Department of Statistics, University of Leeds, Leeds and Alan Turing Institute, London, UK; [2]Department of Biostatistics and Research Support, Julius Center, University Medical Center Utrecht, Utrecht, Netherlands.

We have access to DNA-methylome, miRNome, transcriptome and proteome data measured in cell lines that show protein aggregation and in negative controls. Standard analysis of these data set resulted in significant genes per data set but there was no overlap across the datasets. A combined analysis can detect relevant features shared by all datasets. The challenges are the high dimensionality ($p > N$) of the data and the platform-specific properties.

Several algorithmic approaches have been proposed which decompose the datasets into joint and residual parts, for example multi-group PLS (mg-PLS) and MINT. The joint components capture consistent effects of the molecular measurements on the outcome across all datasets. The optimal components are obtained by iteratively maximising the covariance between the molecular measurements and a dummy matrix based on the binary outcome. The drawbacks of these methods are a lack of platform-specific parts, absence of a proper model for the binary outcome, and overfitting when data are high dimensional. We propose a novel Probabilistic multi-group OPLS (mg-POPLS) model for multiple datasets in terms of joint, platform-specific and residual parts. Systematic differences between the omics data are incorporated by including specific parts. The outcome is modelled via these components by using a latent probit model. The components and coefficients are estimated with maximum likelihood.

An extensive simulation study will be conducted to investigate the performance of mg-POPLS compared to mg-PLS. We apply the mg-POPLS method to the omics data to detect the most relevant features for separating case from control cell lines.

48 | Application of Bayesian networks to rheumatoid arthritis and intermediate biological marker data

Richard Howey*, Heather J. Cordell

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

There is an increasing interest in using causal analysis methods to move beyond initial analysis strategies, such as those applied in the context of genome wide association studies. One approach is to use Bayesian Networks (BN) which allow the relationships between biological and phenotypic data to be investigated in an exploratory manner, and are particularly suited to data sets with many variables. We consider a recent study of rheumatoid arthritis (RA) and the possible influences of methylation and gene expression in CD4+ T cells and B cells. In this study the causal inference test (CIT) was applied using many candidate variable triplets consisting of SNP, methylation and DNA expression data, focussing on SNPs with prior evidence of genetic association with RA. The results from this study suggested many genes where DNA methylation may mediate RA genetic risk. We applied BN analysis to the same data set, exploring the candidate variable triplets as well as additional individual data for sex, age and RA status. We also investigated the benefit of modelling multiple variables simultaneously in a large, complex network. Our results showed replication of only some of the findings of the original study. This is not surprising as analytic methods have different assumptions and may be affected by different features of the data. This example highlights the benefit of exploring different analysis methods even for the same data set.

49 | Haplotype analysis of *BRCA1*/*BRCA2* variants in Korean patients with breast cancer

Won Kyung Kwon[1], Hyeok-Jae Jang[2]*, Ja-Hyun Jang[1], Jeong Eon Lee[3], Yeon Hee Park[4], Jong-Won Kim[1,2]

[1]Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea; [2]Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Korea; [3]Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea; [4]Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Seoul, Korea

Germline mutations of *BRCA1/BRCA2* genes occupies around 10% among patients with breast cancer. These

genes have a lot of polymorphisms and pathogenic variants. Founder effects of *BRCA1/BRCA2* genes have been reported in European, North American and Chinese populations but haplotype analyses with polymorphism rarely have done. We reviewed the *BRCA1/2* variant data from 2012 to 2019 at Samsung Medical center, Seoul, Korea. The total number of patients were 5,090 for *BRCA1* and 5,085 for *BRCA2*, respectively, and pathogenic/likely pathogenic variants were found in 296 and 212, respectively. Among them, we selected the recurrent variants found more than five times and analyzed them with haplotype analyses. Fourteen mutations in *BRCA1* and 7 mutations in *BRCA2* were included. Computational haplotype analysis was constructed with 397 healthy individuals from Korean Reference Genome, using the software PHASEv2.1.1.

The variant *BRCA2* c.1399A > T were detected in unrelated 20 patients which was known as putative founder mutation in Korean. The analysis was confirmed using 24 SNPs and showed the common haplotype block along 5,159 bp which harboring the *BRCA2* mutation c.1399A > T (P value = .0005, OR = 146.49, 95% CI: 8.86-2423.34). Our findings revealed that *BRCA2* c.1399A > T mutation was prevalent (10.3% of the pathogenic variants in *BRCA2* gene) in this clinic based cohort of Korean, and shared founder haplotype. The other recurrent mutations would be presented to reveal the founder mutations in Korean.

50 | Is the association between CYP2A6 and lung cancer mediated through smoking behavior?

Yon Ho Jee[1]*, Suhyun Lee[2], Sun Ha Jee[2,3], Peter Kraft[1,4]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; [2]Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea; [3]Department of Epidemiology and Health Promotion, Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea; [4]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

Genetic variation in cytochrome P450 2A6 (CYP2A6) gene has been found to influence both smoking intensity and lung cancer risk. Genetic studies have demonstrated that the CYP2A6 locus has been most strongly associated with cigarette per day (CPD) in East Asian populations. However, few studies have examined whether the association between CYP2A6 variants and lung cancer is direct or is mediated by pathways related to smoking behavior in these populations. We used inverse odds ratio weighting and logistic regression to calculate odds ratios (OR) and 95% confidence intervals (CI) for estimated total effects (OR^{TE}) of CYP2A6 variants on lung cancer risk, and estimated effects operating through CPD (natural indirect effect; OR^{NIE}) or through paths independent of CPD (natural direct effect; OR^{NDE}) among 209 lung cancer cases and 6882 controls in the Korean Cancer Prevention Study-II Biobank ($N = 16{,}995$). The strongest effect on lung cancer was with rs11878604 ($OR^{TE} = 1.30$ (95% CI: 1.12-1.52)), which decomposed into an OR^{NIE} of 1.01 (95% CI: 0.95, 1.05) and an OR^{NDE} of 1.29 (1.11, 1.51) per C allele. For changes from 0 to 1C allele and 0–2C alleles, respectively, analysis yielded OR^{NDE} of 1.19 (95% CI: 0.79-1.79) and 1.61 (95% CI: 1.09-2.36) and OR^{NIE} of 1.00 (95% CI: 0.86-1.17) and 1.01 (95%CI: 0.86-1.19). These analyses indicate that the association of CYP2A6 variants with lung cancer operates primarily through pathways other than smoking behavior, consistent with previous mediation results on variants near CHRNA5 in European-ancestry populations.

51 | Multi-ethnic genome-wide association study of acute lymphoblastic leukemia

Soyoung Jeon[1]*, Adam J. de Smith[1], Ivo S. Muskens[1], Catherine Metayer[2], Xiaomei Ma[3], Joseph L. Wiemels[1], Charleston Chiang[1]

[1]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; [2]School of Public Health, University of California Berkeley, Berkeley, CA; [3]Yale School of Public Health, New Haven, CT

Acute Lymphoblastic Leukemia (ALL) is a leading cause of childhood mortality in the United States. Genome-wide association studies (GWAS) focusing primarily on European ancestry populations have identified 16 risk loci for childhood ALL. With one exception in *CDKN2A*, these known risk loci are common among populations but do not fully explain heritable risk of ALL, nor the differences in risk between different populations; Latino has the highest incidence of ALL in the world. To investigate potentially low frequency variation affecting the risk of ALL in diverse populations, we conducted a multi-ethnic GWAS using the California Cancer Records Linkage Project case-control study supplemented with additional control individuals from Kaiser Permanente Genetic Epidemiology Research in Adult Health and Aging cohort. In total, we included 76,317 individuals in our study, including 2,191 African Americans, 10,288 Latino Americans, 58,503 Non-Latino Whites and 5,335 East and Southeast Asians. We focused on ~6,500,000 imputed variants that passed QC with minor allele

frequency (MAF) > 0.005. We replicated all previously known loci (max $P = 0.01$) and identified potentially novel loci with lower MAF than most current known loci, including 11 in Non-Latino Whites (mean MAF = 0.095), two in Latinos (mean MAF = 0.148), and one in Asians (MAF = 0.164). Multi-ethnic meta-analysis, laboratory assay validations, and replications are underway to verify these putative loci and assess their contribution to ethnic differences in ALL risk. In summary, our study provides novel associations for childhood ALL risk that are ethnic-specific or shared across populations and highlights the importance of ancestral diversity in GWAS.

52 | A Bayesian hierarchical model for estimating covariate effects on 5-methylcytosine and 5-hydroxymethylcytosine levels in oxy-bisulfite treated DNA

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

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

**Presenting author:* lai.jiang@mail.mcgill.ca

[+]*Corresponding author:* celia.greenwood@mcgill.ca

5-Hydroxymethylcytosine (5hmC) is a transient methylation state associated with demethylation of cytosines from 5-methylcytosine (5mC) to the unmethylated state. Also, this state has been associated with transcription regulation, particularly in the brain.

The presence of 5hmC methylation can be inferred by a paired experiment involving bisulfite treatment and oxidative-bisulfite treatment on the same sample, followed by methylation analysis using a platform such as the Illumina EPIC array. However, existing analysis methods for such data are not ideal. Most approaches ignore the correlation between the two experiments, and also ignore any imprecision associated with damage to the DNA from the additional treatment. We propose a hierarchical Bayesian model to simultaneously estimate 5mC/5hmC signals and any effects on these signals due to covariates or phenotypes, while accounting for potential effect of DNA damage and the dependencies induced by the experimental design. Simulations show that our method has valid type 1 error and better power than a range of alternative methods—including the popular OxyBS method and linear or generalized linear models on transformed proportions. Many alternatives suffer from hugely inflated type 1 error for inference on 5hmC proportions. We also apply this method to explore genome-wide associations between 5mC/5hmC methylation levels and cause of death in post-mortem prefrontal cortex brain tissue samples. Together, these analyses indicate that our method, with a comprehensive modeling framework for DNA damage and covariate effects, is a useful tool to reveal phenotypic associations with 5 mC/5 hmC levels.

53 | The role of brain-derived neurotropic factor (BDNF) genetic variants in exercise training

Rong Jiang[1]*, Kim M. Huffman[2,3], Elizabeth R. Hauser[2,4], Janet L. Huebner[2], Monica J. Hubal[5], Redford B. Williams[1], Ilene C. Siegler[1], William E. Kraus[2,3]

[1]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, USA; [2]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, North Carolina, USA; [3]Department of Medicine, Duke University Medical Center, Durham, North Carolina, USA; [4]Department of Biostatistics, Duke University Medical Center, Durham, North Carolina, USA; [5]Department of Kinesiology, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA

Both acute and chronic exercise stimulate the release of neurotransmitters and neurotrophins, thereby promoting structural and functional brain plasticity. A critical activity-responsive neurotrophin, brain-derived neurotropic factor (BDNF), has a genetic variant Val66Met that is associated with psychiatric disorders and CVD risk factors. However, little is known regarding its effects in response to exercise training. Using STRRIDE (Study of a Targeted Risk Reduction Intervention through Defined Exercise) subjects, we examined the variant's association with training adherence, post-training changes in plasma and muscle BDNF, and training parameters (amount, intensity and mode). As compared to Val/Val, Met was associated with greater training adherence ($n = 431$, $\beta = 3.94$, CI= 1.01-6.86, $P = 0.008$) independent of training parameters. In a subset samples with available data, Val66Met was not associated with muscle *BDNF* transcript or protein concentrations at baseline; yet, Met was related to lower muscle *BDNF* gene expression post-training ($n = 47$, $P < 0.045$). The SNP impact on changes in both muscle and plasma BDNF concentrations varied by training parameters. Irrespective of exercise training program and SNP, greater post-training muscle *BDNF*

gene expression was associated with higher exercise program adherence ($n = 40$, $P < 0.025$). Our results imply that Val66Met doesn't impact BDNF concentrations in the sedentary state, but has variant-specific and mode-specific effects on BDNF concentrations with exercise training. Further, Val66Met is related to exercise training program adherence. Future study with larger sample sizes are needed to understand the complex relationships among *BDNF* genetic variants and BDNF tissue concentrations, and training programs. This may lead to more personalized exercise training programs targeting *BDNF*-dependent outcomes.

54 | Exploring the total and direct effect of 14 triglyceride-containing lipoprotein sub-fraction metabolites and coronary heart disease: A two-sample Mendelian randomisation analysis

Authors: Roshni Joshi[1]*, Pimphen Charoen[2,3], Fotios Drenos[1], S Goya Wannamethee[4], A Floriaan Schmidt[1,5], Aroon D Hingorani[1], on behalf of the UCLEB Consortium

[1]Institute of Cardiovascular Science, University College London, UK; [2]Integrative Computational BioScience (ICBS) Center, Mahidol University, Thailand; [3]Department of Tropical Hygiene, Faculty of Tropical Medicine, Mahidol University, Thailand; [4]Department of Primary Care & Population Health, Faculty of Population Health, University College London, UK; [5]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, the Netherlands

**The presenting author*

Background: Observational studies show triglyceride-containing lipoprotein sub-fractions are associated with an increased risk of coronary heart disease (CHD). It remains unclear whether this association marks a casual, and independent effect, of highly correlated low-density lipoprotein cholesterol (LDL-C) and high-density lipoprotein cholesterol (HDL-C). In this study we examine extent of the total, and direct LDL-C and HDL-C independent associations of 14 triglyceride sub-fractions using genetic evidence.

Methods: We developed genetic instruments for total and 14 triglyceride sub-fractions from a genome wide association study (GWAS) of NMR metabolites in 45,031 individuals. SNPs were selected at significance threshold $P < 5 \times 10^{-6}$ and linkage disequilibrium (LD) clumped at r-squared 0.1. We determined the association of these genetic variants with CHD, and LDL-C and HDL-C from CARDIoGRAM and the Global Lipids Genetics Consortium, respectively. Mendelian randomisation (MR) estimators (with and without correction for horizontal pleiotropy) were used to determine 1) the total effect of 14 triglyceride sub-fractions on CHD, 2) the LDL-C and HDL-C independent effects.

Results: In univariate analysis, total and six triglyceride sub-fractions were associated with CHD. Under Egger regression, total and five triglyceride sub-fractions remained significant (*P value* < 0.05). In multivariable MR analysis, higher total and six triglyceride sub-fractions were associated with CHD independently of LDL-C and HDL-C (OR range: 1.33–1.66). The largest effect was observed for triglyceride in extra small VLDL (OR:1.66 [95% CI 1.11–1.32]).

Conclusion: These findings support a causal direct, and LDL-C and HDL-C independent, effect of triglyceride sub-fractions on CHD risk.

55 | Contribution of rare variant(s) to the genetic risk score

Qing Wu[1,2], Jongyun Jung[1,2]*

[1]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, Nevada, USA; [2]Department of Environmental & Occupational Health, School of Public Health, University of Nevada, Las Vegas, Nevada, USA

A Genetic Risk Score (GRS) aggregate the effects of many genetic variants across the human genome into a single score. It has been shown that this score can have predictive value for multiple common diseases. However, the effects of these genetic variants are from the common variants (e.g. minor allele frequency [MAF] > 1%). It is unknown that how the rare variants (e.g. minor allele frequency [MAF] < 0.5%) are contributing to GRS today. Thus, we implemented the rare variant analysis to The Osteoporotic Fractures in Men Study (MrOS) whether we can find any rare variants and if so, how these rare variants are contributing to the calculation of GRS. To find the rare variants, we implemented the gene annotation with ANNOVAR as well as sequence kernel association test (SKAT) to the phenotype of well-known Bone Mineral Density (BMD). And then we used Plink of the set-based association test to find the effect size and -value for the significant SNPs. We compared GRS_C (common variant only) and GRS_T (common + rare variants) for each individual patient. GRS_T showed the slightly higher score than GRS_C in our analysis (the difference is 0.14 ~ 0.20 for each individual). We also compared these differences with the dependent two pair - test. The -value of this test was significantly lower. Our results demonstrated that at least the rare variants are contributing to GRS_T.

56 | Phenome-wide association study of a comprehensive health check-up database in 10,349 Korean population: Clinical application & transethnic comparison

Eun Kyung Choe, MD, PhD[1,2*§], Manu Shivakumar, MS[1*], Anurag Verma, PhD[3], Shefali Verma, PhD[3], Seung Ho Choi, MD[4], Joo Sung Kim, MD, PhD[4], Dokyoon Kim, PhD[1,5§]

[1]Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [2]Department of Surgery, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, South Korea; [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [4]Department of Internal medicine, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, South Korea; [5]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA

**EKC and MS equally contributed as 1st author*

[§]DK and EKC equally contributed as corresponding author

The rapid growth of genetic data linked to the electronic health records has increasingly leveraged phenome-wide association studies (PheWASs). However, it has been imposed challenges such as using ICD billing codes for phenotype definition, the imbalanced ethnicity in the study population and restrictive application to clinical practice. In this study, we performed the PheWAS using a comprehensive health check-up database (136 phenotypes, including questionnaire, laboratory, imaging, and functional tests) in 10,349 Koreans. We systematically compared loci associated with phenotypes to the PheWAS results from UK Biobank (UKBB) and Biobank Japan Project (BBJ) in aspects of clinically interpretable applications. There were 52 phenotypes overlaps with BBJ (42 phenotypes replicated) and 101 with UKBB (60 phenotypes replicated). In the comparison between Korean and BBJ, whereas activated partial thromboplastin time (23.12% of overlap) and bilirubin (20.61%) had high overlapping loci, loci associated with ophthalmic or cerebrovascular system, smoking, hepatitis C, nephrolithiasis, gastric cancer, bone density were mutually exclusive. In the comparison between Korean and UKBB, the overlap was less than the comparison with BBJ and the highest overlap ratio was 8.783% in fatty liver. Noticeably, Korean had novel variants associated with body mass index (BMI) comparing to both other populations. It suggests that variants associated with TERF2IP, ATRNL1 and BANF1 might be novel Korean specific variants for BMI. In the analysis in ethnicity and nationality difference, our study shows that there are phenotypes that are common or exclusive in genetic associations that should be taken into consideration to perform population based clinical study.

57 | Expanded clustering of type 2 diabetes genetic loci using high throughput approach

Hyunkyung Kim[1,2*], Marcin von Grotthuss[2], Josep Mercader[1,2,3], Jaegil Kim[6], Jose Florez[1,2,3,4], Alisa Manning[2,4,5], Miriam Udler[1,2,3,4]

[1]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, USA; [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; [3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA; [4]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA; [5]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, Massachusetts, USA; [6]GlaxoSmithKline, Cambridge, Massachusetts, USA

Improved understanding of disease-causing pathways for Type 2 diabetes (T2D) may lead to novel therapeutic targets and individualized treatment. Rapid expansion in the number of T2D genetic loci over the past decade can help identify pathways via cluster analysis.

We developed an automated pipeline for clustering of T2D genetic loci starting with genome-wide association study (GWAS) summary statistics from 6 European T2D studies. The pipeline 1) extracts variants reaching genome-wide significance, 2) replaces multi-allelic, ambiguous (A/T, C/G), or low-trait count SNPs with appropriate proxies, and 3) filters for independent signals using $r^2 = 0$. Traits with available summary GWAS data in European populations were included if they met a minimum Bonferroni p-value across the selected variants. This pipeline generated a matrix of associations for 258 independent T2D variants and 107 traits to which we applied Bayesian Non-negative Matrix Factorization (bNMF) clustering.

We identified nine robust clusters of T2D loci, five of which overlap with our previously published analysis of 94 loci. Three clusters indicated variant-trait associations related to insulin deficiency, while another three clusters displayed insulin resistance-related features including: obesity, lipodystrophy, and liver-lipid metabolism. Novel clusters identified in this analysis were related to beta-cell function, lipoprotein-A, and blood traits. Polygenic scores for clusters were associated with distinct clinical outcomes including coronary artery disease, ischemic stroke, and chronic kidney disease.

Our approach expands upon previous clustering work of T2D loci and allows for efficient updating of additional GWAS results. This method can also be readily applied to other diseases to identify key pathways.

58 | Genetic interactions between *ABO* blood group alleles and *FTU2* and *FUT3* modified the risk of pancreatic cancer

Jihye Kim[1]*, Chen Yuan[2], Laufey T. Amundadottir[3], Alison P. Klein[4,5], Harvey A. Risch[6], Brian M. Wolpin[2], and Peter Kraft[1,7]

[1]Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA; [3]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; [4]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA; [5]Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, Maryland, USA; [6]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, USA; [7]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

Pancreatic adenocarcinoma (PDAC) risk has been found to be greater among individuals with non-O blood types than those with O blood type. However, it has not been fully characterized how *FUT2*, determining secretor status, and *FUT3*, determining Lewis antigens, that biologically work together with *ABO* influence the effects of ABO blood groups on PDAC. We examined genetic interactions among 3,904 cases and 3,601 controls in a large pancreatic cancer consortium (PANC4) by utilizing genetic data to ascertain ABO blood groups (rs505922 and rs8176746), secretor status (rs601338), and Lewis antigens (rs812936, rs28362459, and rs3894326). We used haplotypes of the two *ABO* SNPs to determine *ABO* blood alleles (O, A, and B). And, we divided into two groups (O and non-O blood groups) and made a continuous variable by counting the number of non-O alleles (0, 1, and 2). Individuals with the A/A genotype of rs601338 were defined as non-secretors and others were defined as secretors. Using haplotypes of the 3 *FUT3* SNPs, we made two Lewis groups; normal active antigen group and semi or no active antigen group. Multivariable logistic regression was used to estimate ORs and 95% CIs of PDAC adjusting for age and sex. We examined interactions by testing each product term between ABO and secretor or Lewis groups as well as between secretor and Lewis groups individually. We found positive interactions of the semi- or non-active Lewis antigen group with non-O blood group as well as increased number of non-O alleles (nominal $P = .022$ and .008, respectively).

59 | The human urine microbiome in type-2 diabetes mellitus from KARE cohort study

Kang Jin Kim[1], Sang-hum Lee[2], Sang-Chul Park[3] and Sungho Won[1,3,4]

[1]Department of Public Health Science, Graduate School of Public Health, Seoul National University, Seoul, South Korea; [2]Department of Medical Consilience, Graduate School, Dankook University, Seoul, South Korea; Institute of Health and Environment, Seoul National University, Seoul, South Korea; [4]Interdisciplinary Program for Bioinformatics, College of Natural Science, Seoul National University, Seoul, South Korea

Motivation: Advances in sequencing technology have revealed the role of intestinal microflora in the mechanism of type 2 diabetes (T2DM). Research showing the wide distribution of microorganisms throughout the human body, even in the blood, has motivated the investigation of the dynamics of the intestinal microflora throughout the human body. In particular, since intestinal microbial-derived EVs affect glucose metabolism by inducing insulin resistance, extracellular vesicles (EV), a lipid bilayer structure secreted from intestinal microorganisms, have recently attracted attention. Recently, intestinal permeation linked to T2DM is associated with the interaction between intestinal microorganisms and leaky intestinal epithelium, which leads to increased inflammation of macromolecules such as lipopolysaccharides from the membranes of microorganisms, leading to chronic inflammation.

Results: In this article, we first investigate KARE cohorts, the rural community of Ansung and the urban community of Ansan, in South Korea. Both cohorts began in 2001 as a part of the Korean Genome Epidemiology study. From 315 participants in 2013, 2015 and 2017, urine samples are obtained. We found the significant association of GU174097_g, unclassified Lachnospiraceae with T2D group. GU174097_g was positively associated with 60 min insulin level. This implicates that GU174097_g can help the cycle of homeostasis of insulin level in human body. In addition to this, the coefficient of fasting insulin level and HOMA-IR, even if it is not significant, is negatively correlated which can mean that it mitigate insulin resistance and this can decrease the risk of T2D.

60 | Multi-omics association tests with matched samples with replacement

Nam-Eun Kim[1]*, Kangjin Kim[1], Sangcheol Park[2], Sungho Won[1,2,3]

[1]Department of Public Health Sciences, Seoul National University, Seoul, Korea; [2]Institute of Health and Environment, Seoul National University,

Seoul, Korea; [3]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea

Last few decades we experienced the rapid advances in high-throughput technologies and effect of the interplay among the multiple layers of regulations including genetics, epigenetics, transcriptomics, metabolomics, and proteomics on the human disease have been investigated. However multi-omics data have diverse data errors, including sample swapping and mis-labeling, and unless those were correctly adjusted, association analyses become invalidated. In such case, matched case-control study can be utilized to minimize the effect of data errors. However statistical analyses with matched case-control studies suffer from the insufficient sample size, and matched case-control samples with replacement can be often utilized. In this article, we proposed the new statistical methods with multi-omics data for matched samples with replacement. The proposed method was evaluated with simulated data, and we found that the proposed method successfully controlled the type-1 error rates. The proposed method was applied to the microbiome data and we found the candidate operational taxonomic unit associated with the host diseases.

61 | Fast kernel-based rare-variant association tests integrating variant annotations from deep learning

Stefan Konigorski[1,2]*, Remo Monti[1,2,3], Pia Rautenstrauch[1,4], Christoph Lippert[1,2]

[1]Digital Health & Machine Learning Group, Hasso Plattner Institute for Digital Engineering, Potsdam, Germany; [2]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, USA; [3]Computational Regulatory Genomics Group, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany; [4]Department of Computer Science, Institute for Biomedical Informatics, University of Tübingen, Germany
**Presenting author*

In recent years, deep learning has enabled the accurate prediction of the function of DNA- and RNA-sequences based on their nucleotide sequences alone. In another line of research, kernel-based tests have been established as powerful association tests of rare genetic variants. Here, we combine these two streams of research and present *seak* (*se*quence *a*nnotations in *k*ernel-based tests): a fast implementation of flexible set-based genetic association tests that include variant effects on intermediate molecular traits, correcting for family and population structure. This can be interpreted as testing for genetic effects that are mediated or moderated by intermediate molecular traits.

In the analyses, first, we evaluated the proposed tests in Monte Carlo simulation studies of data based on DAGs mimicking different biological pathways. Second, we used the convolutional neural network DeepRiPe to predict the binding affinity to 200 RNA-binding proteins of 4 million single nucleotide variants from whole exome-sequencing experiments of 50,000 individuals in the UK Biobank. Then, we incorporated these predictions as well as predicted splice junctions from SpliceAI in genome-wide association studies with BMI and type II diabetes.

Incorporating variant annotations in the association tests yielded more genome-wide significant loci compared to the standard kernel-based tests SKAT and SKAT-O. Especially integrating disease-relevant RNA-binding proteins such as ELAVL-1 yielded novel candidate genes that would not have been identified otherwise. These results, backed-up by the simulation study results, indicate that *seak* can increase the power of association tests and generate as well as test concrete

hypotheses about underlying biological disease processes.

62 | Confidence intervals and their coverage probabilities for predictions by random forests

Diana Kormilez[1]*, Björn-Hergen Laabs[1], Inke R. König[1,2]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany; [2]German Center for Cardiovascular Research, Germany

Random forests are a popular supervised learning method. Their main purpose is the robust prediction of a phenotype based on a learned set of rules. To evaluate the precision of predictions, their scattering and distributions are important. To quantify this, 95% confidence intervals for the predictions can be generated using suitable variance estimators. However, these estimators may be biased, and the resulting confidence intervals can be evaluated by estimating coverage probabilities through simulations. Therefore, the aim of our study was to examine coverage probabilities for two popular variance estimators for predictions made by random forests, the infinitesimal jackknife (Wager and collaborators, 2014) and the fixed-point based variance estimator (Mentch and Hooker, 2016). We performed a simulation study considering different scenarios with varying sample sizes, various signal-to-noise ratios and differing variable types with a focus on SNP's. Our results show that the coverage

probabilities based on the infinitesimal jackknife are too low for small data sets and small random forests, but too high if based on the fixed-point variance estimator. However, a growing number of trees yields decreasing coverage probabilities for both methods. We additionally apply all methods to a data set predicting coronary artery disease from candidate genetic variants. In conclusion, the relative performance of the variance estimation methods depends on the hyperparameters used for training. Thus, the coverage probabilities can be used to evaluate how well the hyperparameters were chosen and whether the data requires more pre-processing.

63 | Investigating causal effects of genetic variants for Alzheimer's disease in the UK Biobank

Roxanna Korologou-Linden[1,2]*, Emma L. Anderson[1,2], Laura D. Howe[1,2], Louise A. C. Millard[1,2,3], Yoav Ben-Shlomo[2], Dylan M. Williams[4,5], George Davey Smith[1,2], Evie Stergiakouli[1,2†], Neil M. Davies[1,2†]

[1]Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, UK; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Barley House, Oakfield Grove, Bristol, UK; [3]Intelligent Systems Laboratory, Department of Computer Science, University of Bristol, Bristol, UK; [4]MRC Unit for Lifelong Health and Ageing at UCL, University College London, London, UK; [5]Department of Medical Epidemiology & Biostatistics, Karolinska Institutet, Stockholm, Sweden

Introduction: Observational studies for Alzheimer's disease (AD) have reported conflicting evidence for potential modifiable risk factors, possibly due to bias from confounding, selection bias and reverse causation. Genetic studies have identified SNPs associated with late-onset Alzheimer's disease, all exerting low to modest effects (except for those in the apolipoprotein E gene). We perform a phenome-wide association study (pheWAS) to investigate the effects of a polygenic risk score (PRS) for AD on a wide range of phenotypes, minimizing the bias of a prior hypothesis and confounding present in most observational studies.

Methods: We split the UK Biobank sample ($n = 334{,}968$) into three equal age tertiles (39-53, 53-62, and 62-72 years) and calculated a weighted PRS ($p \leq 5\text{x}10^{-8}$), using weights from a meta-analysis of GWAS of Alzheimer's disease. We investigated if there is an age-dependent trend in the association between the PRS for AD and the array of phenotypes. We followed up top hits in a two-sample bidirectional Mendelian randomization (MR) framework, to examine whether these phenotypes have causal effects on risk of AD.

Results: The PRS was associated with medical history, parental health factors, cognitive and brain-related measures, lifestyle factors, as well as biological and physical measures. Using MR, we replicated established factors associated with AD (e.g., fluid intelligence score) and identified novel phenotypes (e.g., sleep).

Conclusions: Our pheWAS found evidence that the PRS for AD was associated with 165 phenotypes, but MR strongly suggests that these are either effects of the disease process (reverse causation) or due to horizontal pleiotropy.

64 | Establishing polygenic risk score reporting standards and a polygenic score catalog to improve validation, interpretation and reproducibility

Peter Kraft[1], Hannah Wand, MS[2], Samuel A. Lambert[3], Cecelia Tamburro[4], Jaqueline MacArthur[3], Michael A. Iacocca[2], Catherine Sillari[4], Jack O'Sullivan[2], Deanna Brockman[5], Birgitt Schuele[2], Eric Venner[6], Iftikhar J. Kullo[7], Robb Rowley[8], Mark McCarthy[9], Antonis C. Antoniou[10], Douglas F. Easton[10], Robert A. Hegele[11], Amit Khera[5], Charles Kooperberg[12], Karen Edwards[13], Kelly E. Ormond[2], Muin J. Khoury[14], A. Cecile J. W. Janssens[15], Katrina A. B. Goddard[16], Michael Inouye[17], Genevieve Wojcik[18]

[1]Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; [2]Stanford University, Stanford, California, USA; [3]European Bioinformatics Institute, Hinxton, UK; [4]National Human Genome Research Institute, Bethesda, Maryland, USA; [5]Massachusetts General Hospital, Boston, Massachusetts, USA; [6]Baylor College of Medicine, Houston, Texas, USA; [7]Mayo Clinic, Rochester, Minnesota, USA; [8]National Human Genome Research Institute, Las Vegas, Nevada, USA; [9]Oxford University, Oxford, UK; [10]University of Cambridge, Cambridge, UK; [11]Western University, London, ON, Canada; [12]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA; [13]University of California, Irvine, California, USA; [14]Centers for Disease Control and Prevention, Atlanta, Georgia, USA; [15]Emory University, Atlanta, Georgia, USA; [16]Kaiser Permanente Center for Health Research, Portland, Oregon, USA; [17]Cambridge Substantive Site, Health Data Research UK, Wellcome Genome Campus, Hinxton, UK; [18]Johns Hopkins School of Public Health, Baltimore, Maryland, USA

Polygenic risk scores (PRS) can bridge the gap between the findings of genome-wide association studies (GWAS) and clinical applications for disease risk estimation. However, there are no accepted standards for the development, reporting, and application of PRS. Reporting standards for both risk model training and independent validation are needed to facilitate PRS translation into clinical care. The ClinGen Complex Disease Working Group—including experts in epidemiology, statistics, disease-specific applications,

implementation, and policy—has developed a PRS reporting framework, updating previous standards (GRIPS statement, 2011). The ClinGen framework includes items for PRS internal validity and reproducibility.

Independent reviewers applied this reporting framework to 30 original research papers on PRS development or validation. We observed substantial heterogeneity in PRS reporting, both in what was reported and how. Overall, there was under-reporting of items needed (a) to assess the internal validity of a PRS; (b) to reproduce a PRS; and (c) to evaluate the intended clinical use or actionability of a PRS. In addition, we observed gaps in external validation of PRS and a lack of diversity in training and validation populations.

The ClinGen framework defines criteria to promote transparency and validity in the clinical application of PRS. Deposition of metadata in the Polygenic Score (PGS) Catalog, a sister-resource to the NHGRI-EBI GWAS Catalog, will provide the information needed to apply and evaluate a PGS in new datasets. A central catalog of consistently annotated PGS will promote the uptake of PRS reporting standards while facilitating data-sharing and reproducible analyses.

65 | Identification of representative trees in random forests based on a new tree-based distance measure

Björn-Hergen Laabs[1]*, Inke R. König[1,2]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany; [2]German Center for Cardiovascular Research

In life sciences random forests are often used to train predictive models, but it is rather complex to gain any explanatory insight into the mechanics leading to a specific outcome, which impedes the implementation of random forests in clinical practice. Typically, variable importance measures are used, but they can neither explain how a variable influences the outcome nor find interactions between variables; furthermore, they ignore the tree structure in the forest in total. A different approach is to select a single or a set of a few trees from the ensemble which best represent the forest. It is hoped that by simplifying a complex ensemble of decision trees to a set of a few representative trees, it is 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. In our work, we developed a new tree-based distance measure for the definition of representative trees and compared it with existing metrics in an extensive simulation study. We show that our new distance metric is superior in depicting the differences in tree structures. Furthermore, we found that the most representative tree selected by our method has the best prediction performance on independent validation data compared to the trees selected by other metrics.

66 | A flexible hierarchical approach for multi-ethnicity or multi-tissue high-throughput omics data for Mendelian randomization or transcriptome analysis

Lai Jiang*, David V. Conti

Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California

We have previously proposed a hierarchical model, *hJAM*, which unifies the framework of Mendelian Randomization (MR) and transcriptome analysis and can incorporate multiple correlated intermediates jointly. Leveraging the information from multiple ethnic groups or across tissues could improve the ability to determine the effects of the intermediates on the outcome. The current model focuses on a relatively small number of intermediates and SNPs. Here, we show the flexibility of *hJAM* and propose two extensions on the *hJAM* model: (a) providing the averaged effects of each intermediate on the outcome across different ethnic groups or tissues; and (b) selecting both SNPs and intermediates for application in high-throughput experiments with a scalable approach. We evaluate the performance of our first extension in extensive simulation scenarios for effect estimation, type-I error and empirical power. The simulation results showed an unbiased averaged effect, maintained correct type-I error and achieved a sufficient power. We evaluate the performance of the second extension in comparison to existing approaches such as multivariable MR based on Bayesian model averaging (MR-BMA) by the area under the curve (AUC) and mean squared error. We apply the two extensions on multi-ethnic data from the Multiethnic Cohort and multi-tissue data from GTEx.

67 | Related pain in patients with breast cancer: A pathway-based analysis

Eunkyung Lee[1]*, Sung Y. Eum[2], Jean L. Wright[3], Susan H. Slifer[4], Eden R. Martin[4,5], Cristiane Takita[6,7], Robert B. Hines[8], James J. Urbanic[9], Carl D. Langefeld[9], Glenn J. Lesser[9], Edward G. Shaw[9], and Jennifer J. Hu[6,10]

[1]Department of Health Sciences, University of Central Florida College of Medicine, Orlando, Florida, USA; [2]Department of Biochemistry and

Molecular Biology, University of Miami Miller School of Medicine, Miami, Florida, USA; [3]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, Maryland, USA; [4]The Center for Genetic Epidemiology and Statistical Genetics, John P. Hussman Institute for Human Genomics, University of Miami School of Medicine, Miami, Florida, USA; [5]Dr. John T. Macdonald Department of Human Genetics, University of Miami Miller School of Medicine, Miami, Florida, USA; [6]Sylvester Comprehensive Cancer Center, University of Miami School of Medicine, Miami, Florida, USA; [7]Department of Radiation-Oncology, University of Miami School of Medicine, Miami, Florida, USA; [8]Department of Population Health Sciences, University of Central Florida College of Medicine, Orlando, Florida, USA; [9]Wake Forest NCORP Research Base, Wake Forest Baptist Comprehensive Cancer Center, Winston-Salem, North Carolina, USA; [10]Department of Public Health Sciences, University of Miami School of Medicine, Miami, Florida, USA

As radiation therapy (RT) can induce DNA damage in cancer cells, interindividual variations in DNA damage repair (DDR) capacity may play a critical role in the variability of normal tissue toxicity. This study aimed to evaluate the association between DDR pathways and RT-related pain. We evaluated 3,876 single nucleotide polymorphisms (SNPs) in 106 genes of five KEGG DDR pathways using the PLINK set-based tests. Among 359 breast cancer patients who underwent adjuvant RT, 81 (22.6%) developed RT-related pain. The nonhomologous end-joining (NHEJ) pathway was most significantly associated with RT-related pain ($P = 0.045$), and it was mainly driven by *XRCC4* ($P = 0.011$) and *XRCC6* ($P = 0.043$). The most significant individual SNP association was found in *XRCC4* rs2089565, which showed that carrying at least one minor C allele was associated with a 1.90-fold elevated risk of reporting RT-related pain (95% CI = 1.30-2.78; $P = 0.001$). In addition, *XRCC4* SNP rs73138171 suggested a potential possibility of replication from a validation cohort ($P = 0.095$) and from a meta-analysis ($P = 0.009$). The variant rs73138171 was observed only among minority populations. With a limited sample size, this pilot study showed a suggestive association between genetic variations in *XRCC4* of the NHEJ pathway and interindividual variations in RT-related pain among breast cancer patients. If validated in a larger population, these findings can be utilized as predictive biomarkers of RT-related pain and normal tissue toxicity. Additionally, the results can provide biological targets for pain management to improve the quality of life of survivors of breast cancer.

Supported by National Cancer Institute Grants No. R01CA135288 to J.J.H. and U10CA081857 to the Wake Forest Research Base CCOP.

68 | Do causal estimates of differential adiposity effects show evidence of impact on the circulating metabolome?

Matthew A. Lee[1,2]*, Kaitlin H. Wade[1,2], Laura J. Corbin[1,2], Nicholas J. Timpson[1,2]

[1]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK; [2]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK

We performed a systematic review of Mendelian randomization (MR) studies where a measure of adiposity was used as the exposure. Data extracted from 179 articles enabled exploration of the downstream effects of increased adiposity, revealing causal associations with a wide array of diseases. Though useful in identifying relationships, intermediate steps remain unclear. Metabolites lie at the interface between genetic and nongenetic factors and provide a read-out of physiological function. Metabolites implicated with pathologies may aid therapy repurposing or deployment.

We used MR to identify the metabolic footprint of increased adiposity. Using common genetic variation associated with body mass index (BMI), waist-hip ratio (WHR) and body fat percentage (BF), we reassessed observational associations with 123 metabolites. Globally, the effects of BMI and WHR on the metabolic profile were similar; the effects of BF on the metabolic profile were more complex, with different directions of effect, and may be a result of inaccurate measurement or suggest a different genetic architecture. Sixty-nine metabolites were associated after multiple-testing correction, with a consistent direction of effect across all exposures observed for: apolipoprotein A-I, phenylalanine, tyrosine. We observe an increase in the essential amino-acid phenylalanine, a pre-cursor to tyrosine (also increased), and a decrease in apolipoprotein A-I, a component of high-density lipoproteins.

In an on-going second stage, which will be available for presentation, we will identify relevant pathways of disease development by: (a) performing MR with metabolites to reassess relationships with identified diseases and (b) exploring these relationships while accounting for the interrelatedness of metabolites.

69 | Two-sample Mendelian randomization study of lipid level and ischemic heart disease

Su Hyun Lee[1]*, Ji Young Lee[1]*, Guen Hui Kim[1], Keum Ji Jung[1], Sun Mi Lee[2], Hyeon Chang Kim[3], Sun Ha Jee[1]

[1]Department of Epidemiology and Health Promotion, Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul,

Korea; [2]Health Insurance Policy Research Institute, National Health Insurance Service, Wonju, Korea; [3]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, Korea

Although high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and triglyceride (TG) concentrations are partially heritable risk factors for cardiovascular disease, their causal role in the development of ischemic heart disease (IHD) is unclear. Genetic variants significantly associated with lipid concentrations were obtained from the Korean Genome and Epidemiology Study (KoGES) ($n = 35{,}000$), and same variants on IHD were obtained from the Korean Cancer Prevention Study-II (KCPS-II) ($n = 13{,}855$). Using Mendelian randomization (MR) approaches, inverse variance weighting (IVW), weighted median, and MR Egger approaches were assessed the causal association between lipid concentrations and IHD. Radial MR identified outliers that became genetic instruments which are subject to pleiotropic bias. Causal association of LDL-C and IHD was observed in IVW method (OR = 1.013, 95% CI = 1.007-1.109). However, HDL-C and TG did not show causal association with IHD. In the Radial MR analysis of the relationship between HDL-C, TG and IHD, outliers were detected. Interestingly, after removing the outlier, a causal association between TG and IHD was found. High levels LDL-C and TG might be associated with increased IHD risk in Korean population, and may be potentially useful as evidence of significant causal relationship.

70 | Deleterious coding variants found among affected family members in the African American Hereditary Prostate Cancer Study (AAHPC) families

Deyana D. Lewis[1]*, Shukmei Wong[2], Angela S. Baker[2], Isaac Powell[3], John D. Carpten[4], Joan E. Bailey-Wilson[1], Cheryl D. Cropp[2,5]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute/National Institutes of Health, Baltimore, Maryland, USA; [2]Integrated Cancer Genomics Division, Translational Genomics Research Institute, Phoenix, Arizona, USA; [3]School of Medicine Urology, Wayne State University, Detroit, Michigan, USA; [4]Keck School of Medicine of the University of Southern California, Los Angeles, California, USA; [5]Department of Pharmaceutical, Social and Administrative Sciences, Samford University, McWhorter School of Pharmacy, Birmingham, Alabama, USA

Prostate cancer (PCa) is the most common cancer in males, but also exhibits a ~1.5-2-fold higher incidence in African American (AA) men compared with whites. Epidemiologic evidence supports a large heritable contribution to PCa, with over 100 susceptibility loci identified to date that can explain ~33% of the familial risk. A portion of this undefined risk may be due to rare susceptibility variants. The African American Hereditary Prostate Cancer (AAHPC) Study, established in 1997, enrolled 77 AA families from seven clinical sites across the United States. The aim of this study is to identify rare, predictive, deleterious variants through exome sequencing of 99 PCa cases selected among 26 AAHPC (two to three affected men sequenced per family) and three female 1000 Genomes controls.

Here we sequenced the exomes of 99 AAHPC cases at a mean coverage of 30x. Post-variant calling quality control (QC) was implemented using Golden Helix SVS 8 software with filters. Mendelian inconsistencies were checked using PLINK. We identified 8 non-synonymous single nucleotide variants (SNVs) that are considered damaging using three predictive scoring tools provided by ANNOVAR. Five genes associated with these SNVs (*CACNA1B, KCNJ1812, KMT2C*, *SULT1A1*, and *SVIL*) have been previously associated with increased risk of PCa. These predicted damaging variants in these five genes represent potentially novel causal candidates in some of the 26 sequenced AAHPC families. Future work will carefully analyze the genomes of additional AAHPC family members who will be sequenced in the near future.

71 | Selection of filtering thresholds on QC measurements in whole genome sequence data in a family-based study

Qing Li[1]*, Stephen Wank[2], Joan E. Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA; [2]Digestive Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, USA

Quality control (QC) is an important step in sequence data analysis. Without a validation sample of the variant calls, common practice is to drop variants based on preset thresholds of several key QC measurements, such as QualitybyDepth, RootMeanSquareMappingQuality, at the variant level. However, those thresholds are heuristically chosen and, sometimes, can be insufficient for certain datasets. In a linkage analysis of small intestinal carcinoid tumors in families, we have DNA sequence data on the same group of 19 individuals from two platforms, 10XGenomics and DRAGEN Bio-IT. In this

project, we were able to examine the characteristics of various QC measurements and the relationship between those measurements and the quality of variant calls. Based on the discordant pairs of variant calls from the two platforms, we were able to check if some of the pre-set thresholds were sufficient. We also applied a random forest algorithm to determine the optimal set of threshold values for key QC measurements. After we excluded the very poor quality variant calls from both platforms based on the most relaxed threshold values on QC variables, there were 6,352,112 variants that were covered by both platforms. While the average concordance rate for those variants was about 75%, we found that we could improve the concordance rate after certain QC variable thresholds were adjusted. Those adjustments helped us to eliminate bad variants calls and reduced Mendelian errors in the downstream analysis. The new QC pipeline will be described in detail.

72 | Computational efficient method to detect genetic interactions associated with age-of-onset in a type 2 diabetes Genome-wide Association Study

Siting Li*, Jiang Gui

Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, USA

Few Genome-Wide Association Study (GWAS) can analyze gene-gene interactions exhaustively due to the huge feature space characterized by SNP-SNP interactions. We developed a new algorithm, Efficient Survival Multifactor Dimensionality Reduction (ES-MDR), to exhaustively search over all pairwise SNP-SNP interactions in GWAS setting. ES-MDR uses Martingale Residuals to replace survival outcomes and identifies significant SNP-SNP interactions associated with age of disease-onset. Simulation results show that ES-MDR is very powerful in identifying SNP-SNP interactions and greatly reduces computational time compared to Survival Multifactor Dimensionality Reduction (SMDR). We applied our novel ES-MDR method to perform GWAS on a genotyped Type 2 Diabetes real data from dbGap to identify gene-gene interaction associated with age-of-onset. We identified a small group of SNPs which are significant after Bonferroni Correction, such as rs5750250, and predicted the phenotype. To test the predictive performance, we divided the samples into training set and test set (2:1). L1 penalized regression was performed on the training set and turning parameters was selected by cross-validation. Time-varying ROC curve was plotted on the test set using the training model. Varying from 30 to 85 years old, the largest area under the ROC curve (AUROC) is 0.8 and the smallest is 0.53.

73 | Integrated multi-ethnicity GWAS and functional analysis identified causal variants in lung cancer

Yafang Li[1]*, Xiangjun Xiao[1], Jinyoung Byun[1], Jun Xia[1], Zhuoyi Song[1], Jihye Yun[2], Younghun Han[1], Christopher Amos[1]. TRICL Consortium

[1]Department of Medicine, Institute of Clinical and Translational Research Center, Baylor College of Medicine, Houston, Texas, USA; [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.

Population-specific genome-wide association studies (GWAS) have been successful in identification of genetic variants associated with lung cancer but lacking the ability to identify true causal variants with effect across multiple populations. We conducted a meta-analysis to combine the information from European ($n = 51{,}961$), Asian ($n = 12{,}434$) and African American ($n = 5{,}766$) populations and identified 9 significant independent variants ($p < 5 \times 10\text{-}8$) with effect across 2 or 3 populations. Five of them are novel variants from gene *ZFP42* in overall lung cancer, *ACTR2* and *RTEL1* in lung adenocarcinoma, and *NECTIN1* and *RORA* in small-cell lung cancer. Four variants from regions closely linked to previously reported *AQP3*, *DLBLD1*, *MTAP* and *MPZL2* gene in single population have been identified with consistent effect in all the three populations. Further eQTL (expression quantitative trait loci) analysis using GTEx lung expression data displayed three of the pan-ethnicity lung cancer variants have cis-effect in AQP3, DCBLD1 and MPZL3 gene expression, suggesting the causal roles of these variants in lung cancer. Experimental study of overproduction of AQP3 in Kidney and colon cell lines showed that AQP3 can induce DNA damage in the cells, in addition to lung cells, implying the functional role of AQP3 in cancer diseases. The functional annotation using information from a variety of sources inferred that six out the nine variants were located in enhancers. And the novel variant in RORA is suggested to have a deleterious effect (CADD-phred score = 13.43) which is consistent with the associated increased risk effect (OR = 3.64, 95% CI = (2.30-5.76)) in small-cell lung cancer.

74 | Improving trans-ethnic portability of polygenic risk scores with predicted expression traits

Yanyu Liang[1]*, Tuuli Lappalainen[2], Ani Manichaikul[3], Abraham A. Palmer[4], Heather Wheeler[5,6], Hae Kyung Im[1]

[1]Section of Genetic Medicine, The University of Chicago, Chicago, Illinois, USA; [2]Department of Systems Biology, Columbia University, New York, New York, USA; [3]Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia, USA; [4]Department of Psychiatry, University of California San Diego, La Jolla, California, USA; [5]Department of Biology, Loyola University Chicago, Chicago, Illinois, USA; [6]Department of Computer Science, Loyola University Chicago, Chicago, Illinois, USA

The success and growth of genome-wide association studies are paving the way for the development of clinically relevant polygenic risk scores (PRS) for complex diseases. However, given the predominantly European-ancestry GWAS available to date, the portability of PRS to other populations is limited. The loss of performance is at least partially due to differences in LD among populations.

We reasoned that prediction approaches that are better anchored in causal mechanisms should be less prone to LD overfitting and will this have better portability. To test this idea, we developed genetic prediction scores based on predicted gene expression. We call these scores predicted transcriptome risk scores (PTRS).

Specifically, we predicted expression for individuals from genotype and expression prediction weights from predictdb.org. Analogous to PRS, which is linear combinations of genotypic dosage with GWAS effect sizes as weights, we calculated PTRS as linear combinations of genetically predicted expression with S-PrediXcan effect sizes as weights.

We surveyed the performance of PTRS across four populations (European, African, East Asian, and South Asian) using 17 quantitative traits from the UK Biobank.

Our preliminary results indicated that:

1) Predicted expression captures about 20% (SE = 2%) of the trait heritability.

2) PTRS is more portable across populations than PRS (for African and East Asian).

2a) Using population-matched genetically predicted expression in PTRS further improves portability.

3) S-PrediXcan-based effect size is still biased towards the population used for training, as expected.

These results indicate that incorporating molecular mechanisms in PRS development can improve portability across populations.

75 | Detecting gene x environment interaction in rare variant analysis for survival outcomes

Elise Lim,[1]* Adrienne Cupples,[1,2] Josée Dupuis,[1] Douglas P. Kiel,[3,4,5] Ching-Ti Liu[1]

[1]Department of Biostatistics, Boston University, Boston, MA, USA; [2]Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA, USA; [3]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA; [4]Department of Medicine, Harvard Medical School, Boston, MA, USA; [5]Hinda and Arthur Marcus Institute for Aging Research, Hebrew SeniorLife, Roslindale, MA, USA

Advanced technology in whole-genome sequencing has offered the opportunity to comprehensively investigate the genetic contribution, particularly rare variants, to complex traits. Numerous methods of rare variant analysis have been developed to detect gene-environment (GE) interactions for continuous and binary phenotypes, but limited work has been done for analyzing time-to-event outcomes. To tackle this challenge, we developed a method in detecting GE interactions of a set of rare variants using mixed effects Cox regression. Under this model, the genetic main effects were treated as fixed, while the GE interaction effects were modeled as random effects. As some variants may be highly correlated due to high linkage disequilibrium, we impose a ridge penalty to estimate the genetic main effects. We adopted a kernel-based method to leverage the joint information across the rare variants and implemented a variance component score test to reduce the computational burden. Our simulation study showed that the proposed method maintains correct type I error and moderate power under various scenarios, such as differing the percentage of censored observation and proportion of causal variants in the model. We illustrated our method to test gene-based interaction with smoking on time-to-fracture with samples from the Framingham Osteoporosis Study and identified two significant bone mineral density-associated genes, *SPTBN1* and *CDKAL1*, which were implicated in lung carcinogenesis and type 2 diabetes respectively. Given the importance of time-to-event outcomes in medical studies, we believe our proposed method can be a significant contribution in genetic association studies as well as time-to-event analyses.

Invited Abstract

76 | Analysis of large-scale biobanks and whole genome sequencing studies: Challenges and opportunities

Xihong Lin

Departments of Biostatistics and Department of Statistics Harvard University and Broad Institute

Big data from genome, exposome, and phenome are becoming available at a rapidly increasing rate with no apparent end in sight. Examples include Whole Genome Sequencing data, smartphone data, wearable devices, and Electronic Health Records (EHRs). A rapidly increasing number of large scale national and institutional biobanks have emerged worldwide. Biobanks integrate genotype, electronic health records, and lifestyle data, and is the trend of health science research. In this talk, I discuss opportunities, analytic tools and resources, and challenges presented by large scale biobanks and population-based Whole Genome Sequencing (WGS) studies for common and rare diseases for analysis of common and rare genetic variants and EHRs. The discussions are illustrated using ongoing large scale whole genome sequencing studies of over 500,000 subjects from the Genome Sequencing Program of the National Human Genome Research Institute and the Trans-Omics Precision Medicine Program from the National Heart, Lung and Blood Institute, and the UK Biobank.

77 | Transcriptome-wide association study of human facial shape identifies potential mediating genes

Dongjing Liu[1]*, Myoung Keun Lee[2], Jacqueline T. Hecht[3], George L. Wehby[4], Lina M. Moreno[5], Carrie L. Heike[6], Mary L. Marazita[1,2], Peter Claes[7,8], Seth M. Weinberg[1,2], John R. Shaffer[1,2]

[1]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [2]Department of Oral Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [3]Department of Pediatrics, University of Texas McGovern Medical Center, Houston, Texas, USA; [4]Department of Health Management and Policy, University of Iowa, Iowa City, Iowa, USA; [5]Department of Orthodontics, University of Iowa, Iowa City, Iowa, USA; [6]Department of Pediatrics, Seattle Children's Craniofacial Center, University of Washington, Seattle, Washington, USA; [7]Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium; [8]Department of Human Genetics, KU Leuven, Leuven, Belgium.

Genome-wide association studies (GWAS) have now identified hundreds of independent genomic loci influencing the normal-range variation in human facial shape. As most of these GWAS peaks are in noncoding intergenic regions, identifying the relevant genes at these loci is often not straightforward. We conducted a transcriptome-wide association study (TWAS) in 2,329 individuals with three-dimensional facial images and genome-wide data with the goal of identifying potentially functional mediators of the SNP-trait associations. We divided the full face into hierarchically arranged facial segments, and generated multi-dimensional phenotypes representing the shape variation within each segment. Gene expression levels were imputed by fitting prediction models trained using PrediXcan and UTMOST, and were then tested for association with the multi-dimensional facial phenotypes. Eleven genes at six loci (*WARS2, RUVBL1/EEFSEC/SEC61A1/GATA2, CRB1, HOXD4, SUPT3H/VEGFA, PDK4, LINC01152*) showed significant associations in four potentially relevant tissues, all of which overlap or are nearby known GWAS loci. We revisited the SNP-trait associations at these loci, while conditioning on predicted gene expression to identify the most likely contributing genes at each locus. In some cases, the nominated genes had an unknown role in craniofacial development or disease, indicating potentially novel associations. We also identified five suggestive loci which yielded a much stronger expression-trait p-value than SNP-trait p-value. These loci were missed in GWAS and therefore represent new candidates, which await follow-up together with the prioritized genes. Taken together, these findings demonstrate the benefits of TWAS in mapping facial genes.

78 | High-throughput reporter assay reveals functional impacts of 3′-UTR SNPs associated with neurological and psychiatric disorders

Andy B. Chen[1], Kriti Thapa[2], Hongyu Gao[1], Jill L. Reiter[1], Junjie Zhang[1], Xiaoling Xuei[1], Hongmei Gu[2], Yue Wang[1], Howard J. Edenberg[1,2], Yunlong Liu[1]*

[1]Department of Medical and Molecular Genetics; [2]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana, USA

Genome-wide association studies (GWAS) can identify noncoding variants associated with specific traits or phenotypes, but cannot distinguish whether such variants are functional or merely in linkage disequilibrium with the causal variants. High-throughput reporter (HTR) assays can be used to experimentally evaluate the impact of genetic variants on gene expression. In this study, our objective was to systematically evaluate the functional activity of 3′-UTR SNPs associated with neurological and psychiatric disorders. We gathered SNPs from the GWAS Catalog that were associated with any neurological or psychiatric disorder trait with P value $< 10^{-5}$. For each SNP, we identified the region that was in linkage disequilibrium ($r^2 > 0.8$) and retrieved all the common 3′-UTR SNPs (allele-frequency > 0.05) within

that region. We further used an HTR assay to measure the impact of the 3′-UTR variants in SH-SY5Y neuroblastoma cells. Of the 13,515 3′-UTR SNPs tested, 371 demonstrated a significant impact on gene expression; many of these variants were also in expression quantitative trait loci (eQTLs) in the brain. These results were then used to train a deep-learning model to predict the impact of novel variants that were not experimentally evaluated and to identify the cis-acting motifs that contribute to the predictions of their functional impacts. In conclusion, this study demonstrates that HTR assays combined with advanced machine-learning models can be used to identify causal noncoding variants associated with complex traits to further understand the etiology of diseases and complex traits.

79 | Testing Cell-type-specific Mediation Effects in Genome-wide Epigenetic Studies

Xiangyu Luo[1], Joel Schwartz[2], Andrea Baccarelli[3], Zhonghua Liu[4]*

[1]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [2]Department of Environmental Health, Harvard University, Boston, Massachusetts, USA; [3]Department of Environmental Health Sciences, Columbia University, New York, NY, USA; [4]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR, China

Epigenome-wide mediation analysis aims to identify DNA methylation CpG sites that mediate the causal effect of genetic/environmental exposures on health outcomes. However, DNA methylations in the peripheral blood tissues are usually measured at the bulk level based on a heterogeneous population of white blood cells. Using the bulk level DNA methylation data in mediation analysis might cause confounding bias and reduce study power. Therefore, it is crucial to get fine-grained results by detecting mediation CpG sites in a cell-type-specific way. However, there is a lack of methods and software to achieve this goal. We propose a novel Method MICS (Mediation In a Cell-type-Specific fashion) to identify cell-type-specific mediation effects in genome-wide epigenetic studies. MICS first estimates the cellular compositions via a reference methylation matrix, and then uses the estimated cell proportions to obtain the cell-type-specific p-values with respect to the effect of an exposure on the DNA methylation CpG sites as well as the effect of DNA methylation on the outcome, and finally combines the two *p*-value matrices using a joint-significance-followed-by-squaring procedure. We conduct simulation studies to demonstrate that our method has correct type I error control, and is powerful and robust under practical settings. We also apply our method to the Normative Aging Study and identify three DNA methylation CpG sites in monocytes that might mediate the effect of cigarette smoking on the lung function.

80 | Statistical model discovering 3D-genetic basis underlying complex diseases: An application to autism spectrum disorder data

Qing Li[1], Jingni He[1], Chen Cao[1], Feeha Azeem[1], Amy Chen[1], Aaron Howe[2], Jun Yan[3], Quan Long[1,4,5]*

[1]Department of Biochemistry and Molecular Biology; [2]Heritage Youth Researcher Summer Program; [3]Department of Physiology and Pharmacology; [4]Department of Medical Genetics; [5]Department of Mathematics and Statistics; University of Calgary, Alberta, Canada

Rationale: Three-dimensional (3D) genome can be assessed by various chromosome conformation capture techniques including Hi-C. However, the contribution of 3D genome to complex traits is yet to be quantified. Meantime, in statistical genetics, identification of genetic interactions underlying complex traits suffers from a challenge in interpreting results. This is partly due to that fact that many statistical methods do not have biological priori knowledge built-in, so that the statistical tests and the biological interpretation are done sequentially.

Methods: To explore the 3D-genetic basis of complex traits, and to address the statistical challenge of interpretation, in this project, we have developed a novel statistical model leveraging the 3D genomic conformation (assessed by Hi-C experiments) and linear mixed models (LMM) to elucidate the 3D-genetic underpinning of complex traits. Using an LMM model, we aggregate the genetic variants in regions that are 3D-interacting into a random term, and then associate the term to phenotypes.

Results: We have thoroughly tested the novel methods using simulations, contrasting to state-of-the-art alternatives. By applying the novel model to benchmark datasets from the MSSNG, world's largest genomic database for autism spectrum disorder (ASD), we have identified novel genes in 3D regions that are associated with the risk of ASD. Interestingly, when applying this method to other psychiatry disorders, we identified shared genetic components (pleiotropic effects) between ASD and Schizophrenia. This project can further understanding of 3D-genetics and pathology of complex diseases.

81 | Patients with a low PRS should be prioritized to rare variant screening

Tianyuan Lu[1,2]*, Sirui Zhou[1], Haoyu Wu[1,3], Vincenzo Forgetta[1], Celia M.T. Greenwood[1,3,4,5] and J. Brent Richards[1,3,4,6]

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

Identifying rare variant carriers is important for optimizing diagnostic and treatment strategies. However, costs of targeted sequencing are high, which limits the clinical utility of rare variants. We posit that individuals with a specific disease, but at low polygenic risk for that disease, may be more likely to carry rare causal variants.

We constructed polygenic risk scores (PRSs) for type 2 diabetes (T2D), short stature, osteoporosis, breast cancer and colorectal cancer using separate cohorts. Among 44,550 exome-sequenced participants in the UK Biobank, we identified carriers of rare variants with a high predicted pathogenicity affecting disease-causing genes. We tested whether PRSs and rare pathogenic variants are both associated with the susceptibility to these diseases. We tested the association between the existence of rare pathogenic variants and the PRS among diagnosed patients.

We found per SD increase in the T2D PRS increased odds of diagnosis 1.60-fold (95% CI: 1.57-1.64) while rare diabetes-causing variants conferred a 2.27-fold (95% CI: 1.41-3.65) increased odds. We observed no distinguishable difference in the distribution of PRS between carriers and non-carriers. Meanwhile, we found among diabetic patients, per SD decrease in T2D PRS was associated with 2.82-fold (95% CI: 1.76-4.54) increased odds of identifying rare variant carriers. We had the same observations for the other four traits.

Common and rare genetic components both contributed substantially to disease pathogenesis. Nevertheless, rare genetic causes were more prevalent among patients with a low PRS. Our study implies patients at low PRS could be prioritized to undergo screening for rare pathogenic variants.

82 | Analysis of the pleiotropy between breast cancer and thyroid cancer

Elise A. Lucotte[1]*, Pierre-Emmanuel Sugier[1,2], Jean-François Deleuze[3], Evgenia Ostroumova[4], Marie-Chistine Boutron[1], Florent de Vathaire[1], Pascal Guénel[1], Benoît Liquet[2,5], Marina Evangelou[6], Thérèse Truong[1]

[1]Inserm, Centre de Recherche en Epidémiologie et Santé des Populations, Université Paris-Saclay, Université Paris-Sud, Villejuif, France; [2]Laboratoire de Mathématiques et de leurs Applications de Pau, Université de Pau et des Pays de l'Adour, Energy Environment Solutions, Centre National de la Recherche Scientifique, France; [3]National Centre of Human Genomics Research, François Jacob Institute of biology, Commissariat à l'Energie Atomique, Paris-Saclay University, Evry, France; [4]Radiation group, International Agency for Research on Cancer, Lyon, France; [5]ARC Centre of Excellence for Mathematical and Statistical Frontiers at School of Mathematical Science, Queensland University of Technology, Brisbane, Australia; [6]Department of Mathematics, Imperial College London, London, UK

Thyroid and breast cancers share a lot of similarities in their biology: both are more frequent in women and are influenced by hormonal and reproductive factors. Individuals diagnosed with breast cancer are more likely to develop thyroid cancer as a secondary malignancy than patient diagnosed with other cancer types, and *vice-versa*. Using GWAS, 313 risk variants were identified for breast cancer. For thyroid cancer, 10 loci were identified and one of them (2q35) was previously reported to increase risk of breast cancer. Thyroid cancer is the only cancer for which genetic factors contribute more than environmental factors. To date, no study has been conducted to identify common genetic factors between breast and thyroid cancer. We have access to GWAS results on thyroid cancer (EPITHYR consortium), which was coordinated by our team, and to the summary statistics of the most recent GWAS conducted by the Breast Cancer Association Consortium (BCAC). In this ongoing study, we aim at studying pleiotropy between both cancers at different scales. First, we will estimate the genome-wide genetic correlation using the LDscore and SumHer methods. Second, we will analyze the association of the polygenic risk scores of breast cancer in association to thyroid cancer risk and *vice-versa*. Third, we will identify the pleiotropic SNPs affecting both cancers. These analyses are still ongoing and the results will be presented. Evidence of carcinogenic pleiotropy will improve our understanding of the diseases etiology and will provide insights on the underlying common biology between both cancers.

83 | Genome-wide gene-environment interaction study for breast cancer risk in European women, using data from the Breast Cancer Association Consortium

Pooja Middha Kapoor[1,2], Xiaoliang Wang[3,4], Marjanka K. Schmidt[5], Montserrat Garcia-Closas[6], Peter Kraft[7,8], Roger L. Milne[9,10,11], Douglas F. Easton[12,13], Sara Lindström[3,4], Jenny Chang-Claude[1,14]* on behalf of the Breast Cancer Association Consortium

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Faculty of Medicine, University Heidelberg, Heidelberg, Germany; [3]Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA; [4]Department of Epidemiology, University of Washington, Seattle, WA; [5]Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands; [6]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA; [7]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [8]Program in Molecular and Genetic Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [9]Cancer Epidemiology Division, Cancer Council Victoria, Australia; [10]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Australia; [11]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia; [12]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; [13]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; [14]Genetic Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Genome-wide association studies may not detect variants that alter the risk of a disease in combination with lifestyle/environmental exposures. In this study, we aimed to identify novel breast cancer susceptibility loci by conducting a genome-wide gene-environment study in women of European ancestry. Using data from 28,176 cases and 32,209 controls genotyped using the iCOGS array and 44,109 cases and 48,145 controls genotyped with the OncoArray array, we assessed interactions between ~7 million genetic variants (MAF > 0.01) and age at menarche, parity, ever use of oral contraceptives (OC), current smoking, and postmenopausal body mass index, for overall breast cancer risk. Standard logistic regression was employed and all models were adjusted for age, study, and ten principal components. Models assessing the role of current smoking were additionally adjusted for former smoking. Array-specific effect estimates were meta-analyzed using the fixed-effect inverse variance method.

Among the variants analyzed for interactions with risk factors, none reached genome-wide statistical significance. After exclusion of regions known to be associated with breast cancer risk, 5 intergenic SNPs in high linkage disequilibrium ($r^2 = 1$) located on chromosome 2 showed interactions with parity at P value $< 5 \times 10^{-7}$. The nearest gene is *LINC01854* (*long intergenic non-protein coding RNA1854*) and contains enhancers which interact with the transcription factor binding sites of *CTCF*, *RAD21*, *ZNF143*, *NFIB*, *NFIC*, and *ATF2*. Further, we found evidence at P value $< 5 \times 10^{-7}$ of interactions for overall breast cancer risk between 1 variant with age at menarche, 2 variants with OC, and 2 variants with current smoking.

We found no strong evidence of interactions between common variants and lifestyle-related factors in overall breast cancer risk. We identified several suggestive gene-environment interactions that might contribute to the understanding of breast cancer etiology, but require replication.

84 | Investigation of genetic risk modifiers of leg ulcer development in sickle cell patients using whole genome sequencing

Candace D. Middlebrooks[1]*, Faith Pangilinan[2], Khadijah E. Abdallah[2], Ashley J. Buscetta[2], Lawrence C. Brody[2], Caterina P. Minniti[3], Joan E. Bailey-Wilson[1], Vence L. Bonham[2]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA; [2]Social and Behavioral Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA; [3]Division of Hematology, Department of Oncology, Montefiore Medical Center, Albert Einstein College of Medicine, New York, USA*

Sickle Cell Disease (SCD) impacts millions of people worldwide and over 100,000 individuals in the United States. There are several serious sub-phenotypes related to SCD which include leg ulcers. SCD-associated leg ulcers can be highly resistant to therapy and severely painful. We sought to determine if there are genetic variants that modify the risk of developing leg ulcers in patients with SCD.

We executed a pilot study where we performed whole genome sequencing (WGS) on blood samples taken from 23 SCD patients who had ($N = 8$) or did not have ($N = 15$) leg ulcers to identify genetic variation that may contribute to increased risk for leg ulcers. Sequencing was performed at the National Intramural Sequencing Center (NISC) at NHGRI using the Illumina NovaSeq 6000 platform. The GATK4 pipeline for germline short variant discovery was used to process the raw files and we used

Plink to perform unadjusted association analysis. Although we did not have enough power to detect a genome-wide significant signal, our top hits include variants in an intergenic region near *PKD1P4* ($P = 1.11$ E-04) and an intronic region within *C4orf* ($P = 1.48$E-04). The latter region is implicated in variation in blood pressure.

The association in the C4orf region is promising since vasculopathy has been implicated in the development of SCD leg ulcers. The results from this study will remain inconclusive until we analyze the additional ~100 samples that were subsequently sequenced by the NISC. We will apply this same pipeline to the full data set and report the results.

85 | On a novel statistical method for integrating multi-omics data

Sarmistha Das, and Indranil Mukhopadhyay*

Human Genetics Unit, Indian Statistical institute, Kolkata, India

Alteration in regulatory activities of a gene may be attributed to various genetic and epigenetic factors, like, mutation in DNA, methylation at various CpG sites and so forth. Traditional methods of analysing omic data separately would incur loss of essential information that might be beneficent in differentiating between the genetic code of a tumor and a normal sample or for case-control data. Recently data integration through joint analysis of various omics data is focused to obtain robust inferences. But it is imperative to understand and include the available downstream information and use more practical assumptions in the model development, to capture diversity in the genomic profiles of patients. We propose a novel statistical method to identify the regulators through integration of different omics data, namely, genotype, gene expression and so forth. Our method includes biological insight from multi-omics data to assess its effect on the gene expression. We derive the asymptotic distribution of our test statistic to calculate p-values fast, perform extensive simulations to analyse the power of the test keeping type I error rate at five percent level. Our proposed method is powerful and consistent as the power of the test increases with the increase in sample size. It is also robust and shows consistent results for different genetic models in our simulation study. We also apply this method to a real data set. Based on the results obtained from simulated and real data, our method looks very promising in differentiating gene expression profiles of patients through integrated analysis of multi-omics data.

86 | Metabolomic signatures of microRNAs in cardiovascular traits: A Mendelian randomization analysis

Rima Mustafa[1]*, Michelle Mens[2], Rui J Pinto[1], Ibrahim Karaman[1], Gennady Roshchupkin[3,4], Jian Huang[1], Paul Elliot[1], Marina Evangelou[5], Abbas Dehghan[1], Mohsen Ghanbari[2]

[1]Department of Epidemiology and Biostatistics, Imperial College London, London, UK; [2]Department of Epidemiology, Erasmus Medical Centre, Rotterdam, The Netherlands; [3]Department of Radiology and Nuclear Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands; [4]Department of Medical Informatics, Erasmus Medical Centre, Rotterdam, The Netherlands; [5]Department of Mathematics, Imperial College London, London, UK

With increasing evidence supporting the association between microRNAs (miRNAs) and cardiovascular traits, a systematic investigation is needed to elucidate underlying pathways. We conducted two phases of two-sample Mendelian Randomization (MR) to identify metabolites that mediate the effect of miRNAs on cardiovascular traits.

Causal estimates were calculated using the inverse variance weighted method and weighted median, weighted mode, and MR-Egger were used as sensitivity analyses. We conducted genome-wide association studies (GWAS) on miRNA expression data ($N = 1{,}687$) in the Rotterdam Study to identify genetic instruments for miRNAs. The association of the instruments with metabolites were examined in the Airwave Health Monitoring Study ($N = 1{,}942$). We conducted MR analysis agnostically on 591 well-expressed miRNAs and 886 metabolites measured by the Metabolon platform. We further examined the causal role of the metabolites on cardiovascular traits. Genetic instruments were identified for metabolites in the Airwave Health Monitoring Study and their associations with cardiovascular traits were extracted from most recent GWAS.

After correcting for multiple testing, our analysis indicated causal associations between 16 miRNAs and 21 metabolites. Of those metabolites, we further found nominally significant association between androsterone sulfate, cysteine sulfinic acid, and N-palmitoylglycine with stroke, and 3-hydroxy-5-cholestenoic acid with coronary artery disease. Our results provided suggestive evidence for metabolites that might link miR-1273h-5p, miR-3937, and miR-4753-5p with stroke, and miR-181a-2-3p with coronary artery disease.

Our approach has added insight into ongoing efforts in studying the role of miRNAs in cardiovascular traits. Further studies with larger sample size are required to validate our findings.

87 | Functional characterization of a *CDKN1B* variant for defining the target genes and their mechanistic underpinnings contributing to SLE susceptibility

Swapan K. Nath

Oklahoma Medical Research Foundation. Oklahoma City, Oklahoma USA

A recent genome-wide association study reported a significant genetic association between rs34330 (−79C/T) of cyclin-dependent kinase inhibitor 1B (*CDKN1B*) gene and risk of systemic lupus erythematosus (SLE) in Han Chinese population. However, its validity and functional mechanisms of action to SLE susceptibility are not yet defined. Here, we performed an allelic association followed by a meta-analysis using 11 independent cohorts ($n = 28{,}828$), *in-silico* bioinformatics, and a series of experimental validations to determine the functional consequence of rs34330. We first replicated the genetic association with rs34330 ($P_{meta} = 1.48 \times 10^{-20}$, OR = 0.84). Following-up with bioinformatics and eQTL analyses, we predicted the rs34330 is located in an active chromatin region that could regulate promoter and/or enhancer activities of the target gene(s). Using luciferase, we observed significant allele-specific promoter activity in HEK293 (kidney) and Jurkat (T-cells). Using ChIP-qPCR, we found allele-specific bindings with three histone marks (H3K27Ac, H3K4Me3, and H3K4Me1) and two transcription factors (Pol-II and IRF-1). Next, we experimentally validated the long-range chromatin interactions between rs34330 and target genes using chromosome conformation capture (3C) assay. Finally, applying CRISPR-based genetic and epigenetic editing, we confirmed the regulatory role of this region containing rs34330 for *CDKN1B* and nearby gene (*APLOD1* and *DDX47)* expressions. Collectively, we replicated genetic association between a potentially functional variant and SLE. The risk allele (C) is likely to influence binding affinities with several histones and regulatory proteins (i.e., IRF1) to regulate the allele-specific expressions of target genes *(CDKN1B/APOLD1/DDX47)*. Hence, their aberrations could be a potential mechanism for exacerbating lupus susceptibility through rs34330.

88 | Genome-wide analysis of copy number variation and normal facial variation in a large cohort of Bantu Africans

Megan Null,[1]* Feyza Yilmaz,[2,3] David Astling,[4] Hung-Chun Yu,[2] Joanne Cole,[2,5] Stephanie A. Santorico,[1,5,6] Richard A. Spritz,[2,5] Tamim H. Shaikh[2,5] and Audrey E. Hendricks[1,5,6]

[1]Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, USA; [2]Department of Pediatrics, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA; [3]Department of Integrative Biology, University of Colorado Denver, Denver, Colorado, USA; [4]Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA; [5]Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA; [6]Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, USA

The face is one of the most distinguishing characteristics of the human body, and similarity between relatives points towards a strong genetic component in normal facial development. However, little is known about the genetic factors underlying normal facial appearance, particularly copy number variations (CNVs). We present a genome-wide association study of the relationship between normal facial variation and CNVs in a sample of Bantu African children. African subjects, as well as structural variations such as CNVs, are arguably understudied areas of genetics; we add to the scientific literature in both understudied areas. We designed two linear mixed effects models that together capture loss and gain CNVs within a region that have the same direction or opposite directions of effect. In our multiple testing adjustment, we consider the correlation of these two models and CNV windows to produce an appropriate adjustment that is not overly conservative. We find suggestive evidence that CNVs play a role in common facial variation with putative novel associations in five regions and evidence of independent CNV association in three regions previously identified in single-nucleotide polymorphism (SNP) GWAS. Additionally, we show that studying CNVs in a sample of Bantu African ancestry provides genetic information not captured by densely imputed SNPs. We find only a small proportion of common CNVs (frequency > 5%) or analysis windows of rare CNVs are well tagged (by SNPs. This supports the need for more resources and research to uncover association between structural variations and complex traits, especially in understudied ancestries.

89 | Omnigenic, polygenic or stratagenic? And why it matters for personalised medicine

Paul F. O'Reilly

Mount Sinai, New York, New York, USA

There has been a recent collective reflection in the field of genetics about how genetic variation leads to complex human disease. This has been important in advancing understanding of the implications of existing data and

because the two major models proposed – the omnigenic and classical polygenic models – have different consequences for the future direction of the field. Here I summarise the two models, describe their appeal and highlight some of their limitations, in particular their lack of consideration of the intimate link between genetics and the environment. I outline an alternative model that I call the "stratagenic model." In the stratagenic model, there are *core functions* coded by a subset of genes, that, together with their interaction with the environment, can each, or as a combination of several, produce disease. Assuming at least partially independent, structured paths to disease implies that there are multiple disease liabilities, not only a single liability. This highlights a key difference between population-level risk variants and individual-level risk variants, that is, not typically considered. Polygenicity is explained by the fact that most core functions are likely to have at least a weak association with disease via the environment in at least some individuals, and population-level risk variants are an aggregation of individual-level risk variants. This, combined with environmental variation across populations, may help to explain weak generalisability of polygenic risk scores across populations. The stratagenic model differs substantially from the two alternatives, could provide explanations for phenomena in the data that the other models do not, and has meaningfully different consequences.

90 | Shared genomic segment analysis via equivalence testing

Sukanya Horpaopan[1], Cathy S. J. Fann[2], Mark Lathrop[3], Jurg Ott[4]*

[1]Department of Anatomy, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand; [2]Institute of Biomedical Sciences, Academia Sinica, Academia Road, Nankang, Taipei, Taiwan; [3]McGill University and Genome Québec Innovation Centre, Montréal, Québec, Canada; [4]Laboratory of Statistical Genetics, Rockefeller University, New York, USA

An important aspect of disease gene mapping is replication, that is, a putative finding in one group of individuals is confirmed in another set of individuals. As it can happen by chance that individuals share an estimated disease position, we developed a statistical approach to determine the likelihood (*P* value) for multiple individuals or families to share a possibly small number of candidate susceptibility variants. Here, we focus on candidate variants for dominant traits that have been identified by our previously developed *Heterozygosity Analysis*, and we are testing the sharing of candidate variants obtained for different individuals. Our approach allows for multiple pathogenic variants in a gene to contribute to disease, and for estimated disease variant positions to be imprecise. Statistically, the method developed here falls into the category of equivalence testing, where the classical null and alternative hypotheses of homogeneity and heterogeneity are reversed. The null hypothesis situation is created by permuting genomic locations of variants for one individual after another. We applied our methodology to the ALSPAC data set of 1,927 whole-genome sequenced individuals, where some individuals carry a pathogenic variant for the BRCA2 gene, but no two individuals carry the same variant.

91 | Effect of AGER-by-smoking Interaction on lung function: A genome-wide interaction study

Boram Park[1], Jaehoon An[1], Wonji Kim[2], Hae Yeon Kang[3], Sang Baek Koh[4], Bermseok Oh[5], Keum Ji Jung[6], Sun Ha Jee[6], Woo Jin Kim[7], Michael H. Cho[8,9], Edwin K. Silverman[8,9], Taesung Park [2,10]*, Sungho Won[1,2,11]*

[1]Department of Public Health Sciences, Seoul National University, Seoul, South Korea; [2]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, South Korea; [3]Department of Internal Medicine, Healthcare Research Institute, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, South Korea; [4]Department of Preventive Medicine, Yonsei University Wonju College of Medicine, Wonju, South Korea; [5]Department of Biochemistry and Molecular Biology, School of Medicine, Kyung Hee University, Seoul, South Korea; [6]Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, South Korea; [7]Department of Internal Medicine and Environmental Health Center, Kangwon National University Hospital, School of Medicine, Kangwon University Chuncheon, South Korea; [8]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA; [9]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA; [10]Department of Statistics, Seoul National University, Seoul, South Korea; [11]Institute of Health and Environment, Seoul National University, Seoul, South Korea

Rationale: Smoking is a major risk factor for chronic obstructive pulmonary function (COPD); however, more than 25% of COPD patients are non-smokers, and gene-by-smoking interactions are expected to affect COPD onset.

Objectives: We aimed to identify the common genetic variants interacting with pack-years of smoking on FEV_1/FVC ratios in individuals with normal lung function.

Methods: A genome-wide interaction study (GWIS) on FEV_1/FVC was performed for individuals with FEV_1/

FVC ratio ≥ 70 in the Korea Associated Resource (KARE) cohort data, and significant SNPs were validated using data from two other Korean cohorts.

Measurements and Main Results: The GWIS revealed that rs10947231 and rs8192575 met genome-wide significant levels (P_{LR} for rs10947231 = 2.23×10^{-12}, P_{LR} for rs8192575 = 1.18×10^{-8}), concurrent with validation analyses. Furthermore, rs8192575 showed significant interaction effects with smoking (beta = −0.02, P_{INT} = 0.0165), and eQTL, TAD, and PrediXcan analyses revealed that both SNPs are significantly associated with *AGER* expression.

Conclusions: SNPs on the 6p21 region are associated with FEV_1/FVC, and the effect of smoking on FEV_1/FVC differs among the associated genotypes.

Keywords: Gene-by-smoking interaction, Genome-wide interaction study (GWIS), COPD, pulmonary lung function, *AGER*

92 | Validation of genetic markers for prognosis in colon cancer patients treated with oxaliplatin-based chemotherapy

Hanla Park[1,2]*, Petra Seibold[1], Dominic Edelmann[3], Axel Benner[3], Lina Jansen[4], Federico Canzian[5], Martin Schneider[6], Michael Hoffmeister[4], Hermann Brenner[4,7,8], Jenny Chang-Claude[1,9]

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Medical Faculty, University of Heidelberg, Heidelberg, Germany; [3]Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany; [4]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; [5]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany; [6]Surgical Oncology Clinic for General, Visceral and Transplantation Surgery, University of Heidelberg, Heidelberg, Germany; [7]Division of Preventive Oncology, National Center for Tumor Diseases (NCT), Heidelberg, Germany; [8]German Cancer Consortium (DKTK), Heidelberg, Germany; [9]Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Several candidate gene studies have reported associations between genetic variants and the efficacy of oxaliplatin treatment, a platinum drug commonly used in colorectal cancer (CRC) patients. However, studies have been limited by small sample sizes, and lack of adjustment for relevant covariates and for multiple testing as well as specific treatment definitions, that is, number of completed cycles of oxaliplatin treatment. Most studies assessed genetic variants as prognostic markers and results were inconsistent. Only two studies investigated predictive genetic markers. This study aimed to validate previously reported associations of prognostic and predictive genetic markers using the largest independent colorectal cancer sample to date and to evaluate further functional variants in relation to survival outcome in relation to oxaliplatin treatment.

Sixty-three candidate single nucleotide polymorphisms (SNPs) were selected based on previous reports (39 SNPs) or regulatory function in candidate genes (24 SNPs). About 1,400 Stage II-IV patients included in a German population-based study (DACHS) received primary adjuvant chemotherapy, 38% of them received oxaliplatin treatment. Multivariable Cox proportional hazards models were used to identify SNPs that are associated with differential overall survival, CRC specific survival, and recurrent-free survival according to the type of chemotherapy (oxaliplatin-based vs. others).

Five of the tested SNPs were associated with prognosis at $p < .05$, but none of the associations were confirmed after adjustment for multiple comparisons. Thus the individual SNPs are unlikely to be of clinical utility and further investigations in well-powered studies with integrated approaches are warranted.

93 | Network analysis with multi-omics data using graphical LASSO

Jaehyun Park[1], Sungho Won[1,2,3]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea; [2]Department of Public Health Sciences, Seoul National University, Seoul, South Korea; [3]Institute of Health and Environment, Seoul National University, Seoul, South Korea

Precision matrix between variables provide their conditional dependencies and can be used to infer noble biological pathways or gene-protein interactions. Recently, as the cost of data generation has decreased, multiple biological omics datasets can be utilized to consider the specific structures of multi-omics data such as different sparsity of dependence within and between omics. In this study, we suggest a new method which can detect the interplay among multi-omics by using different penalization parameters based on graphical LASSO or GRaFo. The parameters can be determined by cross-validation to minimize the penalized likelihood function. The proposed method was evaluated with simulation data and showed that it successfully identify the disease susceptible multi-omics markers. The proposed method was applied to Chronic Obstructive Pulmonary Disease, which illustrates its practical value.

Keywords: network estimation; precision matrix; multi-omics data

94 | Genome-wide analyses identify novel variants associated with degree of depression and reveal importance of HDL cholesterol level

Kyungtaek Park[1], Minji Kim[2], Yongmin Ahn[2] and Sungho Won[1,3]

[1]Interdisciplinary Program for Bioinformatics, College of Natural Science, Seoul National University, Seoul 08826, Korea; [2]Department of Neuropsychiatry, Seoul National University Hospital, Seoul 03080, Korea; [3]Graduate School of Public Health, Seoul National University, Seoul 08826, Korea

Correspondence should be addressed to Sungho Won (sunghow@gmail.com)

Depression is a common psychiatric disorder of which symptoms include lethargy and feeling failure and guilty. It is expected to be the second leading cause of disability by 2020, but genetic and environmental causes of the disease still remain to be elucidated. In this study, genome wide association studies were performed to shed light on the etiology of depressive disorder. We utilized two cohorts from Seoul National University Hospital Gangnam Center, SNUHGC1 and SNUHGC2, which consist of 8,000 and 2,349 samples, respectively. Beck Depression Inventory (BDI) score was used to measure degree of depression of each sample. We found a novel variant, rs1517928 located in 3p14.1, whose interaction with HDL cholesterol level is significantly associated with degree of depression (; p =) in SNUHGC1. Subsequently, it was replicated in SNUHGC2 (; p =). Furthermore, meta-analyses of each BDI question identified 3 novel loci lying on 5q21, 12q21.3 and 16q23, where the most significant variants were rs17159046 (; p =), rs10862690 (; p =) and rs55967401 (; p =), respectively. The three variants were significantly associated with expression level of EFNA5, TMTC2 and CDH13, respectively, in dorsolateral prefrontal cortex and hippocampus. These genes were known to be relevant with clinical features of depression.

95 | Exploring the genetic architecture of the human neurological proteome using whole genome sequencing

Grace Png[1]*, Andrei Barysenka[1], Linda Repetto[2], Xia Shen[2,3,4], Emmanouil Tsafantakis[5], Maria Karaleftheri[6], George Dedoussis[7], Arthur Gilly[1], Eleftheria Zeggini[1]

[1]Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; [2]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK; [3]Biostatistics Group, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen Univeristy, Guangzhou, China; [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [5]Anogia Medical Centre, Anogia, Greece; [6]Echinos Medical Centre, Echinos, Greece; [7]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens, Greece

The human proteome has a stronger genetic component compared to many complex diseases, making it a valuable resource of potential disease biomarkers and drug targets. This is especially so for highly polygenic neurological disorders whose mechanisms remain elusive. Here, we present the first sequence-based protein quantitative trait loci (pQTL) analysis of 92 neurological proteins. We perform a meta-analysis using deep whole-genome sequencing (WGS) data from two isolated Greek cohorts, MANOLIS (22.5 × WGS; $N = 1{,}356$) and Pomak (18.4 × WGS; $N = 1{,}537$). A total of 123 independently-associated variants in 84 loci reach study-wide significance ($P < 1.14 \times 10^{-10}$) for 63 proteins, all of which are at least nominally significant ($P < 3.78\text{x}10^{-4}$) and have the same direction of effect in both cohorts. To further elucidate the genetic architecture, independent variants were classified into 89 (72%) *cis*- and 34 (28%) *trans*-acting pQTLs. Ten variants have consequences equal to or more severe than missense, and 33 overlap regulatory regions. We also discover variants that have previously been linked to psychiatric disorders. For example, an intronic *trans*-pQTL in the *ITIH4* gene is associated with increased NEP levels (*rs2239547;* $P = 1.19 \times 10^{-129}$; BETA = 0.637983; SE = 0.026328), and is an established risk variant for schizophrenia and bipolar disorder. This analysis represents the largest and only WGS-based pQTL study of neurological proteins to date, delivering insight into the rare and common genetic variant landscape underlying the human neurological proteome and its connection to neurological diseases.

96 | Systematic analysis of population and familial effects in developmental stuttering

Hannah G. Polikowsky[1]*, Lauren E. Petty[1], Douglas M. Shaw[1], Shelly J. Kraft[2], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA; [2]Behavioral Speech & Genetics Lab, Department of Communication Sciences and Disorders, Wayne State University, Detroit, Michigan, USA

Developmental stuttering is a speech disorder characterized by prolongation of sounds and interruptions in speech with childhood onset. Despite large heritability

estimates and strong familial trends, the genetic architecture and etiology of developmental stuttering remain elusive. Our study design utilizes a two-pronged approach, leveraging both collected population and familial data. To date, no genome-wide association study of developmental stuttering nor analyses combining both population and familial effects have been published.

Here, we performed a multipoint linkage analysis of a large Australian family showing an autosomal dominant pattern of inheritance for stuttering spanning four generations. We collected genotype data for 38 family members (22 affected). Our preliminary analysis of seven genotyped individuals highlighted regions approaching genome-wide significance (LOD = 2.53) in chromosomes two, 11, and 13. Through collaborative efforts and a social media campaign, we genotyped over 1,000 developmental stuttering cases on the Illumina multi-ethnic genotyping array followed by imputation to the Haplotype Reference Consortium using the Michigan imputation server. Five-thousand-three-hundred ancestry-matched population-based controls were drawn from Vanderbilt's DNA biobank, BioVU. To account for our multi-ethic stuttering case-control population comprised of 41 family networks, we ran a single variant association analysis using SUGEN, a program developed to account for both multi-ethnic populations and related individuals. Initial results identified 47 variants with suggestive significance (P value < 5e-6) located in chromosomes two, four, six, ten, 11, 13, 16, and 19.

Preliminary results suggest both shared and unique genetic architecture between families and population cohorts, implicating protein trafficking pathways in developmental stuttering risk.

97 | Climate change and increasing emergence of Nipah virus: Ecological niche model to understand the current and future risk areas under different scenarios

Malay K. Pramanik

Jawaharlal Nehru University, New Delhi, Delhi-110067, India

Nipah virus is a zoonotic virus, harbored by bats and lethal to humans. Bat to human spill overs occur every winter in Bangladesh, Australia, and Malaysia. In recent, more than 15 people died in India and the distributional areas are continuously increasing worldwide, especially Southeast Asia. However, there is significant heterogeneity in the number of spillovers detected in these countries, and the climatic understanding remains unexplained.

Therefore, the study aimed to predict the role of climate change on Nipah virus distribution worldwide based on MaxEnt and Support Vector Machine modelling. The predictions were made by using HadGEM2-CC and GFDL-CM3 climate model of the 5th assessment report IPCC for the period of 2050 and 2070 with RCP scenario of 4.5 and 8.5.

Bio-climatic variables contributions were assessed using jackknife test and AOC, TSS, and kappa (>0.90) indicate the model performs very high accuracy. The major influencing variables will be Precipitation of Warmest Quarter ($45.01 \pm 0.7\%$), Temperature Seasonality (19.09 ± 0.8). The high risk countries are Bangladesh, Myanmar, Southern and Eastern India, Thailand, Laos, Malaysia, Australia, Somalia, Kenya, Tanzania, China, and there will be a drastic increase in the outbreak for future.

The result provides credible guidelines to different Health Organization worldwide for preparedness, and intervention measures under climate change.

98 | Brain cell-types contributing risk to reading-associated traits

Kaitlyn M. Price[1,2,3]*, Karen G. Wigg[1], Yu Feng[1], Kirsten Blokland[2], Margaret Wilkinson[2], Elizabeth N. Kerr[2,3], Sharon L. Guger[2,3], Maureen W. Lovett[2,3], Cathy L. Barr[1,2,3]

[1]Department of Genetics and Development, Krembil Research Institute, University Health Network, Toronto, Canada; [2]The Hospital for Sick Children, Toronto, ON, Canada; [3]University of Toronto, Toronto, ON, Canada

Reading disabilities (RD) is a neurocognitive trait characterized by difficulties with word recognition, poor spelling, and decoding abilities. It is hypothesized to be caused by subtle disruptions in neuronal migration, which affect connectivity of language-related brain regions. The genes/genetic variants implicated in RD and the cell-types they are expressed in are not well understood. To contribute to understanding the genetics of RD, we previously conducted a genome-wide association study (GWAS) for word reading and identified a significantly associated region on chromosome 12, *LINC00935-CCNT1* ($p \sim 10^{-6}$, aggregated gene-based analysis). Next, to examine what cell-types are contributing to reading, we used Linkage Disequilibrium Score Regression (LDSC). LDSC leverages GWAS data

and cell-type specific gene-expression data, to determine the polygenic signal contributing to heritability from a given cell-type. We performed LDSC on our GWAS using single-cell RNA-sequencing data from brain tissue at different developmental stages (PsychENCODE project). In addition to our GWAS, to increase power, we also performed LDSC on publically available GWAS for attention-deficit/hyperactivity disorder (ADHD), educational attainment and intelligence. These GWAS were chosen because they share significant genetic overlap with reading (previously determined by polygenic risk scores). Using LDSC, we found significant GWAS enrichment for educational attainment/intelligence in fetal excitatory neurons, fetal intermediate progenitor cells, fetal microglial, and adult excitatory/inhibitory neurons (FDR < 5%). We will also perform LDSC using human cell-type data from the Allen Brain Bank. Through these analyses, we will identify cell-types contributing risk to RD and other traits. Identifying what cell-types genetic variants are active in will help us understand how they influence biological pathways and disease.

99 | Imputation of the plasma proteome reveals novel associations with inflammatory diseases

Bram P. Prins*, INTERVAL Study, Adam S. Butterworth

MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, UK

The imputation of gene expression data into large cohorts without such measurements has become a widely used approach to identify novel genes associated with disease risk. Nevertheless, variations in expression levels do not necessarily reflect variations in levels of proteins, the latter being the actual effector units of biological processes and the targets of most drugs. Therefore, utilizing expression-based imputation may result in suboptimal prioritization of risk-associated biomarkers. To this end, we developed prediction models for imputation of proteomic data, and applied these in a set of association analyses for inflammatory diseases to identify disease-associated proteins and novel genomic regions. Protein expression models were constructed by performing elastic net regression using genetic variants within one Megabase of the cis-gene of 2,869 autosomally coded plasma proteins (SomaLogic) in 3,301 participants from the INTERVAL study. Ten-fold nested cross-validations were performed to compute the coefficient of determination (R^2) to assess model performance. For 626 probes (584 unique proteins) we were able to build models with a minimal prediction performance (R^2_{cv}) of at least 0.01, with the largest R^2_{cv} for CLEC12A (0.77). We next used our prediction models to discover 14 previously unknown protein-disease associations, including HP for hip osteoarthritis (P value $= 3.58 \times 10^{-8}$) and PLAU for Crohn's disease (P value $= 1.24 \times 10^{-9}$). This approach demonstrates the potential to identify new biomarkers and drug targets for common complex diseases.

100 | Genome-wide association study of resistance to tuberculosis infection in exposed individuals from various endemic settings

Jocelyn Quistrebert[1,2]*, Marianna Orlova[3], Christophe Delacourt[2,4], Eileen G. Hoal[5], Alexandre Alcaïs[1,2], Lai Thanh[6], Laurent Abel[1,2], Erwin Schurr[3], Aurélie Cobat[1,2]

[1]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM, Paris, France; [2]Imagine Institute, Paris Descartes University, Paris, France; [3]Infectious Diseases and Immunity in Global Health Program, Research Institute of the McGill University Health Centre, Montreal, Canada; [4]Paediatric Pulmonology and Allergology Department, Necker Hospital for Sick Children, Paris, France; [5]South African Medical Research Council Centre for Tuberculosis Research, DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa; [6]Center for Social Disease Control, Binh Duong, Vietnam

The natural history of tuberculosis (TB) is characterized by great interindividual variability after exposure to *Mycobacterium tuberculosis* (Mtb), suggesting a role for host genetics. Certain groups of persons, such as household contacts (HHC) of pulmonary TB patients or individuals in hyperendemic TB areas, are at high risk for becoming infected and developing active disease. However, some individuals, despite repeated strong exposure remain resistant to Mtb infection. We performed a genome-wide association study of resistance to Mtb infection in an endemic region of Southern Vietnam. Resisters were defined as intensely exposed subjects who tested negative in both tuberculin skin test and interferon-gamma release assay ($n = 185$) and were compared to Mtb infected individuals ($n = 353$). We observed a novel genome-wide significant locus on chromosome 10 associated with resistance to Mtb infection (top genotyped variant, OR = 0.42, 95% CI 0.39-0.45, $P = 3.71 \times 10^{-8}$). The locus was replicated in a French multi-ethnic HHC cohort and a familial admixed cohort from a hyperendemic area of South Africa (top genotyped variant, P value $= 1.51 \times 10^{-2}$ and P value $= 1.74 \times 10^{-2}$, respectively). The

associated variants are located upstream and in the intron of a tumor suppressor gene which encodes an ubiquitin ligase that activates a transcription factor known to induce apoptosis of Mtb-infected macrophages. These results demonstrate the value of focusing on the resistance to infection phenotype and including populations from various ancestries and epidemiological settings to reveal genetic determinants of complex infectious diseases such as tuberculosis.

101 | Polygenic risk scores – Is there a need for a more accurate classification within ethnicities?

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 already contains genetic data of 10,259 (51,2%) from 20,000 included very low birth weight infants. Within this database, we construct polygenic risk scores for common complex diseases, based on the GWAS 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.

102 | Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and central nervous system disorders

Xinyuan Zhang[1,2]*, Anastasia M. Lucas[1], Yogasudha Veturi[1], Theodore G. Drivas[1], Anurag Verma[1], eMERGE Consortium, Marylyn D. Ritchie[1]

[1]Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; [2]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Clinical and epidemiological studies have shown that circulatory system diseases and central nervous system disorders often co-occur in patients. However, genetic susceptibility factors shared between these disease categories remain largely unknown given that previous studies generally focused on single diseases for genetic variant discovery. Here, we characterized potential pleiotropy across 107 circulatory system and 40 central nervous system traits using Phenome-wide association studies (PheWAS) and multi-trait joint association (MultiPhen) in 43,015 European ancestry adults from the eMERGE Network. A comprehensive replication analyses were performed in the UK Biobank. Both methods were able to identify previously known disease-associated SNPs in both datasets, such as Alzheimer's disease (rs429358) and coronary artery disease (rs1333049). We characterized 607 SNPs that had a significant association in both datasets via both PheWAS and MultiPhen using an exploratory *P* value threshold. There were 204 of these SNP associations met Bonferroni correction for multiple testing burden. Pleiotropic effects of these SNPs were tested via a formal test of pleiotropy using a sequential multivariate method. We discovered five genomic regions that demonstrated significant evidence of pleiotropy, including *APOE*, *CDKN2B-AS1*, *HLA*, *PRDM8* and *NOTCH4* genes. We observed region-specific patterns of direction of genetic effects for the two disease categories, suggesting potential antagonistic and synergistic pleiotropy. Our findings provide insights into the relationship between circulatory system diseases and central nervous system disorders which can provide context for future prevention and treatment strategies.

103 | Mitochondrial genetic regulation of mitochondrial DNA gene expression in blood

Euijung Ryu[1]*, Gregory D. Jenkins[1], Nicholas B. Larson[1], Joanna M. Biernacka[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic Rochester, Rochester, Minnesota, USA

Genetic variations in mitochondrial DNA (mtDNA) play an important role in various health conditions. However, it is unclear whether those genetic variations affect mitochondrial-encoded gene expression (mtEx). In this study, we aimed to assess whether SNPs in mtDNA (mtSNPs) are associated with mtEx and mtEX-associated nuclear SNPs (nSNPs) interact with mtSNPs.

Using SNP data and RNA-seq gene expression data from 313 subjects from the Genotype-Tissue Expression Project cohort, we tested association of 21 common mtSNPs with expression levels of 15 mtDNA-encoded genes. For testing mtSNP-nSNP interaction effects, we considered 5 selected nSNPs reported to be associated with altered mtEx.

Adjusting for covariates associated with mtEx (top 3 nuclear principal components, gender, study batches, and 45 probabilistic estimation of expression residuals [PEER] factors), MitoG10399A (located in MT-ND3) was associated with altered expression of MT-ND3 gene ($P = 7.2 \times 10^{-6}$) after Bonferroni multiple testing correction. Subjects carrying the G allele of the mtSNP (minor allele frequency [MAF] = 0.28) had lower MT-ND3 expression than those carrying the A allele (12% reduction). None of mtSNP-nSNP interaction association results was significant after correcting for multiple testing, acknowledging low statistical power.

Preliminary analysis of mtSNP data suggests that MitoG10399A regulates expression level of MT-ND3 gene in whole blood samples. Given that this SNP is associated with several complex diseases such as Parkinson's disease and type 2 diabetes, future studies on the disease risk predisposition through mitochondrial genetic regulation of mtDNA gene expression may benefit for better understanding disease etiology.

104 | Strength of polygenic risk score for type 2 diabetes in Arab population

Hossam Almeer[1], Osama Alsmadi[2], Naser Elkum[3], Mohamad Saad[1]*

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [2]King Hussein Cancer Center, Amman, Jordan; [3]Research Department, Sidra Medicine, Doha, Qatar
**Presenting author*

Polygenic risk score (PRS) has recently resurfaced again. Application of PRS would be more important on diseases that are preventable or better treated if diagnosed at early stages. Coronary Artery Disease and Type 2 Diabetes (T2D) are among the traits that PRS was applied the most.

The applicability of PRS across various ethnic groups remains challenging. A recent study showed that PRS developed in one population cannot translate well to other populations. Although interest in studying complex diseases in non-European datasets have increased, Middle eastern populations are still the least studied.

The aim of this study is to evaluate the PRS applicability on an Arab Type 2 Diabetes GWAS data set (~500 cases, ~1200 controls). Type 2 Diabetes prevalence is ~20% in the Gulf countries, where consanguinity plays a major role in high prevalence.

Unlike other studies from the region, our study identified known genes (e.g., *TCF7L2* and *GLIS3*). We applied two PRSs derived from Mahajan et al, 2018 (403 SNPs) and Khera et al, 2018 (~7 M SNPs) on our imputed data. Mahajan's PRS led to slightly better performance. AUC and p-value of the model "T2D~PRS" were 0.636 and 2.07e-21. Adding BMI, Age, and Sex to the model, AUC increased to 0.806. Grouping by decile, individuals whose PRS values fall in the highest decile have a 7.43-fold greater probability to have T2D compared to individuals in the lowest decile. Our results suggest that PRS can be applied to Arab populations and led to higher AUC than European datasets (0.72, Khera et al).

105 | Multiethnic study of genetics of dyslipidemia: No evidence of cardioprotective role of rare variants in the Apo CIII in non-European populations

Sanghera DK, Goyal S, and Bejar C on behalf of Consortium for Genetic Studies of Dyslipidemia

Department of Pediatrics, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Exome sequencing studies on Europeans have identified rare loss of function (LOF) variants in apolipoprotein CIII (Apo CIII) to be associated with reduced triglycerides (TG) and decreased risk of coronary artery disease (CAD). However, there is paucity of data on the cardioprotective role of these and other rare variants in non-European populations.

Here we have evaluated the role Apo CIII variants in 9 diabetic cohorts from Multinational Consortium for Genetic Studies of Dyslipidemia using up to 25,421 individuals. These included individuals from the AIDHS (Sikhs 3964), LOLIPOP (Europeans 2153; Asian Indians 11,330), multiethnic cohorts of South Asians and Chinese from Singapore (Chinese 1826; South Asian 3455), multiethnic population of Oklahoma (European 154, other 280), and Mexican American (SAFS) (2259). Our study could not confirm the role of null variant rs76353203 (R19X) and earlier known splice variants (rs140621530, rs373975305) for their athero-protective effects reported in Europeans and even in Pakistanis. Only one LOF variant (rs138326449) was associated with modestly low TG in 17 carriers in European (LOLIPOP) but 5 of 17 (29%) had CAD and 14 in Asian Indians (LOLIPOP)- 6 of 14 (75%) had CAD; only one carrier was found in AIDHS with modestly low TG (83 mg/dl) but was diabetic. However, this variant was not observed in any other cohort. The vast majority of novel rare variants identified in our data in Apo CIII were associated with high plasma TG. Collectively, these results highlight the challenges of inclusion of rare variant information in clinical risk assessment. Our data also stresses the need for genetic evaluation of diverse ancestries for identifying clinically important therapeutic targets.

Funding Acknowledgements: The Sikh Diabetes Study/Asian Indian Diabetic Heart Study was supported by NIH grants-R01DK082766 (NIDDK) and NOT-HG-11-009 (NHGRI) and grants from Presbyterian Health Foundation. Sequencing services were provided through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute of the NIH.

106 | MixFAR: A multiphenotype association model that detects structure in secondary phenotype space and increases power of association tests

Subrata Paul[1,2], Stephanie A. Santorico[1,3,4,5]*

[1]Department of Mathematical and Statistical Science, University of Colorado Denver; [2]Institute of Behavioral Genetics, University of Colorado Boulder; [3]Human Medical Genetics and Genomics Program; [4]Department of Biostatistics & Informatics; [5]Division of Biomedical Informatics & Personalized Medicine, University of Colorado Denver, CO, USA

Complex disease traits often represent a summarization, and hence dimension reduction, over multiple phenotypes. This summarization process can introduce phenotypic heterogeneity by mixing together disease subtypes, which reduces the power to detect associated genetic variants. Here, we propose Mixture of Factor Analysis Regression (MixFAR) - a multivariate association model that incorporates heterogeneity using a mixture of factor analyzers. MixFAR uses the correlation among phenotypes to simultaneously identify homogeneous subgroups of individuals and perform association testing in a reduced dimensional space.

We compared power of the proposed model with a trait-based association test that uses the extended Simes procedure (TATES) and to a factor mixture analysis (FMA). Simulations were considered for three and six quantitative phenotypes based on two disease subtypes, one of which is associated with a genetic variant. Different levels of heritability, mixture proportion, proportion of associated traits, and correlation structures were considered as well as whether the genetic variant was also associated with the complex disease trait.

In the presence of heterogeneity, our simulations show that MixFAR has higher power compared to TATES and FMA when enough phenotypic information is available (here, six traits) as well as for fewer phenotypes (three traits) when they are positively correlated and all associated with the genetic variant. MixFAR is also robust to the association of the genetic variant with the primary disease trait, which was not the case for TATES. Our new method provides a powerful way to discover both sub-types of disease while also detecting genetic association.

107 | Understanding the contribution of known cardiovascular-related genes to sudden cardiac death in patients undergoing hemodialysis

Tae-Hwi Schwantes-An, Matteo Vatta, Marco Abreu, Leah Wetherill, Howard J. Edenberg, Tatiana M. Foroud, Glenn M. Chertow, Sharon M. Moe

Indiana University School of Medicine, Health Information and Translation Sciences, Indianapolis, Indiana, USA

Cardiovascular diseases (CVDs) are the leading causes of death in patients with chronic kidney disease (CKD). In patients requiring dialysis, CVD accounts up to 60% of all deaths. In the general population, atherosclerotic disease (e.g. myocardial infarction) is the primary cause of cardiovascular death, and fewer than 5% of patients have sudden cardiac death (SCD). In contrast, among patients on dialysis, SCD accounts for 25–30% of all CV deaths.

The increased frequency of SCD as the cause of cardiovascular fatality in CKD begins before the onset of dialysis, indicating that the dialysis procedure is not the only risk. We sought to identify associations between SCD and 174 genes associated with inherited cardiac conditions (ICCs) among patients receiving hemodialysis.

Samples are selected from the Evaluation Of Cinacalcet HCl Therapy to Lower Cardiovascular Events (EVOLVE) trial. A clinical events committee adjudicated all participants to determine outcomes. Total of 126 SCD cases (37 African American [AA] and 89 European Ancestry [Eur]) and 107 controls (34 AA and 73 Eur) matched for age, sex, race, dialysis vintage, and diabetes were identified. NextGen sequencing was done using the TruSight Cardio kit (Illumina, San Diego CA). Variants were annotated using ANNOVAR and filtered based on function and gnomAD minor allele frequency (MAF) <0.05. We conducted burden tests using RVTESTS and RAREMETAL to identify enrichment of variants among cases or controls for SCD. Three distinct models of variant collapsing were tested; (1) individual genes, (2) 17 groups of genes each associated with an ICC (Disease), and (3) 11 groups of genes based on gene function defined by GeneCards (gene function).

We identified 4,014 single nucleotide variants (SNVs) in 143 genes. After filtering for predicted function and MAF, 2,109 variants (1,624 in Eur and 1,526 in AA) were retained for analysis. After multiple testing correction, statistical significance threshold was $p = .0003$ (0.05/143). No statistically significant association between SCD and individual genes was identified. Disease-based burden test did not yield significant associations. Lastly, gene function-based burden test also did not yield any significant associations.

In conclusion, we found no statistically significant associations between rare variants among genes associated with inherited cardiac conditions and sudden cardiac death in patients undergoing hemodialysis in this modest sized study.

108 | Added value of biomarkers and polygenic risk scores as risk factors for coronary artery disease

Natasha Sharapova*[1], Jessye M. Maxwell[1], Kylie Glanville[1], Saskia P. Hagenaars[1], Richard Russell[2], Zina M. Ibrahim[3], Cathryn M. Lewis[1]

[1]Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UK; [2]Global Research and Data Analytics, RGA Reinsurance Company, London, UK; [3]Department of Biostatistics & Health Informatics, King's College London, London, UK

Stratifying the population by risk of disease could allow for targeted approaches to prevention, diagnosis and treatment. Traditional risk factors (e.g. BMI, smoking and family history) are routinely assessed in coronary artery disease (CAD) diagnosis. Biomarkers (e.g. cholesterol and HbA1c) can also be tested to indicate risk but are often only detectable once pathology has set in, so their predictive value may be limited by proximity to age of onset. Polygenic risk scores (PRS), which are a single measure of an individual's genetic liability to disease, remain constant across the lifespan. Thus, they could potentially help identify the groups most at risk and allow for targeted early management and intervention. We used the UK Biobank to build and compare Cox proportional-hazards regression models including traditional risk factors, biomarkers and PRS to explore the value genetic and biomarker data could add as risk factors for CAD. Each model was internally validated through bootstrapping. We found that the area under the curve (AUC) given by our traditional risk factor model (0.734) was modestly improved by including biomarkers (AUC 0.738) and PRS (AUC 0.748), while our full model incorporating all of the potential risk factors achieved an AUC of 0.750. This suggests that biomarkers and PRS hold value alongside traditional risk factors for CAD and could improve stratification of individuals for targeted interventions. Including these data in risk models could potentially allow for better prevention and diagnosis of CAD.

109 | Applying a phenome risk score-based model to identify undiagnosed developmental stuttering cases in a Biobank for genome wide association analysis

Douglas Shaw*[1], Dillon Pruett[2], Hannah Polikowsky[1], Hung-Hsin Chen[1], Lauren E. Petty[1], Robin Jones[2], Jennifer E. Below[1,3]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA; [2]Hearing and Speech Sciences, Vanderbilt University, Nashville, Tennessee, USA; [3]Vanderbilt University Medical Center, Nashville, Tennessee, USA

Developmental stuttering is a speech disorder characterized by a disturbance in fluency and speech pattern, with an adult prevalence of 1–3% in the US. Despite twin-based studies showing ~50% heritability, the genetic etiology of stuttering is still largely unknown. Within Vanderbilt's Electronic Health Record-linked biorepository (BioVU), only 142 cases of stuttering have

diagnostic ICD9/10 codes out of 93,000 genotyped samples, suggesting a large portion of people who stutter are not classically diagnosed within the EHR.

To create a phenome-risk score for stuttering, we identified Phecodes enriched in cases of stuttering within Vanderbilt's ungenotyped EHR (n ~ 2.7 M). We built a Gini Index-based decision tree classifier model using Phecodes enriched in cases as predicting features and diagnosis status as the outcome variable. This model was trained and tested with a set of manually reviewed developmental stuttering cases; sex, age, race, and ethnicity matched controls, and showed 83% positive prediction rate.

We then applied the model in BioVU, identifying >10,000 imputed cases with a sample prevalence of 10.8%. 5977 European cases were selected for a preliminary GWAS of imputed stuttering samples in BioVU. Top suggestive hits include rs12613255 ($\beta = 0.309$; P value$= 2.40 * 10^{-7}$), near *FAM49A*, and rs17553695 ($\beta = 0.208$; P value$= 8.52 * 10^{-7}$), an intronic variant within *RBMS3*, a gene previously associated with osteonecrosis of the jaw and lung squamous cell carcinoma. This model was further validated by comparing results with the GWAS summary statistics of a clinically diagnosed developmental stuttering sample where we found a significant concordance in the direction of effect of all tested variants genome-wide.

Invited Abstract

110 | Polygenic risk scores for lung cancer in Chinese and Caucasian populations

Hongbing Shen

Nanjing Medical University

Lung cancer is the leading type of cancer with the highest incidence and mortality in China and the world. Genetic variation plays an important role in the development of lung cancer. We and other research teams have conducted multiple batches of lung cancer GWAS studies during the past decade, and identified multiple genetic susceptibility loci for lung cancer. Recently, we built a polygenic risk score (PRS) for Chinese populations with 19 SNPs (PRS-19), and then evaluated its utility and effectiveness in predicting high-risk populations of lung cancer in an independent prospective cohort of 95,408 individuals from China Kadoorie Biobank (CKB). The PRS of the risk loci successfully predicted lung cancer incidence in a dose-response manner. Specially, we observed apparently separate lung cancer event curves for low, intermediate and high genetic risk populations respectively during follow-up of the cohort.

However, the above PRS was only applicable to Chinese population. Based on results from the ILCCO Lung cancer OncoArray project (18420 cases and 13977 controls), we derived a PRS of lung cancer for Caucasian Population with 24 SNPs (PRS-24). When applied to the UK Biobank cohort ($N = 0.48$ million), we observed an obvious linear relationship for the positive association between PRS and lung cancer risk. Compared with participants at low genetic risk (the bottom 5%), participants at intermediate (5%-95%) and high genetic risk (the top 5%) had a significantly higher risk of lung cancer, with HRs of 1.82 (95%CI, 1.43-2.32) and 2.73 (95%CI, 2.06-3.62), respectively. Besides, we also observed a joint effect of genetic and smoking on risk of incident lung cancer in a dose-response manner; that is, the overall risk of lung cancer increased as both genetic risk and smoking.

Compared the PRS from Chinese and Caucasian populations, only 13 susceptibility loci were both used in PRS-19 and PRS-24. Seven loci (such as 2p14, 9p13.3, et al) were specifically used for Chinese population, and 11 loci (including 13q13.1-*BRCA2*, 22q12.1-*CHEK2*, et al) were specifically used for Caucasian population, which highlighted genetic heterogeneity of lung cancer in different ethnicities. Although the composition is different, both PRS were proved to be effective tools to predict lung cancer risk and usually with a HR of 2-3 in Chinese and Caucasian populations. Although the effects of smoking and lung cancer risk differ significantly between the two populations, joint effect of genetic and smoking on the risk of incident lung cancer were consistent.

These findings proved that PRS could be effectively used for lung cancer risk prediction and potentially applied in lung cancer screening program for individualized prevention. However, construction of population-specific PRS is necessary.

111 | Development and validation of risk prediction model for lung cancer in Chinese populations: A prospective cohort study of 0.5 million adults

Meng Zhu[1,2], Ci Song[1,2], Hongxia Ma[1,2], Hongbing Shen[1,2]*

[1]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China; [2]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China

Background: With the accumulation of evidence in Europe and America, lung cancer screening was increasingly used for clinical intervention in China.

However, the criteria to enrich high-risk populations of lung cancer remains controversial due to the low incidence and weak association between smoking and lung cancer in the Chinese population.

Methods: A nationwide prospective cohort study, China Kadoorie Biobank, in which 512,585 participants were enrolled from 10 geographically diverse areas across China, was used to develop and validate a risk prediction model of lung cancer. Flexible parametric survival models were used to estimate the 10-year absolute risk of lung cancer accounting for the competing risk of death. The potential clinical benefits of genetic testing for lung cancer screening were evaluated simultaneously.

Findings: A total of 13 variables significantly associated with the risk of lung cancer were used to build a multivariate model, which showed excellent discrimination with C-statistics of 0.78 and 0.77 in the development and validation datasets respectively. The model could identify 25% more lung cancer patients than the NLST criteria (292 *vs.* 232) assuming an equal number of persons were examined ($P = 3.67 \times 10^{-6}$). Joint utilization of the model with PRS could further enrich the high-risk population to be screened from 582.59 per 100,000 person-years to 805.76 per 100,000 person-years ($P = 0.036$).

Interpretation: Joint utilization of the model and genetic testing was capable of improving the enrichment efficiency of high-risk populations for lung cancer. These findings are expected to promote the improvement and formulation of lung cancer screening strategies in China.

112 | Using imputed genotype data in the joint score tests for genetic association and gene–environment interactions in case-control studies

Minsun Song*

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

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

113 | Identifying, testing, and correcting for bias in Mendelian randomization analyses using gene-by-environment interactions

Wes Spiller[1], Fernando Hartwig[2], George Davey Smith[1], Jack Bowden[3]

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK; [2]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil; [3]College of Medicine and Health, University of Exeter, Exeter, UK

Mendelian randomization using Gene-by-Environment interactions (MRGxE) is a sensitivity analysis where instrument by covariate interactions are used to test and correct for pleiotropic bias in summary data Mendelian randomization (MR) studies. This has the notable advantage of allowing the validity of individual instruments to be evaluated, in contrast to alternative pleiotropy robust approaches. However, MRGxE requires the degree of pleiotropic bias to remain constant across levels of the interacting covariate to provide unbiased causal estimates.

To address this limitation, we propose an individual level data modelling framework for testing violations of the constant pleiotropy assumption. This extension of MRGxE allows for candidate instrument by covariate interactions to be readily identified, and facilitates the implementation of novel sensitivity analyses with respect to the underlying assumptions of the approach. These developments have been incorporated into a

comprehensive software package, providing an efficient method through which candidate instrument by covariate interactions can be identified in large scale data and leveraged to provide estimates of causal association.

The utility of the proposed suite of methods is illustrated through an applied analysis re-examining the association between body mass index (BMI) and systolic blood pressure (SBP) using UK Biobank. This includes an initial search for candidate gene-by-environment interactions, the application of a range of MR methods, and finally sensitivity analyses with respect to each candidate interaction covariate. We identify alcohol consumption as the strongest valid interaction, finding evidence of a positive association between BMI and SBP in agreement with previous findings from the MR literature.

114 | Assessment of imputation quality – Comparison of phasing and imputation algorithms in real data

Katharina Stahl[1]*, Damian Gola[2,3], Inke R. König[2,3,4]

[1]Institut für Genetische Epidemiologie, Universitätsmedizin Göttingen, Göttingen, Germany; [2]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany; [3]German Centre for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Lübeck, Germany; [4]Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL)

Despite the widespread use of genotype imputation tools and the availability of different approaches, the latest developments of currently used programs have not yet been compared comprehensively.

We therefore assessed the performance of 35 combinations of phasing and imputation programs, including versions of SHAPEIT, Eagle, Beagle, minimac, PBWT, and IMPUTE, for genetic imputation of completely missing SNPs regarding quality and speed.

We used a data set comprising 1,149 fully sequenced individuals from the German population, subsetting the SNPs to approximate the Illumina Infinium-Omni5 array. 553,234 SNPs across two selected chromosomes were utilized for comparison between imputed and sequenced genotypes.

We found that all tested programs with the exception of PBWT impute genotypes with very high accuracy (mean error rate <0.005). Scores not implementing the true underlying genotypes rate PBWT as the best imputation program, even though the less frequent allele is hardly ever imputed correctly (mean concordance for genotypes including the minor allele <0.0002). For all programs, imputation accuracy drops for rare alleles with a frequency <0.05. Even though overall concordance is high, concordance drops with genotype probability, indicating that low genotype probabilities are rare. The mean concordance of SNPs with a genotype probability <95% drops below 0.9, at which point disregarding imputed genotypes might prove favorable. For fast and accurate imputation, a combination of Eagle2.4.1 using a reference panel for phasing and Beagle5.1 for imputation performs best. Replacing Beagle5.1 with minimac3, minimac4, Beagle4.1, or IMPUTE4 results in a small gain in accuracy at a high cost of speed.

115 | Benefits of phased whole genome sequence: Examples from cystic fibrosis (CF)

Scott Mastromatteo[1], Angela Chen[1], Jiafen Gong[1], Bhooma Thiruv[1], Wilson Sung[1], Zhuozhi Wang[1], Joe Whitney[1], Fan Lin[1], Johanna M. Rommens[1], Lisa J. Strug[1,2]

[1]Program in Genetics and Genome Biology and The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada; [2]Department of Statistical Sciences and Biostatistics, The University of Toronto, Toronto, Canada

Conventional whole genome sequencing (WGS) applications ignore the phase of alleles. Yet knowing which alleles are located on contiguous stretches of chromosome enables interpretation of *cis*-effects of rare disease mutations and of more common regulatory variants (eg expression quantitative trait loci; eQTLs), improving understanding of genotype-phenotype relationships. Here we investigate the impact of phasing in the Canadian CF population.

Population-based phasing is limited by low variant frequencies and incomplete reference catalogues, resulting in frequent switch errors. More accurate *individual-level* phasing can be achieved with long-read sequencing technologies or single molecule barcoding such as by 10x Genomics (10XG). Benchmarking against the Platinum Genomes truth set, we show that 10XG when compared to long-read technologies, produces the longest phased blocks with the fewest switch errors.

Using 10XG followed by WGS, we observed that 97% of genes <100 kb are fully phased within one block across 93% of ~250 individuals with CF, and *CFTR* (189 kb) was phased in a single block in 85% of these individuals with heterozygous CF-causing variants. We also identified distinct *CFTR* backgrounds in individuals carrying variants with reported varying clinical significance such as

D1152H or 5 T, and could derive extended haplotypes for improved interpretation of CFTR eQTLs which could influence CF disease severity. Investigation using haplotype association analysis at CF modifier genes, to determine if distant cis eQTLs display increased association with CF disease severity, is ongoing.

Individual-level phasing approaches provide unprecedented resolution and highlight that analyses that incorporate the full diploid nature of the human genome is achievable.

116 | Combining human and artificial intelligence: Ensemble of convolutional neural networks for disease prediction from microbiome data

Divya Sharma[1]*, Andrew D. Paterson[1,3], Wei Xu[1,2],

[1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [2]Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada; [3]The Hospital for Sick Children, Toronto, Ontario, Canada

Research supports the potential use of microbiome as a predictor of some diseases. Motivated by the findings that microbiome data is complex in nature and there is an inherent correlation due to the hierarchical taxonomy of microbial Operational Taxonomic Units (OTUs), we proposed a novel machine learning method incorporating a stratified approach to group OTUs into clusters based on their phylum. Convolutional Neural Networks (CNNs) were used to train within each of the clusters individually. Further, through ensemble learning approach, features obtained from each cluster were concatenated to improve prediction accuracy. Our two-step approach comprising of stratification before combining multiple CNNs, aided in capturing the relationships between OTUs sharing a phylum efficiently, as compared to using a single CNN overlooking OTU correlations. We used simulated datasets containing 168 OTUs in 200 cases and 200 controls for model testing. Thirty-two causal OTUs were randomly selected and pairwise interactions between three OTUs were used to introduce nonlinearity. We also implemented this novel method in two datasets: (a) Cirrhosis with 118 cases, 114 controls; (b) Type 2 diabetes with 170 cases, 174 controls; to demonstrate the model's effectiveness. Additionally, age and sex were included as factors to examine their role in enhancing the model's performance. Extensive experimentation and comparison against conventional machine learning techniques yielded encouraging results. We obtained mean AUC values of 0.88, 0.94, 0.74, showing a consistent increment of 5%, 12%, and 4% in simulations, Cirrhosis and Type 2 diabetes data respectively, against the next best performing Random Forest technique.

117 | Heterogeneity in obesity and its consequences on health

Jonathan A. Sulc[1,2]*, Anthony Sonrel[1,2], Zoltán Kutalik[1,2]

[1]Swiss Institute of Bioinformatics, Switzerland; [2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

Obesity-associated SNPs have mostly been tested for only one trait in isolation and their joint impact on fat/lean mass accumulation/distribution and downstream effects on health and quality of life remain poorly understood.

We applied principal component analysis on the effect estimates of SNPs on 14 measures of body morphology from the UK Biobank to identify the genetic axes of variation giving rise to differences in body shape and composition. This provided three independent components affecting overall body size, body composition, and body fat distribution, respectively. Our method developed for composite trait Mendelian randomization revealed that these components have both shared and specific effects on health outcomes and quality of life. Of particular interest is the component shifting visceral to subcutaneous fat, which was protective of many obesity-related diseases (such as diabetes, hypertension, hypercholesterolemia, and coronary artery disease) despite being neutral in terms of body mass index and total body fat percentage. A shift in mass from lean to adipose prominently impacted lifestyle, increasing alcohol consumption and smoking. Sex-stratified analyses revealed that increased body size leads to hypothyroidism and decreased socioeconomic status in women alone. Enrichment analyses suggest that brain and nervous tissues contribute most to body size and composition, whereas genes highly expressed in adipose tissue and during development are more likely to affect body fat distribution.

These genetic components provide a basis to better understand the mechanisms underlying inter-individual differences in body fat accumulation and distribution, as well as the consequences they have on health.

118 | Proteomic profiling and protein coregulatory network in plasma, cerebrospinal fluid, and brain tissues for Alzheimer's disease

Yun J. Sung[1,2]*, Chengran Yang[1], Oscar Harari[1], Carlos Cruchaga[1,3]

[1]Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri, USA; [2]Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, USA; [3]Knight Alzheimer Disease Research Center, Washington University School of Medicine, St. Louis, Missouri, USA

Alzheimer's disease (AD), the most common cause of dementia, is a complex neurodegenerative disorder that is characterized by the hallmark pathologies amyloid-beta and tau. Our group and others recently identified several rare variants of triggering receptor expressed on myeloid cells 2 (TREM2) that strongly increases the risk of developing AD. To identify new biomarkers for early diagnosis and treatment, we obtained proteomic profiling of more than 1300 proteins in plasma, cerebrospinal fluid (CSF), and brain tissues from well characterized participants with longitudinal clinical information about disease and cognition. Our high-throughput detection and quantification of proteins was obtained through a SOMAscan, which employs the Slow-Off rate Modified Aptamer (SOMAmer)–based technology. We performed differential analysis and constructed protein coregulation network using weighted correlation network analysis (WGCNA). Pairwise correlation between proteins and network modules of proteins were compared across three tissues. We found multiple proteins with differential abundance levels ($P < 1E\text{-}9$) in AD patients and healthy controls. A subset of these proteins were differentially present ($P < 1E\text{-}9$) in the carriers of TREM2 variants. We evaluated multiple tuning parameters for constructing network to enhance signal-to-noise ratio in the protein adjacency matrix in our data. Sensitivity analyses are underway. Our finding will have implications for the development of biomarkers for clinical diagnostics and can help elucidate the pathways leading to AD brain pathology.

119 | QCprocSE: R package for quality control of processed gene expression data from microarray or RNA-seq experiments

Silke Szymczak[1]*

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

Public repositories such as GEO or ArrayExpress and resources like recount2 provide access to processed gene expression data from thousands of microarray and RNA-seq experiments which can be used as external validation data or jointly analyzed using meta-analysis approaches. However, these data sets might still contain problematic samples or batch effects have not been identified.

The R package QCprocSE integrates several published quality control (QC) steps into a common framework based on SummarizedExperiments objects combining expression values with phenotype information and gene annotation. Samples with a conspicuous distribution of expression values can be detected as outliers in the first two or three principal components (R package gemplot). To detect sample swaps or mix-ups, the reported sex of each sample can be compared to sex predicted based on expression levels of sex specific genes. Duplicated samples have extremely high pairwise correlation (Bioconductor package doppelgangR). Batch effects can be identified by sorting the samples based on processing time and splitting them into batches so that variability between batches is maximized and minimized within (R package BatchI).

Furthermore, the package provides functions to filter lowly expressed genes and to annotate probe sets with gene identifiers, taking care of multiple probe sets mapping to the same gene.

The usefulness of the different QC steps is demonstrated on several examples from different microarray and RNA-seq studies.

120 | A systematic review on the use of methods for left-censored biomarker data

Dominik Thiele[1,2]*, Inke R. König[1,2]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany; [2]Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL), Germany

Classifying patients into subgroups in precision medicine strongly relies on the availability of biomarker data like gene expression profiles. Although there is a huge amount of candidate data, finding suitable profiles still is challenging due to the lack of reproducibility and statistical power. In addition, data is frequently left-censored, and it is yet unclear how best to handle data where a non-negligible proportion has values under a given detection limit. In fact, many approaches have been suggested in the literature that differ with regard to assumptions and application settings. Also, they have been investigated in

different settings as defined, for instance by sample sizes, percentage of non-detects, and the underlying distribution of the data, making the practical choice difficult In this study we therefore target this issue by summarizing published theoretical and simulation studies in which methods for the analysis of left-censored data are suggested and compared with each other. We performed a systematic review following the PRISMA statement to derive an overview of the existing methods and their applicability in different settings. The results will help to guide researchers to select the most suitable method for a specific application.

121 | A Mendelian randomisation method for complex disease exposures

Matthew J. Tudball[1]*, Jack Bowden[1,2], George Davey Smith[1], Kate Tilling[1]

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK; [2]College of Medicine and Health, University of Exeter, Exeter, UK

Background: Mendelian randomisation (MR) is an approach which uses genetic variation in modifiable exposures to identify causal effects in the presence of unmeasured confounding. As MR has become more widespread, an increasing number of studies investigate the effects of complex diseases such as asthma, autism and schizophrenia using diagnosis as the exposure. For such diseases, is common to assume that there is a continuous latent liability, driven by both genetic and environmental factors, and the disease is diagnosed in an individual when this build-up of liability crosses some threshold. However, if liability affects the outcome, rather than diagnosis alone, then the MR assumptions are not satisfied, and the resulting estimates do not have a causal interpretation. This is likely to be a concern when there are a range of subclinical symptoms.

Method: We propose a technique which recovers the effect of latent liability on the outcome in the typical setting where only a binary diagnosis variable is measured. Using commonly available data (e.g. UK Biobank), we can calculate a lower absolute value bound for the effect of latent liability on the outcome on the percentile scale (e.g. the effect of moving from the 10th to 90th percentile of liability). The greater the heritability of the disease, the closer this lower bound is to the true effect. Under some additional assumptions, we can recover the exact liability effect. Our approach is computationally efficient, easy to interpret and provides a considerable improvement over current MR practice for complex disease exposures.

122 | Integration of high-dimensional omics data using sparse orthogonal 2-way partial least squares

Zhujie Gu[1], Said el Bouhaddani[1], Magdalena Harakalova[2], Jeanine J. Houwing-Duistermaat[1,3], Hae-Won Uh[1]

[1]Department of Biostatistics and Research Support, Julius Center, University Medical Center, Utrecht, The Netherlands; [2]Department of Cardiology, University Medical Center, Utrecht, The Netherlands; [3]Department of Statistics, University of Leeds, Leeds, UK

Hypertrophic cardiomyopathy (HCM) is a cardiovascular disease with a prevalence of 1 in 500, which is primarily caused by mutations in several genes, but modulated by other factors, such as epigenetics and gene expression. Our motivating data consists of 15k transcription and 30k methylation levels, respectively, in 13 HCM cases and 10 controls. We aim to construct joint latent components representing the underlying biological system of HCM and identify genes that are involved in it.

We consider unsupervised Partial Least Squares (PLS) related approaches, which decompose correlated two datasets into joint and residual parts. Because omics data are heterogeneous (e.g., different source of variation, scale) and the joint parts estimated by PLS contain data-specific variations, Orthogonal 2-way Partial Least Squares (O2PLS) was proposed to capture the heterogeneity by explicitly including data specific parts and thereby better estimating the joint principal components. As these are linear combinations of all the variables, hampering interpretation, variable selection is needed. An L1 penalty is introduced on the joint loadings of O2PLS to simultaneously perform integration of omics data and variable selection.

An extensive simulation study was conducted to investigate (dis)advantage of sparse O2PLS, which resulted in better performance of sparse O2PLS regarding variable selection and prediction compared to PLS and O2PLS. We applied sparse O2PLS to integrate the HCM datasets. Results showed that the first two sparse joint components of both datasets clearly distinguish the HCM patients and controls, compared to the incomplete separation obtained by applying PCA to each of the omics datasets.

123 | Potential predictive factors for breast cancer subtypes from a North Cyprus cohort analysis

Ayse Ulgen[1]*, Wentian Li[2]

[1]Faculty of Medicine, Girne American University, Kyrenia, North Cyprus via Mersin-10-TURKEY; [2]The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, Northwell Health, Manhasset, New York, USA.

Study conducted to determine predictive factors for breast cancer subtypes for North Cyprus population. More than 300 breast cancer patients with subtype information are surveyed from the State Hospital in Nicosia between 2006 and 2015 for their demographic, reproductive, genetic, epidemiological factors, which represented 40% of total breast cancer cases in the archives during this period. The breast cancer subtypes, Estrogen receptor (ER)+/−, Progesterone receptor (PR)+/−, and human epidermal growth Factor 2 (HER2)+/− status, are determined. Single and multiple variable, regularized regressions, LASSO, with predictive factors as independent variables, breast cancer subtypes as dependent variables are conducted. Despite the fact that our cohort differs significantly from larger cohorts such as the Breast Cancer Family Registry, in age, menopause status, age of menarche, parity, education level, oral contraceptive use, breast feeding, the distribution of breast subtypes is not significantly different. The subtype distribution in our cohort is also not different from another study on a Turkish cohort. Using regularized regressions, we show that the ER+ subtype is positively related to post-menopause and negatively associated with hormone therapy; ER+/PR+ is positively associated with breast feeding, and negatively associated with hormone therapy status. HER2+, which itself is negatively correlated with ER+ and ER+/PR+, is positively related to having first-degree-relative with cancer, and negatively associated with post-menopause. Regressions identify older age to be positively correlated to ER+ and ER+/PR+, negatively correlated to HER2+. To conclude, assuming ER+ and ER+/PR+ to have better prognostic, post-menopause and breast-feeding are beneficial, hormone therapy treatment is detrimental.

124 | Age Prediction in Targeted Whole Genome Methylation Data

Denitsa I. Vasileva[1]*, Ming Wan[1], Allan B. Becker[2], Edmond S. Chan[3], Celia M. Greenwood[4], Catherine Laprise[5], Andrew J. Sandford[1] and 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]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [5]Université du Québec á Chicoutimi, Saguenay, Canada.

Background: The Horvath epigenetic clock is a popular age prediction algorithm widely utilized in adult studies. However, its accuracy in children has not been extensively assessed.

Objective: Evaluate differences in biological age vs. predicted epigenetic age in samples from two Canadian asthma studies.

Study Design: This study was conducted using 812 samples selected from two Canadian samples: Canadian Asthma Primary Prevention Study (CAPPS, n = 632 samples) and the Saguenay-Lac-Saint-Jean (SLSJ, n = 180 samples). The CAPPS study is a longitudinal birth cohort which follows 549 children at high-risk for developing asthma from birth to age 15 with biological samples collected at birth, age seven and 15. The SLSJ study consists of multigenerational families of French Canadian descent, from which three generational triads were selected to evaluate generational effects of methylation.

Methods: Targeted methylation sequencing was conducted using Illumina's MethylCapture (San Diego, California) sequencing library. The Horvath age prediction algorithm was used for age prediction. It consists of 353 age-informative CpGs, of which 324 were targeted in the MethylCapture library, identified using Illumina (San Diego, California) 27K and 450K array data from 8000 mostly adult samples.

Results: In the CAPPS study, relative differences between biological and chronological age were 1.83±1.77 (cord blood), 0.95 ±1.09 (age seven), 0.46 ±0.39 (age 15) and 0.26 ± 0.22 (age >18). The decreasing variance associated with increasing age, may be indicative of limitations of the Horvath algorithm in school aged children, thus illustrating the need for more age specific algorithms for estimating the epigenetic clock in children.

125 | Testing for the absence of causal effects – Mendelian randomization turned around

Maren Vens*, Michelle Kretschmer, Inke R. König

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein Campus Lübeck, Lübeck, Germany

Mendelian randomization is a robust approach for assessing causal relationships using data from observational studies. Genetic variants are used as instrumental variables for causal inference about the effect of a putative causal risk factor on an outcome. Critically, the genetic variants have to fulfill three fundamental assumptions that are strong, whether for estimation or for testing, but at least two of these cannot be tested directly. Since the number of applications of Mendelian randomization is expanding quickly, more attention needs to be paid to these assumptions, and positive and negative results from Mendelian randomization studies are potentially subject to biases from violations.

As a solution, it has been suggested that the strength of Mendelian randomization might not be in the proof but the exclusion of causality, arguing that only in very specific settings would the evidence for no association lack robustness. We therefore propose a statistical procedure following non-inferiority testing for exclusion of causality using Mendelian randomization. Extensive simulation studies show in which settings this approach is robust and can be recommended for practical application.

126 | Efficient simulation of ancestry in large datasets

Georgia Tsambos[1,2], Peter Ralph[3], Jerome Kelleher[4], Stephen Leslie[1,2,5], Damjan Vukcevic[1,2]*

[1]School of Mathematics and Statistics, University of Melbourne, Parkville, Australia; [2]Melbourne Integrative Genomics, University of Melbourne, Parkville, Australia; [3]Department of Mathematics, University of Oregon, Eugene, USA; [4]Big Data Institute, University of Oxford, Oxford, UK; [5]School of BioSciences, University of Melbourne, Parkville, Australia

To assess the performance of methods in population genetics, as well as enable inference via techniques such as approximate Bayesian computation (ABC), we need the ability to simulate realistic genetic datasets while retaining detailed information about the history of the simulated genomes. This is especially important when the primary interest is on patterns of ancestry.

Many existing methods can infer the ancestral origin of chromosomal segments. However, it is difficult to simulate chromosomes for which the true origin of those segments is known; existing approaches are approximate and ad-hoc. We are often interested in the ancestral population that particular genomic segments have been inherited from ("local ancestry"), the amount of recently shared genomic material across individuals ("identity-by-descent"), or patterns of mixed ancestry between multiple populations ("admixture"). Recovering this information from the overall genealogies is challenging. Recent advances allow us to efficiently record genetic information using a succinct tree sequence data structure, which provides unprecedented detail about the genealogy of the sample. However, for the purposes of studying ancestry, this detail can be overwhelming and difficult to analyse.

We present a method that extends these existing state-of-the-art tools to allow efficient extraction of information on local ancestry, identity-by-descent and admixture. We show that this is possible in large simulations under realistically complex demographic scenarios, with minimal computational overhead. To illustrate the usefulness of this procedure, we show how it can be used to benchmark the performance of ancestry inference methods and an example of using ABC to infer demographic models.

127 | Using off-target data from whole-exome sequencing to improve genotyping accuracy, association analysis, and polygenic risk prediction

Jinzhuang Dou[1,2], Degang Wu[1,2], Lin Ding[1], Kai Wang[1], Minghui Jiang[1], E Shyong Tai[3,4,5], Jianjun Liu[5,6], Xueling Sim[3], Shanshan Cheng[1], Chaolong Wang[1]*

[1]Department of Epidemiology and Biostatistics, Key Laboratory for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China; [2]Computational and Systems Biology, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore; [3]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore; [4]Duke-NUS Medical School, National University of Singapore, Singapore, Singapore; [5]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [6]Human Genetics, Genome Institute of Singapore, Singapore, Singapore

Whole-exome sequencing (WES) has been widely used to study the role of protein-coding variants in genetic diseases. Noncoding regions, typically covered by sparse off-target data, are often discarded by conventional WES analyses. Here, we develop a genotype calling pipeline named WEScall to analyse both target and off-target data. We leverage linkage disequilibrium shared within study samples and from an external reference panel to improve genotyping accuracy. In a WES data set of 2,527 Chinese and Malays, WEScall can reduce the genotype discordance rate from 0.26% to 0.08% across 1.1 million SNPs in the deeply sequenced target regions. Furthermore, we obtain genotypes at 0.70% discordance rate across 5.2 million off-target SNPs. Using this data set, we perform genome-wide association studies of 10 metabolic traits. Despite our small sample size, we identify 10 loci at genome-wide significance ($p < 5 \times 10^{-8}$), including

eight well-established loci. The two novel loci, both associated with glycated haemoglobin levels, are *GPATCH8-SLC4A1* (rs369762319, $p = 2.56 \times 10^{-12}$) and *ROR2* (rs1201042, $p = 3.24 \times 10^{-8}$). Finally, using summary statistics from UK Biobank and Biobank Japan, we show that polygenic risk prediction can be significantly improved for six out of nine traits by incorporating off-target data. These results demonstrate WEScall as a useful tool to facilitate WES studies.

128 | A novel powerful eQTL weighted gene based association test using GWAS summary data

Jianjun Zhang[1], Sicong Xie[2], Jianguo Liu[1], Xuexia Wang[1]*

[1]Department of Mathematics, University of North Texas, Denton, Texas, USA; [2]Beijing National Day School, Beijing, China

Although genome-wide association studies (GWASs) have identified many genetic variants underlying complex traits, a large fraction of heritability still remains unexplained. Integrative analysis that incorporates additional information such as expression quantitative trait locus (eQTL) data into sequencing studies (denoted as transcriptome-wide association study [TWAS]) can aid the discovery of trait associated genetic variants. However, general TWAS methods only incorporate one eQTL-derived weight (e.g. cis-effect), and thus can suffer a substantial loss of power when the single estimated cis-effect is not predictive for the effect size of a genetic variant or when there are estimation errors in the estimated cis-effect. In this study, we propose an omnibus test which utilizes a Cauchy association test to integrate association evidence demonstrated by three different traditional tests (Burden test, Quadratic test, and Adaptive test) using GWAS summary data with multiple eQTL-derived weights. The *P* value of the proposed test can be calculated analytically, and thus it is fast and efficient. We applied our proposed test to two schizophrenia (SCZ) GWAS summary datasets and two lipids trait (HDL) GWAS summary datasets. Compared to the three traditional tests, our proposed omnibus test can identify more trait-associated genes. For example, the gene ZNF184 was uniquely identified by the proposed method in the SCZ1 data, and validated by the results from the SCZ2 data set. ZNF184 encodes for a zinc-finger protein involved in transcriptional regulation, may regulate the expression of schizophrenia associated genes, and influence therapeutic response, supporting a plausible biological mechanism underlying the disease.

129 | Whole exome sequencing of severe asthma identifies novel gene association candidates

Zuoheng Wang[1], Nayang Shan[1], Xiting Yan[1,2], Shrikant Mane[3], Jose L Gomez[2], Geoffrey L. Chupp[2]

[1]Department of Biostatistics, Yale School of Public Health; [2]Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, Yale School of Medicine; [3]Department of Genetics, Yale School of Medicine

Rationale: Severe asthma is a heterogenous, chronic inflammatory disease of the airways that is driven by genetic risk and environmental exposure. Severe asthma patients are at a high risk of exacerbations which may lead to hospitalization and death. The advent of new sequencing technology has enabled us to deep sequence the human genome to identify rare and common genetic variants, genes and pathways influencing the risk of severe asthma. From the Yale Center for Asthma and Airway Disease Asthma Cohort, we identified patients with a history of severe asthma, irreversible airflow obstruction, near-fatal asthma, and family history of asthma in 3 or more first degree relatives. We hypothesized that this extreme asthma phenotype will be enriched for similar genetic variants even though they are not related. We sought to examine this hypothesis by performing whole exome sequencing (WES) in this subset of patients.

Methods: We performed WES on 39 patients with severe asthma and ≥ 3 first degree relatives with asthma. Genotypes were obtained from the mapped sequence reads to human reference genome hg38. Rare allele counts on the patients were compared with the 1000 Genomes using the binomial proportion test. Association tests were performed with clinical phenotypes of asthma. Pathway enrichment analysis was applied to the top lists of variants and genes using MetaCore to identify significant pathways associated with both rare and common genetic variants identified in these analyses.

Results: 112,241 variants, including both common and rare, were obtained from the WES data. Among them, 429 variants showed enrichment of rare alleles in patient samples compared to the 1000 Genomes (FDR < 0.05). Ninety-three were nonsynonymous variants. Pathway analyses on the variants demonstrated that genes involved in airway smooth muscle contraction in asthma (FDR = 0.0045), and immune response CCR3 signaling in eosinophils (FDR = 0.032) were enriched in the 93 nonsynonymous variants. These genes include Telokin, the IP3 receptor, MyHC and MYLK. Among the immune response pathways were the IL4 signaling pathway and Th2 cytokine induced mucous metaplasia were found to

be associated with hospitalization, pre- and post- FEV1/FVC ratio, sputum YKL40 level, and eosinophil counts.

Conclusions: WES of severe asthma with a strong family history demonstrated enrichment of non-synonymous variants in genes involved in smooth muscle function and eosinophil signaling. This suggests that combined perturbations in airway smooth muscle contraction and T2 inflammation underlie risk of developing severe near fatal asthma.

130 | Incorporating multiple sets of eQTL weights into gene-by-environment interaction analysis identifies novel susceptibility loci for pancreatic cancer

Tianzhong Yang[1,2], Hongwei Tang[1,3], Chris I Amos[4], Donghui Li[1], Peng Wei[1]*

[1]The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [2]University of Minnesota, Minneapolis, MN, USA; [3]University of Colorado, Denver, CO, USA; [4]Baylor College of Medicine, Houston, TX, USA

**Presenting author*

It is of great scientific interest to identify the interactions between genetic variants and environmental exposures that may modify the risk of complex diseases. However, larger sample sizes are usually required to detect gene-by-environment interaction (GxE) than required to detect genetic main association effects. To boost the statistical power and improve the understanding of the underlying molecular mechanisms, we incorporate functional genomics information, specifically, eQTL, into a data-adaptive GxE interaction test, called aGEw. This test not only addresses eQTL weights from multiple tissues but also provides an extra layer of weighting at the genetic variants level. Using extensive simulations, we show that the aGEw test can control the Type 1 error rate, and the power is resilient to the inclusion of neutral variants and non- informative external weights. We applied the proposed aGEw test to the Pan- creatic Cancer Case-Control Consortium GWAS (discovery cohort of 3,585 cases and 3,482 controls) and the PanScan II GWAS (replication cohort of 2,021 cases and 2,105 controls) with smoking as the exposure of interest. We identified two novel putative smoking-related pancreatic cancer susceptibility genes, TRIP10 and KDM3A. We have also implemented the aGEw test in an R package "aGE".

Keywords: Data-adaptive association testing, eQTL, Gene-environment interaction, Multiple functional weights, Pancreatic Cancer, PrediXCan

131 | Genetic colocalisation networks to inform and validate biological protein interactions

Hannah F. Wilson*, Eleanor Sanderson, Gibran Hemani

Medical Research Council Integrative Epidemiology Unit, University of Bristol, Population Health Sciences, University of Bristol

A better understanding of the human proteome can lead to new drug targets and genetic therapies for disease. Protein quantitative trait loci (pQTLs) are genetic variants that influence protein abundance. Undertaking genetic colocalisation to determine which proteins have a shared causal variant, then using this to build a protein interaction network, can increase our understanding of genetic regulation on the protein pathway.

To investigate protein interactions, we use 1,699 proteins that have at least one pQTL, and undertake colocalisation analyses to determine if they share a causal genetic variant with any of the 3,369 proteins for which we have summary statistics. Using only pQTLs, and not the full genome of variants, reduced our multiple testing burden allowing for the possibility of more associations to be found. Testing if trans effects arise due to shared causal variants, we distinguish between statistical pleiotropy, the causal variants for two proteins are in linkage disequilibrium, and biologically pleiotropy, there is one variant which has a causal effect on the two proteins. Using this knowledge, we determine the specificity of these pQTLs as instruments in Mendelian randomisation analyses, perform pathway enrichment analyses to determine if statistical or biological pleiotropy is more prevalent for certain gene sets, and compare the proteins we found to be causally associated, to known biological protein interactions. Combining the results from these analyses we create a protein interaction network that increases our understanding of genetic influences on protein interactions and informs future studies of the specificity and usefulness of pQTLs.

132 | Using polygenic, APOE and familial risk for Alzheimer's to identify potential mediators of Alzheimer's disease in the UK Biobank

Hei Man Wu*, Paul F. O'Reilly

Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, USA

Establishing lifestyle factors on the causal path to Alzheimer's Disease (AD) provides one way of reducing AD prevalence. Here we investigate the downstream effects of AD genetic risk to identify potential mediators. Three

forms of AD genetic risk were investigated: polygenic risk score for AD (excluding the APOE gene), APOE e4 gene dosage, and family risk of AD.

Among 275 traits tested, we identified 77 traits associated with at least one form of AD genetic risk. Contrary to expectations, we observed that AD genetic risk is associated with a range of better health traits, for example, reduced red meat intake, increased oily fish intake, and lower BMI. However, when we accounted for parental age, most of the associations between family risk for AD and these traits were eliminated, indicating "survivor effects" impacting the original results. Nevertheless, the association between the APOE-e4 risk allele and positive health traits remained. To account for the possibility of lifestyle changes due to family history or diagnosis of cardiovascular disease explaining this, we re-ran the analyses in the sample of individuals who were reportedly free of cardiovascular problems, statin use, and family history of high blood pressure, stroke and heart disease. With these factors accounted for the associations were no longer significant, suggesting lifestyle changes due to awareness of cardiovascular problems. Moreover, we observed highly significant associations (P-value $< 1 \times 10^{-100}$) between the APOE-e4 allele and blood lipid traits (P-value (Apolipoprotein) $= 0$; p-value (LDL direct) $= 3.09 \times 10^{-265}$; P-value (Cholesterol) $= 1.32 \times 10^{-200}$;) Finally, there were significant interactions between the APOE-e4 allele and age, including, interestingly, interactions that appear to be sex-specific (Apolipoprotein B: p-value (female) $= 5.91 \times 10^{-5}$; P-value (male) $= 0.352$)).

133 | Exploiting polygenic risk scores and family data in the UK Biobank to infer de novo or rare deleterious alleles

Hei Man Wu*, Shing Wan Choi, Paul F. O'Reilly

Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, USA

Extreme values of a trait are often subject to purifying selection and thus we may expect individuals in the tails of trait distributions to be enriched for de novo or rare deleterious alleles. Therefore, individuals with extreme trait values may have an average polygenic risk score (PRS), since PRS are based on common variant effects, even if there is a strong linear trend between PRS and trait in most of the population. This "polygenic regression-to-the-mean" thus provides a potential way to identify individuals harbouring de-novo/rare alleles. Moreover, such evidence can be verified with availability of sibling data; individuals with extreme trait values caused by a de novo/rare allele are more likely to have siblings with typical trait values.

We utilized the UK Biobank data to test for enrichment of non-polygenic aetiology at the extremes of a range of continuous traits, using these two orthogonal approaches. We also tested whether extreme trait individuals had older fathers and lower fecundity.

We found significant results for both the PRS deviation test and sibling analysis, and observed significant consistency between the results of these two approaches ($R = 0.280$, $P = 1.59 \times 10^{-5}$). This suggests that rare/de novo mutations may be enriched at the extreme ends of the distributions of certain traits. No significant associations were observed in the paternal age or fecundity tests.

Our approach may enable the selection of traits and samples likely to harbour de novo/rare alleles, maximizing the power of sequencing studies. The same approach could be applied to disease traits with continuous measures of severity or age-of-diagnosis data.

134 | Variable selection in nonparametric additive quantile regression for genetic or genomic data with a priori information

Peitao Wu, Josée Dupuis, Ching-Ti Liu

Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

A priori information, such as biological pathways, is a useful supplement in identifying risk factors of a trait using genomic data. In addition, investigating the whole spectrum of the traits distribution may enhance the discovery of the genetics or genomics markers and provide better insights into the underlying biological mechanism. However, the commonly used methods to incorporate prior information evaluate the mean function of the outcome only and rely on unmet assumptions. To address these concerns, we propose a nonparametric additive quantile regression with network regularization to incorporate known biological networks or pathways. To account for nonlinear associations, we approximate each predictor's additive functional effect with the expansion of a B-spline basis. We implement the group Lasso penalty to obtain a sparse model. We define the network-constrained penalty by the total l2 norm of the difference between the effect functions of any two predictors that are linked in the known network. We further propose an efficient computation procedure to optimize our objective function. Simulation studies show that our proposed method performs well in identifying more truly

associated variables/genes and less falsely associated variables/genes than alternative approaches. We apply the proposed method to analyze the microarray gene-expression data set in the Framingham Heart Study and identify several novel body mass index associated genes in addition to some previously identified genes. In conclusion, our proposed approach efficiently identifies the trait-associated variables/genes in a nonparametric additive quantile regression framework by leveraging known network or pathway information.

135 | Genome-wide DNA methylation profiling reveals diagnostic biomarkers for esophageal squamous cell carcinoma

Yiyi Xi[1], Xinyu Wang[2], Jianzhong Su[2], Chen Wu[1,3,4], Dongxin Lin[1]

[1]Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; [2]Institute of Biomedical Big Data, Wenzhou Medical University, Wenzhou, China; [3]Jiangsu Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China; [4]CAMS Key Laboratory of Genetics and Genomic Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Esophageal squamous cell carcinoma (ESCC) is a common malignancy with early detection crucial to patient survival. And the aberrant DNA methylation in ESCC serves early diagnosis but lacks comprehensive profiling. Through genome-wide profiling of 91 paired samples from Chinese ESCC patients, we identified 35,577 differentially methylated CpG sites (DMCs) with high confidence. Our integrated analysis of gene expression and methylation profiles revealed expression of 344 genes associated with gene-body DNA methylation. We further established a methylation-based model to identify malignant samples with eight markers of diagnosis relevance. The model achieved high level of accuracy in distinguishing tumor samples from normal ones in our training set ($AUC = 99\%$) and validation set ($AUC = 98\%$). We also integrated samples of esophagus mucosa from healthy individuals into the entire ESCC cohort; the model can separate tumors clearly from adjacent tissues or normal esophagus mucosae ($AUC = 98\%$). Application of the classifier in pan-squamous cell carcinoma from TCGA proved comparable performance in patient identification in ESCC ($AUC = 96\%$), HNSC ($AUC = 97\%$) and LUSC ($AUC = 96\%$). In conclusion, we present the comprehensive profiling of aberrant DNA methylation in 91 ESCC patients within underlying biological process leading to tumorigenesis. And our results provide promising potentials of methylation-based diagnosis for ESCC.

Keywords: Esophageal squamous cell carcinoma; DNA methylation; Biomarker; Diagnosis

136 | Prism vote: A stratified statistical framework to perform prediction for complex diseases

Xiaoxuan Xia*, Rui Sun, Maggie Haitian Wang#

[1]The Jockey Club School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, Hong Kong SAR; [2]The Chinese University of Hong Kong Shenzhen Research Institute, Shenzhen, China
#Correspondence
**Presenting author*

Accurate risk prediction using genetic data is essential for informing future health status, disease targeted prevention and intervention. Ethnicity-specific trait loci or heterogeneous genetic effects across subpopulations pose a challenge to complex traits prediction. In this paper, we introduced a statistical framework, Prism Vote (PV), to perform phenotype classification. In PV, training samples are stratified to several strata, and each test sample will be assigned probabilistically to each stratum or mixed across multiple strata according to its genetic structure. The predicted phenotype of a subject is calculated by a weighted sum of predictions by all strata based on a Bayesian formula. An index is derived to balance the tradeoff between a decreased sample size and a tailored prediction model within each stratum. The PV framework could be implemented with most prediction models, such as logistic regression (LR) and polygenic risk score. We applied PV to two real GWAS data sets, Alzheimer's disease (AD) within the Caucasian population, and schizophrenia (SCZ) across European and African ancestries. For AD, when using genome-wide significant SNPs from literature, the stratified analysis improves the mean AUC from 0.775 to 0.782. For SCZ, PV + LR improves LR with the top 10 principal components (PCs) from 0.664 to 0.676, when using associated SNPs by the target data. The PV could serve as a generalized prediction model to include heterogeneous genetic ancestry in risk prediction.

137 | A fast clustering algorithm and sampling strategy implemented in preparing samples for a large multiethnic genome-wide meta-analysis to identify lung cancer susceptibility

Xiangjun Xiao[1,2], Yafang Li[1,2], Christopher Amos[1,2]

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX; [2]Departement of Medicine, Epidemiology and Population Sciences, Baylor College of Medicine, Houston, TX

In genome-wide analysis (GWAS) studies, there are many issues that challenge the performance of valid statistical analyses, such as the presence of duplicates, sample contamination, sample swapping, cryptic relatedness and pedigree structure. Mixed models have been developed to resolve the relatedness problem, however, for the large datasets we are analyze that have over 100,000 samples, the computing efficiency and reliability are not very satisfying. In this study, we developed a fast clustering algorithm to cluster 101,821 samples using pair-wise identity by descent (IBD) values and generated a maximum sample set using a novel weighted strategy.

Pair-wise IBD values of 101,821 samples were estimated based on around 190,000 random markers of low LD pattern using PLINK in a multithread setting. We used an empirical value of IBD = 0.15 as cutoff to define samples as having related status, and then a sequential query algorithm was used to grow clusters through dynamically querying all samples left. In this way, 15,884 unique clusters were identified. We used a novel sampling method of step-wise querying to generate a list of candidate less dependent samples for each cluster, we put different weights on these samples with properties, such as disease status and study specific measurements, and so forthto assist in samples to retain. Finally, we chose an entry with maximum score in each cluster. We subsequently generated a data set with 70,639 samples using above strategies in contrast with 46,811 samples with default unrelated setting in KING.

Our method was implemented through a Java/R pipeline with running time of several seconds for 101,821 samples, and it can be extended to use various types of existing pair-wise genetic measurements. We next compared a mixed effects analysis, which can integrate data from multiple racial populations with meta-analysis stratified by ethnic background. Genome-wide analysis of a case-control study of lung cancer and using around 6 million markers identified more significant results with 70,639 sample set than ones with 46,811 sample set systemically (R2 = 0.23). We further compared the performance between meta-analysis and mixed model implemented in GMMAT method, meta-analysis has more significant results than GMMAT. There was a strong correlation between the results for GMMAT and the fixed effects meta-analysis (R2 = 0.7). Although there were more significant results, t even though the GMMAT analysis retained more samples. This result suggests that our method can provide statistically valid and more robust analysis.

138 | Incorporation of rare genetic variants improved the prediction performance of polygenic risk score

Hongyan Xu*

Department of Population Health Sciences, Augusta University, Augusta, Georgia, USA

Polygenic risk score combines the genetic contribution from multiple genetic variants across the genome and has the potential clinical utilities in predicting disease risk for common human diseases. Current polygenic risk scores are based on the results from genome-wide association studies, where the information are from common genetic variants. With the availability of genome sequencing data, many rare genetic variants have been shown to be associated with the risk of common human diseases. In this study, we build a polygenic risk score based on both common and rare genetic variants. We incorporate the effects of rare genetic variants by dividing them into subgroups according to the signs of the effect and combine the rare genetic variants in one subgroup with a genetic load. Results from our simulation study show that our polygenic risk score has improved predictability measured by the C-statistic. Our polygenic risk score also has lower prediction error from cross-validation than the risk score without rare genetic variants. We applied our polygenic risk score method to the breast cancer data from the Cancer Genome Atlas to predict the risk of developing breast cancer.

139 | Clustering of human microbiome sequencing data: A distance-based unsupervised learning model

Dongyang Yang[1]*, Wei Xu[1,2]

[1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [2]Department of Biostatistics, Princess Margaret Cancer Centre, UHN, Toronto, Ontario, Canada

Analysis of the human microbiome allows the assessment of the microbial community and its impacts on human health. Microbiome composition can be

quantified using 16 S rRNA technology into sequencing data which are usually skewed and heavy-tailed with excess zeros. Clustering approaches are useful in personalized medicine by identifying subgroups for patients stratification. However, there is currently a lack of standardized clustering method for such complex microbiome sequencing data. We propose a clustering algorithm with a specific beta diversity measure that can address the presence-absence bias encountered for sparse count data and effectively measure the sample distances for stratification. A parametric based mixture model is developed to model the microbiome composition distribution by incorporating a set of mixture components. Our distance matrix used for clustering is obtained from sample-specific distributions conditional on the observed operational taxonomic unit (OUT) counts and estimated mixture weights. The method can provide accurate estimates of the true zero proportions and thus construct a precise beta diversity measure. Extensive simulation studies have been conducted and suggest that our proposed method achieves substantial improvement compared with some widely used distance measures for clustering analysis when a large proportion of zeros is presented in the data. Specifically, accuracy rates improve 9.1% and 9.2% compared to Manhattan and Euclidean distances for the 2-clusters scenario and 9.4% and 7.9% for the 3-clusters scenario. We apply our method to human microbiome study to identify distinct microbiome states and compare our method to the other clustering methods.

140 | Population pharmacogenomics: Enrichment of ancestry-informative markers in pharmacogenetic loci

Hsin-Chou Yang[1,2,3]*, Chia-Wei Chen[1], Yu-Ting Lin[1], Shih-Kai Chu[1]

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan;* [2]*Institute of Statistics, National Cheng Kung University, Tainan, Taiwan;* [3]*Institute of Public Health, National Yang-Ming University, Taipei, Taiwan*

Recent studies have pointed out the essential role of genetic ancestry in population pharmacogenetics. In this study, we analyzed the whole-genome sequencing data from The 1000 Genomes Project (Phase 3) and the pharmacogenetic information from Drug Bank, PharmGKB, PharmaADME, and Biotransformation. We found that ancestry-informative markers were enriched in pharmacogenetic loci (PGx), suggesting that trans-ancestry differentiation must be carefully considered in population pharmacogenetics studies. Ancestry-informative PGx were located in both protein-coding and non-protein-coding regions, illustrating that a whole-genome analysis is necessary for an unbiased examination over PGx. Finally, those ancestry-informative PGx that targeted multiple drugs were often a functional variant, which reflected their importance in biological functions and pathways. This study developed an efficient algorithm for an ultrahigh-dimensional principal component analysis, created genetic catalogues of ancestry-informative markers and genes, established a high-accuracy prediction panel of genetic ancestry, explored pharmacogenetic patterns, and constructed a genetic ancestry pharmacogenomic database "Genetic Ancestry PhD" (http://hcyang.stat.sinica.edu.tw/databases/genetic_ancestry_phd/).

141 | Whole genome sequencing of skull-base chordoma reveals genomic alterations associated with local recurrence and chordoma-specific survival

Jiwei Bai[1], Jianxin Shi[2], Chuzhong Li[1], Shuai Wang[1], Tongwu Zhang[2], Bin Zhu[2], Hela Koka[2], Alisa M. Goldstein[2], Yazhuo Zhang[1], Xiaohong R. Yang[2]*

[1]*Beijing Neurosurgery Institute, Capital Medical University, Beijing, China;* [2]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD, USA*

Chordoma is a rare bone tumor with an unknown etiology and high recurrence rate. Currently, no validated clinical or molecular prognostic panel is available to predict recurrence. In addition, treatment options are limited for patients with advanced disease. Genomic analyses of chordoma are scarce and have been limited to sacral chordoma. Here, we conducted whole genome sequencing of 80 skull-base chordomas and identified *PBRM1*, a SWI/SNF (SWItch/Sucrose Non-Fermentable) complex subunit gene, as a significantly mutated driver gene. Genomic alterations in SWI/SNF genes, including mutations and structural variations in *PBRM1* and *SETD2*, were the most prevalent events (16%). SWI/SNF gene alterations and the chromosome 22q deletion, which involved another SWI/SNF gene (*SMARCB1*), showed strong associations with poor chordoma-specific survival (CSS) and worse recurrence-free survival (RFS). Despite the low mutation rate, extensive somatic copy number alterations (SCNAs) frequently occurred. Most SCNAs were clonal and showed highly concordant profiles between paired primary and recurrence/metastasis samples, indicating the importance of SCNAs in chordoma initiation. Our results provide important biological

and clinical insights into skull-base chordoma and demonstrate the potential of designing a multi-marker panel in prognostic prediction.

142 | Effect of dimension reduction using local principal components in regression based multi-SNP analysis

Fatemeh Yavartanoo[1]*, Shelley B. Bull[2,3], Andrew D. Paterson[3,4], Myriam Brossard[2], Delnaz Roshandel[4], Yun Joo Yoo[1,5]

[1]Department of Mathematics Education, Seoul National University, Seoul, South Korea; [2]Prosserman Centre for Population Health Research, The Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [3]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [4]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada; [5]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea

High density SNP arrays and next generation sequencing (NGS) create datasets consisting of large numbers of markers with complex correlation structures. Regional regression can be used to model the relationship between multiple genotypes and a phenotype. However, multicollinearity due to linkage disequilibrium (LD) can cause instability in the regression and difficulties in detection of association, requiring strategies for efficient dimension reduction. We developed a method of Dimension Reduction using Local Principal Components (DRLPC) in which the highly correlated SNPs within a region are clustered and replaced by the first principal component constructed locally among the SNPs in the cluster before the regression analysis. DRLPC can also remove some SNPs to help resolve multicollinearity and reduce dimension, under the assumption that the effect of a removed SNP can be captured by the remaining SNPs due to high linear dependency.

To evaluate power gains achieved by dimension reduction, in this study, we compared the power of several multi-SNP statistics constructed for regional analyses. These statistics include: multiple *df* generalized Wald tests, single and multiple linear combination (MLC) tests, a global principal component regression test, SSB, SKAT and minimum P statistics. In simulation studies based 100 genes with less than 500 SNPs on chromosome 22 data of 1000 Genomes Project, DRLPC effectively reduces dimension up to 45~80% and has higher power than the Wald test. However, MLC tests which already reflect the cluster structure between SNPs showed little power benefit.

143 | Identifying differentially methylated regions via sparse conditional Gaussian graphical models

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*

Motivation: Identifying differentially methylated regions (DMRs) between two or more conditions is of interest, because regional changes in methylation could be associated with disease traits and other covariates such as age and sex. As methylation status at nearby CpG sites have been shown to be more dependent than those that are far away in distance, one possible analytic method that captures this local dependence structure while adjusting for potential covariates could be a conditional Gaussian graphical model (CGGM). We propose to develop a modified CGGM incorporating a local dependence structure for modelling differential methylation.

Methods: We propose a model with conditional spatial local dependence among CpG sites in the DMRs that estimates (a) the parameter matrix for mapping covariate inputs to methylation outputs; and (b) the network over outputs conditioning on inputs based on estimating the Cholesky factor of the inverse covariance matrix. The method involves solving a penalized Gaussian likelihood problem with a hierarchical group penalty on the Cholesky factor and a penalty on the parameter matrix. The problem of estimating the Cholesky factor can be decomposed into several independent subproblems which can be solved efficiently in parallel. The method yields a sparse, block-diagonal, symmetric and positive definite estimator of the inverse covariance matrix, which leverage the fact that methylation at CpG sites within clusters are spatially locally dependent.

Results and Conclusions: Simulation results and implementation on targeted bisulfite sequencing data will be presented at IGES 2020.

144 | Platforms comparison and bias correction for measuring DNA-methylation: The Illumina beadchip versus custom capture bisulfite sequencing

Ting Zhang[1,2]*, Khaoula Belahsen[2,3], Kathleen Oros Klein[2], Antonio Ciampi[1,4], David Stephens[5], Aurelie Labbe[1,4,6,9], Marie Hudson[2,8], Celia MT Greenwood[1,2,4,7,10]

[1]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada; [2]Lady Davis Institute for Medical Research, Montreal, Quebec, Canada; [3]École Polytéchnique, Palaiseau, France; [4]Ludmer Centre for Neuroinformatics and Mental Health, Montreal, Quebec, Canada; [5]Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada; [6]Department of Psychiatry, McGill University, Montreal, Quebec, Canada; [7]Department of Oncology, McGill University, Montreal, Quebec, Canada; [8]Department of Medicine, McGill University, Montreal, Quebec, Canada; [9]Douglas Mental Health University Institute, Montreal, Quebec, Canada; [10]Department of Human Genetics, McGill University, Montreal, Quebec, Canada.

DNA methylation can be measured with several platforms, each with its own inherent strengths, limitations and biases. To illustrate and better understand the biases associated with two platforms, we analysed and compared the estimated methylation levels from the Illumina 450 K array and targeted custom capture bisulfite sequencing (TCCBS) on 42 samples from 20 individuals.

In this study, Illumina 450 K data were normalized with SWAN and funtooNorm to remove probe type and colour channel-related biases. For TCCBS data, we built zero inflated negative binomial models for the methylated and unmethylated counts separately, as a function of GC content. Measurement of bias and its corresponding correction method are proposed based on the model estimated counts. We then model the discrepancy between estimated methylation levels from two platforms and ranked the importance of predictors by a random forest model.

The absolute difference between the estimates of methylation level depends strongly on the average level of methylation at the probe. At high mean methylation levels (>60%), the TCCBS platform tends to give higher estimates than the Illumina platform, but the opposite is true when mean methylation levels are below 60%. Other important factors include GC content, methylation quality, and cell type. The normalization method for the Illumina data did not make a large impact. These results will be further compared before and after correction of TCCBS data for GC content. Overall, this study can help to form guidelines for methylation data preprocessing.

145 | hESCCs express albumin to assist metastasis through JunD phosphorylation

Ce Zhong*, Xiang Jie Niu, Yuling Ma, Chuanwang Miao, Qionghua Cui, Yuqian Wang, Xinjie Chen, Wen Tan, Chen Wu, Dongxin Lin

Department of Etiology and Carcinogenesis, National Cancer Center/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 17 Panjiayuan Nanli, Chaoyang District, Beijing, China

Human serum albumin (HSA) is the most abundant plasma protein that binds fatty acids, bilirubin, thyroxine and hemin. HSA is expressed specifically in healthy liver cells regulated by gene *ALB*. HSA plays a vital role in substance exchange. On RNA and protein level, we find that both human esophageal squamous cell carcinoma (hESCC) tissue and hESCC precursor tissue express HSA. We then explore how hESCC tissue breaks through the heterogeneity of organs. Using MS, EMSA and online promoter site prediction tool, we find that JunD is a promoter of gene *ALB*. With the GOF(gain-of-function) mutation of gene *map2k7*, an upstream gene of JunD, phosphorylation of JunD is enhanced, as a result, hESCC cells express gene *ALB*. To explore how HSA produced in hESCC works and whether it works in the same way as it does produced by healthy liver tissue, we use flow cytometry and mouse model. And we find that HSA assist tumor metastasis through ROS cleavage. With the online prognosis data that has already been published and our single cell hESCC gene expression data, we reveal that hESCC express HSA to assist tumor metastasis.

146 | Precision improvement for Mendelian randomization median method

Yineng Zhu*, Qiong Yang

Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA

Mendelian Randomization (MR) is a method that uses genetic variants as instrumental variables when inferring causal relationships between an exposure and an outcome, which overcomes the inability to infer such relationship in observational studies due to unobserved confounders. However conventional MR is sensitive to pleiotropy, which renders the instrumental variables invalid. Recently median-based MR method was proposed that provides consistent causal estimates when 50% of

genetic variants are invalid instrumental variables. Similarly, weighted median method and penalized weighted median method provide more efficient and less biased estimates under certain conditions. However, when some of the genetic variants are correlated, the standard errors and confidence intervals of median-based MR are not valid since these methods use bootstrap to calculate standard errors, which assumes independent resampling units. Here we proposed a quasi-bootstrap method that uses Cholesky decomposition to deal with the correlated data problem. We show that the new method has valid type one errors and power increases tremendously from 0.63 (weighted median method, other median methods have very close powers) to 0.95 for some scenarios using simulation studies when our new method was used to calculate standard errors and confidence intervals.