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

The 2015 Annual Meeting of the International Genetic Epidemiology Society

Published online 14 September 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21916

1

A comparison of methods for inferring causal relationships between genotype and phenotype using multi-omics data

Holly F. Ainsworth (1) So-Youn Shin (2) Heather J. Cordell (1)
(1) Institute of Genetic Medicine, Newcastle University (2) MRC Integrative Epidemiology Unit (IEU), Bristol University

Many novel associations between common genetic variants and complex human disease have been successfully identified using genome-wide association studies (GWAS). However, a typical GWAS gives little insight into the biological function through which these associated genetic variants are implicated in disease. Indeed, rather than finding variants which directly influence disease risk, the variants implicated by GWAS are typically in linkage disequilibrium with the true causal variants. Understanding the causal role of the genetic variants in disease etiology and moving towards therapeutic interventions is not simple. Integration of additional data such as transcriptomic, proteomic and metabolomic data, measured in relevant tissue in the same individuals for whom we have genomic (i.e. GWAS) data, could potentially provide further insight into disease pathways. Yet, open questions remain on how to assess the causal direction of association between these variables.

We review currently available statistical methods for inferring causality between variables that use a genetic variant as a directional anchor. We consider Mendelian Randomisation, Structural Equation Modelling, a Causal Inference Test and several Bayesian methods. We present a simulation study assessing the performance of the methods under different conditions, assuming throughout that we have a single genetic variant and two phenotypic variants that are associated with one another, although the underlying causal relationship may vary. In particular, we consider how the causal inference is affected by the presence of common environmental factors influencing the observed traits.

2

Shared polygenic effects of FEV1 in the first genetic study in UK Biobank

Richard J. Allen (1) Louise V. Wain (1) Nick Shrine (1) Suzanne Miller (2) Victoria E. Jackson (1) Ioanna Ntalla (1) Maria Soler Artigas (1) James P. Cook (1) Andrew P. Morris (3) Eleftheria Zeggini (4) Jonathan Marchini (5,6) Panos Deloukas (7) Anna Hansell (8) Richard Hubbard (9) Ian Pavord (10) Neil C. Thompson (11) David P. Strachan (12) Ian P. Hall (2) Martin D. Tobin (1,13), UK BiLEVE Consortium
(1) Department of Health Sciences, University of Leicester (2) Division of Respiratory Medicine, University of Nottingham (3) Department of Biostatistics, University of Liverpool (4) Wellcome Trust Sanger Institute (5) Department of Statistics, University of Oxford (6) Wellcome Trust Centre for Human Genetics, University of Oxford (7) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University London (8) Faculty of Medicine, School of Public Health, Imperial College London (9) Faculty of Medicine and Health Sciences, School of Medicine, University of Nottingham (10) Respiratory Medicine, University of Oxford (11) Institute of Infection, Immunity & Inflammation, University of Glasgow (12) Population Health Research Institute, St George's University of London (13) National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital

Forced Expiratory Volume (FEV1), i.e. how much air an individual can forcibly exhale in one second, is reduced in obstructive lung diseases. The aims of this study were to investigate polygenic effects of FEV1 and how these are shared between different subsets of the population.

50,008 subjects of European ancestry with high quality spirometry measurements, half of whom were never smokers and the rest heavy smokers, were sampled from the extremes and middle of the percent predicted FEV1 distribution in UK Biobank. The dataset was split into a discovery group and a target group. Polygenic risk scores were generated for each individual in the target group based on single SNP analysis in the discovery group. These scores were regressed against FEV1 to identify shared polygenic effects between the two groups. The dataset was split into a discovery and target group 1) arbitrarily, 2) by discovery group of those with high FEV1 and target of low FEV1, 3) by discovery group of heavy smokers and target of never smokers and 4) by discovery group of those with doctor diagnosed asthma and target group of those without doctor diagnosed asthma.

There are shared polygenic effects between high FEV1 and low FEV1 ($p = 1.64 \times 10^{-22}$), for low FEV1 between never and heavy smokers ($p = 2.29 \times 10^{-16}$) and for low FEV1 between those with and without doctor diagnosed asthma ($p = 6.06 \times 10^{-11}$).

These results show that FEV1 is affected by many variants, most having very small effect sizes, explaining why

The Article title is changed to "The 2015 Annual Meeting of the International Genetic Epidemiology Society" from "Annual Meeting of the International Genetic Epidemiology Society" on 21st September 2015 after original publication date.

they were not identified in GWAS to date. The whole of UK Biobank is currently being genotyped giving future analyses greater power to detect variants with modest effect sizes.
This research was conducted using the UK Biobank Resource.

3

Using methylation quantitative trait loci to enhance GWAS results for autism spectrum disorder across developmental stage and tissue type

Shan V. Andrews (1,2) Shannon E. Ellis (3) Kelly M. Bakulski (1,2,4) Christine Ladd Acosta (1,2) Andrew P. Feinberg (4) Dan E. Arking (3) M. Daniele Fallin (2,4,5)
(1) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health (2) Wendy Klag Center for Autism and Developmental Disabilities, Johns Hopkins Bloomberg School of Public Health (3) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine (4) Center for Epigenetics, Johns Hopkins University School of Medicine (5) Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

Many of the SNPs previously associated with autism spectrum disorder (ASD) are intragenic and/or do not have a clear functional consequence. A potential function may be expression regulation via epigenetics. Examining methylation quantitative trait loci (meQTLs), or SNPs that appear to control DNA methylation (DNAm) levels at particular CpG sites, with respect to previously reported ASD-related variants may provide a functional context for their ASD associations. We defined ASD-related loci using the autism results of the Psychiatric Genomics Consortium (PGC), a large mega-analysis of ASD GWA studies. We defined meQTLs using joint genotype and peripheral blood DNAm data from the Study to Explore Early Development (SEED), a national multi-site autism case-control study of children aged 2–5 years. We found that ASD-related variants were enriched for meQTLs at a p-value of 0.029. We will present results detailing the extent to which this genome-wide enrichment is observed in meQTLs derived from post-mortem brain tissue. We have also identified novel ASD candidate genes via interrogation of the nature and extent of DNAm control at specific PGC-identified ASD loci. We will present several of these loci discovered via the SEED peripheral blood meQTLs, the post-mortem brain meQTLs, and cord blood meQTLs using samples from an enriched risk birth cohort (Early Autism Risk Longitudinal Investigation; EARLI). The utility of these analyses will be to further understand the contribution of meQTLs towards ASD etiology across tissue type and developmental stage.

4

On Approaches to Improve Power in Analyses of Genetic Effects on Time-To-Event Outcomes in Longitudinal Studies

Konstantin G. Arbeev (1) Liubov S. Arbeeva (1) Olivia Bagley (1) Hongzhe Duan (1) Igor Akushevich (1) Alexander M. Kulminski (1) Mikhail Kovtun (1) Irina V. Culminskaya (1) Deqing Wu (1) Svetlana V. Ukraintseva (1) Anatoliy I. Yashin (1)
(1) Duke University

Traditional methods to estimate the effects of genetic markers on time-to-event outcomes can be enhanced if we complement them with the demographic approach that takes into account the demographic structure of the population under study at the time of biospecimen collection or genotyping. We present such an approach, the longitudinal genetic-demographic model, and the results of a simulation study in this model which illustrate that taking into account the demographic structure in addition to follow-up data improves power in analyses of genetic effects on time-to-event outcomes. We show that the effect is especially noticeable in the studies with short follow-up periods in which case the demographic structure can play a more substantial role in distinguishing between the allele- or genotype-specific risks compared to the information from the follow-up period. Joint analyses with follow-up data for non-genotyped participants provide an additional reserve for improving power. We applied the approach to analyses of data from the NHLBI's Candidate Gene Association Resource (CARe) project such as the Cardiovascular Health Study CARe and the Framingham Cohort CARe data. We performed GWAS of longevity traits and constructed the polygenic risk scores from the selected top SNPs affecting these traits. The results illustrate that application of the longitudinal genetic-demographic model can reveal additional signals in the data compared to the use of traditional approaches which do not take into account the demographic structure and follow-up data from non-genotyped individuals.

5

Playing musical chairs in multi-phenotype studies improves power and identifies novel associations

Hugues Aschard (1) Noah Zaitlen (2) Peter Kraft (1)
(1) Harvard School of Public Health (2) University of California San Francisco

Variability in complex human traits is associated with many factors, including exposures, biomarkers and genetic variants. Identifying the genetic variants that are causally associated with human phenotypes among the tens of millions of variants that are typically tested remains a challenge. Current strategies to improve power to identify modest genetic effects mostly consist of applying univariate statistical approaches such as linear or logistic regression (LR) and increasing study sample sizes. While successful, these approaches do not leverage the environmental and genetic factors shared between the phenotypes typically collected in contemporary cohorts. Here we develop a method called Musical Chairs (MC) that improves identification of small effects in studies where a large number of correlated variables have been measured on the same samples. MC is a data-driven approach that

leverages our previous work (Zaitlen et al., 2012, Aschard et al., 2015) to select covariates that will increase power for each SNP-phenotype pair considered. Simulations provide direct support that MC can achieve dramatic increases in statistical power equivalent to a two or even three or four fold increase in sample size. To demonstrate the power of our approach in real data, we performed a genome-wide screen for cis expression QTLs in the GEUVADIS cohort. We examined the expression of 12,167 genes in 375 individuals of European descent. At a stringent FDR of 0.1% standard LR identified 1,660 genes with at least one cis-eQTL while our MC approach identified 2,154 genes, an increase of 30%. As cohorts move toward large-scale phenotypes collections, MC will improve the ability of the community to identify associated genetic variants.

6

Integrating null data: a family-based (epi)genetic study of TMAO

Stella Aslibekyan (1) Ryan Irvin (1) Bertha A. Hidalgo (1) Elias Jeyarajah (2) Irina Shalaurova (2) Erwin Garcia (2) Hemant K. Tiwari (1) Devin A. Absher (3) Donna K. Arnett (1)
(1) University of Alabama at Birmingham (2) LipoScience Inc. (3) Hudson Alpha Institute for Biotechnology

Trimethylamine-N-oxide (TMAO), a pro-atherogenic metabolite species, has recently emerged as a promising new risk factor for cardiovascular disease. Animal studies have shown that circulating TMAO levels are regulated by genetic and environmental factors, however, these findings were never replicated in humans. We used data from the family-based Genetics of Lipid Lowering Drugs and Diet Network (n = 991, all European Americans) to investigate genetic and epigenetic determinants of TMAO in humans. We first estimated TMAO heritability at 22.9%, indicating a modest-to-moderate genetic influence. We subsequently used 1000 Genomes imputed data to estimate genome-wide associations with TMAO levels, adjusting for age, sex, and study site. The genome-wide study yielded no results that were significant at the genome-wide level. To further explore potential heritable determinants of TMAO, we quantified genome-wide DNA methylation using the Illumina Infinium array at ~450,000 CpG sites on CD4+ T-cells. We then tested for association with circulating TMAO, adjusting for T-cell purity, age, sex, and study site. Upon adjusting for multiple testing, none of the epigenetic findings were statistically significant. An integrated analysis of genetic and epigenetic data yielded an intergenic region on chromosome 7 containing a CpG that was suggestively linked to TMAO ($P < 10^{-6}$), which was in turn significantly associated with local sequence variation ($P < 10^{-8}$) as a methylation quantitative trait locus. Our findings suggest a weak influence of genetic factors on TMAO levels and highlight the promise of integrating genetic and epigenetic data to understand complex traits.

7

The effect of phenotypic outliers and non-normality on rare-variant association testing

Paul M. Auer (1) Suzanne M. Leal (2)
(1) University of Wisconsin-Milwaukee (2) Baylor College of Medicine

Rare-variant association studies have made important contributions to human complex trait genetics. These studies rely on specialized statistical methods for analyzing rare-variant associations, both individually and in aggregate. Many of these tests assume a normally (or approximately normally) distributed phenotype. However, many quantitative phenotypes are not normally distributed in healthy populations. We show that rare-variant association tests are uniquely susceptible to biases caused by outliers and non-normality and how these biases can be prevented. We investigated the impact that phenotypic outliers and non-normality have on the performance of rare-variant association testing procedures, e.g. combined multivariate collapsing (CMC) method and the sequence kernel association test (SKAT) as well as testing individual rare variants. Ignoring outliers or non-normality can significantly inflate type I error rates. We found that rank-based inverse normal transformation and trait winsorisation were both effective at maintaining type I error control without sacrificing power in the presence of outliers. Rank-based inverse normal transformation of quantitative trait values was the optimal method for non-normally distributed traits and the power to detect associations using either CMC or SKAT was considerably greater (e.g. > 50%) than using permutation to obtain empirical p-values or natural logarithm transformation of quantitative trait values before association testing. For rare variant association studies of quantitative traits with outliers or non-normality, we recommend using rank-based inverse normal transformation in order to transform phenotypic values prior to association testing.

8

Linkage methods used to evaluate *EYA4* as a candidate risk locus in GELCC familial lung cancer families linked to 6q

Joan E. Bailey-Wilson (1) Claire L. Simpson (1) Anthony M. Musolf (1) Susan M. Pinney (2) Mariza de Andrade (3) Colette R. Gaba (4) Ping Yang (3) Ming You (5) Ann G. Schwartz (6) Diptasri Mandal (7) Yanhong Liu (8) Margaret R. Spitz (8) Elena Y. Kupert (9) Christopher I. Amos (10) Marshall W. Anderson (9)
(1) National Human Genome Research Institute, National Institutes of Health, Baltimore, MD (2) University of Cincinnati College of Medicine, Cincinnati, OH (3) Mayo Clinic, Rochester, MN (4) University of Toledo Dana Cancer Center, Toledo, OH (5) Medical College of Wisconsin, Milwaukee, WI (6) Karmanos Cancer Institute, Wayne State University, Detroit, MI (7) Louisiana State University Health Sciences Center, New Orleans, LA (8) Baylor College of Medicine, Houston, TX (9) Medical College of Wisconsin, Milwaukee,

WI (10) Geisel School of Medicine, Dartmouth College, Lebanon, NH

We published evidence of linkage of familial lung cancer (FLC) to a region on 6q (PMID: 15272417, 20215501). Wilson et al. (doi:10.1038/onc.2013.396) found that *EYA4* (in our linkage region) was frequently inactivated biallelically in sporadic LC, displays tumor suppressor gene-like properties, affects DNA repair and that 5 of 17 common SNPs in this 0.3Mb region showed nominal association to FLC ($p < 0.05$; not significant after multiple test correction). Here, we sequenced 75 people in 9 strongly 6q-linked families for 37Mb of 6q (130-167Mb). To detect sharing of variants among affecteds in each family while allowing for phenocopies, two-point affected-only linkage analysis was performed on LC affection status for all sequence variants using the Elston-Stewart algorithm in the R package paramlink with penetrance of 0.01, 0.1 and 0.1 for dd/Dd/DD and D allele frequency of 0.01, at $\theta = 0$.

The highest LOD score summed across families in the 37Mb targeted region was not near *EYA4*. No family had a variant in or near *EYA4* that was on the linked haplotype or had a LOD score close to the maximum family-specific LOD score in the region. The highest total LOD score in the region was found at a missense coding variant, rs41267809, at 6:160953642-SNV (LOD = 1.099, allele frequency 1–3% in 1000Genomes) in the *LPA* gene. This signal was driven almost entirely by a single family. Three heterozygotes for this variant were observed in a replication sample of 55 independent FLC cases (5.4%). Other families have maximum LODs in other non-coding regions. Annotation of these non-coding variants is ongoing. Thus, linkage analyses have shown that *EYA4* is unlikely to contain functional variants that explain the linkage to 6q in these families. *LPA* is a candidate in one family.

9

Identity-By-Descent detection among over 850,000 present-day Americans in the AncestryDNA cohort

Mathew J. Barber (1) Keith Noto (1) Ross E. Curtis (1) Julie M. Granka (1) Yong Wang (1) Jake K. Byrnes (1) Peter Carbonetto (1) Eunjung Han (1) Amir Kermany (1) Natalie M. Myres (1) Catherine A. Ball (1) Kenneth G. Chahine (1)
(1) AncestryDNA

Discovering identical by descent (IBD) regions for pairs of genomes is an important and difficult part of many genetic analyses. Analysis of large cohorts is especially hard given that new samples are always arriving and the number of possible pairs to compare grows quadratically with sample size. GERMLINE is an algorithm in common use that derives computational efficiency from locality-sensitive hashing, and we have developed our own more scalable implementation. We have addressed accuracy concerns of the GERMLINE algorithm with a post-processing filter: TIMBER. TIMBER is computationally efficient, effective, and utilizes the large amount of putative IBD segments that are discovered by running GERMLINE on a large data set (over 100K people). TIMBER evaluates the confidence we have in a putative IBD segment depending on the frequency with which the individuals involved have putative IBD in the same location with the rest of the database. TIMBER attempts to correct for biases in the IBD data in a personalized way without requiring us to model the biases explicitly. We have evidence that TIMBER is a very useful post-processing filter from both real and simulated data.

10

Estimation of Cell Type Specific DNA Methylation Effects using Whole Blood Methylation Data

Richard T. Barfield (1) Andrea Bacarelli (1) Xihong Lin (1)
(1) Harvard University T.H. Chan School of Public Health

Association analysis of DNA methylation data is challenged by cell type heterogeneity, as DNA methylation collected in studies is typically a mixture of different cell types. For example, studies using whole blood measure DNA methylation from a mixture of different lymphocytes. This cell type heterogeneity can potentially bias results, as DNA methylation is known to be an important mechanism of tissue differentiation, and thus may differ by cells. Current methods to adjust for this involve including observed or estimated cell type counts as additional covariates in the analysis. These methods however do not estimate the exposure effects on cell type specific DNA methylations. Direct measurements of cell type specific methylations involve intensive lab work and are costly. We develop in this paper a statistical method to estimate exposure effects on cell-specific methylations using whole blood methylation data when cell type counts are available but cell-specific methylations are unavailable. Specifically, we assume cell type specific regression models of the exposure effects on cell type specific methylations. We treat cell specific methylations as missing data, and develop an EM algorithm to estimate the exposure effects on cell type specific methylations adjusting for covariates by using whole blood methylation data and cell type counts. To the best of our knowledge, this is the first method to estimate these effects without assaying each cell type. We applied our method to the Normative Aging Study to study the smoking effect on cell type specific methylations.

11

IBD estimation, segmental sharing detection, and pedigree reconstruction in non-human primates

Jennifer E. Below (1) Muthuswamy Raveendran (2) Ronald A. Harris (2) Cheng Xue (2) Betsy Ferguson (3) Sree Kanthaswamy (4) David G. Smith (4) Clifford J. Jolly (5) Jane Phillips-Conroy (6) Donna Muzny (2) Richard Gibbs (2) Fuli Yu (2) Jeffery Rogers (2)
(1) University of Texas Health Science Center at Houston (2) Human Genome Sequencing Center, Baylor College of Medicine (3) Oregon National Primate Research Center, Oregon Health and Science University (4) Californial National

Primate Research Center (5) New York University (6) Washinton University

The field of human genetics is rooted in the study of transmission of DNA sequences through families. As a result, computational strategies to detect allele sharing and identity by descent (IBD) have evolved alongside sequencing technologies to account for patterns of linkage disequilibrium, recent admixture, and population substructure. Today, whole genome and whole exome sequencing (WGS, WES) are being applied in cohorts of nonhuman species in which characterization of the inherent complexity of population genetics is exacerbated by incomplete information, complex mating patterns, and often high levels of nucleotide diversity. The subjects of primate studies range from highly controlled, pedigreed populations, such as research colonies, to wild caught, presumably outbred animals with no available information on kinship relationships. We present results of genome-wide IBD estimation, segmental sharing, and pedigree reconstruction using PRIMUS in WGS from 20 rhesus macaques (Macaca mulatta) from two US research colonies and WES from 29 wild caught baboons (14 Papio cynocephalus, 15 Papio kindae). The applicability and accuracy of tools developed to estimate segmental sharing (GERMLINE, PLINK, IBDseq), relatedness (ERSA, PLINK, KING) and reconstruct pedigrees (PRIMUS) were evaluated using colony records and field observations. Nonhuman primates are crucial models for studies of infectious disease, neurobiology, and psychobiology that are central to biology and medicine. Detecting cryptic relatedness and validating pedigrees are essential to genetic studies of primates; determining kinship among wild nonhuman primates increases our ability to investigate fundamental aspects of primate population biology.

12

A comparison of polygenic contribution to Autism Spectrum Disorder for Common, Rare and Copy Number Variants

Kelly S. Benke (1) Brooke Sheppard (1) Kelly Bakulski (1) Alison Singer (1) April Shu (1) Christine Ladd-Acosta (1) Danielle M. Fallin (1)
(1) Johns Hopkins

Current evidence suggests that the genetic contribution to autism spectrum disorder (ASD) is polygenic, consisting of small, accumulating effects of many variants. While some of the contribution may come from rare single variant mutations or rare copy number variation, the summation of modest effects among common variants is still likely to be important. We have created a polygenic risk score, derived from discovery results available from the Psychiatric Genomics Consortia (PGC) mega-analysis for ASD, using measured genotypes in the Study to Explore Early Development (SEED), a multi-site case-control study to investigate the interplay of genetic and environmental risk factors that underlie disease risk. The association of the polygenic score was significant, but appeared to be due to confounding by genetic ancestry. We will further characterize these associations in SEED by creating a score reflective of the burden of rare variation (using 1000 genomes imputed data) and rare copy number variation burden (using Penn CNV calls). Correlations among the three scores, their potential interactions, and their ability to classify membership into case status will be explored. This is the first effort to associate a polygenic score from the PGC discovery effort for ASD in an independent study, and the first study to investigate multiple polygenic scores in this context. Our results can shed light on whether rare, common and CNV burden are correlated or represent independent contributions to the risk of autism.

13

Mixed models for time-to-event outcomes with large-scale population cohorts and genome-wide data

Christian Benner (1) Matti Pirinen (1) Veikko Salomaa (2) Juni Palmgren (1,3) Samuli Ripatti (1,4)
(1) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland (2) National Institute for Health and Welfare, Helsinki, Finland (3) Karolinska Institutet, Stockholm, Sweden (4) Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Tens of thousands of Finns have been genotyped as part of the national biobanking effort. Combining genomic profiles of tens of thousands of Finns with health registry data provides a unique resource in the world to accomplish breakthroughs in human genomics. In particular, the information can help determine to what extent phenotypic variation in age of onset of disease is attributable to genetic effects. Recent methodological development has enabled heritability estimation for quantitative traits in population cohorts by estimating genetic sharing among pairs of individuals from genome-wide marker data. However, connecting age of onset to big genomics data form large-scale population cohorts has so far not been feasible with hitherto existing methods.

We introduce a method that deals with prospective information from health registry data and permits heritability estimation and association testing. Our approach implements a piecewise constant hazard model that contains an individual-specific Gaussian random effect with arbitrary covariance structure. Computationally, we analyze a Bayesian model using a Markov Chain Monte Carlo algorithm. We compared our method in simulations to a mixed effects Cox model, as implemented in the R package 'coxme'. Our method produced lower mean squared error of variance and regression parameters under varying proportion of variance explained in the genotype, different number of causal variants, sample size and high censoring rates. Using over 10,000 unrelated Finns, we also estimated that the heritability of cardiovascular disease (defined as coronary heart disease and ischaemic stroke) in terms of relative risk of disease due to genetic effects is 27% (95% CI: 6–51%).

14

Latino does not equal Latino: Major components of Native American Ancestry and Mortality due to Gallbladder Cancer in Chile

Justo Lorenzo Bermejo (1) Felix Boekstegers (1) Rosa González Silos (1) Katherine Marcelain (2) Macarena Fuentes Guajardo (3) Francisco Rothhammer (3)
(1) Statistical Genetics Group, Institute of Medical Biometry and Informatics, University of Heidelberg, Germany (2) Program of Human Genetics, Institute of Biomedical Sciences, Medical Faculty, University of Chile, Chile (3) Instituto de Alta Investigación, Tarapacá University, Chile

Latino heterogeneity regarding individual proportions of African, European and Native American ancestry is widely recognized. However, major Native American subcomponents are rarely separated. Results from the present study based on genome-wide single nucleotide polymorphism data from 1,872 Chileans and 9,641 gallbladder cancer deaths illustrate that this separation may be crucial.
The ADMIXTURE software was used for supervised ancestry estimation. Surrogates of European and African ancestry were 80 Utah residents with Northern and Western European ancestry and 87 Yoruba in Ibadan, Nigeria from the 1000 Genomes Project, respectively. Surrogates of Native American ancestry were either 64 Human Genome Diversity Project sample from the Americas (HGDP), or 9 Mapuche and 10 Aymara persons representing the two largest indigenous peoples in Chile. The relationship between genetic ancestry and aggregated gallbladder cancer mortality data was investigated by multiple linear regression to estimate expected regional ancestries, followed by multiple Poisson regression to quantify the association between regional mortalities and expected ancestries.
Estimated components of Native American ancestry based on HGDP and on the whole "Mapuche plus Aymara" collective were practically identical (Pearson r = 0.9998). However, the average "Mapuche plus Aymara" ancestry component (42%) was smaller than the sum of the separated Mapuche (39%) and Aymara (7%) components. A 1% increase in "Mapuche plus Aymara" ancestry translated into a 1% increased mortality risk of gallbladder cancer (P = 0.08). In contrast, a 1% increase in the Mapuche ancestry component represented a 3% increased mortality risk ($P = 5 \times 10^{-22}$), and a 1% increase in Aymara ancestry resulted in a 3% decreased mortality risk ($P = 10^{-11}$).
These results underline the importance of suitable surrogates for ancestry estimation which reflect the actual composition and genetic heterogeneity of study individuals.

15

Lung Cancer Environmental Exposure Network

Emily C. Bih (1) Christian Darabos (1) Jason H. Moore (2)
(1) Dartmouth College (2) University of Pennsylvania

Along with genetic factors, environmental exposure to harmful chemical substances in our immediate environment may contribute to many complex diseases, such as cancers, diabetes or autism. This study investigates disorders related to lung cancer through exposure chemicals. Through this association we reveal the common environmental factors involved in seemingly unrelated disorders. To research these connections, we have conducted a thorough PubMed literature survey, collecting data that matches environmental chemical substances to lung cancer as well as other substances and diseases. To best study the data, we use a bipartite network made of two distinct sets of vertices, substances and diseases, and use edges to represent "highly probable" causal relationships. Our network is made of 15 different substances associated with lung cancer and 90 disorders of which 33 are cancer-related phenotypes. We analyze the projections of the bipartite network to find commonalities between disorders. The most disorders were associated with tobacco smoke chemicals NNAL, cotinine, and PCB respectively. Gastric cancer, colorectal cancer, and cardiovascular disease were found to be associated with the most environmental chemicals that are associated with lung cancer. This study reinforces tobacco smoke chemicals as potent carcinogens but also investigates possible relationships between cancer and autoimmune diseases and behavioral disorders that are related through tobacco smoke chemicals. Future studies will integrate genetic data into the network to generate an "exposome" that will be used to identify common cancer contributing factors and disease pathways to prevent and treat complex diseases and cancer progression.

16

Alternative study designs identify genes associated with variation in lung function among patients with cystic fibrosis

Elizabeth E. Blue (1) Mary J. Emond (1) Tin L. Louie (1) Jessica X. Chong (1) Ronald L. Gibson (1) Michael J. Bamshad (1)
(1) University of Washington

Primary ciliary dyskinesia (PCD) and cystic fibrosis (CF) are autosomal recessive disorders with abnormalities in mucociliary clearance and lung disease. PCD is caused by genetic defects in ciliary structure and function; CF is caused by variants in *CFTR* which cause abnormal airway mucus features. We hypothesize that genetic variants in PCD genes may influence lung disease severity among subjects with CF.
From the Exome Sequencing Project (ESP) cohort, we identified extremes in adjusted lung disease severity: 62 HIGH subjects and 60 LOW subjects. Targeted association testing of PCD genes using SKAT-O found that *DNAAF1* was nominally significant. A SKAT-O genome scan defined a set of 10 candidate genes.
We compared HIGH and LOW subjects with 3,096 ESP controls. Association testing of common variants found strong

association to *CFTR* ($p < 10^{-36}$ at rs10229820, $\lambda < 1.01$ for both HIGH and LOW). SKAT-O analyses found *DNAAF1* trended toward significance with HIGH ($p = 0.0587$). SKAT-O analysis of only "functional" variants found HIGH to be nominally associated with DNAH8, while LOW is associated with OFD1 ($p = 2.18 \times 10^{-05}$). Analyses of all variants in the 10 candidate genes from the HIGH vs. LOW genome scan found HIGH is significantly associated with *SELPLG* ($p = 0.00246$), while LOW is significantly associated with *SLC7A7* ($p = 0.00064$). Both *SELPLG* and *SLC7A7* have published evidence for their role in lung disease.
Several HIGH and LOW subjects were heterozygous for rare PCD variants; the genes implicated were not detected by association tests.
Small sample sizes with extreme phenotypes generated novel hypotheses with alternative study designs. We conclude that variants in cilia genes and other pathways may impact the severity of lung disease in CF patients.

17

The distribution of *ABCA4* variants in Stargardt disease from the ProgStar studies

Samantha M. Bomotti (1,2) Rupert W. Strauss (2,3) Hendrik P. Scholl (2) Robert Wojciechowski (1,2)
(1) Johns Hopkins Bloomberg School of Public Health (2) Johns Hopkins Wilmer Eye Institute (3) Medical University Graz

Type 1 Stargardt disease (STGD1; OMIM 248200) is the most common juvenile retinal dystrophy. It follows an autosomal recessive mode of inheritance at the *ABCA4* locus. There is no cure and it remains difficult to diagnose due to its high allelic and phenotypic heterogeneity. We have collected genotype and clinical data from 365 STGD1 patients recruited for the ProgStar studies, which aim to characterize the natural history of STGD1 (http://progstar.org/). This is the largest cohort of STGD1 patients collected to date. We report here on the distribution of *ABCA4* variants in this STGD1 population. Data were compiled from nine clinical centers across the United States and Europe. The biological assays used to identify *ABCA4* variants in genetic testing laboratories depended on the technologies available at the time of patient diagnosis (from 2004 to present). Collection of the clinical data is ongoing and will allow for in-depth phenotypic and genetic analyses once completed.
The majority (89.3%) of the 365 STGD1 patients were heterozygous for *ABCA4* mutations. In 23 cases (6.3%), only one *ABCA4* mutation was identified, suggesting incomplete coverage of *ABCA4* or locus heterogeneity. The most common variant was c.5882G > A (G1961E), found in 97 (26.6%) patients. The second most common variant, c.2588G > C (G863A), was observed in 43 (11.8%) patients. The third most common variant (c.5461-10T > C) appeared in 31 (8.5%) patients. These variants are also the most frequently found in previously published cohorts.
Initial variant analyses in the ProgStar studies confirm the high allelic heterogeneity of STGD1. These data will help characterize the clinical impact of variants (including rare variants) on STGD1 progression and severity.

18

Comparison of Illumina Infinium 450K Methylation BeadChip preprocessing methods in an Epigenome Wide Association Study

Martha Brucato (1) Nara Sobreira (2) L. Zhang (2) Christine Ladd-Acosta (1) Chrissie M. Ongaco (3) Jane Romm (3) Maggie Baker (2) Kimberly Doheny (3) Débora R. Bertola (4) Chong Ae Kim (4) Ana B. A. Perez (5) Maria I. Melaragno (5) Vera A. Meloni (5) David Valle (2) Hans Bjornsson (2)
(1) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD (2) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD (3) Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD (4) Unidade de Genética, Instituto da Criança, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo (5) Genetics Division, Department of Morphology and Genetics, Universidade Federal de São Paulo

Kabuki Syndrome (KS) is a Mendelian disorder that involves the histone methylation machinery. KS is characterized by growth retardation, intellectual disability, immunological problems, and facial dysmorphology. We collected a cohort of clinically diagnosed KS patients ($n = 29$) and age-and-sex-matched controls ($n = 9$) from Brazil. We compared DNA methylation patterns of KS patients with histone machinery mutations (*KMT2D* and *KMT2A*) to those of matched normal controls with the Illumina Infinium HumanMethylation450 BeadChip platform, a reliable and reproducible technology for assaying DNA methylation at 485,517 loci across the genome. We applied four different preprocessing methods to the raw data, including quantile normalization, functional normalization, noob (normal-exponential using out-of-band probes), and functional normalization plus noob. An epigenome wide association study for KS phenotype was then conducted on each preprocessed dataset. Differentially methylated positions (DMPs) and differentially methylated regions (DMRs) were identified after adjustment for patient sex, blood sample cell composition, and ancestry. KS associated DMPs and DMRs were consistently discovered regardless of the preprocessing method. No preprocessing method emerged as clearly superior and each yielded comparable results in this empirical dataset. Our study conclusion is therefore robust to the preprocessing method chosen: individuals with a genetic abnormality in histone methylation have shared changes in their DNA methylation, suggesting the occurrence of crosstalk between histone and DNA methylation.

19

Extension of a rare variant sharing exact test to sharing patterns involving a subset of affected relatives

Alexandre Bureau (1) Joshua Sampson (2) Lisa R. Yanek (3) Rasika Mathias (3) Ingo Ruczinski (4)
(1) Université Laval, Canada (2) National Cancer Institutes, USA (3) Johns Hopkins University School of Medicine, USA (4) Johns Hopkins Bloomberg School of Public Health, USA

We previously proposed using the probability that a set of related affected subjects share a rare variant (RV) to test the null hypothesis of absence of linkage and association to a disease, without requiring a control group to estimate variant frequency. The fact that different RVs, common variants or non-genetic causes may influence a trait in different members from large extended families may result in causal RVs being shared by only a subset of affected relatives, and be missed by a test considering only RVs shared by all affected relatives. We extended our algorithm to compute exact probabilities of sharing a RV among a subset of affected family members instead of all of them. We studied factors influencing power of this exact test statistic analytically, and compared it by simulation to the unified linkage and case-control test implemented in the pedigree extension of Variant Annotation, Analysis and Search Tool (pVAAST). Causal RV genotypes were simulated conditional on a platelet aggregation phenotype in 6 European-American families from the GeneSTAR study, and then the causal allele in each family (if any) was assigned to a different coding RV in genes with various numbers of RVs. Control and founder haplotypes spanning each gene were sampled from 1000 Genomes Project European ancestry phased sequences. Under single-locus and two-locus heterogeneity models, power was mostly driven by causal RVs relative risk of disease. The test allowing sharing by a subset of affected relatives was only slightly more powerful than the test requiring all relatives to share, and had similar power to the pVAAST test which also account for case and control frequency differences. The sharing approach is suitable in populations where sequencing data is scarce.

20

Prostate Cancer in African American Men

Sarah G. Buxbaum (1) Carrie L. Snyder (2) Ellastine Buckner (1) Mark L. Stacey (2) Rasaki Aranmolate (1) Omofosolade Kosoko-Lasaki (2) Olugbemiga T. Ekundayo (1) Henry T. Lynch (2)
(1) Jackson State University (2) Creighton University

Prostate cancer (PCa) is the second leading cause of cancer death in American men. Both incidence and death rates have trended downward since 1999; however, in 2011, African American men's death rates were still double those of white men's.
Without having genotype data to predict relative risk, a detailed family history is known to be useful and inexpensive means of predicting an individual's risk of PCa. This study was designed to assess environmental factors which may improve risk assessments.
We undertook a study of possible socioeconomic, occupational, built environment, and psychosocial determinants of prostate cancer, interviewing 77 African American men who had been diagnosed with prostate cancer in two cities: Jackson, Mississippi in the deep south, and Omaha, Nebraska in the midwest.
The two states differ substantially in cancer incidence. In 2011, among men, Mississippi had the highest incidence rate in the country for all cancers at a rate of 257.8, according to the CDC. In Nebraska, among men, the incidence rate for all cancers was 202.6. For prostate cancer, in Mississippi, the rate was 147.5 and in Nebraska, 123.9.
Questionnaire data was entered into an Access program developed for this project which automatically flagged data entries that were outside the expected range for a particular field. Pedigree information was obtained during the interviews and blood samples were collected and stored for future genetic analyses. A heritability analysis was performed and we found that one family in particular with multiple affected siblings was driving the estimate. Pedigree charts showing affection status with prostate cancer and other cancers were generated using the Madeline 2 program.

21

Detecting patient subgroups using reduced set of disease-related markers with iterative pruning Principal Component Analysis (ipPCA)

Kridsadakorn Chaichoompu (1) Isabelle Cleynen (2) Ramounac Fouladi (1) David Ellinghaus (3) Matthias Hübenthal (3) Kristel Van Steen (1) on behalf of the International inflammatory bowel disease genetics consortium (IIBDGC)
(1) Montefiore Institute, University of Liege, Belgium (2) Department of Clinical and Experimental Medicine, University of Leuven, Belgium (3) Institute of Clinical Molecular Biology, University of Kiel, Germany

Genetic markers such as Single Nucleotide Polymorphisms (SNPs) can be used to find subgroups of populations or patients with carefully selected clustering algorithms. The iterative pruning principal component analysis (ipPCA) has been shown to be a powerful tool to identify fine substructures within general populations based on SNP profiles. Usually, SNPs contributing to such profiles have passed rigorous quality control procedures, similar to the ones used for GWAs. Alternatively, attention is restricted to a smaller subset such as PCA-correlated SNPs.
Here, we applied ipPCA on real-life data consisting of the 163 known inflammatory-bowel disease (IBD) associated loci in 13,400 healthy individuals and 29,500 IBD (16,902 Crohn's disease (CD), and 12,598 ulcerative colitis (UC)) patients from the IIBDGC. Prior to clustering by ipPCA, in each

group separately, we regressed out the first five Principal Components (PCs) that were computed from a filtered panel of genome-wide SNPs, to account for general population strata. Next, we applied ipPCA on the healthy group, to learn about the presence of a population-specific partitioning in controls. Then we performed three subphenotype analyses: CD only, UC only and the combined group of CD and UC patients (IBD). For each patient subgroup analysis and for the ipPCA analysis on controls, we highlighted and compared the key SNP drivers.
CD patients could be molecularly reclassified in two groups, and similar for UC patients. The combined patient group could be subdivided in four groups. Finally, we compared demographic and clinical features among the different groups and looked for meaningful characterizations of adjusted patient clusters by performing pathway analysis on driver genes.

22

Whole exome sequencing and linkage analysis of patients with pulmonary nontuberculous mycobacterial infection

Fei Chen (1) Eva P. Szymanski (2) Kenneth N. Olivier (3) Xinyue Liu (4) Hervé Tettelin (4) Steven M. Holland (2) Priya Duggal (1)
(1) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD (2) Laboratory of Clinical Infectious Diseases, NIAID, NIH, Bethesda, MD (3) Cardiovascular and Pulmonary Branch, NHLBI, NIH, Bethesda, MD (4) Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

Pulmonary nontuberculous mycobacterial (PNTM) infection is a rare syndrome that often affects women over 50 years of age. These women have a distinct body morphology (tall and lean with scoliosis) and clinical features including mitral valve prolapse and pectus excavatum, suggestive of an underlying genetic mechanism. To identify genetic regions harboring PNTM associated loci, we performed whole-exome sequencing (WES) on 17 cases and 21 unaffected individuals from 10 families and 57 sporadic cases recruited at the NIH Clinical Center in 2001–2013. One family was excluded from analysis due to quality control issues. Ninety-two percent of the PNTM cases were female. We conducted a genome-wide linkage study using 8,209 independent, common genetic variants on autosomal chromosomes from the WES. Using multi-point parametric linkage analysis, we identified a significant linkage region on chromosome 6q12-q16 (HLOD = 3.90) under a common recessive model with a 5% allele frequency and 100% penetrance, and a suggestive loci on chromosome 3q25-q27 (HLOD = 2.99) under a rare recessive model with a 1% allele frequency and 40% penetrance. These identified linkage regions will be examined for homozygous and compound heterozygous variants that fit the suggested recessive model in both the PNTM families and sporadic cases.

23

GMMAT: logistic mixed models to control for population stratification and relatedness in genetic association studies with binary traits

Han Chen (1) Chaolong Wang (2) Matthew P Conomos (3) Adrienne M Stilp (3) Zilin Li (1) Tamar Sofer (3) Adam A. Szpiro (3) Wei Chen (4) John M. Brehm (4) Juan C. Celedón (4) Susan S. Redline (5) George J. Papanicolaou (6) Timothy A. Thornton (3)
(1) Harvard T.H. Chan School of Public Health (2) Genome Institute of Singapore (3) University of Washington (4) Children's Hospital of Pittsburgh (5) Brigham and Women's Hospital (6) National Heart, Lung, and Blood Institute

Population stratification and relatedness can result in spurious association findings in Genome-Wide Association Studies (GWAS). Although principal component analysis has been widely used to adjust for population structure in unrelated samples, it fails in the presence of relatedness. Recently, linear mixed models have become a popular alternative to account for both population structure and relatedness, and have been used to analyze both continuous and binary outcomes in genetic association studies. However, issues in their use with binary outcomes have often been overlooked. We show that applying linear mixed models to binary traits may lead to incorrect type I error rates in the presence of population stratification, which could occur even when no inflation is observed from the overall quantile-quantile plot. As an alternative, we developed the Generalized linear Mixed Model Association Test (GMMAT), a computationally efficient method for mixed model analysis of binary traits using score statistics. Fitting generalized linear mixed models using GMMAT is almost two orders of magnitude faster than existing approaches. In addition, GMMAT can be applied to large samples as it performs score tests by fitting the mixed model only once per GWAS under the null hypothesis of no genetic association. We show in both simulation studies and empirical data from the Wellcome Trust Case-Control Consortium that GMMAT is effective for controlling population structure and relatedness.

24

Association Analysis of Longitudinal Genetic Data

Jo-Yao Chien (1) Shin-Hua Lin (1) Wendy Yi-Ying Wu (1) Li-Shang Chen (2) Amy Ming-Feng Yen (2) Tzeng-Ying Liu (3) Hsiu-His Chen (4) Ming-Wei Lin (1)
(1) Institute of Public Health, National Yang-Ming University, Taipei, Taiwan (2) School of Oral Hygiene, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan (3) Lienchiang County Government, Matsu, Taiwan (4) Institute of Epidemiology and Preventative Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

Quantitative traits, such as blood pressure and levels of cholesterol, generally change with age or time. Longitudinal genetic studies with multiple measurements offer a

valuable resource to examine how genetic and environmental factors that may affect complex traits over time. However, most of genetic studies of quantitative traits using longitudinal data did not take into account genes may have a time varying effect. It may result in loss of power in detecting genetic effects of quantitative traits. Fan et al. (2012) proposed a flexible non-parametric model by employing penalized splines to estimate non-parametric functions for longitudinal genetic data. Nonetheless, they did not consider the time varying effect of the genes. In this study, we conducted a simulation study to investigate the time varying effect of the genes by adapting non-parametric models with different genetic effects as well as different minor allele frequencies and compared the performance of the method with other three parametric methods. We found that the power and bias rate of both non-parametric method and the parametric cubic model are close to the correctly specified model under the genes with time invariant effect. However, under the genes with time varying effect, the non-parametric penalized linear model remains the best choice in fitting real data. Our results provide a basis for practical methodology application to longitudinal genetic data.

25

Rare Variant Association Tests for Longitudinal Family Studies

Li-Chu Chien (1) Yen-Feng Chiu (1) Fang-Chi Hsu (2) Donald Bowden (3)
(1) Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli 35053, Taiwan, ROC (2) Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA (3) Center for Diabetes Research, Center for Genomics and Personalized Medicine Research, Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

Functional variants are likely to be aggregated in family studies enriched with affected members, and this aggregation increases the statistical power for rare variant detection. Longitudinal family studies provide additional information for identifying genetic and environmental factors associated with disease over time. However, methods for the analysis of rare variants in longitudinal family data remain fairly limited. These methods should be capable of accounting for different sources of correlations and handling large amounts of sequencing data efficiently. To identify rare variants in longitudinal family studies or family members with multiple phenotypes, we extended the powerful pedigree-based burden and kernel association tests to genetic longitudinal studies. Generalized estimating equation (GEE) approaches were used to generalize the pedigree-based burden and kernel tests to multiple correlated phenotypes under the generalized linear model framework. Adjustments were made for the fixed effects of confounding factors. These tests accounted for the complex correlations between multiple correlated phenotypes and between individuals within the same family. Comprehensive simulation studies were conducted to compare the proposed tests with mixed-effects models and marginal models using GEEs under various configurations. When the proposed tests were applied to the Diabetes Heart Study, exome variants of *POMGNT1* and *JAK1* geneswere found to be associated with type 2 diabetes.

26

Estimating penetrance in the context of competing mortality: Application to Familial Pancreatic cancer

Erica J. Childs (1) Bryan Lau (2) Amanda Blackford (1) Giovanni Parmigiani (3,4) Alison P Klein (1,5)
(1) Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA (2) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA (3) Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MD, USA (4) Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA (5) Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, MD, USA

Understanding the interplay between inherited genetic risk, environmental factors and competing risk is a necessary step to provide accurate risk counseling to families seeking clinical risk assessment due to a strong family history of disease. Pancreatic cancer is an increasing public health burden; it is currently the 4th leading cause of cancer death in the United States but projected to rise to 2nd by 2020. It has been estimated that up to 10% of newly diagnosed pancreatic cancer patients report a family history of the disease and current cigarette smoking approximately doubles pancreatic cancer risk. In addition, cigarette smoking is strongly associated with increased all-cause mortality. In order to better understand how familial risk and smoking impact risk of pancreatic cancer in families we formulated a semi-parametric model incorporating autosomal dominant inheritance to estimate the risk of pancreatic cancer while simultaneously modeling the effects of a major gene, cigarette smoking, and the competing risks of death due to other causes. Full analysis and results will be presented. These data provide the foundation for more detailed risk-assessment for high-risk pancreatic cancer families.

27

Examination of established cancer risk variants in putatively high-risk pancreatic cancer patients: A PACGENE study

Erica J. Childs (1) Kari G. Chaffee (2) Steven Gallinger (3) Sapna Syngal (4) Ann G. Schwartz (5) Michele L. Cote (5) Melissa Bondy (6) Michael G. Goggins (7) Ralph H. Hruban (7) Stephen Chanock (8) Robert Hoover (8) Charles Fuchs (9,10) David N. Rider (2) Laufey Amundadottir (8) Rachael

Stolzenberg-Solomon (11) Brian Wolpin (9,12) Harvey A. Risch (13) Gloria M. Petersen (2,14) Alison P. Klein (15,16)
(1) Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA (2) Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA (3) Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada (4) Population Sciences Division, Dana-Farber Cancer Institute, and Gastroenterology Division, Brigham and Women's Hospital, Boston, MA, USA (5) Department of Oncology, Karmanos Cancer Institute and Wayne State University, Detroit, MI, USA (6) Baylor College of Medicine, Dan L. Duncan Cancer Center, Houston, TX, USA (7) Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, MD, USA (8) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA (9) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA (10) Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA (11) Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD, USA (12) Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA (13) Yale School of Public Health, New Haven, CT, USA (14) These authors contributed equally to this work (15) Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA (16) Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, MD, USA

Individuals from pancreatic cancer families are at increased risks, not only of pancreatic cancer, but also of breast, ovarian and colon cancer. While some of these increased risks may be due to alterations in high-penetrance genes, including *BRCA1*, *BRCA2*, and *PALB2*, and mismatch-repair genes, common genetic variants may also be involved. We sought to examine the role of variants previously associated with risk of pancreatic, breast, ovarian or prostate cancer in a high-risk population of cases with either a family history of pancreatic cancer or a diagnosis of the disease prior to age 50. We genotyped 985 cases (79 early onset cases, 906 cases with family histories) and 877 controls on the iCOGS array with custom content. In total, 216,072 SNPs, including 4,917 custom SNPs (to replicate pancreatic cancer GWAS results), were genotyped. Analyzing these data using a log-linear additive model, we replicated several of the previously reported pancreatic cancer susceptibility loci, including variants on 2p13.3 and 7p13 that had not yet been independently replicated (2p13.3, rs1486134: OR = 1.36, 95%CI, 1.13-1.63, $p = 9.29 \times 10^{-4}$; and 7p13, rs17688601: OR = 0.76, 95%CI, 0.63-0.93, $p = 6.59 \times 10^{-3}$). In general, the magnitudes of association observed in our high-risk patient samples were similar to those seen in population- and hospital-based studies. In addition, we identified two SNPs, from genomic regions implicated in other cancer sites with suggestive evidence of association to pancreatic cancer ($p < 10^{-6}$). Common variants may indeed be involved in pancreatic cancer susceptibility even in "high-risk" populations.

28

A Comparison Study of Fixed and Mixed Effect Models for Gene Level Association Studies of Complex Traits

Chi-yang Chiu (1) Jeesun Jung (2) Daniel Weeks (3) Alexander F. Wilson (4) Joan Bailey-Wilson (4) Christopher I. Amos (5) Momiao Xiong (6) James Mills (1) Ruzong Fan (1)
(1) Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH) (2) National Institute on Alcohol Abuse and Alcoholism, NIH (3) University of Pittsburgh (4) National Human Genome Research Institute, NIH (5) Geisel School of Medicine, Dartmouth College (6) University of Texas, Houston

In association studies of complex traits, fixed effect regression models are usually used to test for association between phenotypic traits and major gene loci. In recent years, variance-component tests based on mixed models were developed for region-based genetic variant association tests. In the mixed models, the association is tested by a null hypothesis of zero variance via a sequence kernel association test (SKAT) and its optimal unified test (SKAT-O). Although there are some comparison studies to evaluate the performance of mixed and fixed models, there is no systematic analysis to determine when the mixed models perform better and when the fixed models perform better.

Here we evaluated, based on extensive simulations, the performance of the fixed effect and mixed model statistics, using genetic variants located in 6, 9, 12, and 15 kb simulated regions. We compared the performance of three models: (1) mixed models which lead to SKAT and SKAT-O, (2) traditional fixed effect additive models, and (3) fixed effect functional regression models. We performed simulation analyses for two scenarios: (1) all causal variants are rare; (2) some causal variants are rare and some are common.

We found that the fixed effect tests have accurately controlled false positive rates. In most cases, either one or both of the fixed effect models performed better than or similar to the mixed models, when some causal variants are rare and some are common, or when all causal variants are rare except for the 12 and 15 kb region cases. We argue that the fixed effect models are useful in most cases. In practice, it makes sense to perform analysis by both the fixed and mixed effect models and make a comparison, and this can be readily done using our R codes and the SKAT and SKAT-O packages.

29

metaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis

Anna Cichonska (1,2) Juho Rousu (2) Pekka Marttinen (2) Antti J. Kangas (3,4) Pasi Soininen (3,4) Terho Lehtimäki (5) Olli T. Raitakari (6,7) Marjo-Riitta Järvelin (8,9,10) Veikko Salomaa (11) Mika Ala-Korpela (3,4,12) Samuli Ripatti (1,13,14) Matti Pirinen (1)
(1) Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland (2) Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland (3) Computational Medicine, Institute of Health Sciences, University of Oulu and Oulu University Hospital, Oulu, Finland (4) NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland (5) Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere School of Medicine, Tampere, Finland (6) Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, Turku, Finland (7) Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku and Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland (8) Department of Epidemiology and Biostatistics, MRC Health Protection Agency (HPA) Centre for Environment and Health, School of Public Health, Imperial College London, London, United Kingdom (9) Institute of Health Sciences, University of Oulu, Oulu, Finland (10) Biocenter Oulu, University of Oulu, Oulu, Finland (11) National Institute for Health and Welfare, Helsinki, Finland (12) Computational Medicine, School of Social and Community Medicine and the Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom (13) Public Health, University of Helsinki, Helsinki, Finland (14) Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom

We introduce metaCCA, a computational framework for multivariate analysis of a single or multiple genome-wide association studies based on univariate regression coefficients. To our knowledge, it is the first summary statistics-based approach that allows multivariate representation of both phenotype and genotype.
metaCCA extends the statistical technique of canonical correlation analysis to the setting where original individual-level data are not available. Instead, metaCCA operates on three pieces of the full covariance matrix: Cxy of univariate genotype-phenotype association results, Cxx of genotype-genotype correlations, and Cyy of phenotype-phenotype correlations. Cxx is estimated from a reference database matching the study population, e.g. the 1000Genomes, and Cyy is estimated from Cxy. We employ a covariance shrinkage algorithm to add robustness to the method.
Multivariate meta-analysis of two Finnish studies of nuclear magnetic resonance metabolomics by metaCCA, applied to standard univariate output from the program SNPTEST, shows an excellent agreement with the pooled individual-level analysis of original data. Root mean squared error between metaCCA's and original -$\log_{10}$ p-values is 0.02 when 455,521 SNPs from chromosome 1 are tested for an association with a cluster of 9 correlated lipid measures, and 0.45 when also genotypes are treated multivariately. We observed that moving from univariate to multivariate analyses of 9 lipids changes the top p-value from 10^{-10} to 10^{-24} for *CETP*, and from 10^{-9} to 10^{-12} for *APOE*. Motivated by the examples of known lipid genes, we envisage that multivariate association testing using metaCCA has a great potential to provide novel insights from published summary statistics.

30

Estimating clinical outcomes and classifying *CFTR* variants of unknown significance in children with a positive newborn screening for Cystic Fibrosis

David V. Conti (1) Colleen Azen (1) Duncan C. Thomas (1) Danieli Salinas (1)
(1) University of Southern California

Each year ~3,700 infants in the US are identified through newborn screening to be at risk for developing Cystic Fibrosis (CF). While ~70% of CF cases of European ancestry are due to carrying two copies of F508del, other combinations of any two CF-causing variants can cause disease. With information on over 39,000 patients, the Clinical and Functional TRanslation of *CFTR* mutation database has classified only 174 as CF-causing of the near 2,000 *CFTR* variants identified. Thus, there is a great need to predict the clinical outcomes in children with positive newborn screening and to predict the penetrance of variants of unknown significance (VUS). We present a Bayesian hierarchical model, with one level for the phenotypes (e.g. immunoreactive trypsinogen, sweat chloride concentration), a level for the risk of CF given the disease-causing status of the joint genotype, and a level for each variant as a function of external information (e.g. genomic annotation). We apply our model to data from 848 children with a positive screening test from the California Newborn Screening Program. We estimate that ~7-10% of the VUS have an elevated posterior probability of being classified as CF-causing variant. We demonstrate sensitivity to various models forms (e.g. with and without phenotype information) and prior specifications (e.g. with and without genomic annotation). We discuss the use of this model to classify previously unobserved VUS, predict clinical status for newborns with a positive screening test, and to impact clinical care protocols.

31

Methodology for the analysis of multi-ethnic genome-wide association studies

James P. Cook (1) Andrew P. Morris (1)
(1) Department of Biostatistics, University of Liverpool, UK

Traditional genome-wide association studies (GWAS) have primarily used collections of individuals from homogeneous population groups because: (i) geographical confounding between the trait and genetic variation can inflate type I error rates; and (ii) there may be reduced power due to heterogeneity in allelic effects on the trait between ethnicities. Confounding is typically accounted for by including principal components (PCs) as covariates in a regression framework, and has been demonstrated to control type I error rates within ancestry groups. However, by including genetic data from diverse populations, the first two PCs also generate axes of genetic variation anchored by individuals of African, European and East Asian ancestry.

In this study, we performed simulations of binary outcomes and GWAS data from diverse ancestries to evaluate: (i) power and type I error in multi-ethnic analyses with adjustment for PCs; and (ii) power to detect heterogeneity in allelic effects between populations by including an interaction between a SNP and the first two PCs. The study demonstrated that type I error rates were controlled by including 10 PCs as covariates in the regression analysis, and that power to detect heterogeneity was strong even at modest effect sizes.

We applied these methods to an imputed multi-ethnic GWAS of Type 2 diabetes (T2D) using 71,604 unrelated participants (9,747 cases) in the Resource for Genetic Epidemiology Research on Adult Health and Aging. Lead SNPs at ten loci attained genome-wide significant evidence ($p < 5 \times 10^{-8}$) of association with T2D, including a novel signal mapping to *TOMM40* ($p = 2.8 \times 10^{-9}$), a gene previously implicated in Wolfram Syndrome, a neurodegenerative disorder characterised by diabetes.

32

An exploration of known type 2 diabetes susceptibility variants: informative heterogeneity revisited

Laura J. Corbin (1) Christopher R. Boustred (2) Kimberley Burrows (1) George Davey Smith (1) Debbie A. Lawlor (1,3) Ruth J. F. Loos (4,5,6,7) Massimo Mangino (8) Susan M. Ring (1,3) Andrew J. Simpkin (1) Nicholas J. Timpson (1)
(1) MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK (2) North East Thames Regional Genetics Service, Great Ormond Street Hospital NHS Foundation Trust, London, UK (3) School of Social and Community Medicine, University of Bristol, Bristol, UK (4) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK (5) The Charles Bronfman Institute for Personalized Medicine, The Mindich Child health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA (6) The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, New York, USA (7) The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA (8) DTR Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

The number of common genetic loci associated with complex disease is constantly expanding but the biological mechanisms underlying these associations are often unknown. We present an approach to explore the aetiology of known genetic signals for type 2 diabetes in healthy individuals. Analyses were performed in collections from the Avon Longitudinal Study of Parents and Children (7,011 subjects) and the UK Adult Twin Registry (2,482 subjects). A series of composite metabolic phenotypes were derived by subtracting a standardized measure of adiposity (body mass index (BMI)) from indicators of glycaemic health, such as standardized fasting plasma glucose. These composite phenotypes capture metabolic disparity as a quantitative trait and were analysed in a series of genome-wide association analyses.

Using composite metabolic disparity variables, we were able to provide insight into the aetiology of complex associations by examining contributions to multiple traits in an informative manner. We were able to delineate what appear to be single-trait genetic signals from those with potentially complex pleiotropic contributions. We identified a number of loci that show evidence for an opposite effect on BMI and fasting plasma glucose (*TCF7L2*, *ADCY5*, *CDKAL1*) and which coincide with those previously implicated as being involved in defective beta cell function. There was also evidence for an association between the *IRS1* locus and the composite phenotype of insulin minus BMI that may be the result of bidirectional effects on the contributing phenotypes. We propose that this approach can provide insights into the links between metabolic traits and disease, and is complementary to both pathway-based analyses and functional studies.

33

Regional IBD Analysis (RIA): a new method for linkage analysis in extended pedigrees using genome-wide SNP data

Heather J. Cordell (1) Jakris Eu-ahsunthornwattana (1) Jakris Eu-ahsunthornwattana (2) Richard A. J. Howey (1)
(1) Newcastle University, UK (2) Mahidol University, Thailand

The study of rare variants has revived interest in linkage analysis. However, exact calculations for traditional linkage analysis are computationally impractical in large, extended pedigrees. Although simulation-based methods can be used, they require significant computation and are not exact. We propose Regional IBD Analysis (RIA), a non-parametric linkage method based on comparison of locally and globally estimated identity by descent (IBD) sharing in affected relative pairs (ARPs). In RIA, genome-wide SNP data are used to calculate the "global" expected IBD sharing probabilities for each ARP, against which a "local" set of IBD probabilities, estimated using SNP data within windows of pre-specified width, are compared. The global and local IBD probabilities are used to construct a non-parametric maximum likelihood statistic (MLS)-like test of linkage in each window. We illustrate our method with real nuclear-family data from a study

of vesicoureteral reflux and simulated data based on large extended pedigrees from a study of visceral leishmaniasis. RIA successfully detected the linkage signals with significant reduction in computational time (e.g. 2 hours vs 66 hours on 3,626 individuals from 308 extended families, genotyped at 545,433 SNPs) compared to traditional methods. RIA should be useful in studies involving large extended families, in which traditional linkage analysis is not feasible. Additionally, because it does not rely on prior knowledge about familial relatedness, RIA has an additional advantage of being robust to pedigree misspecification and can be used even in the absence of pedigree information.

34

Immunochip analysis identifies amino acid residues in five separate HLA genes driving the association between the MHC and primary biliary cirrhosis

Rebecca Darlay (1) Kristin L. Ayers (1,2) George F. Mells (3) Victoria A. Money (4) Jimmy Z. Liu (5,6) Mohamed A. Almarri (5,7) Graeme J. Alexander (8) David E. Jones (1) Richard N. Sandford (3) Carl A. Anderson (5) Peter T. Donaldson (1) Heather J. Cordell (1)
(1) Newcastle University, UK (2) Icahn School of Medicine at Mount Sinai, USA (3) Cambridge University, UK (4) Durham University, UK (5) Wellcome Trust Sanger Institute, UK (6) New York Genome Center, USA (7) Department of Forensic Science and Criminology, Dubai Police HQ, United Arab Emirates (8) Cambridge University Hospitals National Health Service (NHS) Foundation Trust, UK

Primary biliary cirrhosis (PBC) is a classical autoimmune liver disease characterized by progressive auto-immune destruction of intrahepatic bile ducts. The strongest genetic association is seen with *HLA-DQA1*∗04:01, with at least three additional independent *HLA* haplotypes contributing to susceptibility. To identify functional, potentially causal variants within the HLA region, we used dense SNP data in 2861 PBC cases and 8514 controls to impute classical *HLA* alleles and amino acid polymorphisms using the software packages HIBAG and SNP2HLA. Through stepwise analysis we demonstrate that association in the HLA region is largely driven by variation at five separate amino acid positions: position 11 of HLA-DPβ1 ($p = 4.72 \times 10^{-60}$), position 74 of HLA-DRβ1 ($p = 2.95 \times 10^{-40}$), position 57 of HLA-DQβ1 ($p = 1.73 \times 10^{-21}$), position 156 of HLA-C ($p = 2.83 \times 10^{-12}$) and position −13 in the signal peptide of HLA-DQβ1 ($p = 7.64 \times 10^{-12}$). Three dimensional modelling and calculation of electrostatic potentials was performed to explore the effect of these key residues on these molecules. Two of the associated residues were shown to affect the electrostatic charge of the peptide binding groove. An aspartic acid at residue 57 of HLA-DQβ1, which protects against PBC, induces a negative charge in pocket 4 of the peptide binding groove, whereas an arginine at residue 156 of HLA-C, which is associated with an increased risk of PBC, induces a positive charge within the peptide binding groove.

35

Comparison of Haplotype-based Statistical Tests for Disease Association with Rare and Common Variants

Ananda S. Datta (1) Swati Biswas (1)
(1) University of Texas at Dallas

Recent literature has highlighted the advantages of haplotype association methods for detecting rare variants associated with common diseases. As several new haplotype association methods have been proposed in the last few years, a comparison of new and standard methods is important and timely for guidance to the practitioners. We consider nine methods: Haplo.score, Haplo.glm, Hapassoc, Bayesian hierarchical GLM (BhGLM), Logistic Bayesian LASSO (LBL), regularized GLM (rGLM), Haplotype Kernel Association Test (HKAT), wei-SIMc-matching, and Weighted Haplotype and Imputation-based Tests (WHaIT). These can be divided into two types — individual haplotype-specific tests and global tests depending on whether there is just one overall test for a haplotype region (global) or if there is an individual test for each haplotype in the region. Haplo.score is the only method that tests for both; Haplo.glm, Hapassoc, BhGLM, and LBL are individual haplotype-specific while the rest are global tests. For comparison, we also apply a popular collapsing method — Sequence Kernel Association Test (SKAT) and its two variants — SKAT-O (Optimal) and SKAT-C (Combined). We carry out an extensive comparison on our simulated data sets as well as on the Genetic Analysis Workshop (GAW) 18 simulated data. Further, we apply the methods to two real datasets — GAW18 real hypertension data and Dallas Heart Study sequence data. We find that LBL, Haplo.score, and rGLM perform well over a wide range of scenarios. Also, haplotype methods are more powerful than SKAT and its variants in situations considered here, reinforcing the usefulness of haplotype-based approaches.

36

Characterizing an inverse axis between orthogonal sources of genetic risk

Lea K. Davis (1) S. Hong Lee (2) Eric R. Gamazon (3) Hae-Kyung Im (1) Dongmei Yu (4) Stephanie Williams (5) Patrick F. Sullivan (5) Carol Mathews (6) James Knowles (7) Jeremiah Scharf (4) Naomi Wray (2) Nancy J. Cox (3)
(1) University of Chicago (2) Queensland University (3) Vanderbilt University (4) Massachusetts General Hospital (5) University of North Carolina (6) University of Florida (7) University of Southern California

Significant roles for both polygenic risk and risk from large effect highly penetrant, but rare, variants has been established across many complex traits. Based on these observations, we hypothesize that if an individual may develop disease by crossing either a polygenic liability threshold or a 'penetrant variant' liability threshold, we should detect an inverse correlation between these two orthogonal sources of genetic risk among cases. We tested this hypothesis with

polygenic risk scores (PRS) calculated using the best linear unbiased prediction method for the estimation of random effects drawn from previously published genetic relationship matrices (Davis et al., 2013; Lee et al., 2013) and copy number variation (CNV) data (McGrath et al., 2014; Sanders et al., 2011) in multiple psychiatric phenotypes including Tourette Syndrome (TS), obsessive-compulsive disorder (OCD), and autism spectrum disorders (ASDs). We show that there is a modest negative association between polygenic risk and presence of rare (< 1% in the Database of Genomic Variants), large (> 500 kb), genic, CNVs among individuals with TS ($N = 516$; $p = 0.02$), but no statistically detectable difference in PRS between CNV carriers and non-carriers among individuals with OCD ($N = 919$; $p = 0.12$). Additionally, we found lower PRS scores in individuals with ASD and rare, large, genic, CNVs ($N = 777$; $p = 0.05$), however, no difference was detected among unaffected siblings harboring such CNVs ($N = 622$; $p = 0.21$). Taken together, the results from TS and ASD suggest that both sources of genomic risk are critically important and orthogonal, and should be considered jointly.

37

Admixture analyses of phenotypes related to the metabolic syndrome in a Brazilian population

Mariza de Andrade (1) Luting Xue (2) Julia M. P. Soler (3)
(1) Mayo Clinic, Rochester, MN, USA (2) University of Boston, MA, USA (3) University of São Paulo, SP, Brazil

We conducted admixture analyses of metabolic syndrome in the Baependi family study from the state of Minas Gerais, Brazil to better understand the basis of ethnic differences in the metabolic syndrome. The Baependi family study consists of 80 families and 1,109 subjects with complete clinical information and genotype data from Affymetrix 6.0 SNP chip. Genome-wide admixture analysis was performed to test whether local ancestry in this data was associated with any of the phenotypes in the metabolic syndrome using a mixed linear model adjusted for age, sex, principal components 1 and 2 taking into account the family structure. Admixture analysis was performed using PCAdmix, a principal components based algorithm for determining ancestry along each chromosome from a high-density genome-wide set of phased single nucleotide polymorphism (SNP) genotypes of admixed individuals. It returns a local ancestry estimate for haplotypes for each reference population. We used HapMap Phase 3 CEU and YRI samples and the HGDP Native American (NA) samples as the reference samples for PCAdmix. We observed an average local ancestry to be 0.69/ 0.13/ 0.18 across the genome for CEU/ YRI/ NA, respectively. We observed a peak on chr2p due to African ancestry for systolic blood pressure, on chr3 due to Native American ancestry for diastolic blood pressure, and on chrs 2q and 5q due to African ancestry for truncal obesity. In summary, by performing association analysis using local ancestry help to understand the ethnic differences in phenotypes related to the metabolic syndrome better than the standard association analysis.

38

Identification of Gene-Environment Interactions in Venous Thromboembolism (VTE) using time to event is more effective than using case-control approach

Mariza de Andrade (1) Sebastian M. Armasu (1) Bryan M. McCauley (1) Tanya M. Petterson (1) John A. Heit (1)
(1) Mayo Clinic, Rochester, MN, USA

VTE is a major public health problem with over 500,000 events per year in the U.S. The average age- and sex-adjusted annual VTE incidence is 132 per 100,000 person-years; VTE incidence increased by ~5% per decade over the last 35 years. We suspect this incidence increase is largely due to our limited understanding of which individuals should be targeted for prophylaxis. While several exposures (hospitalization, pregnancy, etc.) are strongly associated with VTE, these exposures have poor predictive value for the individual. We had showed that VTE is highly heritable and follows a complex polygenic inheritance mode with environmental interaction. Here, we focus on pregnancy as the environmental exposure using data from our Mayo VTE study that included 472 incident VTE events among 1,146 women. Treating these women as a historical cohort and date of birth as time zero, we performed time-to-VTE analyses. Pregnancy was associated with an increased hazard of VTE (HR = 2.49; 95%CI: 1.84, 3.36; $p = 3.2 \times 10^{-9}$). Of these 1,146 women, 634 had genome-wide genotype data imputed to 1000G, and 277 had pregnancy-related VTE. Pregnancy remained associated with VTE in this cohort subset (HR = 3.3; $p = 1.4 \times 10^{-8}$). The proportional hazards assumption held in both data sets ($p = 0.73$). In a GWA study, 2 chr 7 SNPs, rs10215876 and chr7:44909852:D, at 5–6kb 3' of *PURB* and 1 chr 9 SNP in *LINGO2* were associated with reduced hazards of VTE in pregnancy (HRs = 0.40 [0.28-0.58, $p = 1.2 \times 10^{-8}$], 0.41[0.29-0.59, $p = 3.3 \times 10^{-8}$] and 0.63 [0.52-0.75, $p = 3.3 \times 10^{-7}$], respectively). *PURB* encodes for a DNA-binding protein that preferentially binds the purine-rish element, PUR; deletion of this gene is associated with myelodysplastic syndrome and acute myeloid leukemia, both of which are associated with VTE.

39

Leveraging gene regulatory data in hypothesis-driven GWAS: The importance of shared tissue specificity

Jessica Dennis (1) Alejandra Medina (2) Vinh Truong (3) Lina Antounians (4) Pierre Morange (5) David-Alexandre Trégouët (6) Michael Wilson (4) France Gagnon (3)
(1) Dalla Lana School of Public Health, University of Toronto (2) International Laboratory for Research in Human Genomics, Universidad Nacional Autónoma de México (3) Dalla Lana School of Public Health, University of Toronto (4) Department of Molecular Genetics, University of Toronto (5) Faculty of Medicine, Aix-Marseille University (6) Université Pierre et Marie Curie

Disease-associated SNPs are enriched for tissue specific gene regulatory regions. We asked whether incorporating gene

regulatory information into GWAS via a hypothesis-driven (HD) GWAS would improve discovery. Plasma levels of von Willebrand factor (VWF) and tissue factor pathway inhibitor (TFPI) are markers of endothelial cell (EC) dysfunction, a precursor to thrombosis, and are heritable. We conducted HD-GWAS' of TFPI and VWF plasma levels, prioritizing SNPs in EC regulatory regions, which were experimentally ascertained using ChIP-seq for epigenetic histone modifications. Our GWAS sample included 255 subjects from 5 families ascertained on thrombosis. We tested the significance of individual SNPs via the stratified (s) FDR q-value and calculated an empirical p-value for the hypothesis itself by comparing the sum of the Wald values for the prioritized SNPs to the sum of the Wald values from 1,000 GWAS of prioritized SNPs and permuted phenotypes. 183,415 of 6,143,330 SNPs were in one of 19,163 EC regulatory regions. The top 5 VWF-associated SNPs were from the prioritized stratum, with sFDR q-values < 0.05, and the hypothesis was highly significant ($p < 0.001$). Although the top 11 TFPI-associated SNPs were from the prioritized stratum, the minimum sFDR q-value was 0.30 and the hypothesis p-value was 0.05. VWF is expressed exclusively by EC, whereas TFPI is expressed by multiple tissues. Our results suggest that an HD-GWAS with epigenetic data is most promising when the epigenetic data and phenotype share the same tissue specificity, and we are seeking replication in an independent sample. We are also developing our analysis framework into a software package to be used with any phenotype, which will leverage public epigenetic databases.

40

An extension of Conditional Inference Forest methodology for predictive biomarkers and personalized medicine applications

Benjamin Dizier (1,2) Kristel Van Steen (2,3)
(1) Systems and Modeling Unit, Department of Electrical Engineering and Computer Science (Montefiore Institute), University of Liège, Quartier Polytech 1, Allée de la Découverte 10, 4000 Liège, Belgium (2) Systems Biology and Chemical Biology, GIGA-R, University of Liège, Quartier Hôpital, Avenue de l'Hôpital 11, 4000 Liège, Belgium (3) Systems and Modeling Unit, Department of Electrical Engineering and Computer Science (Montefiore Institute), University of Liège, Quartier Polytech 1, Allée de la Découverte 10, 4000 Liège, Belgium

Heterogeneity of treatment effect is a major hurdle for new drug development. When treatment effect is heterogeneous, average treatment effect estimation of new drug treatment effect is biased and requires very large trials to demonstrate a significant effect. This increases costs, time and risk of drug development.
There is big hope that personalized medicine would bring solutions to this issue by enabling identification of patients more likely to benefit from the new drug to demonstrate efficacy. Generating actionable information from all the data generated by high-throughput technologies requires robust methodologies to avoid the pitfalls of exhaustive subgroup analysis.
Here, we present a novel machine learning approach based on recursive partitioning for estimation of individual treatment benefit through counterfactual framework in time to event outcomes. The estimate of individual treatment benefit is used to assess heterogeneity of treatment effect and potential presence of subgroups that may derive benefit/harm from the new drug. Conditional on the presence of treatment effect heterogeneity, predictive biomarkers that could potentially become companion diagnostics for patient stratification or help understanding drug mode of action are identified through variable importance. The methodology allows for prognostic index or propensity score adjustments in case of strong prognostic effects (independent of treatment) and deviation from randomization between treatment arms.
The performance of the novel methodology is further evaluated and validated via application to synthetic and real-life data.

41

Different Genomic Subsets and Cell Types Contribute to the Polygenicity and Heritability for Coronary Artery Disease

Ron Do (1) Hong-Hee Won (2) Pradeep Natarajan (2) Amanda Dobbyn (3) Kasper Lage (4) Soumya Raychaudhuri (5) Eli Stahl (6)
(1) Genetics and Genomic Sciences, Icahn School of Medicine, New York, New York, USA (2) Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA (3) Institute for Genomics and Multiscale Biology, Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, New York, USA (4) Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA (5) Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA (6) Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Coronary artery disease (CAD) and myocardial infarction (MI) is one of the leading causes of death and disability worldwide. Genome-wide association studies (GWAS) have identified up to 45 loci associated with CAD/MI. However, the molecular consequences of these loci remains largely unknown. Efforts by the National Institutes of Health (NIH) Roadmap Project, have allowed for systematic assessment of the molecular functions of the genome. By leveraging data from the NIH RoadMap Project we investigated links between single nucleotide polymorphisms (SNPs) associated with MI/CAD and their biological impacts on gene function. First, we stratified MI/CAD according to genomic compartments and observed that SNPs that were proximal to the protein-coding region of genes exhibited significant polygenicity and heritability (> 59%) for MI/CAD. Next, we showed that the polygenicity and heritability of MI/CAD

are enriched in histone modification regions in specific cell types. By focusing on 45 MI/CAD-associated SNPs identified from GWAS, we showed that a higher number of these loci are enriched within specific regulatory elements, including active enhancer and promoter regions. Finally, we observed significant heterogeneity across cell types, with particularly strong signal observed in adipose nuclei, brain substantia nigra, brain angular gyrus, spleen cell types and smooth muscle tissues. These results suggest that the genetic etiology of MI/CAD is largely explained by tissue-specific regulatory perturbation within the human genome.

42

Differential expression of transcript isoforms in schizophrenia

Eugene I. Drigalenko (1) Winton Moy (2) Jubao Duan (2) Harald H. H. Göring (3) Pablo V. Gejman (2) Alan R. Sanders (2)
(1) Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX (2) Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, IL (3) South Texas Diabetes and Obesity Institute, University of Texas Health Science Center, San Antonio, TX

Transcriptome sequencing (RNA-seq) enables investigation of differential expression of transcript isoforms and genes, with expression level of the former likely more meaningful than of the latter since transcript isoforms often encode different proteins with different biological function.
We profiled the transcriptome (> 9M single-end RNA-seq reads per subject) in lymphoblastoid cell lines of 1,278 European ancestry subjects (550 cases, 728 controls) from the MGS collection. We used programs Tophat and Cufflinks2 for read alignment, transcript assembly, and quantification (FPKM) of each transcript isoform and gene. We restricted our analyses to the 108,346 transcripts and 21,215 genes from Gencode v.20 annotation that showed detectable expression (FPKM > 0) in at least 80% (1,022) of subjects.
We used square root as the variance stabilizing transformation for the FPKM, and then adjusted for 12 covariates. After inverse normalization of residuals, we evaluated the association of expression levels with schizophrenia status by linear regression.
We found 1,265 transcripts and 1,252 genes to be differentially expressed by affection status (Bonferroni corrected $p < 0.05$). Of these transcripts, 1,015 were in genes also showing significant differential expression in the gene-level analysis, while 250 were in genes not significantly differentially expressed. 447 differentially expressed genes did not show differential expression at the transcript isoform level. We found a number of significant transcripts and genes in HUGO Gene Nomenclature Committee gene families, such as histocompatibility complex, histones, interferons, and interleukins.

43

Characterisation of the metabolic impact of rare genetic variation within *APOC3*: Proton NMR based analysis of rare variant gene effects

Tom Dudding (1) Fotios Drenos (1) Johannes Kettunen (2,3,4) Peter Wurtz (2) Pasi Soininen (2,3) Antti Kangas (2) Aroon Hingorani (5) Tom Gaunt (1) Juan P. Casas (6) Mika Ala Korpela (1,2,3,7,8) George Davey Smith (1) Nicholas J. Timpson (1)
(1) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK (2) Computational Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland (3) NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland (4) Public Health Genomics Unit, Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland (5) Genetic Epidemiology Group, University College London, London, UK (6) London School of Hygiene and Tropical Medicine, London, UK (7) Computational Medicine, School of Social and Community Medicine, University of Bristol, Bristol, UK (8) Computational Medicine, Oulu University Hospital, Oulu, Finland

We previously reported rare variant in *APOC3* (rs138326449) associated with plasma TG levels (-1.43SD (s.e. = 0.27),$p = 8.0 \times 10^{-8}$) discovered using low read-depth whole genome sequence. Here we aimed to characterise in greater detail the impact of variation at this locus.
A high-throughput serum nuclear magnetic resonance (Proton NMR) metabolomics platform was used to quantify ≥ 225 metabolic measures in > 10,000 participants from the Avon Longitudinal Study of Parents and children and > 3500 from the British Women's Heart and Health Study. We analysed the effect of the *APOC3* variant on the measured metabolites and used the common *LPL* rs12678919 polymorphism to test for LPL independent effects.
In testing all 225 metabolites measured in all samples for association with rs138326449, 142 showed evidence of association ($p < 0.05$). rs138326449 was associated with TG ($p = 6.7 \times 10^{-6}$) and HDL ($p = 1.1 \times 10^{-11}$). We found additional associations with VLDL and HDL composition, other total cholesterol measures and fatty acids. Extra resolution showed that the effects of rs138326449 on VLDL and HDL are across the entire spectrum of their particle size and are not specific. Comparison of the *APOC3* and *LPL* association revealed that of the 225 metabolites tested, 3 had no overlapping effects between rs138326449_*APOC3* effects and those predicted by rs12678919_*LPL*. Specifically, the composition of medium and very large VLDL is not predicted by the action of *APOC3* through *LPL*.
We characterised the effects of *APOC3* rs138326449 loss of function mutation in lipoprotein metabolism. Results are consistent with recent clinical trials of *APOC3* inhibition and should be used to inform future Mendelian randomisation and recall by genotype studies.

44

Examining the causal effect of Vitamin D on childhood caries: A Mendelian Randomization study

Tom Dudding (1,2) Steve J. Thomas (2) Karen Duncan (2) Debbie A. Lawlor (1) Nicholas J. Timpson (1)
(1) MRC Integrative Epidemiology Unit, University of Bristol
(2) School of Oral and Dental Science, University of Bristol

Dental caries is the localized destruction of susceptible tooth tissues by acidic by-products from bacterial fermentation of dietary products. Previous studies have reported an inverse association between vitamin D and childhood dental caries, but whether this is causal is unclear.

Mendelian randomisation (MR) is an analytical method used to provide evidence for causal associations between potentially modifiable risk factors and health outcomes. We undertook an MR study, using genetic variants known to influence circulating 25-hydroxyvitamin D (25 (OH)D) levels, in 5,544 European origin children from the South West of England, to determine the causal effect of circulating 25 (OH)D on dental caries. Data on caries and related characteristics were obtained from parental completed questionnaires between 38 and 91 months and clinical assessments in a random 10% sample at 31, 44 and 61 months.

In multivariable confounder adjusted analyses we found no strong evidence for an association of 25 (OH)D with caries experience, severity or onset, or having a general anesthetic (GA) for dental problems. Known relationships between vitamin D and *CYP2R1*, *DHCR7* and *GC* were replicated and when combined these SNPs were confirmed as a strong genetic instrument for vitamin D with per allele changes of a clinically relevant magnitude. In MR analysis the odds ratios per 10 nmol/L increase in 25 (OH)D were 0.88 (95% confidence interval (CI): 0.78, 1.00; P = 0.054) for caries experience and 0.94 (95% CI: 0.73, 1.21; P = 0.62) for GA. Our MR study suggests that there might be an inverse causal effect of 25 (OH)D on dental caries. However, our estimates are imprecise and a larger study is required to determine a robust effect. This study highlights the opportunity to apply genetics based causal methods to dental epidemiology.

45

A Framework for the Behavior of Rare Variant Tests in the Presence of Large Numbers of Variants

Mackenzie Edmondson (1) Reginald Lerebours (2) Katie McKenzie (3) Nathan Tintle (4)
(1) St. Michael's College (2) Harvard University (3) Duke University (4) Dordt College

As we continue in the era of next-generation sequencing data, questions about how to most appropriately analyze genetic sequence data still remain, particularly with regard to the influence of rare genetic variants in human disease. Numerous rare variant testing methods have been proposed, but as sample size and coverage increase, along with a shift towards whole genome sequencing, the number of variants being considered continues to increase. Current gene-based rare variant testing methods have been well-tested for sets of less than 100 variants, but as the number of variants increases into the hundreds and thousands, anecdotal evidence suggests current industry standard methods may no longer perform optimally. Alternatively, two-stage methods (e.g., pathway methods) offer an alternative approach which may be better in some cases. This framework (1) classifies proposed variant set tests and explains observed differences in performance which can be used to directly connect genetic disease models with statistical power, (2) guides researchers in prospective test selection, and (3) provides the opportunity to analytically evaluate novel set-based rare variant tests.

46

MVtest: a method to flexibly model the genetic determinants of trait variability

Todd L. Edwards (1) Eric S. Torstenson (1) John Gilbertson (1) Michael J. Bray (1) Krystal S. Tsosie (1) Sarah C. Stallings (1) Ayush Giri (1) Mollie Bodin (1) Suzette J. Bielinski (2) Jyotishman Pathak (2) Maureen E. Smith (3) Abel Kho (3) Geoff Hayes (3) Jennifer A. Pacheco (3) Marylyn D. Ritchie (4) Shefali Setia (4) Gerard Tromp (5) Martha J. Shrubsole (1) Reid M. Ness (1) Douglas K. Rex (1) Thomas M. Ulbright (6) Mariza de Andrade (2) David R. Crosslin (7) Gail P. Jarvik (7) Eric B. Larson (8) David S. Carrell (8) Xiao-Ou Shu (1) Wei Zheng (1) Dan M. Roden (1) Ruth J. F. Loos (9) Digna R. Velez Edwards (1) Chun Li (10)
(1) Vanderbilt University (2) Mayo Clinic (3) Northwestern University (4) Penn State University (5) Geisinger Health System (6) Indiana University (7) University of Washington (8) Group Health (9) Mount Sinai School of Medicine (10) Case Western Reserve University

Genetic association studies of complex traits have successfully identified many heritable determinants of quantitative traits. Interactions with genetic exposures may be important aspects of relationships between genomes and phenotypes. However, these relationships are difficult to detect with standard modeling and study designs. Our goal is to detect SNPs with effect modifiers by evaluating the main effect on trait variance that is created by interaction. We developed the mean and variance test (MVtest) to simultaneously model the mean and log-variance of a quantitative trait as linear functions of genotypes and covariates with estimating equations. The advantages of our strategy are that the effect modifier does not need to be known, multiple testing is limited to conventional thresholds, and meta-analysis can combine evidence for association across studies. In simulations we showed control of type I error and power compared with several alternatives. In meta-analysis of 45,348 subjects, we detected novel associations with body mass index (BMI) variability at the *ACER2* gene on chr9 (p-value = 1×10^{-11}), previously reported to be associated with lipogranulomatosis, and in the chr1q41 region nearby the *LYPLAL1* gene (p-value = 1.3×10^{-8})

previously reported by several large consortia to have sexually dimorphic associations with adiposity. We also replicated the previously reported association with BMI variability at the *FTO* locus (p-value = 0.001). Known genes for mean BMI were also detected as well as a novel finding at a SNP 51kb upstream of the *LEP* gene (p-value = 8×10^{-7}), which is known to act on energy homeostasis and satiety. These results show that studies of main effects on trait variability can detect SNPs with contextual effects.

47

Investigating Imprinting As A Mechanism For The Development Of Asthma and related phenotypes In Two Canadian Birth Cohorts

Aida Eslami (1) Loubna Akhabir (1) George Ellis (1) Allan B. Becker (2) Anita L. Kozyrskyj (2) Catherine Laprise (3) Peter D. Paré (1) Andrew J. Sandford (1) Denise Daley (1)
(1) Centre for Heart Lung Innovation, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada (2) Department of Pediatrics and Child Health, Faculty of Medicine, University of Manitoba, Winnipeg, MB, Canada (3) Université du Québec á Chicoutimi, Saguenay, QC, Canada

Asthma is a complex disease caused by a combination of genetic and environmental factors. To date, 23 genome-wide association studies (GWAS) have been completed for asthma as a primary phenotype. Combined, these studies identified 44 SNPs at $p < 1 \times 10^{-6}$. The consensus is that main genetic effects of these common SNPs do not fully explain the heritability of asthma. Genomic imprinting is a mechanism, which may explain some of the missing heritability. Imprinting is an epigenetic phenomenon where the expression of genes depends on their parental origin (parent-of-origin effect).
Specific genomic imprinted regions are associated with asthma and related phenotypes (atopy and airway hyperresponsiveness (AHR)).
To identify candidate genomic regions for imprinting we used GWAS data from two family-based studies (trios): the Canadian Asthma Primary Prevention Study (CAPPS) and the Study of Asthma Genes and Environment (SAGE). We used a likelihood-based variant of the Transmission Disequilibrium Test. Parent-of-origin effects were tested by including a modifier (the sex of parents) in the analysis.
In the joint analysis with 148 asthmatic trios, 13 SNPs showed significant parent-of-origin effects with $p < 10^{-5}$. 3 SNPs remained significant after 100,000 permutations. Notably, we showed a parent-of-origin effect at a known imprinted gene, *CTNNA3*. Six SNPs were in or near Long non-coding RNA genes. The analyses for atopy (237 trios) and AHR (231 trios) yielded 3 and 2 significant SNPs after permutation, respectively.
To increase statistical power and confirm our results, we will perform a meta-analysis using 3 family studies: CAPPS, SAGE and the Saguenay-Lac-Saint-Jean study.

48

Two-Phase Designs for Joint Quantitative-Trait-Dependent and tag-SNP-Dependent Sampling

Osvaldo Espin-Garcia (1,2) Radu V. Craiu (1) Shelley B. Bull (1,2)
(1) Universty of Toronto (2) Lunenfeld-Tanenbaum Research Institute

Regional sequencing to follow up findings from a genome-wide association study (GWAS) can be cost effective for fine mapping. Under a case-control design, for example, sampling within strata defined by both disease status and genotype is more powerful than sampling from cases and controls only. For quantitative traits (QT), sampling strategies that select informative individuals for sequencing according to extreme-trait or tag-SNP values are known to have good properties, but designs based on joint sampling are not as immediate. To fill this gap we propose an extension of the available methods that account for joint QT-tag SNP dependent sampling. The proposed method corresponds to a formulation of semiparametric maximum likelihood estimation in which the tag SNP is an auxiliary covariate in inferring association between a sequence SNP and a normally distributed QT.
The purpose of this work is twofold. First, we assess improvements associated with QT-tag SNP joint sampling compared to QT or tag SNP marginal sampling. Second, we compare alternative designs, i.e. strata definition and sample allocations, with respect to estimation efficiency, test validity and power. To this end, we are performing a simulation study across levels of linkage disequilibrium (LD) between the tag and sequence SNPs for various minor allele frequencies, and across a range of effect sizes; Wald, score and likelihood ratio tests are implemented. We find that a joint sampling design can yield better power for the same sample size compared to either trait-dependent or tag-SNP sampling. Preliminary results under a balanced allocation demonstrate that efficiency and power of the proposed method improves as LD increases while type I error remains at adequate rates.

49

Meta-analysis of Complex Diseases at Gene Level by Functional Regression Models

Ruzong Fan (1) Christopher I. Amos (2) Yifan Wang (3) Daniel Weeks (4) Yun Li (5) Haobo Ren (6) Iryna Lobach (7) Momiao Xiong (8) Jason H. Moore (2) Michael Boehnke (9)
(1) Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), NIH (2) University of Dartmouth (3) Food and Drug Administration (FDA) (4) Pittsburgh University (5) University of North Carolina (6) Regeneron Pharmaceuticals, Inc. (7) University of California San Francisco (8) University of Texas - Houston (9) University of Michigan

We develop functional regression models (FRMs) to perform meta-analysis of multiple studies to evaluate the

relationship of genetic data to quantitative or dichotomous traits adjusting for covariates by using functional data analysis techniques. By treating multiple genetic variants of an individual in a human population as a realization of an underlying stochastic process, the genome of an individual in a chromosome region is a continuum of sequence data rather than discrete observations. The genome of an individual is viewed as a stochastic function which contains both genetic position and linkage disequilibrium (LD) information of the genetic markers. Unlike the previously developed MetaSKAT methods, which are based on mixed effect models to model the major gene contributions from loci as a random effect, FRMs are fixed effect models and genetic effects of multiple genetic variants are assumed to be fixed. The FRMs extend traditional population genetics model naturally and can fully utilize LD and genetic position information, while MetaSKAT only models pair-wise LD between each individual marker and the trait locus and cannot model LD among genetic markers.

Based on the FRMs, we developed test statistics to test for an association between a complex trait and multiple genetic variants in one genetic region. We then performed extensive simulations to evaluate the empirical type I error rates and power performance of the proposed models and tests. The proposed test statistics control the type I error very well and are conservative, and have higher power than MetaSKAT in most cases. Specifically, we show: (1) the proposed test statistics have higher power than MetaSKAT for quantitative traits no matter some causal variants are rare and some are common or all causal variants are rare, (2) the proposed test statistics have higher power than MetaSKAT for dichotomous traits when some causal variants are common and some are rare, and (3) when the causal variants are all rare (i.e., minor allele frequencies less than 0.03), the proposed test statistics have similar or slightly lower power than MetaSKAT for dichotomous traits.

The proposed methods were applied to analyze genetic data of 22 gene regions of type 2 diabetes data from a meta-analysis of eight European studies. These 22 genes are all from literature and each of them contains SNPs/variants which are related to the trait by single SNP/variant analysis in various studies. Hence, they can be treated as candidate genes and can be used to test the performance of gene based tests. The proposed methods detected significant association for 18 genes (p-values $< 3.10 \times 10^{-6}$) and tentative association for 2 genes (p-values around 10^{-5}) and no association for 2 genes, while MetaSKAT detected none. The models and related test statistics can analyze rare variants or common variants or a combination of the two, and can be useful in whole genome-wide and whole exome association studies.

50

A novel gene-based analysis method based on MB-MDR

Ramouna Fouladi (1,2) Claudia Schurmann (3) Kyrylo Bessonov (1,2) Jean-Philippe Vert (4,5,6) Ruth J. F. Loos (3) Kristel Van Steen (2,7)

(1) Systems and Modeling Unit, Department of Electrical Engineering and Computer Science (Montefiore Institute), University of Liège, Quartier Polytech 1, Allée de la Découverte 10, 4000 Liège, Belgium (2) Systems Biology and Chemical Biology, GIGA-R, University of Liège, Quartier Hôpital, Avenue de l'Hôpital 11, 4000 Liège, Belgium (3) Icahn School of Medicine at Mount Sinai, New York, NY, USA (4) MINES Parisaech, PSL-Research University, CBIO-Centre for Computational Biology, 35 rue St Honoré, 77300 Fontainebleau, France (5) Institut Curie, 75248 Paris Cedex, France (6) INSERM U900, 75248 Paris Cedex, France (7) Systems and Modeling Unit, Department of Electrical Engineering and Computer Science (Montefiore Institute), University of Liège, Quartier Polytech 1, Allée de la Découverte 10, 4000 Liège, Belgium

Here we present a novel gene-based test that builds upon Model-Based Multifactor dimensionality reduction (MB-MDR) which was initially developed to investigate SNP-based interactions from GWAS data. It relies on an organization of individuals in multi-locus genotype categories, followed by a labeling of these categories using trait information and appropriate association tests.

In a gene-based setting, any set of features (e.g., SNPs, structural variants, epigenetic markers) that can be mapped to a gene is submitted to a clustering algorithm to find groups of individuals (clusters) with similar "gene-based" profiles. The resulting profile types for a gene can be considered to be category levels of a gene-based summary variable. These categories can be labeled by MB-MDR as before.

Earlier, we obtained promising results taking rare and common variants as features and genes as regions of interest derived from exome sequencing data from GAW17. Here we show the utility of the approach on common variants derived from GWAS data, provided optional parameters are optimized to the new context, including those related to kernel selection prior to clustering. In particular, we explore the pros and cons of using diffusion kernels which allows the incorporation of biological information via defining a meaningful graph structure on the input features to each gene. In addition, special attention is given to the automatic detection of the optimal number of aforementioned clusters and the optimal minimal cluster size.

The newly proposed gene-based tool is evaluated on synthetic data via extensive simulations. We furthermore show its practical use on real-life data from the BioMe Biobank (Mount Sinai school of medicine, USA) for Type II diabetes.

51

A hierarchical model for differential isoform analysis with application to ovarian cancer

Brooke L. Fridley (1) Rama Raghavan (1) Junqiang Dai (1) Kimberly R. Kalli (2) Ellen L. Goode (2)

(1) University of Kansas Medical Center (2) Mayo Clinic

Using next-generation sequencing we assessed the transcriptome of tumors from epithelial ovarian cancer patients with

the goals to compare common high-grade serous (HGS) with the rarer histotypes and to compare a novel hierarchical differential isoform analyses to pairwise differential gene expression analyses. RNA was extracted and sequenced from 55 HGS, 42 endometrioid (ENDO), and 3 mucinous (MUC) fresh frozen tumors, followed by alignment (TopHat2) and abundance estimation (RSEM). Several lessons were learned from handling batch effects attributed to difference in library preps. Using a mixed model with fixed histotype, isoform, and histotype∗isoform effects and random subject effects for 10,141 genes we completed hierarchical differential isoform analyses. The hierarchical model detected 164 genes with histotype effects (and no interactions) and 347 genes with interactions ($p < 5 \times 10^{-6}$). Comparing these findings to results from pairwise differential gene expression analyses using edgeR, the majority of the genes with histotype main effects were also detected by the pairwise edgeR analyses on gene abundance estimates. However, 1,874 genes were only detected with the mixed model for a significant interaction ($p < 0.05$). For example, isoforms of the autophagy gene *ATG4B* that were highly expressed in MUC were expressed at a much lower level in ENDO and HGS tumors with no significant overall difference in level of expression across all histologies (Mixed Model: histotype p = 0.97, interaction p = 2.5×10^{-11}; pairwise edgeR: ENDO vs HGS p = 0.93, HGS vs MUC p = 0.96, MUC vs ENDO p = 0.98). This work suggests that a combination of analytical approaches is needed to best inform interpretation of somatic RNA sequencing data.

52

Kernel-based Pathway Meta-Analysis in ILCCO / TRICL Genome-wide Association Studies

Stefanie Friedrichs (1) Christopher I. Amos (2) Paul Brennan (3) David Christiani (4) Rayjean J. Hung (5) Angela Risch (6,7,8) Irene Brüske (9) Neil Caporaso (10) Maria T. Landi (10) Thorunn Rafnar (11) Heike Bickeböller (1)
(1) Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Göttingen, Germany (2) Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA (3) International Agency for Research on Cancer (IARC), Lyon, France (4) Harvard School of Public Health, and Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA (5) Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada (6) Division of Molecular Biology, University Salzburg, Salzburg, Austria (7) Translational Lung Research Center Heidelberg (TLRC-H), Member of the German Center for Lung Research (DZL), Heidelberg, Germany (8) Div. Epigenomics and Cancer Risk Factors, DKFZ German Cancer Research Center, Heidelberg, Germany (9) Institute of Epidemiology I, Helmholtz Zentrum München-German Research Center for Environment Health, Neuherberg, Germany (10) Department of Health and Human Services, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH), Bethesda, MD, USA (11) deCODE genetics, Reykjavik, Iceland

Testing SNP sets is a promising strategy to improve power and interpretation in the analysis of genome-wide association studies (GWAS). We defined SNP sets representing genes and whole pathways using knowledge on gene membership and interaction within 284 pathways taken from the Kyoto Encyclopedia of Genes and Genomes database. To evaluate these pathways with a logistic kernel machine test (LKMT) efficiently, we enlarged gene boundaries to include the whole linkage disequilibrium (LD) block at the start and end of a gene. To date, we have analyzed five lung cancer GWAS of the ILCCO/TRICL consortia: from Central Europe, Germany, Harvard, Texas, and Toronto. Additional TRICL GWAS not yet analyzed are from the NCI and Iceland. In the LKMT, we employed the linear kernel as well as two more advanced kernels either adjusting for size bias in the number of SNPs/genes/pathways or incorporating the network structure of genes and the correlation structure of investigated SNP sets (Freytag et al., 2012, 2014). Analyses were all completed with our R-package "Kernel Approaches for Non-linear Genetic Association Regression" (kangar00). Thus far, in the five GWAS already analyzed, we have found the Oxytocin signaling pathway in one of the studies to prove significant when using the network based kernel. The resulting pathway-level p-values across studies were combined in a meta-analysis, initially using Fisher's p. No significant p-value emerged in meta-analysis. P-values were somewhat heterogeneous across kernels and studies. The linear kernel appears to overestimate pathway impact, whereas the size-adjusted kernel tends to underestimate it. The network kernel lies in between, with 17 pathways reaching nominal significance.

53

A Methods Comparison: In silico prioritization of genetic risk variants using functional genomic information

Sarah A. Gagliano (1) Reena Ravji (2) Michael R. Barnes (3) Michael E. Weale (4) Jo Knight (2)
(1) University of Toronto (2) Centre for Addiction and Mental Health (3) Queen Mary University of London (4) King's College London

Complex diseases are caused by many genetic risk variants, and environmental factors, but only a proportion of the risk variants have been identified with current sample sizes and techniques. By incorporating functional genomic annotations (e.g. histone modifications, transcription factor binding sites) as predictors, statistical learning has been widely investigated for prioritizing those variants that are likely to be associated with complex disease. We compared three published prioritization procedures, which use different definitions of risk variants (classifier), different statistical learning algorithms, and different predictors with regard to the quantity, type and coding. We explored different combinations of algorithm and annotation set for various classifiers. As an application, we tested which methodology

performed best for prioritizing variants using data from a large schizophrenia meta-analysis by the Psychiatric Genomics Consortium. Results suggest that all methods have considerable (and similar) predictive accuracies (AUCs 0.64-0.71) in test data set, but there is more variability in the application to the schizophrenia GWAS. Different annotations came up as most important for predicting risk variants depending on the algorithm and/or the classifier. Depending on the classifier, one can target common variants (e.g. from genome-wide association studies) or variants from sequencing studies. In conclusion, a variety of algorithms and annotations seem to have a similar potential to effectively enrich true risk variants in genome-scale datasets, but how those risk variants should be defined is unclear.

54

Mendelian Randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer

Chi Gao (1) Chirag J. Patel (2) Sara Lindstrom (1) Brandon Pierce (3)
(1) Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA (2) Center for Biomedical Informatics, Harvard Medical School, Boston, MA (3) Department of Health Studies, The University of Chicago, Chicago, IL on behalf of GAME-ON initiative

Adiposity traits have been associated with risk of many cancers but whether these associations are causal is unclear. We carried out Mendelian Randomization analyses to assess the causal relationship of birth weight, childhood obesity, adult body mass index (BMI), and waist-hip-ratio (WHR) on risks of five different cancers, including breast, ovarian, prostate, colorectal and lung. We considered weighted genetic risk scores including 7 birth weight associated variants, 9 childhood obesity associated variants, 77 BMI associated variants, and 14 WHR associated variants (all GWAS significant). We tested the association between these genetic scores and risk of cancers using summary statistics from the Genetic Associations and Mechanisms in Oncology (GAME-ON) Consortium, which included 51,537 cancer cases and 61,600 controls from 33 participating studies.
We found a significant inverse association between the genetic risk score for childhood obesity and risk of breast cancer (OR = 0.80 per unit increase in log odds of childhood obesity; 95%CI: 0.72,0.89; $p = 5.6 \times 10^{-5}$) and a positive association with ovarian cancer (OR = 1.19; 95%CI: 1.00,1.41; p = 0.048). Similarly, we found a significant inverse association between the genetic risk score for high adult BMI and risk of breast cancer (OR = 0.66 per s.d. increase in BMI; 95%CI: 0.57,0.77; $p = 2.5 \times 10^{-7}$), and a positive association with ovarian cancer (OR = 1.35, 95%CI: 1.05,1.72; p = 0.017). Positive associations of genetic scores for adult BMI were also observed for lung cancer (OR = 1.27, 95%CI: 1.09,1.49; $p = 2.9 \times 10^{-3}$) and colorectal cancer (OR = 1.39, 95%CI: 1.06,1.82; p = 0.016). We observed a marginally significant inverse association between WHR and breast cancer (OR = 0.73, 95%CI: 0.53,1.00; p = 0.051). Findings from this study help better understand the relationships between adiposity and cancer risks. Our results for breast and lung cancer are particularly interesting, given previous reports of effect heterogeneity by menopausal status and smoking status.

55

Replication effort for common variants associated with carotid intima media thickness within four independent samples

Marie H. Geisel (1,2) Nicole Heßler (3) Stefan Coassin (4) Susanna Moskau-Hartmann (5) Gudrun Nürnberg (6) Lewin Eisele (2) Frauke Hennig (7) Marcus Bauer (8) Amir-Abbas Mahabadi (8) Susanne Moebus (2) Raimund Erbel (8) Barbara Hoffmann (7,9) Peter Nürnberg (6) Thomas Klockgether (10) Karl-Heinz Jöckel (2) André Scherag (1) Florian Kronenberg (4) Andreas Ziegler (3,11,12)
(1) Clinical Epidemiology, Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany (2) Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University of Duisburg-Essen, Essen, Germany (3) Institute of Medical Biometry and Statistics (IMBS), University of Lübeck, Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Germany (4) Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria (5) Department of Epileptology, University of Bonn, Bonn, Germany (6) Cologne Center of Genomics, University of Cologne, Cologne, Germany (7) IUF-Leibniz Institute for Environmental Medicine, Düsseldorf, Germany (8) West-German Heart and Vascular Center, Department of Cardiology, University Hospital of Essen, Essen, Germany (9) Medical Faculty, Heinrich Heine University of Düsseldorf, Düsseldorf, Germany (10) Department of Neurology, University Hospital Bonn, Bonn, Germany (11) Center for Clinical Trials, University of Lübeck, Lübeck, Germany (12) School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Carotid intima media thickness (cIMT) is used as marker for subclinical atherosclerosis. In a genome-wide association meta-analysis (GWAMA), lead by the CHARGE consortium (Bis et al., 2011), three genomic regions associated with cIMT and one region with suggestive evidence were reported. Here, we aimed at replicating the best four CHARGE cIMT loci within a meta-analysis of four independent studies: the Bonn IMT Family Study, Heinz Nixdorf Recall Study, Salzburg Atherosclerosis Prevention Program in Subjects at High Individual Risk and the Bruneck study. Secondly, we present unbiased estimates of the CHARGE discoveries updating their replication results by our meta-analysis findings. Thirdly, we performed a fine-mapping within a ±50kb

region centered on the four lead SNPs.Our meta-analysis included 5,233 participants, and 15,313 participants when updating the CHARGE replication results. Effect size estimates were comparable to those of the GWAMA, but were formally not replicated ($p > 0.01$). Updating the CHARGE replication results, the three best CHARGE cIMT loci were replicated with updated effect size estimates that were closer to the discovery GWAMA estimates. Regarding the fine-mapping, the association signal was scattered around the original lead SNPs. We observed stronger estimates for alleles at variants other than the four lead SNPs, mostly with a nominal p-value > 0.05 and without known functional implications. Despite the absence of a formal replication and no new information by fine-mapping, our replication effect size estimates indicate that the cIMT loci are very likely true.

56

A functional polymorphism in miRNA-1229 influences the risk of Alzheimer's disease

Mohsen M. Ghanbari (1) Mohammad Arfan M. Ikram (1) Hans H. de Looper (1) Albert A. Hofman (1) Stefan S. Erkeland (1) Oscar O. Franco (1) Abbas A. Dehghan (1)
(1) Erasmus University Medical Center

Genome-wide association studies (GWAS) have enabled us to identify a large number of genetic variants associated with Alzheimer's disease (AD). However, the vast majority of the identified variants are non-genic that their biological relevance to the disease remain to be elucidated. MicroRNAs (miRNAs) serve as key post-transcriptional regulators of gene expression and are involved in various biological processes. Genetic variation in miRNA-related sequences has been shown to interfere with miRNA gene regulation and subsequently affect disease risk. Here, we investigated the extent to which variants fall in miRNAs and miRNA-binding sites could constitute a part of the functional variants associated with AD. Using data from the thus far largest GWAS on late-onset AD, we found that rs2291418 (Chr5;179798324:G > A) within the pre-miR-1229 sequence is associated with an increased risk of AD (p-value = 6.8×10^{-5} and $\beta = 0.18$). In silico analysis showed that rs2291418 affect the processing of pre-miR-1229 and in vitro assays demonstrated that the mutant allele enhance the level of mature miR-1229-3p. Subsequently, we found a number of miR-1229-3p target genes that may mediate the miRNA-effect on AD. Furthermore, we identified 11 variants in the 3'UTR of 10 genes linked with AD that would potentially interfere with miRNA-mediated regulation of the host genes by disrupting, creating or modifying miRNA binding sites. With this approach, we further found two new genes, *DMWD* and *HBEGF*, that are associated with AD. These findings may improve our understanding of the role of miRNAs in the pathophysiology of AD and contribute to better annotation of GWAS findings.

57

Genotyping of Trichomonas vaginalis in symptomatic women in Shahrekord city (southwestern Iran), 2011

Payam Ghasemi-Dehkordi (1) Bahman Khalili (2) Gholamreza Pourshahbazi (2) Hossein Yousofi-Darani (3) Morteza Hashemzadeh-Chaleshtori (1) Abbas Doosti (4)
(1) Cellular and Molecular Research Center, Shahrekord University of Medical Sciences, Shahrekord, Iran (2) Department of Parasitology and Mycology, Faculty of Medicine, Shahrekord University of Medical Sciences, Shahrekord, Iran (3) Department of Mycology and Parasitology, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran (4) Biotechnology Research Center, Islamic Azad University, Shahrekord Branch, Shahrekord, Iran

Trichomonas vaginalis is causative agent of vaginitis in female and urithritis in men. It is primarily transmitted by sexually route. It is known that each geographical area has its own set of Trichomonas vaginalis strain. The aim of this study was to determine the genotyping of Trichomonas vaginalis strains in the Shahrekord city (Chaharmahal Va Bakhtiari province, southwest Iran). A total of 1725 vaginal samples were taken from clinically suspected women for Trichomonas vaginalis infection and 21 specimens were diagnosed as positive by direct smear wet mount and culture repeated passage of the parasite in the modified TYI-S-33 medium. The genomic DNA was extracted from each sample and the nested polymerase chain reaction was applied using specific oligonucleotide primers for actin gene amplification. Finally, the restriction fragment length polymorphism using RsaI, MseI, and HindII restriction enzymes were done on PCR products for genotyping. PCR-RFLP analysis on 21 positive cases (1.22%) were showed that the most frequent genotype was H (8 cases), followed by G (4 cases), E (3 cases), and P (2 cases). N and I genotypes were detected in each 1 case. Also, there were 2 cases mix (E and H) genotype. The findings of the present work showed that 7 different genetic strains in isolated Trichomonas vaginalis from symptomatic women in Shahrekord city. For suggestion it would be better in further studies that the accurate determination of genetic diversity of this parasite will be done in Chaharmahal Va Bakhtiari province and other parts of the country.

58

A Multi-Ethnic Genotyping Array for the Next Generation of Association Studies

Christopher R. Gignoux (1) Genevieve L. Wojcik (1) H. Rich Johnston (2) Christian Fuchsberger (2) Suyash Shringarpure (1) Alicia R. Martin (1) Stephanie Rosse (3) Niha Zubair (3) Daniel Taliun (4) Ryan Welch (4) Carsten Rosenow (5) Jared O'Connell (5) Luana McAuliffe (5) Jay L. Kaufman (5) John Picuri (5) Jane L. Romm (6) Charles Kooperberg (3) Kari E. North (7) Christopher A. Haiman (8) Ruth J. F. Loos (9) Tara C. Matise (10) Noura S. Abdul-Husn (9) Gillian Belbin (9) Hyun M. Kang (4) Goncalo Abecasis (4) Michael Boehnke (4) Zhaohui S. Qin (4) Christopher Carlson (3) Kathleen C.

Barnes (11) Carlos D. Bustamante (1) Eimear E. Kenny (9) The PAGE Study
(1) Department of Genetics, Stanford University, Stanford, CA (2) Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta GA (3) Fred Hutchinson Cancer Research Center, Seattle, WA (4) Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI (5) Illumina Inc., San Diego, CA (6) Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD (7) Carolina Center for Genome Sciences, Chapel Hill, NC (8) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA (9) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY (10) Department of Genetics, Rutgers University, Piscataway, NJ (11) Division of Allergy and Clinical Immunology, Johns Hopkins University, Baltimore, MD

In the past decade, genome-wide association studies (GWAS) have been extraordinarily successful albeit primarily in populations of European descent. To address this disparity, a collaboration between Illumina and the PAGE, CAAPA, and T2D-Genes Consortia developed the 1.9M SNP Multi-Ethnic Genotyping Array (MEGA). MEGA is a single platform to interrogate diverse variation and screen for prior genetic discoveries.
The GWAS backbone leverages whole genomes from the 1000 Genomes Project (TGP) and CAAPA, with additional compatibility from the Illumina HumanCore array. We developed a novel cross-population tag SNP strategy to maximize imputation accuracy across six continental populations, with improved performance from previous generations of arrays. Importantly the performance of the array is high across all continental TGP superpopulations (> 90% accuracy for MAF $\geq$ 1%). We chose rare, functional candidates from > 36,000 multi-ethnic exomes, prioritizing loss-of-function and predicted damaging sites. We also curated variants with domain experts, including boosting coverage in regions of interest (e.g. MHC), over 5,000 ancestry informative markers, uniparental markers, and over 25,000 variants of clinical, prior GWAS, pharmacogenetic, and eQTL importance.
Here we will present data on MEGA genotyping of > 60,000 African-American, Hispanic/Latino, East Asian and Native Hawaiian individuals. A centralized community repository at www.pagestudy.org/mega will house relevant information, including reference data generated from HGDP and other panels. We intend MEGA to be a platform and an analytical resource for researchers interested in large-scale studies of diverse populations from across the globe.

59

How low can you go: cohort-wide 1× whole genome sequencing in a Greek isolate reveals multiple quantitative trait

Arthur Gilly (1) Lorraine Southam (2,3) Rachel Moore (1) Aliki-Eleni Farmaki (4) Jeremy Schwartzentruber (1) Petr Danecek (1) Emmanouil Tsafantakis (5) George Dedoussis (3) Eleftheria Zeggini (1)
(1) The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK (2) The Wellcome Trust Sanger Institute, Wellcome Trust Centre for Human Genetics, Wellcome Genome Campus, Hinxton, UK (3) The Wellcome Trust Centre for Human Genetics, Oxford, UK (4) Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece (5) Anogia Medical Centre, Crete, Greece

Very low-depth sequencing has been proposed as a potential means to increase sample sizes in next-generation association studies. Isolated populations offer power gains in detecting associations in rare and low-frequency variants. Here, we conduct sequence-based association studies in an isolated cohort from Crete, Greece using very low-depth (1×) whole genome sequencing (WGS). After having established a robust pipeline for the analysis of 1× WGS data, we carried out single-point association of 29,616,352 variants with 42 medically-relevant quantitative traits, including haematological, biochemical and anthropometric measurements. We find 57 independent genome-wide significant ($p < 1 \times 10^{-8}$) signals (binomial $p < 1 \times 10^{-80}$) in the HELIC-MANOLIS cohort of 1,239 individuals. Two thirds of these signals arise from variants with frequency <1%; 86% of all associated variants are intronic or outside of genes. Initial analysis indicates a mixture of known and novel loci. For example, we replicate association of the rare R19X variant in *APOC3* with blood triglyceride levels ($\beta = -1.09$, $\sigma = 0.164$, $p = 1.01 \times 10^{-10}$). We find it to contribute to a rare variant burden in *APOC3* ($p = 3.0 \times 10^{-18}$), which remains genome-wide significant after excluding R19X ($p = 6.15 \times 10^{-10}$). This strong burden signal was not recapitulated in an analysis of GWAS data in the same cohort imputed up to a combined reference panel of 4,873 sequenced individuals from the 1000 Genomes Project and UK10K, as well as 250 individuals from HELIC-MANOLIS with WGS at 4× depth. We present one of the first population-based next-generation association studies based on very low depth WGS, and demonstrate the advantages of this approach over a mixed GWAS and imputation study design.

60

Meta-analysis of summary statistics from quantitative trait association studies with unknown sample overlap

Arthur Gilly (1) Ioanna Tachmazidou (1) Eleftheria Zeggini (1)
(1) The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

Current meta-analysis methods for GWAS assume independence of samples included in individual studies. However, association studies increasingly sample from the same populations or cohorts, or use identical control datasets. Individual participant data are often inaccessible, which makes the degree of relatedness or overlap difficult to calculate. This

sample overlap can be very high, especially if the source population is small.

Lin and Sullivan provide a theoretical framework for case-control and quantitative trait meta-analysis that accounts for a known number of overlapping samples. Province and Borecki use tetrachoric correlation to estimate sample relatedness or overlap from summary statistics. However this p-value based method does not account for differences in genetic effect directions, nor does it produce a summary of these effects. We adapt this estimator and integrate it with Lin and Sullivan's inverse-variance based method to provide a meta-analysis of both effect sizes and p-values.

Using simulations based on GWAS genotypes from 10,000 individuals from the UKHLS cohort, we show that this method maintains the type-I error rate under an average 7% at a 5% significance threshold, even with very large sample overlaps, as opposed to a linear increase of 0.12% per 1% sample overlap using an uncorrected meta-analysis. Both the p-value correction and overlap estimation were robust to sample size variation and to MAF filtering of the input dataset. We demonstrate that tetrachoric correlation can estimate sample overlap with 95% accuracy.

We implement our method in a software package that scales to genome-wide sequencing data, and can control for unknown sample relatedness or overlap in meta-analysis of up to 15 studies.

61

Targeted genomic screening in the general adult population

Katrina A. Goddard (1)
(1) Center for Health Research - Kaiser Permanente Northwest, Portland, Oregon, USA

The application of genome or exome technology in healthy adult populations has promise as a predictive tool to guide preventive medicine strategies, but the technology has not yet been rigorously assessed for this purpose. Debate about recent calls for widespread expansion of screening for individual genes in populations has highlighted the numerous knowledge gaps that remain. Several efforts are underway to explore emerging opportunities for this application of genomic medicine, but these efforts are also revealing the challenges that must be addressed and overcome. I will discuss early findings from these exploratory programs, and clarify the implications of population level screening with genomic approaches from a variety of stakeholder perspectives.

62

Whole exome sequencing in high-risk chronic lymphocytic leukemia families: do rare germline variants in somatically altered genes or GWAS genes contribute to susceptibility?

Lynn R. Goldin (1) Melissa Rotunno (1) Mary L. McMaster (1) Meredith Yeager (2) Belynda D. Hicks (2) Laurie Burdette (2) Alisa M. Goldstein (1) Laura Fontaine (3) Margaret A. Tucker (4) Gerald E. Marti (5) Stephen J. Chanock (6) Neil E. Caporaso (1)
(1) Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD (2) Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD (3) Westat Inc., Rockville, MD (4) Human Genetics Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD (5) National Heart Lung and Blood Institute, NIH, Bethesda, MD (6) Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD

Chronic lymphocytic leukemia (CLL) shows strong familial aggregation and co-aggregates with other lymphoid malignancies. A number of genes with small to moderate effects have been identified by GWAS but no high-risk genes have been identified. We used whole exome sequencing to look for an enrichment of rare variants in 23 genes known to be recurrently somatically mutated in CLL as well as 45 genes identified from recent GWAS studies. We sequenced 106 patients/obligate carriers in 37 CLL families, 19 of which had 3–5 patients/ obligate carriers sequenced and 18 with two patients sequenced. We also analyzed 47 individuals in a set of 16 comparable "control" families defined as other cancer families sequenced in our lab where a single segregating mutation had been identified. We used a Nimblegen v.3 exome capture library and paired end sequencing on the Illumina HiSeq2000. Reads were aligned using Novalign. Variant discovery and calling of substitutions, insertions and deletions were performed using standard methods. We identified all rare ($\leq$ 1% frequency in European populations) exonic nonsynonymous or UTR variants in the selected genes shared among all patients and obligate carriers within a family. For the somatic analysis, 12/37 (32%) CLL families had shared variants in one or more of the 23 genes compared to 4/16 (25%) control families (p = ns). Very few rare variants in the 45 GWAS genes were found in either the CLL or the control families. We conclude that CLL families are not enriched for carrying rare variants in either somatically mutated or GWAS genes.

63

Gene discovery obstacles in familial melanoma, a complex disease

Alisa M. Goldstein (1) Mary C. Fraser (1) Xiaohong R. Yang (1) Margaret A. Tucker (1)
(1) Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD

Linkage analysis was critical for the discovery of *CDKN2A*, the first and most common high-risk cutaneous melanoma (CM) susceptibility gene identified with mutations occurring in about 20% of familial CM kindreds. In contrast, other strategies including candidate gene, tumor loss of heterozygosity, and next generation sequencing (NGS) studies were

needed to identify six rarer high-risk CM genes (mutations found in < 1-2% of CM-prone families). To examine gene discovery complexities, we compared 23 families with *CDKN2A* mutations to 5 families with mutations in "rarer" CM genes (2 in *CDK4*, 1 in *POT1*, 2 deletions in *ARF*) from an ongoing familial CM genetics study. Families with mutations in rarer CM genes were significantly more likely to include sporadic patients (4/5) than families with *CDKN2A* mutations (4/23), p = 0.015. In addition, although early age at cancer diagnosis is a hallmark of genetic susceptibility, sporadic patients in two rarer gene mutation families were among the youngest cases in their families. Selection of the youngest melanoma patient (s) for NGS or related studies would not have resulted in disease gene identification in these families. Similarly, linkage analysis was not effective in the rarer gene searches because of the phenotypically indistinguishable sporadic CM patients, the limited number of families with mutations in the rarer genes, and the presence of deletions in *ARF* that distorted the LOD score evaluation. Although based on a single study, the comparisons reveal complications in gene discovery and suggest the need to consider familial genetic complexities, to investigate multiple patients in a family, and to increase sample size (potentially through collaborations) to aid disease gene discovery.

64

Evolutionarily Derived Networks to Inform Disease Pathways

Britney E. Graham (1) Christian Darabos (1) Minjun Huang (1) Louis Muglia (2) Jason H. Moore (1,3) Scott M. Williams (1)
(1) Dartmouth College, Hanover, NH (2) Cincinnati Children's Hospital Medical Center, Cincinnati, OH (3) University of Pennsylvania, Philadelphia, PA

Identifying pathways underlying human phenotypes is critical in understanding disease etiology. However, for complex diseases methodologies rarely explain a majority of risk. Our previously proposed human phenotype networks (HPN) identify connections between diseases based on shared genetics and reveal common processes among unrelated phenotypes. This study integrates HPN with Evolutionary Triangulation (ET), a novel statistical method comparing allele frequencies among three populations and their relationship to phenotype prevalences such that two populations are similar while a third is divergent. To build our network, we used ET to identify SNPs by triangulating for a target disease, while relying on prior association data to map the SNPs to their related phenotypes. We selected YRI, GIH and CEU as sample populations based on melanoma prevalence, high in CEU and low in the others. The ET analysis extracted 733 SNPs mapped to 25 phenotypes linked by 236 edges. Four of the phenotypes identified follow the same distribution as melanoma in these populations. Although melanoma, the target phenotype, is missing from the network, we found tanning and vitamin D, both associated with melanoma. Additionally, we identified 18 traits related to Type 2 diabetes, which also follows the same distribution. In conclusion, applying network analysis to ET datasets can extract subnetworks of traits related to multiple phenotypes that follow the same distributions across populations, demonstrating the utility of combining these two analytical approaches to further expose underlying etiological pathways of complex disease.

65

A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architecture

Brian Greco (1) Allison Hainline (2) Jaron Arbet (3) Kelsey Grinde (4) Alejandra Benitez (5) Nathan Tintle (6)
(1) University of Texas Austin (2) Vanderbilt University (3) University of Minnesota (4) University of Washington (5) University of California, Berkeley (6) Dordt College

The widespread availability of genome sequencing data made possible by way of next-generation technologies has yielded a flood of different gene-based rare variant association tests. Most of these tests have been published because they have superior power for particular genetic architectures. However, for applied researchers it is challenging to know which test to choose in practice when little is known a priori about genetic architecture. Recently, tests have been proposed which combine two particular individual tests (one burden and one variance components) to minimize power loss while improving robustness to a wider range of genetic architecture. In our analysis we propose an expansion of these approaches, yielding a general method that works for combining any number of individual tests. We demonstrate that running multiple different tests on the same dataset and using a Bonferroni correction for multiple testing is never better than combining tests using our general method. We also find that using a test statistic that is highly robust to the inclusion of non-causal variants (Joint-infinity) together with a previously published combined test (SKAT-O) provides improved robustness to a wide range of genetic architecture and should be considered for use in practice. Software for this approach is available. We support the increased use of combined tests in practice– as well as further exploration of novel combined testing approaches using the general framework provided here–to maximize robustness of rare-variant testing strategies against a wide range of genetic architecture.

66

Epistasis associated to inflammatory bowel disease (IBD) in humans

Elena S. Gusareva (1,2) Zhi Wei (3) James A. Traherne (4) Jean-Pierre Hugot (5,6) Isabelle Cleynen (7) Judy H. Cho (8) Hakon Hakonarson (9,10,11) Kristel Van Steen (1,2), the International IBD Genetics Consortium
(1) Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium (2) Bioinformatics and Modeling,

GIGA-R, University of Liege, Belgium (3) Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA (4) Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge, UK (5) Service des maladies digestives et respiratoires de l'enfant, Assistance Publique Hopitaux de Paris, Hopital Robert Debré, Paris, France (6) Inserm et Université Paris Diderot Faculté Xavier Bichat, Paris, France (7) Department of Clinical and Experimental Medicine, TARGID, KU Leuven, Leuven, Belgium (8) Department of Internal Medicine, Section of Digestive Diseases, Yale University, New Haven, CT, USA (9) Center for Applied Genomics, Abramson Research Center, Children's Hospital of Philadelphia, Philadelphia, Pa, USA (10) Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pa, USA (11) Department of Pediatrics, Perelman School of Medicine Philadelphia, University of Pennsylvania, Philadelphia, Pa, USA

Gene-gene interactions underlie biochemical pathways and have been well demonstrated in model organisms. Very few examples exist on replicated epistasis in humans. Here, we performed genome-wide scans to detect epistasis associated to Crohn's disease (CD) and ulcerative colitis (UC). We used extensive data of the IIBDGC consisting of 18,277 and 14,224 CD and UC patients, respectively, and ~34,050 healthy controls from 15 European countries typed on the Immunochip. At first, we removed rare variants at MAF < 0.05 and filtered common variants at linkage disequilibrium (LD) of $r^2 > 0.75$. To limit our results to independent effects, SNPs on chromosome 6 (which contains the HLA locus), were furthermore pruned to ensure an LD of $r^2 < 0.35$. We adjusted the binary traits, CD and UC, for population stratification by regressing out the first 5 principal components in R-3.0.1. The study cohorts were randomly stratified into two subgroups (referred as discovery and replication). We then performed screenings for epistatic interactions with new adjusted trait values in the two subgroups using multidimensional reduction tool MB-MDR with permutation-based (step-down MaxT) multiple testing correction and significance assessment at 0.05. We identified 14 and 6 SNP-pairs associated to CD and UC, respectively, which were concordant between the discovery and replication groups. All SNP-pairs involved concomitant variants located on the same chromosomes (for CD at 1p31.3, 5p13.1, 16q12.1 and for UC at 1p31.3, 6p21.3). A more detailed investigation of these findings, as well as the implementation of different analysis protocols, will further increase our understanding of possible epistatic mechanisms underpinning IBD.

67

Variations in *TPH2* gene are associated with the Metabolic Syndrome and Obesity in the HeartSCORE study

Indrani Halder (1) Victoria D. Causer (1) Steven E. Reis (1)
(1) University of Pittsburgh

Depression and metabolic syndrome (MS) have a bidirectional association but causes underlying this relationship are unknown. The serotonin system affects both outcomes. Tryptophan hydroxylase 2 (*TPH2*) gene encodes a protein that catalyzes the rate limiting step in the serotonin pathway & *TPH2* SNPs predict depression. We examined if *TPH2* variation is associated with MS and its components.

Subjects were 771 European Americans recruited in Pittsburgh; 45–74 years; 64% female, from all Framingham risk strata with data on MS components, demographics and genotypes for 24 *TPH2* SNPs available for all. We pruned SNPs based on HWE, minor allele frequency (> 0.05) and linkage disequilibrium and identified five tag SNPs: rs7963803, rs11179002, rs4760750, rs4760820, and rs12231356 that capture 99% of *TPH2* variation and generated five haplotypes (frequencies > 0.05).

MS was modeled as a latent variable unifying individual syndrome components. In structural equation models adjusted for age and gender two haplotypes, CGCCG (Hap1) and CGCGG (Hap2), were associated with this MS variable ($p = 0.003, 0.013$). Hap1 was also associated with a continuous metabolic risk score (MRS; defined as mean of standardized MS components) adjusted for other appropriate covariates ($p = 0.001$) and with covariate adjusted waist circumference ($p < 0.001$). rs4760820 was associated with MRS (< 0.001) and waist circumference ($p = 1.62 \times 10^{-6}$). In four-SNP moving window haplotypes using all 24 SNPs, all haplotypes containing rs4760820 were associated with MRS ($P = 0.0003$) & waist ($P = 0.0001$).

Our results show that a *TPH2* variant which has previously been associated with depression is also associated with MS primarily through effects on central adiposity.

68

Large-scale phenome-wide scan in twins using electronic health records

Scott J. Hebbring (1) Zhan Ye (1) Jyotishman Pathak (2) John Mayer (1) Yijing Cheng (3) Steven J. Schrodi (1)
(1) Marshfield Clinic Research Foundation (2) Mayo Clinic (3) Mayo Clinic

Challenges in population-based genetic research have resulted in a re-awakening of family-based studies. However, difficulties arise with identifying the most interesting diseases and families for family-based research. Use of large patient populations linked to an electronic health record (EHR) may alleviate such challenges. Using readily available EHR data, we identified 28,888 twins from Marshfield and Mayo Clinic. In the twins, we measured familial aggregation phenome-wide (5,598 phenotypes). Not surprisingly, the top associations in both cohorts included numerous perinatal phenotypes including those related to birth weight (e.g., $P \cong 5.0 \times 10^{-659}$) and jaundice (e.g., $P \cong 7.9 \times 10^{-468}$). It is not unexpected that individuals born together will likely have comparable birth weights and similar infant-related co-morbidities. There were also numerous common phenotypes previously characterized with high heritability (e.g., myopia, $P \cong 2.1 \times 10^{-158}$; and developmental delays, $P \cong 2.2 \times 10^{-188}$) along with numerous rare Mendelian diseases. In total, 1,406 of the 5,598

phenotypes were statistically enriched for concordance ($P < 8.9 \times 10^{-6}$), many with unknown genetic etiologies. With our novel phenome-wide methodologies highly translatable to other EHR systems, and potentially to other familial relationships, this study may pave the way to biotechnologically smart EHR systems that integrate family data to predict, prevent, and treat many diseases for the advancement of "precision medicine." Lastly, this study provides an intriguing perspective for the future of genetic epidemiologic research. Specifically, the future when large patient populations with sequenced genomes are unified by familial relationships in an integrated EHR system.

69

Incorporating between-pedigree co-ancestry in variance-components linkage analysis

James E. Hicks (1) Michael A. Province (1)
(1) Washington University School of Medicine

As rare variants become an increasing focus in genetic epidemiology, family-based study designs will return to the forefront. Tools, such as linkage analysis, used for family data will need to be updated to accommodate this hypothesis.
In conventional linkage analysis pedigrees are assumed to be independent, and only identity-by-descent (IBD) states within each pedigree are used. However, cryptic relatedness is present in populations and haplotypes shared by cryptically related individuals can harbor rare variants that influence phenotypes.
With the development of long range phasing-based methods for detection of shared genomic segments using dense genotypes, IBD states across the genome can be inferred, without use of pedigree information. This is done by identifying long runs of genotypes identical-by-state (IBS) which are unlikely to be identical without being IBD.
This method allows for cryptic relatedness to be incorporated into linkage analysis. In variance-components linkage analysis, IBD states are modeled in a covariance matrix. Conventionally, these states are computed within pedigrees by modeling recombination rates between genotypes in a sparse set of markers. Replacing that covariance matrix with one determined from shared segment methods can increase the accuracy and power of linkage analysis. Power is increased in the scenario when there is a haplotype shared IBD between members of different pedigrees. If there is no between-pedigree IBD, the analysis reduces to conventional variance-components analysis. By determining IBD states by long runs of dense IBS genotypes, linkage signals can be determined from their physical position, allowing more precise localization.

70

r2VIM: A variable selection method for identifying complex genetic models associated with human traits

Emily R. Holzinger (1) James Malley (2) Qing Li (1) Joan E. Bailey-Wilson (1)
(1) National Human Genome Research Institute, National Institutes of Health, Baltimore (2) Center for Information Technology, National Institutes of Health, Bethesda

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models (e.g. multivariable models with non-linear interaction effects), which likely contribute to the heritability of complex human traits. Machine learning methods, such as Random Forests (RF), are an alternative analysis approach that may be more optimal for identifying these effects. One caveat to RF is that there is no standardized method of selecting a set of variables with a low false positive rate (FPR) while retaining adequate power. We have developed a variable selection method called r2VIM. This method incorporates recurrency and variance estimation into RF to guide optimal threshold selection. We assess how this method performs in simulated SNP genotype data with a variety of complex effects (multiple loci with interactions and main effects). Our findings indicate that the optimal selection threshold can identify interactions with adequate detection power while maintaining a low FPR in the selected variable set. For example, the optimal VIM threshold had an average detection power of 0.80 and an average FPR of 0.11 for a model with a two-locus interaction and no main effects. However, the optimal threshold is highly dependent on the simulated genetic model, which is unknown in biological data. To address this, we permute the phenotype and re-run r2VIM to generate a null distribution of VIMs. The results from the permuted data are used to choose a selection threshold in the non-permuted analysis by comparing FPR estimates at different VIM thresholds. Our initial results show that the best balance between FPR and detection power is produced by selecting the VIM threshold with an FPR of close to 0.05 in the permuted data.

71

Large scale genome-wide association study for birth weight identifies 13 novel loci and reveals genetic links with a variety of adult metabolic and anthropometric traits

Momoko Horikoahi (1,2) Felix R. Day (3) John R. B. Perry (3) Jouke-Jan Hottenga (4,5) Ruifang Li-Gao (6) Robin Beaumont (7) Nicole M. Warrington (8,9) Nicholas J. Timpson (10), EGG Consortium
(1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK (2) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK (3) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK (4) Netherlands Twin Register, Department of Biological Psychology, VU University, Amsterdam, the Netherlands (5) EMGO Institute for Health and Care Research, VU University and VU University Medical Center, Amsterdam, the Netherlands (6) Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands (7) Institute of Biomedical and Clinical Science, University of Exeter Medical School, Royal Devon and Exeter Hospital, Exeter, UK (8) Faculty of Medicine and Biomedical Sciences,

The University of Queensland Diamantina Institute, Brisbane, Australia (9) School of Women's and Infants' Health, The University of Western Australia, Perth, Australia (10) MRC Integrative Epidemiology Unit at the University of Bristol, University of Bristol, UK

Lower birth weight (BW) is associated with adult metabolic diseases including type 2 diabetes (T2D) and coronary artery disease (CAD), but the underlying causes are poorly understood. We conducted expanded analyses with genome-wide association studies (GWAS) of 75,924 European (EUR) and 9,960 non-EUR singletons and imputation up to the phase 1 integrated 1000 Genomes Project reference panel. We aimed to (i) discover novel BW loci, and (ii) explore the shared genetic overlap between BW and adult traits.

We combined association statistics between sex-specific BW Z-scores and each of 10.3M SNPs across studies via fixed-effects meta-analysis. We detected 13 novel loci at genome-wide significance ($P < 5 \times 10^{-8}$): near *EPAS1*, *ESR1*, *YKT6*, *CLDN7*, *ZBTB7B*, *HHEX/IDE*, *SREBF2*, *C20orf203*, *FOXA2*, *TBX20*, *CCND1*, *INTS7* and *WNT4-ZBTB40*. The lead SNPs at all loci were common except at *YKT6* (minor allele frequency EUR: 1%, African-American: 0.2%). Several novel BW signals overlapped those associated with adult traits: *HHEX/IDE* in T2D, *FOXA2* in hyperglycemia, *ZBTB7B* in waist-to-hip ratio and *INTS7* in height. Using LD score regression, we estimated the genome-wide genetic correlation (RG) between birth weight and GWAS of other traits and found strong correlations with height (RG = 0.43, $P = 7 \times 10^{-43}$), T2D (RG = –0.35, $P = 4 \times 10^{-8}$), systolic blood pressure (BP, RG = –0.30, $P = 1 \times 10^{-8}$) and, CAD (RG = –0.32, $P = 2 \times 10^{-8}$).

In summary, we extended the number of BW associated loci from 7 to 20, and provided strong genome-wide genetic links between lower BW and higher risk of T2D, high BP and CAD in later life, which partially explain the complex relationships between genetic variation, early growth and adult metabolic disease.

72

Increased power for detection of parent-of-origin effects via the use of haplotype estimation

Richard Howey (1) Chrysovalanto Mamasoula (1) Ana Töpf (1) Ron Nudel (2) Dianne F. Newbury (2) Simon E. Fisher (3) Judith A. Goodship (1) Bernard D. Keavney (1,4) Heather J. Cordell (1)

(1) Newcastle University (2) University of Oxford (3) Max Planck Institute for Psycholinguistics and Radboud University (4) University of Manchester

Parent-of-origin (or imprinting) effects relate to the situation where traits are influenced by the allele inherited from only one parent, with the allele from the other parent having little or no effect. Given SNP genotype data from case/parent trios, the parent-of-origin of each allele in the offspring can often be deduced unambiguously; however this is not true when all three individuals are heterozygous. Most existing methods for investigating parent-of-origin effects operate on a SNP by SNP basis and either perform some sort of "averaging" over the possible parental transmissions or else discard ambiguous trios. If the correct parent-of-origin at a SNP could be determined, this would provide extra information and increase the power to detect effects of imprinting. We propose making use of the surrounding SNP information, via haplotype estimation, to improve estimation of parent-of-origin at a test SNP for case/parent trios, case/mother duos and case/father duos. This extra information is then used in a multinomial modelling approach to estimate parent-of-origin effects at the test SNP. We show through computer simulations that our approach has increased power over previous approaches, particularly when the data consist only of duos. We apply our method to two real data sets and find a decrease in significance of p-values in genomic regions previously thought to possibly harbour imprinting effects, thus weakening the evidence that such effects actually exist in these regions, although some regions remained more significant than expected. Software is available at www.staff.ncl.ac.uk/richard.howey/emim.

73

Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations

Yi-Juan Hu (1) Yun Li (2,3) Paul Auer (4) Dan-Yu Lin (2)

(1) Department of Biostatistics and Bioinformatics, Emory University (2) Department of Biostatistics, University of North Carolina, Chapel Hill (3) Department of Genetics, University of North Carolina, Chapel Hill (4) Joseph J. Zilber School of Public Health, University of Wisconsin, Milwaukee

In the large cohorts that have been used for genome-wide association studies (GWAS), it is prohibitively expensive to sequence all cohort members. A cost-effective strategy is to sequence subjects with extreme values of quantitative traits or those with specific diseases. By imputing the sequencing data from the GWAS data for the cohort members who are not selected for sequencing, one can dramatically increase the number of subjects with information on rare variants. However, ignoring the uncertainties of imputed rare variants in downstream association analysis will inflate the type I error when sequenced subjects are not a random subset of the GWAS subjects. In this article, we provide a valid and efficient approach to combining observed and imputed data on rare variants. We consider all commonly used gene- level association tests, all of which are based on the score statistic for assessing the effects of individual variants on the trait of interest. We show that the score statistic based on the observed genotypes for sequenced subjects and the imputed genotypes for non-sequenced subjects is unbiased. We derive a robust variance estimator that reflects the true variability of the score statistic regardless of the sampling scheme and imputation quality, such that the corresponding association tests always have correct type I error. We demonstrate through extensive simulation studies that the proposed tests are substantially

more powerful than the use of accurately imputed variants only and the use of sequencing data alone. We provide an application to the Women's Health Initiative (WHI). The relevant software is freely available.

74

Phylogenetic analysis of mitochondrial genomes with n-grams

Hsin-Hsiung Huang (1)
(1) University of Central Florida

The alignment-free n-gram based method with the out-of-place measures as the distance has been successfully applied to automatic text or natural languages categorization in real time. However, it is not clear about its performance and the selection of n for comparing genome sequences. In this study, the author proposed a symmetric version of the out-of-place measure and an approach for finding the optimal range of n to construct a phylogenetic tree with the symmetric out-of-place measures. This approach then is applied to four mitochondrial genome sequence datasets. The resulting phylogenetic trees are similar to the standard biological classification. It shows that the proposed method is a very powerful tool for phylogenetic analysis in terms of both classification accuracy and computation efficiency.

75

An ensemble distance measure of Natural Vector and k-mer methods for the phylogenetic analysis of multiple-segmented viruses

Hsin-Hsiung Huang (1)
(1) University of Central Florida

The Natural Vector combined with Hausdorff distance has been successfully applied for classifying and clustering multiple-segmented viruses. Additionally, k-mer methods also yield promising results for global genome comparison. It is not known whether combining these two approaches can lead to more accurate results. The author proposes a combination of the Hausdorff distances of the 5-mer counting vectors and natural vectors which achieves the best classification without cutting off any sample. Using the proposed method to predict the taxonomic labels for the 2,363 NCBI reference viral genomes dataset, the accuracy rates are 96.95%, 94.37%, 99.41% and 93.82% for the Baltimore, family, subfamily, and genus labels, respectively. We further applied the proposed method to 48 isolates of the influenza A H7N9 viruses which have eight complete segments of nucleotide sequences. The resulting single-linkage clustering trees and statistical analysis both indicate that the proposed method can lead to the most reasonable phylogenetic relationships.

76

Identifying rare genetic variants associated with risk and severity of airflow limitation

Victoria E. Jackson (1) Louise V. Wain (1) Ian Sayers (2) Ian P. Hall (2) Martin D. Tobin (1), UK COPD Exome Chip Consortium UK BiLEVE
(1) Departments of Health Sciences and Genetics, Adrian Building, University of Leicester, Leicester, UK (2) Division of Respiratory Medicine, University Hospital of Nottingham, Nottingham, UK

Chronic obstructive pulmonary disease (COPD), a leading cause of disability and mortality, is characterised by fixed airflow limitation. Severity of airflow limitation may further be categorised by percent predicted forced expiratory volume in the first second (FEV1). Cigarette smoking is the most significant risk factor for COPD and airflow limitation; however there is also a genetic component.
We carried out discovery case-control analyses using 3226 cases with airflow limitation (based on spirometry indicative of GOLD stage 2 or worse), and 4784 population based controls, in addition to analyses of percent predicted FEV1 within cases. Samples were genotyped using the llumina Human ExomeBeadChip which is enriched for low frequency, functional exonic variants. We undertook both single SNP and gene-based association analyses.We followed up SNPs and genes of interest ($P < 10^{-5}$) in independent samples from the UK BiLEVE consortium, a study of a subset of 50,000 individuals from UK Biobank. We additionally undertook a meta-analysis of the discovery and UK BiLEVE samples, for a subset of variants genotyped in both populations.
The case-control analyses identified associations with low frequency, nonsynonymous SNPs in *MOCS3* (MAF = 1.1%,) and *IFIT3* (MAF = 0.7%, $P = 7.49 \times 10^{-6}$), and a rare splice variant in *SERPINA12* (MAF = 0.03%,$P = 8.53 \times 10^{-6}$) was associated with percent predicted FEV1 in cases. We shall present these novel loci and highlight some of the challenges of working with exome array data, in particular problems with combining data from multiple sample collections, including batch effects and differential missingness, and choices of appropriate statistical tests.

77

Turning Publicly Available Gene Expression Data into Discoveries Using Gene Set Context Analysis

Zhicheng Z. Ji (1) Steven S. A. Vokes (2) Chi C. V. Dang (3) Hongkai H. Ji (1)
(1) Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health (2) Department of Molecular Biosciences, The University of Texas at Austin (3) Abramson Cancer Center, University of Pennsylvania

Gene Set Context Analysis (GSCA) is an open source software package to help researchers use massive amounts of publicly available gene expression data to make discoveries. Users can interactively visualize and explore gene and gene set activities in 25,000+ consistently normalized human and mouse gene expression samples representing diverse biological contexts (e.g., different cells, tissues and disease types, etc.). By

providing one or multiple genes or gene sets as input and specifying a gene set activity pattern of interest, users can query the expression compendium to systematically identify biological contexts associated with the specified gene set activity pattern. In this way, researchers with new gene sets from their own experiments may discover previously unknown contexts of gene set functions and hence increase the value of their experiments by making more discoveries. GSCA has a graphical user interface (GUI). The GUI makes the analysis convenient and customizable. Analysis results can be conveniently exported as publication quality figures and tables. GSCA is available at https://github.com/zji90/GSCA. This software significantly lowers the bar for biomedical investigators to use PED in their daily research for generating and screening hypotheses, which was previously difficult because of the complexity, heterogeneity and size of the data.

78

Genetic Effect and Association Test for Covariance Heterogeneity in Multiple Trait Comorbidity

Yuan Jiang (1) Yaji Xu (2) Heping Zhang (3)
(1) Oregon State University (2) Food and Drug Administration (3) Yale University

Genes and environmental factors may not only contribute to the prevalence of diseases, but can also create a "ripple effect" on multiple disorders. A concrete example includes the case that the population with a particular gene tend to develop multiple drug abuses simultaneously. To answer this question, we propose a new concept of genetic effects on multiple trait covariance, in contrast to the usual definition of genetic effects on disease prevalence. In addition to the new concept, we develop an association test for the covariance heterogeneity among multiple traits, unlike most existing methods that assume homogeneity of the covariance. Preliminary results show that ignoring the genetic effect on multiple trait covariance can result in loss of power when performing genetic association test for comorbidity. We also provide evidences for the importance of the new genetic effect from investigation of a real data set.

79

A powerful allele based test for rare markers in case-control association studies

Marianne A. Jonker (1) Connie R. Bezzina (2) Michael W. T. Tanck (2)
(1) VU Medical Center, Amsterdam (2) Academic Medical Center, Amsterdam

In a case-control study aimed at localizing disease variants, association between markers and the disease status is often tested by comparing the marker allele frequencies among cases and controls. These marker allele frequencies are expected to be different if the corresponding marker is associated with the disease. It is known that the power of the commonly used allele based test is also based on the marker allele frequency; markers with a low minor allele frequency have less power to be detected if they are associated with the disease, than markers with high minor allele frequency. We propose an allele based test that is more powerful for rare markers in many situations, and may, therefore, be more effective when searching for rare causal variants. The test is applied to data of patients with a first acute ST-elevation myocardial infarct with (cases) or without (controls) ventricular fibrillation and the results are compared with the results obtained with the traditionally used allele-based test. More interesting locations and markers have been found for follow-up study.
In the light of the current interest in detecting association between complex phenotypes and low-frequency variants and localizing causal variants with small minor allele frequencies, the implications are expected to be of relevance.

80

Genome-Wide Association Study of Postmenopausal weight change: The Women's Health Initiative Study

Anne E. Justice (1) Misa Graf (1) Sachiko Yoneyama (1) Marian L. Neuhouser (2) Geetha Chittoor (3) JoAnn E. Manson (1,4) Penny Gordon-Larsen (5) Annie Green Howard (6) Kari E. North (1)
(1) Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC (2) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA (3) Department of Nutrition, and Nutrition Research Institute, University of North Carolina at Chapel Hill, Kannapolis, NC (4) Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA (5) Department of Nutrition, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC (6) Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC

The discovery of genetic variation influencing weight change (WTc) has the potential to identify important pathways for disease prediction and treatment; yet, the bulk of GWAS are conducted in cross-sectional studies. One of the challenges of working with WTc is that it varies across life stages. Our study will address inherent limitations to GWAS of WTc by maintaining focus on WTc in females following menopause. Our analyses include up to 6,367 African American (AA) and 2,942 Hispanic/Latinos (HL) from the Women's Health Initiative (WHI) study who have provided informed consent, have multiple weight measures after menopause (mean # of visits = 5.3), have WTc slopes within 4 SD of the mean, and have been genotyped and imputed to HapMap release 22. Using a linear mixed model approach, individual WTc slopes were calculated based on BLUP (best linear unbiased predictor) divided by years of follow-up. We performed association analyses with WTc slope adjusted for age at visit 1, clinical center and principal components, using an additive genetic model. While no associations reached genome-wide

significant (GWS) ($p < 5 \times 10^{-8}$) for the AA participants, one locus near *KANK1* reached GWS for WTc in HL individuals. To further limit heterogeneity of WTc, we restricted analyses to those participants who gained weight (WTc > 0). Despite a decrease in our sample size (N = 4,413 AA and N = 1,484 HL), we identified one GWS locus near *CNTN4* for AA and one additional locus that reached GWS near *CBLN2* for HL. These results highlight the complex nature of WTc across this life stage and ancestries, and so underscore the importance of GWAS on longitudinal phenotypes that reduce phenotype heterogeneity by focusing on subpopulations of interest, like postmenopausal women.

81

MARV: A novel method and software tool for genome-wide multi-phenotype analysis of rare variants

Marika Kaakinen (1) Reedik Mägi (2) Krista Fischer (2) Marjo-Riitta Järvelin (3) Andrew P. Morris (4) Inga Prokopenko (5)
(1) Imperial College London, London, UK (2) Estonian Genome Center, University of Tartu, Tartu, Estonia (3) Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK (4) Department of Biostatistics, University of Liverpool, Liverpool, UK (5) Department of Genomics of Common Disease, Imperial College London, London, UK

Recently, genome-wide association studies (GWAS) have been expanded to analysis of low-frequency and rare variants (MAF≤5%, both denoted by RVs). Power for variant detection could also be increased by jointly analysing multiple correlated phenotypes. We have developed a method and software for genome-wide Multi-phenotype Analysis of RVs (MARV), combining features from both RV burden tests and multi-phenotype analyses. Specifically, the proportion of RVs at which an individual carries minor alleles within a gene region is modelled on linear combinations of phenotypes in a regression framework. MARV also implements model selection via the Bayesian Information Criterion (BIC). Preliminary simulation work on 1,000 individuals with 10,000 replicates and two continuous phenotypes with a correlation ranging between –0.9 and 0.9 show good control of type I error rate. Power is increased when the genetic effect is in the opposite direction than the correlation between the phenotypes. Application to empirical data, namely to fasting insulin (FI), triglycerides (TG) and waist-to-hip ratio (WHR) from 4,788 individuals from the Northern Finland Birth Cohort 1966, shows the ability of MARV to detect multi-phenotype associations. We identified RV associations, at genome-wide significance ($P < 1.7 \times 10^{-6}$, Bonferroni correction for 30,000 genes) in *APOA5*, a known common and RV locus for TG and lipids, with the best fitting model including TG only ($P = 2.2 \times 10^{-7}$), and in *ZNF259*, a known common variant locus for TG and lipids, with TG and FI providing the best fit ($P_{model} = 3.1 \times 10^{-9}$), and stronger associations than in univariate analyses (PTG = 6.7×10^{-8}; PFI = 0.13). For the first time we show RV associations in *ZNF259* with cardiometabolic traits.

82

Genetic Epidemiology Survey in Pakistan-A Case Report

Shamsa Kanwal (1) Muhammad Shoaib Akhtar (1) Asif Hanif (1) Muhammad Aslamkhan (1)
(1) University of Health Sciences Lahore

In a country like Pakistan where no epidemiology data is available for local population, it is tough to study genetic epidemiology. Our research group, recently, conducted a big survey to collect data of different diseases for genetic epidemiology in districts of Bahawalnagar and Chiniot of Punjab Province. District Bahwalnagar has five tehsils (an administrative subdivision) and 118 union councils. Major castes of district are Arains, Butts, Bhattis, Chotis, Hotis, Joya, Jutts, Kambohs, Pathans, Qureshis, Rajpoots, Syeds and Wattoos. The total targeted population was 739,371 (females with age 15 years and above) and our sample size was 1920. The primary study was to find out the genetic epidemiology of breast cancer and impact of consanguinity on it. District Chiniot has its three tehsils (an administrative subdivision) and 42 union councils. The total targeted population was 1,316,620 and our sample size was 579. Major isonym groups and castes of district are Ansaris, Arains, Balochs, Lalis, Mughals, Pathans, Rajpoots, Rehmanis, Sayals, Sheikh, Syeds and Thaheems. Population is studied primarily for Genetic Epidemiology of colour-blindness and to find out incidence of consanguinity and its impact upon colour-blindness. However, other diseases studies in both districts were Cataract, Diabetes, Eye Trauma, Hypertension, Kallmark's Syndrome, Liver Cirrhosis, Myocardial Infarction, Night Blindness, Parkinson's Syndrome, Tuberculosis and Vision Impairments. Moreover, smoking status and tea consumption among study individuals were also studied. During the field work we faced lot of challenges e.g., cultural gaps, funding issues, travelling and biggest of all the denial of permission to work in female institutions where men are not allowed. Because of very high consanguinity, ca 80%, within various isonym groups/castes, Pakistan provides unique opportunities for genetic epidemiological studies.

83

A Highly Adaptive Test for Gene- or Pathway-Multivariate Trait Association with Application to Neuroimaging Data

Junghi Jkim Kim (1) Wei Wpan Pan (1)
(1) University of Minnesota

Testing for genetic association with multivariate traits has become increasingly important, not only because of its potential to boost statistical power, but also for its direct relevance to some applications. For example, there is accumulating evidence showing that some complex

neurodegenerative and psychiatric diseases like Alzheimer's are due to disrupted brain networks, for which it would be natural to identify genetic variants associated with disrupted brain networks. In spite of its promise, testing for multivariate trait associations is challenging: if not appropriately used, its power can be much lower than testing on each univariate trait separately (with a proper control for multiple testing). Furthermore, differing from most existing methods for single SNP–multiple trait association, we consider gene- and pathway-based association testing for multiple traits, due to well known genetic heterogeneity and small effect sizes of individual SNPs. Because the power of a test critically depends on several unknown factors such as the proportions of associated SNPs, genes and traits among those tested, we propose a highly adaptive test that data-adaptively determines some optimal parameters in the test to yield high power across a wide spectrum of situations. We compare the performance of the new test with several existing tests using both simulated and real data. We apply the proposed test to structural MRI data drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI) project to identify genetic variants associated with the human brain default mode network (DMN).

84

X chromosome-wide analysis identifies DNA methylation sites influenced by cigarette smoking

Daniella Klebaner (1) Qin Hui (1) Jacquelyn Y. Taylor (2) Jack Goldberg (3) Viola Vaccarino (1) Yan V. Sun (1)
(1) Emory University (2) Yale University (3) University of Washington

Cigarette smoking is a major cause of chronic disease worldwide. Smoking may induce cellular and molecular changes including epigenetic modification, with both short-term and long-term modification patterns. Recent epigenome-wide association studies (EWAS) have identified a number of smoking-related DNA methylation (DNAm) sites. However, the X chromosome DNAm sites have been largely overlooked due to a lack of an analytical framework dealing with sex-dimorphic distribution. To identify novel smoking-related DNAm sites on the X chromosome, we examined the sex dimorphism of X chromosomal DNA sites and conducted a sex-specific association study of cigarette smoking using a discovery sample of 140 middle-age twins, and two replication samples of 78 twins, and 327 unrelated individuals including 47,15 and 26 current smokers respectively. The top smoking-related DNAm sites in *BCOR* and *TSC22D3* were significantly associated with current smoking after multiple testing correction. These smoking-associated sites were replicated in the two replication samples with meta-analysis p-values of 2.28×10^{-10}, and 1.56×10^{-8}. The X chromosome harbors hundreds of genes and thousands of epigenetic markers important to cellular and molecular function. DNAm sites on the X chromosome may provide important regulation mechanisms linking the environmental exposures and disease pathophysiology, particularly for sex-specific effects. Existing EWAS of human diseases need to include the X chromosomal sites to complete the epigenomic scan.

85

Do little interactions get lost in dark random forests?

Inke R. König (1) Marvin Wright (1)
(1) Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany

Untangling the genetic background of complex diseases requires identifying interaction effects and genetic variants involved in interactions. Random forests (RF) and other machine learning techniques have repeatedly been heralded to be suitable for this endeavor. Specifically, it is mostly argued that RF variable importance measures (VIM) take interaction effects naturally into account. However, it has been shown that especially in the high-dimensional situation, standard VIM fail to detect interaction effects if the interaction partners do not have strong marginal effects (e.g., Winham et al., 2012). Alternatively, paired VIM may be utilized which have not investigated in simulation experiments regarding their ability to detect interaction effects (Ishwaran, 2007). In addition, previous simulation studies only investigated specific simple interaction settings. In contrast, many interaction scenarios are conceivable in reality including, for example, synergistic, modifying, or redundant interaction effects (Lanktree & Hegele, 2009).
In this project, we aim to fill these gaps by a comprehensive simulation study. We simulate different realistic interaction scenarios as described in the literature (Lanktree & Hegele, 2009; Musani et al. 2007). Utilizing a variety of VIM from RF including those defined as paired importance measures (Ishwaran, 2007), we will investigate whether these identify genetic variants involved in interactions, and whether interaction effects per se can be characterized.

86

A Poisson Regression Approach for Association Mapping of Count Phenotypes

Hemant Kulkarni (1) Saurabh Ghosh (1)
(1) Indian Statistical Institute, Kolkata

Complex disorders are usually governed by quantitative precursors. While there has been extensive development of statistical methods for association mapping of these quantitative endophenotypes, development of such methodologies for discrete phenotypes (e.g., symptom counts) remains a challenging area of current research as standard approaches such as ANOVA or linear regression may not be appropriate to analyze count phenotypes. We propose a novel approach based on a Poisson generalized linear model (PGLM) for association analysis of a count phenotype using data on a random sample from a population as well as on nuclear families. For the population-based design, we consider two tests

for association and compare the type-I errors and the powers of these tests with ANOVA, Kruskal-Wallis (KW) and a regression based on a standard Poisson model. For the family based design, which is restricted to informative trios, we consider a PGLM that models the dependence of the phenotype of the offspring on the alleles transmitted by both parents. We evaluate the performance of four different test procedures based on the alleles transmitted by the two parents. We also compare these tests with the commonly used TDT and FBAT procedures. Our simulations reveal that the standard Poisson regression does not provide appropriate type-I errors in the presence of over dispersion in the model but the scale adjusted PGLM maintains the proper sizes of the tests. The PGLM model also yields more power compared to ANOVA and KW. For the family based design, inclusion of both parents in the model results in additional power but possibly at the expense of being susceptible to population stratification.

87

Unfolding heterogeneity of complex traits has strong potential for advancing GWAS

Alexander M. Kulminski (1) Yury Loika (1) Irina Culminskaya (1) Konstantin G. Arbeev (1) Anatoliy I. Yashin (1)
(1) Duke University

Genome-wide association studies (GWAS) have been invented to accelerate the progress in improving healthspan. It is increasingly recognized, however, that the traditional GWAS strategy faces serious difficulties. These difficulties raise concerns that GWAS have exhausted their potential, particularly for complex traits. To advance the progress, currently prevailing GWAS heavily rely on increasing sample size. An inherent property of complex traits is their heterogeneity. Accordingly, just increasing the sample size without unfolding the inherent heterogeneity of these traits may be inefficient. To examine the efficiency of the traditional GWAS strategy on increasing sample size in case of inherently heterogeneous traits, we selected 36 loci which were associated with body mass index in two recent large-scale GWAS meta-analyses (PMC3014648 and PMC4382211). The efficiency of these meta-analyses was characterized by gaining p-value per individual in each sample, i.e., p/N. The results show that the efficiency of the analyses in the larger study (PMC4382211, N = 320K) was better than in the smaller study (PMC3014648, N = 240K) by at least 20% for two loci only (5.6%). For 21 loci (58%), the efficiency was better by at least 20% in the smaller study. For the remaining loci the efficiency was about the same in each study. The results are qualitatively the same when we compare subsamples in each study. These results imply that increasing the sample size relying only on genomic approaches in handling heterogeneity across cohorts in meta-analyses substantially underuses GWAS potential. Unfolding an inherent heterogeneity of complex traits will substantially benefit GWAS.

88

Computationally Efficient Solutions for Functionalizing Common Variants in Three-Dimensional Models

Caleb A. Lareau (1,2) Colby F. DeWeese (1,3) Bill C. White (3) Brett A. McKinney (3) Courtney G. Montgomery (1)
(1) Oklahoma Medical Research Foundation (2) Harvard University (3) University of Tulsa

A significant challenge of the post-genome wide association study (GWAS) era has been the functionalization of associated variants with complex phenotypes. While expression quantitative trait loci (eQTL) have elucidated the effects of some variants, a majority of associations lack functional characterization. Here, we discuss the implementation of two software tools, epiQTL and dcVar, which have been computationally optimized to enable the functionalization of variants in three-dimensional models. While epiQTL efficiently computes the effect of variant epistasis leading to differential expression of quantitative traits, dcVar is designed to uncover variants that generate or destroy correlation structures between pairs of expression probes. These tools have been successfully applied to a variety of eQTL datasets and demonstrated functional effects in complex traits, including variance in human height, Alzheimer's disease risk, and variable response to smallpox vaccine. Our publically available tools can be readily applied to a variety of datasets to supplement existing research efforts in uncovering functional effects of variants and elucidating underlying disease mechanisms.

89

Novel Application of Beta-binomial Models to Assess X Chromosome Inactivation Patterns in RNA-Seq Expression of Ovarian Tumors

Nicholas B.Larson (1) Stacey Winham (1) Zach Fogarty (1) Melissa Larson (1) Brooke Fridley (2) Ellen L. Goode (1)
(1) Mayo Clinic (2) University of Kansas Medical Center

In females, X-chromosome inactivation (XCI) epigenetically silences transcription of one of the two homologous X chromosomes to achieve similar expression levels to males. Some genes are known to escape XCI under normal conditions, and aberrant XCI patterns may occur in female-specific cancers, such as ovarian cancer (OVCA). Which homolog is silenced in a given cell is randomly selected in early development and transmitted mitotically, and tissues that are skewed toward a specific homolog can inform the XCI status of individual genes. We conducted a two-stage analysis to estimate XCI in 453 X genes in 114 OVCA tumor samples using allele-specific expression read counts derived from genome-wide SNP and RNA-Seq expression data. We first applied a composite likelihood-ratio test using a single parameter beta-binomial model, identifying 89 skewed XCI samples to use for gene-level XCI evaluation. We then assessed genic XCI via a two-component beta-binomial mixture model fit using a Bayesian MCMC approach, accommodating extra-binomial

variation and sample-specific skewness. Posterior samples of XCI mixture component variables were used to estimate the posterior probability of XCI escape per gene. Overall, our estimates of genic escapee patterns conformed well to previous LCL studies. However, a mean 5.4% of genes per sample thought to escape silencing showed evidence of XCI, while 8.5% indicated the opposite pattern. Moreover, 22% of genes demonstrated heterogeneity of escape status across samples. These results may indicate inter-tissue XCI differences or cancer-related aberrant XCI, and further research on paired tumor-normal tissues is necessary to evaluate somatic XCI alteration in cancer.

90

A practical guide to study design, sample size requirements and statistical analysis methods for rare variant complex trait association studies

Suzanne M. Leal (1) Gao T. Wang (1) Di Zhang (1) Zongxiao He (1) Hang Dai (1) Biao Li (1)
(1) Baylor College of Medicine

We evaluated rare variant association (RVA) study designs and methods using data from NHLBI-Exome Sequencing Project (ESP) as well as exome data that was simulated using demographic models and the empirical distribution of functional variants in ESP. Our simulated data is highly consistent with real world data distribution of singleton, and doubleton variants as well as the variant cumulative minor allele frequency (MAF). Using resampled genotypes from ESP, we evaluated relative power of 10 RVA methods by analyzing 16,568 genes across the genome, and we demonstrated that a method most powerful for one gene is not necessarily the most powerful for another, simply due to differences in the genomic sequence context, rather than phenotypic model assumptions. Using simulated data of European samples we evaluated impact of phenotypic model, missing data, non-causal variants and choice of empirical MAF cutoff in RVA analysis. We found that the assumption of variable effects model favors variable threshold tests (e.g. VT) greatly, but the power gain of weighted burden tests (e.g. WSS) are marginal compare to the constant effect. SKAT perform poorly when most variants increase disease risk and there are few protective variants. The impact of non-causal variants and missing data are more significant than the choice of RVA methods, and the enrichment of functional variants is most crucial to the success of most RVA methods. Sample sizes are highly dependent on gene size, for example for odds ratio 2.0, for genes with short coding region lengths (~400bp), > 90,000 samples are required to achieve a power of 80% to detect an association for an exome significant level of $\alpha = 2.5 \times 10^{-6}$ while for average sized genes (~1,400bp) the required sample size is > 50,000.

91

Statistical analysis of RNA-seq data at scale

Jeff T. Leek (1)
(1) Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA

RNA-seq is now the primary technology used to measure transcriptional abundance. The analysis of RNA-seq data can be done at multiple levels (genes, regions, or transcripts) and at multiple scales (small experiments or large population cohorts). I will discuss statistical challenges in developing and applying software for the analysis of RNA-seq data at multiple scales including reproducibility, statistical power, trust in genomic annotations, and detection and removal of artifacts. These issues are critical in the analysis of data from genomic experiments in general, but are particularly acute in the analysis of dynamic data from transcriptomes.

92

Modified Random Forest Algorithm to Identify gene-gene Interaction in Case-Parent Trios Studies of Oral Cleft

Qing Li (1) Emily Holzinger (2) Jacqueline B. Hetmanski (3) Mary L. Marazita (3) Terri H. Beaty (3) Joan E. Bailey-Wilson (2)
(1) Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA (2) Computational and Statistical Genomics Branch (NHGRI/NIH) 333 Cassell Dr., Suite 1200, Baltimore, MD, USA (3) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Non-syndromic cleft lip with or without cleft palate (CL/P) is a common birth defect with strong genetic components. Genome-wide association studies (GWAS) for CL/P have identified multiple genes as influencing risk of CL/P. Previously, we have modified the random forest (RF) algorithm to identify gene-gene interaction in trios. Our modified RF uses a sampling scheme with a regression-based splitting criterion to account for the relationship among the cases and 'pseudo controls' (genotypic combinations possible given the parental genotypes). In this project, to further explore possible gene-gene interactions, we applied our method to case-parent trio data from a previous GWAS study of CL/P with multiple ethnic/racial groups from an international consortium. To control for the randomness in the search algorithm, we employed the 'recurrency variable importance metric' to finalize the list of top SNPs possibly involved in gene-gene interactions. We have applied our method to a subset of ~400 markers, including those belonging to WNT pathway and those with large marginal effects. Our method identified markers with large marginal effects, (in genes *IRF6* and *C1orf107* among the Asian trios; and in *hcG_1814486* and *UNC5C* among the European trios), as the top SNPs possibly involved in gene-gene interactions. We did not find any evidence of gene-gene interaction within the WNT pathway. We are currently applying this method to the entire genome-wide marker panel using these GWAS data, since this method has the potential to identify possible gene-gene interactions among markers without large marginal effects. The results

will be compared with the results from another ensemble method, trioFS, which is part of the R package, trio.

93

The variation of DNA methylation at vast the majority of CpG sites are due to individual factors but not genetic or shared environment factors

Shuai Li (1) E. E. Ming Wong (2) Carmel Apicella (1) Jennifer Stone (3) Gillian S. Dite (1) Jihoon E. Joo (2) Graham G. Giles (4) Melissa C. Southey (2) John L. Hopper (1)
(1) Centre for Epidemiology and Biostatistics, The University of Melbourne, Victoria 3010, Australia (2) Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Victoria 3010, Australia (3) Centre for Genetic Origins of Health and Disease, The University of Western Australia, Western Australia 6009, Australia (4) Cancer Epidemiology Centre, Cancer Council Victoria, Victoria 3004, Australia

To study the causes of the variation of DNA methylation, we used the Illumina HumanMethylation450 BeadChip and measured DNA methylation at 468,406 CpG sites on autosomal chromosomes from dried blood spots for 479 Australian women (mean age 56 years) comprising 66 MZ and 66 DZ twin pairs, and 215 of their sisters (130 families). For each site, we applied multivariate normal pedigree analysis to fit variance components models, with A referring to additive genetic factors, Ct to environmental factors shared equally by twins alone, Cs to environmental factors shared equally by siblings (including twin-sister pairs), and E to individual factors. Reduced and nested models were fitted and compared using the likelihood ratio test. The significance of the deleted component was assessed at $P < 0.017$. The best fitted model was decided by only including significant components and E component.
The number of sites (proportion in parentheses) whose best fitted model was ACsCtE, ACtE, CsCtE, AE, CtE, CsE and E was: 1 (0%), 2,920 (0.6%), 23 (0%), 8,831 (1.9%), 20,395 (4.4%), 21 (0%), and 436,215 (93.1%), respectively. In total, 11,752 (2.5%) sites included the A component (mean estimated heritability 62%) and 23,339 (5.0%) sites included the Ct component (mean estimated Ct 30%). For the 86,305 sites with known SNPs, these percentages were 4.8% and 6.5%, respectively; while for those sites without known SNPs, they were 2.0% and 4.7%, respectively.
We conclude that variation in DNA methylation at the vast majority of CpG sites interrogated by the HM450K is best explained by individual factors alone, even for sites containing known SNPs; and non-genetic factors shared by twins alone explain variation at more sites than genetic factors.

94

FastPop: a rapid principle component derived method to infer intercontinental ancestry using genetic data

Yafang Li (1) Jinyoung Byun (1) Guoshuai Cai (1) Xiangjun Xiao (1) Joe Dennis (2) Douglas Easton (2) Ivan Gorlov(1) Michael Seldin (3) Christopher I. Amos (1)
(1) Dartmouth College (2) Cambridge University (3) UC Davis

PCA has become a standard procedure in population genetics study for substructure analysis. The eigenvectors from PCA are easy to use for population adjustment in GWA studies. However it lacks the ability to provide clear information for ancestral origin, and usually does not yield an approach that can be generalized from study to study. To fill this gap, we developed FastPop, an efficient R package for inference of ancestry with PCA scores as the input. FastPop will first compute scores for individuals based on eigenvectors from PCA analysis, and then estimate the proportional ancestry of each individual based on the scores. We demonstrate the use of our software using markers that are present on the Core Content of Illumina products. We selected 2,318 SNPs across the whole genome based on having a large fixation index value among European, African and Asian populations for PCA analysis. We conducted PCA analysis of 505 Hapmap samples with European, African or Asian ancestry along with a collection of 19,661 additional samples of unknown ancestry. We also ran Structure program on the same dataset to benchmark the performance of FastPop. The results from FastPop are highly consistent with that from structure across the 19,661 samples. The correlations of the results between FastPop and structure are 0.99, 0.97 and 0.99 for European, African and Asian, respectively. Compared with Structure, FastPop is more efficient as it can finish ancestry inference for 19,661 samples in 16 minutes compared with 21–24 hours required by structure. FastPop can also provide PCA scores based on SNP weights so the scores of reference population such as Hapmap samples can be applied to other studies providing same set of markers are used.

95

Comparison of Heritability Estimation and Linkage Analysis for Multiple Traits Using PC Approaches

Jingjing J. Liang (1) Tao T. Feng (1) Brian B. Cade (2) Xihong X. Lin (3) Richa R. Saxena (4) Kevin K. Gleason (2) Xiaofeng X. Zhu (2) Susan S. Redline (2)
(1) Department of Epidemiology and Biostatistics, Case Western Reserve University (2) Division of Sleep Medicine, Harvard Medical School, Harvard University (3) Department of Biostatistics, School of Public Health, Harvard University (4) Center for Human Genetics Research, Harvard Medical School, Harvard University

A disease trait is often measured by multiple phenotype measurements based on the availability of the phenotype in large samples, measurement reliability, or biological relevance. Given that multiple phenotypes may be correlated and reflect common underlying genetic mechanisms, the use of multivariate analysis of multiple traits may improve statistical power to detect genes and variants underlying complex traits.

The literature, however, has been unclear as to the optimal approach for analyzing multiple correlated traits. In this study, heritability and linkage analysis were conducted for multiple Obstructive Sleep Apnea Hypopneas Syndrome (OSAHS) related phenotypes using data from Cleveland Family Study that comprises 1,301 African and European American individuals. Two principal component analysis (PCA) approaches were used for combining six OSAHS related phenotypes: the apnea hypopnea index, three sleep study metrics (average hypopnea duration, average oxygen saturation, and percent time at oxygen saturation of < 90%) and two symptoms (habitual snoring and excessive sleepiness). The two PCA approaches are the traditional PCA by maximizing the variance and the method by maximizing the heritability (Rabinowitz and Ott, 1999). Our study demonstrates that PCs generally result higher heritability and linkage evidence than individual traits. Furthermore, the PCs obtained by maximizing heritability instead of variance can be transferred across populations, strongly suggesting the common underlying genetic mechanisms for OSAHS across populations. Thus, PCs of maximizing heritability can be better traits for trans-ethnic population analysis in genetic studies.

96

A New Method for Joint Analysis of Multiple Traits in Association Studies

Xiaoyu Liang (1) Qiuying Sha (1) Shuanglin Zhang (1)
(1) Michigan Technological University

Currently, the analysis of most of genome-wide association studies (GWAS) have been performed on a single trait. As evidenced in neuroimaging genetics and other studies, a complex disease may exhibit its occurrence or progression in several syndromes. Therefore, using only one single trait may lose statistical power to identify the underlying genetic mechanism. There is an increasing need to develop and apply powerful statistical tests to detect association between multiple traits and the genetic variant. In this paper we have developed a new method for joint analysis of multiple traits in association studies. The newly proposed method combines p-values obtained in standard univariate GWAS to obtain one trait-based p-value. It allows researchers to test their genetic associations using standard GWAS software. In addition, traits of different types (e.g., dichotomous, ordinal, continuous) can easily be analyzed simultaneously. We performed extensive simulations to evaluate the performance of the proposed method and compare the power of our method with the powers of TATES, SUM-SCORE based on univariate score test statistics and the standard MANOVA. Our simulation studies showed that the proposed method has correct type I error rates and is either the most powerful test or comparable with the most powerful tests.

97

Statistical Analysis of Massive Genetic and Genomic Data in Genetic Epidemiology

Xihong Lin (1)
(1) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

The human genome project in conjunction with the rapid advance of high throughput technology has transformed the landscape of health science research. The genetic and genomic era provides an unprecedented promise of understanding genetic underpinnings of complex diseases or traits, studying gene-environment interactions, predicting disease risk, and improving prevention and intervention, and advancing precision medicine. A large number of genome-wide association studies conducted in the last ten years have identified over 1,000 common genetic variants that are associated with many complex diseases and traits. Massive targeted, whole exome and whole genome sequencing data as well as different types of omics data have become rapidly available in the last few years. These massive genetic and genomic data present many exciting opportunities as well as challenges in data analysis and result interpretation. They also call for more interdisciplinary knowledge and research, e.g., in statistics, machine learning, data curation, molecular biology, genetic epidemiology and clinical science. In this talk, I will discuss analysis strategies for some of these challenges, including rare variant analysis of whole-genome sequencing association studies; analysis of multiple phenotypes (pleiotropy), and integrative analysis of different types of genetic and genomic data.

98

Network-based analysis of genome-wide association data identifies a gene sub-network underlying childhood-onset asthma

Yuanlong Liu (1) Myriam Brossard (1) Chloé Sarnowski (1,4) Patricia Margaritte-Jeannin (1,4,3) Felipe Llinares-Lópe (2) Amaury Vaysse (1) Marie-Hélène Dizier (1,3) Emmanuelle Bouzigon (1) Florence Demenais (1)
(1) INSERM UMR-946, Paris, France (2) ETH Zürich, Basel, Switzerland (3) Université Paris Diderot, Paris, France (4) Université Paris Sud, Paris, France

Genome-wide association studies (GWASs) have identified 21 loci associated with asthma. However, these loci account for a small part of asthma susceptibility. To identify novel asthma genes, we performed a network-based analysis that integrates information of the Human Protein Interaction Network (HPIN) and GWAS data. We used two datasets from the GABRIEL Asthma Consortium that consisted of outcomes of meta-analyses of 9 childhood asthma GWASs each (3,031 cases/2,893 controls and 2,679 cases/3,364 controls, respectively). GWAS signals were overlaid to HPIN by assigning SNPs to genes and using gene-wise P-values obtained through circular genomic permutations (CGP). Modules enriched with asthma-associated genes were generated by a dense module search strategy.
We identified 10 module pairs that had high similarity between the two datasets. By merging these modules within each dataset and intersecting the two gene lists, we identified a sub-network of 91 genes. This sub-network was significantly

associated with childhood asthma ($P < 10^{-4}$ using 10,000 CGP). Among the sub-network genes, 14 were reported associated with asthma by previous GWASs and 22 with nominally significant gene-wise P-values were novel candidates. Three KEGG immune-related pathways were found significantly enriched in these genes. Moreover, the number of connections (14) among the known and novel candidate genes was significantly higher than expected by chance ($P = 3 \times 10^{-4}$). These results show the benefit of integrating GWAS data and HPIN to identify novel functionally related genes underlying childhood asthma. Funding: FP7-316861, ANR11BSV1-027, ANR-USPC2013.

99

Comparison of Performance of Genotype Imputation: Population-based Imputation and Family-based Imputation

Ching-Ti Liu (1) Xuan Deng (1) Virginia Fisher (1) Yanhua Zhou (1) L. Adrienne Cupples (1)
(1) Boston University

The era of genome-wide association studies with population-based imputation (PBI) has been successful in identifying common variants associated with complex diseases; however, much genetic variation remains to be explained and less frequent variants (LFV) may contribute. To identify LFV, a population study of unrelated individuals is no longer as an efficient study as a family study, where a variant rare in the population may be frequent in a family. Thus, family-based imputation (FBI) may provide an opportunity to evaluate LFV and their effects on traits of interest. To compare the performance of PBI and FBI, we conducted extensive simulations under several scenarios on the availability of genotypes. We simulated 30,000 variants using SeqSIMLA for 1000 three-generation families of size 12 as true sequence data. We masked genotype information for variants unavailable in Framingham 550K GWA genotype data for subjects who are less informative for imputation based on ExomePicks. We implement IMPUTE2 with duoHMM in SHAPEIT (IMPUTE) for PBI, and MERLIN as well as GIGI along with MORGAN for FBI. We used correlation coefficients (r^2) and imputation quality scores (IQS) to assess imputation quality. All approaches performed well with $r^2 > 0.99$ or IQS > 0.98. For common variants, IMPUTE outperforms others. For variants with MAF < 5%, GIGI and IMPUTE are superior and compatible in general. Our findings indicated that PBI provides better quality in general while FBI also offers compatible imputation quality for less frequent variants. The development of genotype imputation, incorporating inheritance pattern within families and linkage disequilibrium (LD) information from large available genotype resources, may yield more gains than current imputation approaches.

100

The Genetics of Obesity - Going beyond common variation and common traits

Ruth J. F. Loos (1,2,3)
(1) The Charles Bronfman Institute for Personalized Medicine, The Mindich Child health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA (2) The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, New York, USA (3) The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

Large-scale genome-wide association studies (GWAS) have identified > 150 loci associated with adiposity traits, predominantly in European, but also in Asian and African ancestry populations. While the explained variance remains small, each of the loci may harbor genes that are involved in pathways relevant to obesity. However, for most loci the causal gene (s) or variant (s) remain to be determined. Pinpointing the causal gene/variant, however, is a critical, yet challenging, step towards translation of GWAS-discoveries into functional follow-up experiments, which are needed to elucidate the mechanisms that regulate body weight and energy homeostasis. The reasons why translation of GWAS-identified loci has been challenging are, at least in part, because phenotypes most often studied are heterogeneous and genotypes are common.

So far, most large-scale GWAS meta-analyses have focused on body mass index (BMI), as a proxy of overall obesity, and waist-to-hip ratio (WHR), as a proxy for abdominal obesity. However, both BMI and WHR are heterogeneous phenotypes; for example, body fat percentage can differ widely across individuals with the same BMI. In addition, these GWAS have so far focused on common variants and BMI/WHR-associated loci are typically predicted to be non-functional, as they often lie in non-coding regions of genes or even in-between genes. Thus, while large-scale GWAS meta-analyses for BMI and WHR have been extremely successful in identifying new loci, heterogeneous outcomes and non-functional variants have hampered the translation of these GWAS discoveries.

I will review how studying more homogenous adiposity outcomes and/or functional variants might facilitate translation. For example, in a recent GWAS for body fat percentage, a more accurate estimate of adiposity, we integrated association data of a wide range of cardiometabolic traits, revealing fascinating association patterns with estimates of growth and maturation, and also (paradoxically) with favorable lipid and glycemic profiles. The identified loci for body fat percentage only partially overlapped with those for BMI, and their cross-phenotype association signature provided informative insights that point towards the potential candidate genes in the loci. A GWAS of circulating leptin levels, a biomarker of adiposity, revealed six loci, one of which near *LEP*. To locate the causal gene in the other five loci, we developed a knockdown transplant strategy in adipose tissue of mice, revealing potential new genes that regulate circulating leptin levels.

Most recently, the GIANT (Genetic Investigation of ANthropometric Traits) consortium has started the meta-analyses of data from > 400,000 individuals genotyped using the

ExomeChip, which contains around 240,000, predominantly low-frequency and rare, coding variants. Preliminary analyses confirmed mutations in the well-known *MC4R* gene, and have identified functional variants in genes involved in thyroid disease and food metabolism.
Taken together, while GWAS of common outcomes have been successful in identifying many loci, translation of these observations has been difficult. Studying more accurate measures of adiposity and potentially coding variants promises to result in loci that are easier to interpret. Such loci might provide the insights needed to carefully design experimental follow-up studies that will help elucidate the pathways involved.

101

Integrating multidimensional omics data for cancer outcomes

Shuangge Ma (1)
(1) Yale University

In cancer research, multidimensional studies are gaining popularity. In such studies, multiple types of (epi)genetic measurements are collected on the same subjects. It remains a challenging problem how to identify cancer markers and link multidimensional (epi)genetic measurements with cancer outcomes. In our study, the goal is to identify cancer markers from one type of (epi)genetic measurement, especially gene expression, and build a cancer outcome model, with the assistance of information from other types of (epi)genetic measurements (for example, multiple regulators of gene expressions). The proposed approach is realized in two steps. In the first step, we develop the "regulatory modules", which have been motivated by gene co-expression analysis and network analysis. These modules are constructed using regularized estimation and describe the regulation relationships among different types of (epi)genetic measurements. We are then able to make a proper decomposition of gene expression based on the regulation mechanisms. In the second step, a regularization approach is applied for identifying markers and building a cancer outcome model. Compared to the existing alternatives, the proposed method can better accommodate the complex regulations among different types of measurements and extract more information on cancer outcomes. The proposed method is applied to analyze the TCGA (The Cancer Genome Atlas) data on the prognosis of cutaneous melanoma. The constructed regulatory modules contain hallmarks of cancer as well as new discoveries. The prognostic model has superior prediction performance over the competing alternatives.

102

Trans-ethnic meta-analysis reveals novel loci and effector genes for kidney function in diverse populations

Anubha Mahajan (1) Jeffrey Haessler (2) Yukinori Okada (3) Adrienne Stilp (4) John Whitfield (5) Cathy Laurie (4) Nora Franceschini (6) Andrew P. Morris (1,7)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK (2) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, USA (3) Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan (4) Department of Biostatistics, University of Washington, Seattle, WA, USA (5) QIMR Berghofer Medical Research Institute, Brisbane, Australia (6) University of North Carolina, Chapel Hill, NC, USA (7) Department of Biostatistics, University of Liverpool, Liverpool, UK

Genome-wide association studies (GWAS) have identified several loci for reduced estimated glomerular filtration rate (eGFR), a measure of kidney function used to define chronic kidney disease (CKD). However, these loci map to large genomic intervals, limiting progress in identifying causal transcripts and understanding pathogenesis of CKD. We performed trans-ethnic meta-analysis to: (i) discover novel eGFR loci; and (ii) fine-map eGFR loci by leveraging differences in linkage disequilibrium between diverse populations.
We considered 9 GWAS (71,638 individuals of European, African American, Hispanic, and East Asian ancestry), each imputed up to the 1000 Genomes reference panel. We identified 20 loci at genome-wide significance ($p < 5.0 \times 10^{-8}$), including two not reported previously: LRP2 ($p = 5.6 \times 10^{-10}$) and NFATC1 ($p = 1.3 \times 10^{-8}$).
We constructed "credible sets" at each locus and resolved fine-mapping to less than 10 variants at 11 loci. At two loci, the credible set included coding variants: *GCKR P446L* and *CPS1 N1412T*. At the remaining 9 loci, the credible set mapped only to non-coding sequence, suggesting these association signals impact through eGFR regulatory mechanisms. At the *RGS14-SLC34A1* locus, the two variants in the credible set overlap promoter histone marks and DNAse hypersensitivity sites. The lead SNP is an eQTL for *RGS14*, highlighting this gene as the likely effector transcript at this locus.
Our findings provide evidence that trans-ethnic GWAS are useful for discovery of novel eGFR loci and for prioritisation of potential causal variants that can be taken forward for experimental validation, enhancing our understanding of the pathophysiology of kidney function.

103

Inverse regression of genotype on phenotypes versus ASSET: competing strategies for pleiotropy analysis

Arunabha Majumdar (1) Tanushree Haldar (1) John Witte (1)
(1) University of California, San Francisco

Pleiotropy analysis comprises two main components: testing for an overall association of a genetic variant with multiple phenotypes and exploring the optimal subset of non-null traits that trigger a genome-wide (GW) signal of pleiotropic association. However, most of the methods for assessing pleiotropy are designed for only testing the global association. Inverse regression of genotype on phenotypes (MultiPhen; O'Reilly et al., 2012), is a flexible framework that

allows for assessing overall pleiotropy and which traits are underlying any observed signals. For a SNP associated with multiple phenotypes, we evaluate how to best determine an optimal subset of non-null traits using different model selection criteria (AIC, BIC, adaptive LASSO, etc.) in the inverse regression framework. We compare this approach to ASSET (Bhattacharya et al., 2012), which uses a subset-based meta-analysis to provide an optimal subset of non-null traits in addition to the p-value of global association. In particular, we use simulations to compare these two approaches with respect to the power of detecting association and accuracy of the selection of non-null traits that is measured in terms of specificity (discarding true null traits) and sensitivity (including true non-null traits). We observed that MultiPhen is more powerful than ASSET when a subset of traits is associated with a SNP and ASSET is more powerful when all the traits are associated. With respect to the selection, for four phenotypes, inverse regression coupled with adaptive LASSO performs consistently better than ASSET, whereas for eight phenotypes, ASSET performs better than inverse regression when trait correlation is high. Both of the approaches are applied to a large cohort data to give an empirical example of their performance.

104

Prognostic models for melanoma using integrated clinical and genomic data

Ernest Mangantig (1) Mark M. Iles (1) Jérémie Nsengimana (1) Jon Laye (1) Julia A. Newton Bishop (1) Timothy D. Bishop (1) Jennifer H. Barrett (1)
(1) Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK

Gene expression levels within the tumour are associated with survival in melanoma patients, and there is also preliminary evidence that the patient's genotype may influence survival. Combining clinical data with gene expression data may improve prediction but no studies so far have analysed the combined effects of the three different types of factor on melanoma-specific survival (MSS). This study aims to compare methods of combining clinical and genomic data to build prognostic models for melanoma. Clinical and genome-wide genetic data were available from a cohort of ~2,000 melanoma patients from Leeds, UK, with genome-wide gene expression data available on a subset of 204 participants (mRNA extracted from formalin-fixed primary tumours). Two approaches were used to combine the data. In the first (agnostic) method, Lasso penalized Cox regression was applied separately to genotype and gene expression data to select SNPs and gene expression levels related to MSS. A polygenic risk score was created from the set of 20 selected SNPs, and this was included in a multivariable Cox regression of MSS on 15 gene expression levels, with and without clinical predictors. The second approach is to combine in a single model polygenic scores for intermediate traits, such as gene expression levels, telomere length and pigmentation. For each gene whose expression is related to MSS, a genome-wide expression-quantitative-trait-loci analysis was conducted to identify SNPs from which to derive a polygenic score, while the score for the other traits is based on previously published risk estimates. The performance of the different models will be compared to identify the best approaches to integrating different types of data to build survival models.

105

Results from a genome-wide association study of red-blood cell fatty acids in the Framingham Heart Study

Katie McKenzie (1) Kristin Koch (2) Jacob O'Bott (3) Jason Westra (4) Nathan Tintle (4)
(1) Duke University (2) Baylor University (3) University of Maryland, Baltimore (4) Dordt College

Most genome-wide association studies interested in fatty acids have explored relationships between genetic variants and plasma phospholipid fatty acid proportions, but few have utilized membrane fatty acid profile of red blood cells (RBC) and even fewer have accounted for lifestyle covariates such as diet. Recently, using RBC fatty acid data from the Framingham Offspring Study, over 2.5 million single nucleotide polymorphisms (SNPs) were tested for association with 14 RBC fatty acids. In that analysis, 191 different SNPs were found to be associated with at least one fatty acid. We now report on an updated analysis which includes adjustments for dietary covariates and additional fatty acids, which has identified novel loci. In additional to single marker analyses, we also report on the results of gene based testing approaches and another novel approach using a mixture of normal distributions to assist in directly tying fatty acid levels to particular genotypes.

106

Up For A Challenge (U4C) - Stimulating Innovation in Breast Cancer Genetic Epidemiology

Leah E. Mechanic (1) Sara Lindström (2) Huann-Sheng Chen (1) Kenneth Daily (3) Eric J. Feuer (1) Tiffany Green (1) Jay Hodgson (3) Christine M. Kaefer (1) Anna Kern (1) Kevin McTigue (1) Thea Norman (3) Solly Sieberts (3) Audrey Wellons (1) Elizabeth M. Gillanders (1)
(1) National Cancer Institute, Division of Cancer Control and Population Sciences (2) Harvard T.H. Chan School of Public Health (3) Sage Bionetworks

Breast cancer is the most commonly occurring cancer, and the second most common cause of cancer deaths in women in the United States. Epidemiologic studies suggest that genetic factors play a key role in determining who is at increased risk of developing breast cancer. To date, genome-wide association studies (GWAS) helped researchers identify more than 90 common genetic variations. Although GWAS greatly increased our understanding of the genetic components of breast cancer risk, results to date explain only a small proportion of the estimated genetic contribution to the risk of

breast cancer. Shifting the focus of analysis from individual single nucleotide polymorphisms (SNPs) to pathways, could lead to the identification of novel gene sets involved in breast cancer risk. Therefore, in June 2015, the National Cancer Institute (NCI), in collaboration with Sage Bionetworks, launched "Up For A Challenge (U4C) - Stimulating Innovation in Breast Cancer Genetic Epidemiology" to encourage unique approaches to more fully decipher the genomic basis of breast cancer. Utilizing innovative approaches, the goal of this Challenge is to identify new genes or combinations of genes, genetic variants, or sets of genomic features involved in breast cancer risk. In addition, NCI aims to advance innovation in the field of genetic epidemiology by making data more widely available, increasing the amount and diversity of minds approaching a difficult scientific problem, and promoting broader collaborations' NCI will award up to $50,000 in prizes based on identification of novel findings, replication of findings, innovation of approach, evidence of novel biological hypotheses, and collaboration. Updates and current status of the U4C will be presented.

107

Cross-validated BLUPs for linear mixed models with multiple variance components and repeated measures: eQTL and longitudinal studies

Joel A. Mefford (1) John Witte (1) Noah Zaitlen (1)
(1) University of California, San Francisco

Linear mixed models (LMM) are used in GWAS to improve power for detecting associations and to control for confounding by population or family structure. The power gains arise by accounting for the ancillary contributions to the phenotype, or structure in the data – such as shared genetics or repeated measures. We model these contributions through a leave-one-out cross validation approach to generate cross-validated BLUPs (cvBLUPs) – estimates of the realized values of random effects for individuals. The cvBLUPs are used in models with multiple variance components: an eQTL analysis with two components to capture genetic and expression covariances; and a longitudinal analysis of CD4 cell count in HIV+ subjects, with components to account for genetic and family covariance and repeated measures.
Existing LMM approaches over-fit to the data, causing the BLUPs to be correlated with un-modeled contributions to the phenotype, including the effects of test-SNPs in an LMM-GWAS. To address this, we generate cvBLUPs, which we show are independent of un-modeled factors. Using cvBLUPs in association tests allows unbiased estimates of the test-SNP effect size, in analyses with one or more variance components. To evaluate our method we first simulate GWAS under different settings of number of subjects, genetic markers, fraction of causal markers, and population structure. We then apply our methods to an eQTL analysis of data from the GEUVADIS cohort.There, cvBLUP cuts correlation of genetic predictions with expression to.11 from an inflated.96 with naïve BLUP. Finally we analyze CD4 cell count recoveries after initiation of ART in the UARTO and ARKS HIV+ cohorts, where cvBLUPs from multiple variance components control inflation of test statistics.

108

Parental age, birth order, and neurodevelopmental disorders

Alison K. Merikangas (1) Louise Gallagher (1) Aiden P. Corvin (1) Elizabeth A. Heron (1)
(1) Trinity College Dublin

Advanced paternal age is a well-established risk factor for the development of schizophrenia (SZ); however, the mechanism for this association remains unclear. Other explanations, such as maternal age, birth order, and family size have been implicated. The aims of this study are: (1) to investigate the association between SZ proband birth order and paternal age, controlling for effects of both maternal age and family size; and (2) to examine the specificity of the paternal age effect by testing whether these effects extend to Autism Spectrum Disorder (ASD) as an index of other Neurodevelopmental Disorders (NDDs).
Samples included 264 probands with SZ from an Irish SZ collection (69% male) and 2,539 probands with ASD from the Simons Simplex Collection (87% male). In both collections, higher proband birth order was associated with both maternal and paternal older age categories. Ordinal regression models were used to examine the association between parental age and proband birth order, controlling for family size, and the opposite parent's age. In both SZ and ASD, higher proband birth order was associated with both maternal and paternal older age after controlling for family size.
This work adds to the growing body of knowledge that probes the association between paternally derived de novo mutations and SZ, and extends this work to ASD. Direct examination of de novo mutations and other genetic variants, together with phenotype information, parental age information, and other relevant perinatal factors using translational epidemiological approaches may be a more fruitful line of investigation. Moreover, other factors such as birth interval, and proband sex should also be considered in future studies

109

Systematic meta-analyses and field synopsis of genetic association studies in colorectal adenomas

Zahra Montazeri (1) Evropi Theodoratou (2) Christine Nyiraneza (1) Maria N. Timofeeva (3) Harry Campbell (2) Julian Little (1)
(1) University of Ottawa, Ottawa, Canada (2) University of Edinburgh, Edinburgh, UK (3) International Agency for Research on Cancer, Lyon, France

Colorectal cancer (CRC) constitutes a major global public health challenge. Most CRCs develop from preneoplastic asymptomatic lesions known as colorectal adenoma (CRA). We have previously summarized the associations between

common genetic variants and CRC in a field synopsis of genetic association and GWAS, but the genetic basis of CRA is less well documented. We now present the first synthesis of all published genetic association data for CRAs and the results of meta-analyses to summarise risk estimates.
Using Medline and the HuGENet phenopedia^{TM}, we identified and synthesized all published genetic association data for CRAs. We conducted meta-analyses of the identified studies and data from two GWAS to summarise risk estimates. We applied the Venice criteria and Bayesian False Discovery Probability (BFDP) to assess the levels of the credibility of associations.
9,750 titles and abstracts were initially screened, and 1750 publications were identified for full text screening of which 130 articles met the inclusion criteria. Data were extracted for 181 SNPs in 74 genes. The variant at 8q24.21 (rs6983267) was considered as "highly credible" and *MTHFR* (C677T), *NAT1*, *NQO1* (Pro187Ser), and *TP53* (Arg72Pro) as "less credible". The identification of genetic variants with influence on CRA risk may provide new insights into the fundamental biological mechanisms involved in early CRC development and help to inform future research. Further, CRA risk-associated SNP variants may also show utility in contributing to future risk scores for accurate population risk stratification which could be of potential value in improving CRC screening modalities.

110

GAMETES 2.0: Expanding the complex model and data simulation software to generate heterogeneous datasets, custom models, and quantitative traits

Jason H. Moore (1) Ryan J. Urbanowicz (1) Peter Andrews (1)
(1) University of Pennsylvania

Increasing acknowledgement of the complexity of common diseases, particularly with regards to complex multivariate patterns of association, has led to an increased interest in the development of new analytical methodologies, algorithms, and software able to detect, model, and characterize such patterns. In order to properly develop, test, and evaluate such methodologies, a variety of representative simulated datasets for simulation studies are required. Previously we developed the GAMETES software for the rapid, deterministic generation of strict, purely epistatic single nucleotide polymorphism (SNP) models based on user defined parameters such as heritability, minor allele frequencies, prevalence, and the order of interaction (e.g. 2-way, or 3-way). GAMETES 2.0 expands the capabilities of this software, allowing users to (1) combine multiple genetic models for the simulation of datasets with patterns of genetic heterogeneity, (2) generate custom 2 or 3-way SNP models with a report of the model's characteristics and predicted relative detection difficulty, (3) generate datasets with a quantitative trait/endpoint as opposed to a binary discrete class endpoint. Quantitative endpoints are generated from GAMETES genetic models by using the penetrance values of specific genotype combinations as a centroid for selecting a continuous trait value for each subject in the dataset. We test this new simulation software by generating simulation studies with heterogeneity and quantitative traits respectively and demonstrate that we can identify these simulated patterns using advanced machine learning approaches (i.e. QMDR and EXSTRACS), and feature selection approaches (ReliefF, SURF, SURF*, and MultiSURF).

111

Large-scale exome chip association analysis identifies novel type 2 diabetes susceptibility loci and highlights candidate effector genes

Andrew P. Morris (1)
(1) Department of Biostatistics, University of Liverpool, Liverpool, UK

To evaluate the contribution of low-frequency and rare coding variants to type 2 diabetes (T2D), we combined exome array data in 56,597 cases and 176,024 controls from five ancestry groups (European, South Asian, African American, East Asian and Hispanic). We tested variants for association with T2D using a linear mixed model to account for relatedness and population structure and combined summary statistics across studies in a fixed-effects meta-analysis. A total of 44 coding variants, mapping to 26 loci, were associated with T2D at exome-wide significance ($p < 5 \times 10^{-7}$). All but three were common, with minor allele frequency (MAF) > 5%. Thirteen variants were located outside established T2D loci, including common alleles in *PNPLA3* (I148M, $p = 1.6 \times 10^{-9}$) and *POC5* (H36R, $p = 1.1 \times 10^{-7}$), and a European-specific low-frequency mutation in *FAM63A* (Y285N, $p = 6.5 \times 10^{-9}$, MAF = 1%). Within established T2D loci, 22 variants have not been formerly implicated in disease risk, so we investigated their relationship with previously reported non-coding lead SNPs through conditional analyses. At the *CILP2* locus, the association of *TM6SF2* E167K ($p = 3.6 \times 10^{-12}$) was indistinguishable from the non-coding lead SNP (rs16996148, $p^{COND} = 0.052$), suggesting that the signal is mediated through this gene. Conversely, the association of *GIPR* E354Q ($p = 1.1 \times 10^{-8}$) was not eliminated after conditioning on the non-coding lead SNP (rs8108269, $p^{COND} = 4.0 \times 10^{-6}$), implying that these signals are distinct. Our results indicate that low-frequency and rare coding variants of large effect do not make a major contribution to T2D risk. However, these analyses implicate several novel genes in T2D pathogenesis and provide direct insight into the underlying biology of the disease.

112

Using Bayes Model Averaging to Identify GxE Interactions in Genome-wide Association Studies

Lilit C. Moss (1) David V. Conti (2)
(1) University of Southern California (2) University of Southern California

GWAS typically search for marginal associations between a SNP and a disease trait while gene-environment (GxE) interactions remain generally unexplored. Numerous approaches exist to increase power for testing GxE interaction by leveraging either SNP marginal effects or case-control ascertainment. However, these potential gains are accomplished under certain assumptions and it is often unclear if these are applicable a priori. Here, we use simulations to highlight performance across various methods as a function of marginal and interaction effect sizes, direction of effects, and the correlation of the two factors in the source population. Substantial variation in performance leads to uncertainty as to which approach is most appropriate for any given analysis. Bayes model averaging offers a statistical foundation for incorporating model uncertainty and we present a framework that: (1) balances the robustness of a case-control approach with the power of the case-only approach; (2) leverages marginal SNP effects; (3) allows for the incorporation of prior information; and (4) allows the data to determine the most appropriate model. We average over the inclusion of parameters corresponding to the marginal SNP and GxE interaction effects and the G-E association in controls. The resulting method exploits the joint evidence for marginal SNP and GxE interaction effects while gaining power from a case-only equivalent analysis. We demonstrate that this method detects SNPs within a wide range of scenarios with the potential for increased power over current methods. We apply this approach to a scan for asthma in the USC Children's Health Study.

113

Genome-wide association analysis with multivariate ECG traits

Martina Müller-Nurasyid (1,2,3) Carolina Roselli (1) Sonja Greven (4) Moritz Sinner (2) Melanie Waldenberger (3,5,6) Annette Peters (3,5,7) Konstantin Strauch (1,8) Stefan Kääb (2,3)
(1) Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany (2) Department of Medicine I, Ludwig-Maximilians-Universität Munich, Munich, Germany (3) DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany (4) Department of Statistics, Ludwig-Maximilians-Universität Munich, Munich, Germany (5) Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany (6) Research unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany (7) German Center for Diabetes Research, Neuherberg, Germany (8) Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany

Complex diseases are characterized by multifaceted phenotypes with different etiology. Therefore, it is often necessary to analyze a variety of phenotypes in order to elucidate the pathophysiological mechanism of the underlying disorder. Especially in genome-wide association analysis, the high correlation between phenotypes is often ignored.
The ECG is a valuable source for a collection of highly correlated biomarkers with high impact on prognosis, diagnosis and prevention of various diseases. Several ECG traits are known to have high heritability estimates, which implicates strong genetic components. We evaluated ~3,800 persons with 10sec 12-lead ECG measurements in the KORA S4 study (Cooperative Health Study in the Region of Augsburg) in Southern Germany. After exclusion of known abnormal ECGs, we derived standardized residuals for several ECG-intervals and the respective principle components (PCs). Subsequently, we investigated the difference between (a) the combination of separate genome-wide association screens (b) multivariate genome-wide association screens and (c) a combination of separate genome-wide association screens on independent PCs derived from the correlation structure of phenotypes.
The differences between the results for the three analysis types were marginal. Deriving PCs from the correlation matrix can help reduce the number of genome-wide screens. However, this procedure implicates a new challenge: the clinical interpretation of cross-phenotype definitions of statistically relevant phenotypes based on PCs and the interpretation and clinical relevance of significant genetic association signals with these derived new phenotypes.

114

Reproducibility in MCMC-based Linkage Analyses using Dense Marker Maps

Anthony Musolf (1) Claire L. Simpson (1) Susan M. Pinney (2) Mariza de Andrade (3) Collette Gaba (4) Ping Yang (3) Ming You (5) Ann G. Schwartz (6) Diptasri Mandal (7) Elena Y. Kupert (5) Christopher I. Amos (8) Marshall W. Anderson (5) Joan E. Bailey-Wilson (1)
(1) National Human Genome Research Institute, National Institutes of Health, Baltimore, MD (2) University of Cincinnati College of Medicine, Cincinnati, OH (3) Mayo Clinic, Rochester, MN (4) University of Toledo Dana Cancer Center, Toledo, OH (5) Medical College of Wisconsin, Milwaukee, WI (6) Karmanos Cancer Institute, Wayne State University, Detroit, MI (7) Louisiana State University Health Sciences Center, New Orleans, LA (8) Geisel School of Medicine, Dartmouth College, Lebanon, NH

Markov-Chain Monte Carlo (MCMC) algorithms are one of the more commonly employed methods in multi-point linkage software. MCMC algorithms estimate complex posterior distributions by creating a steady-state chain of samples assumed to be drawn from the posterior distribution. Though the nature of this approach means that the results (LOD scores in linkage analyses) will vary slightly depending on when the

chain is terminated, LOD scores have been shown to be reproducible when marker loci are in linkage equilibrium.
However, some programs offer the ability to correct for linkage disequilibrium (LD) internally, allowing researchers to avoid LD pruning of the marker map and perform multipoint linkage analyses on highly dense maps containing multiple SNPs per cM. We have observed that MCMC-based approaches that correct for LD exhibit excess false positives with very dense SNP maps and the results of the MCLINK-LD program are often not reproducible (after allowing for a sufficient burn-in period) even after pruning the marker map to 1cM minimum intermarker distance. Using real lung cancer data from extended families, we show that after ten identical analyses on the same data with a 1cM map, LOD scores often vary by 1-1.5 LOD units. Thus significant LOD scores in one analysis are much reduced in other analyses of the same data. We have developed or are in the process of developing methods to increase reproducibility in these LD correcting MCMC algorithms including: 1) binning the chromosomes into 1, 2, or 5 cM segments and selecting the most polymorphic marker within the bin for the analysis and 2) using an average of linkage results over multiple replicate analyses. We will then compare our results to MCMC programs which require complete LD pruning.

115

A new and scaleable Bayesian framework for joint re-analysis of marginal SNP effects

Paul J. Newcombe (1) David V. Conti (2) Sylvia Richardson (1)
(1) MRC Biostatistics Unit, Cambridge, UK (2) Division of Biostatistics, Department of Preventive Medicine, Zilkha Neurogenetic Institute, University of Southern California, USA

Recently, large scale GWAS meta-analyses - accumulating information over tens of thousands of people - have boosted the number of known signals for some traits into the tens and hundreds. However, the availability of many correlated single nucleotide polymorphisms (SNPs) presents analytical challenges and typically variants are only analysed one-at-a-time. This complicates the ability of fine-mapping to identify a small set of SNPs for further functional follow up.
We describe a new and scaleable algorithm for the re-analysis of published marginal associations under joint multi-SNP models, in which correlation is accounted for according to estimates from a reference dataset. SNPs which best explain the joint pattern of effects are highlighted via an integrated Bayesian penalized regression framework. Through a realistic simulation study, including an application to 10,000 SNPs, we demonstrate substantial gains in the proportion of true signals among top ranked SNPs (positive predictive value) using our multivariate framework. We also present a real data application to published results from MAGIC (Meta-Analysis of Glucose and Insulin Related Traits Consortium) - a GWAS meta-analysis of more than 15,000 people, in which we re-analyse several top loci associated with glucose levels two hours after oral stimulation. Our algorithm was able to rule out many SNPs as false positives and for one gene, *ADCY5*, joint modeling of the pattern of effects across the locus highlighted an alternative, and more plausible, SNP to the reported index.

116

Missing heritability: is the gap closing? An analysis of 32 complex traits in the LifeLines Cohort Study

Ilja M. Nolte (1) Peter J. van der Most (1) Behrooz Z. Alizadeh (1) Paul I. de Bakker (2,1) H. Marike Boezen (1) Marcel Bruinenberg (3) Lude Franke (4) Pim van der Harst (5) Gerjan Navis (6) Dirkje S. Postma (7) Marianne G. Rots (8) Ronald R. P. Stolk (1,3) Morris A. Swertz (4) Bruce H. Wolffenbuttel (9) Cisca Wijmenga (4) Harold Snieder (1)
(1) Department of Epidemiology, University of Groningen, University Medical Center Groningen, The Netherlands (2) Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, The Netherlands (3) LifeLines Cohort Study, University of Groningen, University Medical Center Groningen, The Netherlands (4) Department of Genetics, University of Groningen, University Medical Center Groningen, The Netherlands (5) Department of Cardiology, University of Groningen, University Medical Center Groningen, The Netherlands (6) Department of Nephrology, University of Groningen, University Medical Center Groningen, The Netherlands (7) Department of Pulmonology, University of Groningen, University Medical Center Groningen, The Netherlands (8) Department of Medical Biology, University of Groningen, University Medical Center Groningen, The Netherlands (9) Department of Endocrinology, University of Groningen, University Medical Center Groningen, The Netherlands

In recent years, thousands of single nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies as robustly associated with various complex traits and diseases, but they often explain only a small part of the variance. Recently, a new method was proposed to determine the proportion of dominance genetic variance at all common SNPs, which might account for part of the missing heritability.
In this study, we selected 1,442 SNPs that were genome-wide significantly associated with 32 continuous traits from five disease areas (musculoskeletal, cardiovascular/renal, metabolic, hematologic/inflammatory, and pulmonary). We used 13,436 subjects from the LifeLines Cohort Study, a population-based prospective cohort examining the health of 167,729 persons from the northern Netherlands, to estimate the proportions of phenotypic variance that could be attributed to additive and dominance genetic variance at all common SNPs as well as to perform replication analyses and to calculate the percentage of phenotypic variance explained by all selected associated SNPs combined in genetic risk scores (GRSs).

The additive genetic variance at all common SNP explained a significant proportion of the phenotypic variance for all traits ranging from 7.5% to 52.2%, but none of the traits showed significant dominance genetic variance. A total of 66.0% of all high-quality SNPs were significantly associated with the trait of interest and all GRSs associated significantly with their respective trait with variances explained between 0.02% and 15.5%.

In conclusion, a considerable part of the common SNP heritability for complex disease traits remains to be explained and detected. Such variants will have mostly additive effects.

117

No evidence for genetic risk score (GRS)-energy intake interaction for body mass index or waist-to-hip ratio

Heather M. Ochs-Balcom (1) Leah Preus (1) Jing Nie (1) Jean Wactawski-Wende (1) Marian Neuhouser (2) Lesley Tinker (2) Mary Pettinger (2) Cheng Zheng (3) Rasa Kazlauskaite (4) Rami Nassir (5) Lihong Qi (6) Lara E. Sucheston-Campbell (7)

(1) Department of Epidemiology and Environmental Health, School of Public Health and Health Professions, University at Buffalo, Buffalo, NY (2) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (3) School of Public Health, University of Wisconsin, Milwaukee, Milwaukee, WI (4) Rush University Medical Center, Chicago, IL (5) Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, Department of Internal Medicine, University of California, Davis, CA (6) Department of Public Health Sciences, University of California, Davis, CA (7) Department of Cancer Prevention and Population Sciences, Roswell Park Cancer Institute, Buffalo, NY

We conducted a gene-environment interaction study to evaluate whether dietary energy intake modifies the effect of known genetic variation on BMI and WHR (genetic risk score) and if the genetic variation contributing to variance of BMI, WHR and total energy intake are correlated.

Summary GRS for BMI (32 SNPs) and WHR (12 SNPs) informed by meta-GWAS were constructed in 3809 women of European ancestry from WHI and GARNET. Main effects of BMI and WHR genetic risk scores and biomarker-calibrated energy intake (which can account for dietary measurement error) on BMI and WHR were measured with regression models; interactions were measured on the multiplicative and additive scales. Heritability of and genetic correlation between BMI, WHR and energy intake were measured with Genome Complex Trait Analysis (GCTA).

In BMI models, the BMI GRS ($p_{adjusted} = 6.5 \times 10^{-10}$) and calibrated energy intake ($p_{adjusted} = 2.0 \times 10^{-16}$) were significant. WHR GRS ($p_{adjusted} = 1.5 \times 10^{-4}$) and calibrated energy intake ($p_{adjusted} = 2.3 \times 10^{-15}$) in WHR models were also significant. There was no evidence for interaction on multiplicative or additive scales. GCTA show heritability is significant for BMI ($h^2 = 0.29$, SE = 0.10, p = 0.002), WHR ($h^2 = 0.24$, SE = 0.10, p = 0.006) and FFQ-estimated energy intake ($h^2 = 0.30$, SE = 0.01, p = 0.001); however neither BMI nor WHR is significantly genetically correlated with FFQ-estimated energy intake ($p > 0.2$).

GRS have independent effects on BMI and WHR from those conferred via total energy intake, and our study found no evidence for modification of the effect of genetic risk of obesity by energy intake. However, our study may be inadequately powered to gain necessary insights at this time. GCTA analyses suggest that while BMI, WHR and energy intake have significant heritability, there are no shared underlying genetic influences on obesity and energy intake.

118

Novel method to estimate regional genetic associations improves genetic scores performance

Guillaume Pare (1) Shihong Mao (1) Wei Q. Deng (2)

(1) McMaster University, Department of Pathology and Molecular Medicine (2) University of Toronto, Department of Statistical Sciences

Despite considerable efforts, known genetic associations only explain a small fraction of predicted heritability. We have previously shown that large regions joint associations, where multiple contiguous genetic variants are included in regression models, can improve the variance explained by established association loci as compared to genome-wide significant SNPs alone (Pare et al., 2015). However, such regional joint associations are not easily amenable to estimation using summary data because of sensitivity to linkage disequilibrium (LD). Since only large genetic meta-analyses are likely to be powered to identify weak associations, we propose a novel method to estimate regional variance explained using summary GWAS data, accounting for LD while remaining robust to misspecification.

We first divided the genome into SNP blocks of approximately 1 Mb minimizing inter-block LD. We then derived a method to estimate regional genetic effects using summary data, and showed it is asymptotically equivalent to multiple regression models when genetic effects are strictly additive (i.e. no interaction nor haplotype effect). We next extended our method to provide corresponding gene scores (GS) incorporating all genotypes within a given region.

We first compared our method to multiple regression on height and BMI using individual-level (N = 7,776) data from the Health Retirement Study (HRS). While strong correlations were noted ($R^2 \sim 0.25$), we estimated the loss of variance explained due to the additivity assumption at 0.15 and 0.26, respectively. Using summary GWAS statistics from GIANT for height (N = 253,288) and BMI (N = 236,231), we estimated variance explained by each SNP block and derived corresponding GS. We found that only the top 20% of blocks contributed to height variance in HRS, with GS prediction R^2 of 0.055 and 0.021 before and after adjustment for genome-wide SNPs. Indeed, polygenic contribution of

the 20% top blocks by REML (using GCTA) was 0.63 (SD = 0.03) in HRS, while it was 0.00 (SD = 0.04) in the remaining 80%. In contrast, all blocks contributed to BMI variance, with GS prediction R^2 of 0.050 and 0.044 before and after adjustment for genome-wide SNPs. In summary, our results show that 1) strictly additive models do not fully capture complex traits associations, 2) traits differ in terms of the presence and distribution of regional associations, and 3) GS derived from regional associations can improve trait prediction.

119

Examining the Effect of Sequencing Depth on the Stability of Allele-Specific Epigenetic Effects

Richard C. Pelikan (1) Graham Wiley (1) Patrick Gaffney (1) Courtney G. Montgomery (1)
(1) Oklahoma Medical Research Foundation

Recent research has shown that distinct cell types are uniquely described by their epigenetic repertoires. Characterization of these repertoires through ChIP-Seq and similar methods has shown that epigenetic interactions are sensitive to allele-specific variation. As we begin to consider the dependencies of ChIP-Seq and similar data on allele-dependent effects, it is important to ask whether the sequencing depth for these experiments is sufficient to represent such an effect dependably and how stable the result is expected to be across replicate experiments when the sequencing depth changes.
To address the reproducibility of results from techniques that assess allele-specific ChIP-Seq peak coverage, we have developed a methodology that demonstrates the stability of an analytical result as a function of sequencing depth. We use the irreproducible discovery rate (IDR) method, in combination with knowledge about cell type-specific expectations of chromatin modification, to establish to what degree a particular sequencing depth will result in sufficient coverage in areas of interest. We applied this methodology to several rarefaction experiments in which the same biological sample is sequenced at varying depths. Our results show that a tradeoff boundary exists to illustrate the potential improvements in recovery and stability of informative peak areas with increasing sequencing depth. This tradeoff boundary differs as the ChIP antibody changes and implies that different experiments benefit more from increased sequencing depth than others.
We expect this methodology to be useful to those budgeting for large-scale sequencing studies across many individuals, while providing reasonable expectations of the results achieved.

120

Bayesian hierarchical model for joint estimation of SNP effects with integration of prior biological knowledge

Miguel Pereira (1) John R. Thompson (2) Christian X. Weichenberger (3) Cosetta Minelli (1)
(1) National Heart and Lung Institute - Imperial College London, UK (2) University of Leicester, UK (3) EURAC, Bolzano, Italy

Genome-Wide Association Studies are classically analyzed by estimating SNP effects individually and adjusting for multiple testing. However, SNPs identified so far explain a small proportion of the predicted heritability for most traits. The joint analysis of SNPs has been suggested to improve SNP detection in GWAS but its implementation is limited by computational efficiency. Another way to improve SNP detection is through integration of prior information that gives more weight to SNPs supported by biological evidence, which can be achieved using the Bayesian framework.
By extending the computationally efficient approach proposed by Yi et al. (2011), we used a Bayesian hierarchical model to exploit SNP correlation structure, defined by linkage disequilibrium D', and incorporate prior biological knowledge. SNPs in strong LD were grouped in LD blocks, and prior knowledge was retrieved with a bioinformatic tool (DINTOR). We piloted the model by testing the association of BMI with 2,614 SNPs within 30 LD blocks, of which 6 contained one "true" SNP, in 1,829 individuals. SNP effects were estimated both with and without LD block structure and prior knowledge.
The Bayesian joint estimation of SNP effects ranked the true SNPs and true LD blocks higher than the classical approach. Adding the LD structure to the model improved the ranking of true LD blocks, which was further improved by integration of prior knowledge, with 5 of the true LD blocks ranking in the top 6.
These preliminary results suggest that both LD block structure and integration of prior knowledge can improve SNP detection. We are currently working on: 1) improving the incorporation of prior knowledge; 2) using LD r^2 instead of D'; 3) scaling up the analysis to genome-wide level.

121

Impact of reference population relatedness on imputation quality

Lauren E. Petty (1) Lindsay S. Tucker (1) Heather M. Highland (1) Naveen Ramesh (1) Mandar Karhade (1) Craig L. Hanis (1) Jennifer E. Below (1)
(1) University of Texas Health Science Center at Houston

Genotype imputation is commonly employed to leverage population-based haplotype structure to estimate uninterrogated genotypes. 1000 Genomes (1KG) and HapMap 3 (HM3) are common reference populations for imputation. Both are well-studied and efforts have been made to characterize cryptic relatedness within them. Extended shared segments due to close relatedness in reference datasets will inflate estimates of r^2, and may bias imputation results.
To interrogate these effects, 1,890 Mexican-American individuals from Starr County, Texas were phased in SHAPEIT and imputed in IMPUTE2 using HM3 as a reference, and using a HM3 maximum unrelated subset excluding $\leq$ 3rd

degree relationships (removing 97 individuals). The same was also done using 1KG reference; 38 people were removed for the maximum unrelated subset.
We found little effect of reference data relatedness in comparisons of 1KG imputation results. Difference in dosage of the minor allele was 0 at 98.9% of dosages. Of the variants with non-zero dosage difference, 98.0% of the differences were ≤ 0.1. In HM3, we found a dosage difference of 0 at 93.7% of dosages. Of the non-zero dosage differences, 97.5% were ≤ 0.1. HM3 has about 2.5 times more related individuals with a similar overall sample size, so we observe greater dosage differences in the cohort that exhibited more relatedness.
The trend in dosage differences detected in comparisons of 1KG and HM3 suggests that reference datasets harboring more relatedness may create problematic bias in imputation. To confirm these effects are driven by reference sample relatedness, we have simulated reference data of varying degrees of relatedness and connectivity, and present recommended thresholds for controlling for effects of relatedness.

122

Detecting putative causal genetic variants using linkage disequilibrium in distinct ethnic background genome-wide association studies

Achilleas N. Pitsillides (1) Josée Dupuis (1)
(1) Boston University

Multiple methods for discovering putative causal variants in the context of Genome Wide Association Studies (GWAS) have been proposed. However, most methods are geared to single genetic background studies, and only evaluate possible causal variants that were part of the study. We propose a method to rank potentially causal variants exploiting distinct ethnic background GWAS. Our method takes advantage of the difference in linkage disequilibrium (LD) structure between different populations. Using LD information allows our method to consider putative causal variant candidates that have not been genotyped or imputed in any of the ethnic-specific GWAS.
The proposed method takes as inputs the significantly associated variants within a given region, their statistical significance (P-value) in the two distinct ethnic background GWAS, and the LD information from the two population and outputs a ranked set of putative causal variants. Each variant in the LD map that has non-zero r^2 with any of the significantly associated variants is considered as a putative causal variant. We use the LD information to build a weighted graph. The degree of the putative causal variants in the subgraph induced by the significant variants and the putative causal variant serves as the ranking weight. We evaluate our method by simulating two ethnically different populations of size 1000, one based on the HapMap CEU and one on the hapmap YRI population. We generate multiple uncorrelated traits based on different "causal" variants; we then use the hapmap LD maps and the computed p-values as the input to our method. We assess the ability of the algorithm to identify the correct causal variants among the top ranked variants.
In our simulations our method ranked the causal variant as the top variant 30% of the time. In the same simulations our algorithm ranked the real causal variant in the top 10 putative causal variants 80% of the time.
The proposed method of ranking putatively causal variants works well in our simulations and has the added benefit of being able to discover causal variants that are not necessarily part of the evaluated set of variants. The method is easily expandable to multiple ethnic populations.

123

DNA Methylation Score as a Biomarker in Newborns for Sustained Maternal Smoking During Pregnancy

Sarah E. Reese (1) Michael Wu (2) Shanshan Zhao (1) Siri Haberg (3) Per Ueland (4) Roy M. Nilsen (5) Øivind Midttun (6) Stein E. Vollset (3,4) Bonnie Joubert (1) Shyamal Peddada (1) Wenche Nystad (3) Stephanie J. London (1)
(1) National Institute of Environmental Health Sciences (2) Fred Hutchinson Cancer Research Center (3) Norwegian Institute of Public Health (4) University of Bergen (5) Haukeland University Hospital (6) Bevital AS

Maternal smoking during pregnancy leads to numerous adverse outcomes in offspring. We previously identified highly reproducible methylation signals using a genome wide methylation platform (Illumina 450K) that reflect sustained, rather than transient, smoking during pregnancy. Questionnaire assessment of smoking in pregnant women may not detect all smokers and timing of assessments may be insufficient to assess sustained smoking. There is no biomarker of sustained smoking during pregnancy. We used data on a short-term smoking biomarker, cotinine, measured in maternal plasma during pregnancy and Illumina 450K newborn cord blood DNA methylation in a pregnancy cohort (N = 1,279) to develop a biomarker of sustained maternal smoking during pregnancy. We used logistic least absolute shrinkage and selection operator (LASSO) regression with area under the curve (AUC) cross-validation to train a model predictive of sustained maternal smoking in pregnancy. We used the LASSO regression coefficients of CpGs most strongly associated with cotinine to develop a score that reliably predicted smoking status in the train set (N = 1,058; AUC = 0.98, Sensitivity = 85%, Specificity = 98%). As expected, predictive performance was lower on the much smaller test set (N = 221; AUC = 0.87, Sensitivity = 57%, Specificity = 93%). This score is a promising novel biomarker in newborns of sustained maternal smoking during pregnancy that should be useful where information on time course of smoking is limited or missing. This quantitative biomarker, which incorporates duration as well as dose, might improve the ability to detect health effects of maternal smoking during pregnancy.

124

Application of joint models in genetic association studies

Ghislain Rocheleau (1,2,3) Mickaël Canouil (1,2,3) Loïc Yengo (2,3) Philippe Froguel (1,2,3,4)
(1) Lille 2 University, Lille, France (2) UMR 8199 - Pasteur Institute, Lille, France (3) FR 3508 - European Genomic Institute for Diabetes, Lille, France (4) Department of Genomics of Common Disease, Imperial College London, London, UK

New statistical methods need to be proposed as an alternative to the current cross-sectional design predominantly used in genome-wide association studies (GWAS), especially when longitudinal (repeated) measures of a trait are available for analysis. It is known that efficient modeling of temporal longitudinal trajectories will increase statistical power to detect genetic loci associated with that trait. To optimize use of existing phenotypic data, we propose a joint model approach aimed at identifying genetic markers simultaneously associated to temporal trajectories of a trait and an event outcome. Joint models are mostly used in clinical trials in order to reduce bias in estimating the overall treatment effect on survival and longitudinal biomarker. Some simulations have further shown that power for testing the treatment effect is increased when the longitudinal trajectory and the time-to-event outcome are highly correlated.
Standard formulation of the joint model involves two components: a longitudinal component and a time-to-event component. Our proposed model consists in a linear mixed model combined with a Cox proportional hazards model. We illustrate the application of the joint model approach in genetic epidemiology by exploiting the strong link between temporal variation of blood glucose levels and onset of type 2 diabetes (T2D). Using genotypes assayed with the Metabochip DNA arrays (Illumina) from 4,500 subjects recruited in the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance), we reexamine previous GWAS findings for some confirmed glycaemia and T2D loci. Power estimation, sample size calculation and effects of missing data are also discussed.

125

Meta-analysis of gene-set analyses based on genome wide association studies, method development and application within ILCCO/TRICL consortia

Albert Rosenberger (1) Stefanie Friedrichs (1) Christopher I. Amos (2) Paul Brennan (3) Gordon Fehringer (4) Irene Brüske (5) Rayjean J. Hunh (4) Martina Müller-Nurasyid (6) Angela Risch (7) Heike Bickeböller (1)
(1) Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Göttingen, Germany (2) Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA (3) International Agency for Research on Cancer, Lyon, France (4) Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada (5) Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany (6) Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany (7) Division of Molecular Biology, University Salzburg, Salzburg, Austria

Genome-wide association studies (GWASs) are typically based on single nucleotide polymorphisms (SNPs). The interplay of genes in the etiology of the phenotype in question is not considered. Gene-set analysis (GSA) methods aim to combine markers belonging to a pre-defined SNP set. In general GSA provides no estimate of the strength of association but only p-values (pGS), indicating some accumulation in significance of observed associations with a phenotype for genes or markers within the gene set (GS). For a meta-analysis concordance in the observed patterns of single marker association estimates should be taken into account. We propose an enhanced version of Fisher's inverse X^2-method META-GSA, weighing each study to account for imperfect correlation between patterns.
We simulated 500 small GWASs (500 cases, 500 controls, 100 SNPs) and applied Wilcoxon's rank sum test was used as GSA method. META-GSA has greater power to discover truly associated gene sets compared to simply pooling the p-values. It also outperformed pooledGWAS-GSA, based on a single GSA after pooling marker-specific associations.
Applying META-GSA to four case-control GWASs of lung cancer (Central European and Toronto/SLRI Studies; German Lung Cancer and MDACC Studies) revealed the pathway GO0015291 ("transmembrane transporter activity") as significantly enriched with associated genes (GSA-method: EASE, $p = 0.0315$ corrected for multiple testing).

126

A comparison of statistical methods for the discovery of genetic risk factors using longitudinal family study designs

Marie-Hélène Roy-Gagnon (1) Kelly M. Burkett (2) Jean-François Lefebvre (1) Cheng Wang (1) Bénédicte Fontaine-Bisson (3) Lise Dubois (1)
(1) School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada (2) Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada (3) Nutrition Sciences Program, University of Ottawa, Ottawa, Ontario, Canada

The etiology of immune-related traits is often complex, involving many genetic and environmental factors and their interactions. Methodological approaches focusing on an outcome measured at one time point fail to capture complex disease mechanisms that fluctuate over time. It is increasingly recognized that longitudinal studies have great potential to shed light on complex disease mechanisms involving genetic factors. However, longitudinal data requires specialized statistical methods, especially in family studies where multiple sources of correlation must be modeled. Using simulated data with known gene-disease relationship, we examined

the performance of different analytical methods for investigating associations between genetic factors and longitudinal phenotypes in twin data. The simulations were modeled on data from the Québec Newborn Twin Study, an ongoing population-based longitudinal study of twin births with multiple phenotypes, such as cortisol levels and body mass index, and with sequencing data on immune-related genes. We compared approaches that we classify as (1) family-based methods applied to summaries of the observations over time, (2) longitudinal-based methods with simplifications of the familial correlation and (3) Bayesian family-based method with simplifications of the temporal correlation. We found that for estimation of the genetic main and interaction effects, all methods gave estimates close to the true values and had similar power. If heritability estimation is desired, approaches of type (1) also provide estimates close to the true value. Our work shows that simpler approaches are likely adequate to detect genetic effects; however, interpretation of these effects is more challenging.

127

Getting the most from your data: Comparing analyses using qualitative traits and related quantitative traits

Jeremy A. Sabourin (1) Alexa J. M. Sorant (1) Alexander F. Wilson (1)
(1) Genometrics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD

Quantitative genetics theory assumes that a qualitative trait is determined by an underlying quantitative liability, which may be a combination of lower-level quantitative traits. In this study, we used simulation to compare the power to detect a causal SNP with a qualitative trait, its underlying liability (usually unmeasurable), and a quantitative trait responsible for a portion of the liability. We generated a population of 10,000 unrelated individuals and defined a quantitative trait Q driven by a single causal SNP with heritabilities h^2 of 0.03, 0.05 and 0.1. This trait and additional sources of variation were combined to define a quantitative liability L, where the trait Q explains a proportion f of 50% and 75% of the liability's variation. A threshold on the liability was established to define a binary trait B (case/control status) with 20% of the population affected. Samples of size 500, 1,000, and 2,000 were selected both at random and ascertained for a balance of cases and controls in the sample. The power to detect the effect of the single causal SNP present was evaluated in 1,000 replicates of each model for all three simulated traits, using simple linear regression for the quantitative traits and logistic regression for the binary trait, with a genome-wide significance criterion of 5×10^{-8}. For the models considered, the power for the low-level quantitative trait Q was generally greater that for the liability L, which in turn was greater the power for the binary trait B for the same samples.

128

Investigating the causal relationship between atopic dermatitis and childhood mental health using Mendelian randomization

Hannah M. Sallis (1) Hannah M. Sallis (2) Jonathan Evans (1) George Davey Smith (2) Lavinia Paternoster (2)
(1) Centre for Academic Mental Health, School of Social and Community Medicine University of Bristol, Bristol, UK (2) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, UK

Although childhood atopic dermatitis (AD, or eczema) is common, there is little research into its impact on children's mental health. We expanded on this using a Mendelian randomization (MR) approach which uses genetic instruments to proxy a modifiable risk factor and is largely free from issues such as confounding and reverse causality, thus strengthening causal inferences using observational data.
AD was classified according to data collected at 11 times points from birth to 11 years for participants enrolled in the Avon Longitudinal Study of Parents and Children (ALSPAC). Parent-reported adolescent mental health was measured using the Strengths and Difficulties Questionnaire (SDQ) completed when the child was 12 and 16 years old.
Observational analyses found an association between AD and increased hyperactivity scores at age 12 and also with increased conduct problems at age 16.
Instruments were constructed using 1) immune related SNPs robustly associated with AD in a recent genome-wide association study, 2) mutations in the *FLG* gene, associated with skin barrier function.
We found no evidence of an association between hyperactivity scores and either instrument. However, at age 16, we found evidence of an association between the conduct disorder subscale and both genetic instruments. Contrasting effects were found, with immune SNPs associated with decreased conduct problems, and the *FLG* mutation associated with an increase. We also investigated directional pleiotropy using Egger regression. Our results suggest that the causal relationship is complex. Pleiotropic mechanisms are likely to exist and it is not clear what effect early successful treatment and/or prevention of AD may have on these outcomes.

129

Optimized genetic risk prediction for vitiligo and its use to define disease subtypes

Stephanie A. Santorico (1,2) Ying Jin (2,3) Daniel Yorgov (1) Subrata Paul (1) Tracey Ferrara (2) Richard Spritz (3,8)
(1) Mathematical and Statistical Sciences, University of Colorado, Denver, USA (2) Human Medical Genetics and Genomics Program, University of Colorado School of Medicine, Aurora, Colorado, USA (3) Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, USA

Generalized vitiligo (GV) is an autoimmune disease in which white patches of skin and hair result from destruction of melanocytes. GV patients have ~20% risk of other autoimmune diseases, pointing to shared genetic risk factors and perhaps environmental triggers. In previous work, we discovered and replicated 26 GV loci via two GWAS and 2 via candidate gene studies. We have just completed a third GWAS and combined analysis over all GWAS. After imputation using the 1KGP Phase I reference panel, QC, and ancestry matching, ~660,000 genotyped and ~20,000,000 imputed SNPs were compared in 2,853 European-derived white (EUR) cases and 37,412 EUR controls, providing power for common variants to OR 1.2. The combined analysis yielded 72 novel loci, several with multiple independent association signals, which are undergoing replication in an independent set of 2,138 EUR cases and 2,262 controls. Building off this success and extensive existing data, we compare a polygenic risk score for vitiligo, optimizing the set of SNPs for inclusion, to a major loci risk score. Disease subtypes will be explored by clustering cases based on a general genetic risk score or a set of biological pathway-specific risk scores. These will be correlated with vitiligo sub-phenotypes, occurrence/non-occurrence of other autoimmune diseases, age of onset, and self-reported treatment histories. Work will be presented through the lens of our long-term goal: optimized prediction of vitiligo risks and subtypes to facilitate clinical application of optimal therapies based on genetic subtyping of disease.

130

Sunburn, sun exposure, and sun sensitivity in the study of nevi in children

Jaya M. Satagopan (1) Susan A. Oliveria (1) Arshi Arora (1) Michael A. Marchetti (1) Irene Orlow (1) Stephen W. Dusza (1) Martin A. Weinstock (2) Alon Scope (1) Alan C. Geller (3) Ashfaq A. Marghoob (1) Allan C. Halpern (1)
(1) Memorial Sloan Kettering Cancer Center (2) Brown University (3) Harvard University

To examine the joint effect of sun exposure and sunburn on nevus counts (on the natural logarithm scale; log nevi) and the role of sun sensitivity.

We describe an analysis of cross-sectional data from 443 children enrolled in the prospective study of nevi in children. To evaluate the joint effect, we partitioned the sum of squares due to interaction between sunburn and sun exposure into orthogonal components representing: (i) monotonic increase in log nevi with increasing sun exposure (rate of increase of log nevi depends upon sunburn), and (ii) non-monotonic pattern.

In unadjusted analyses, there was a marginally significant monotonic pattern of interaction (p-value = 0.08). In adjusted analyses, sun exposure was associated with higher log nevi among those without sunburn ($p < 0.001$), but not among those with sunburn ($p = 0.14$). Sunburn was independently associated with log nevi ($p = 0.02$), even though sun sensitivity explained 29% (95% CI: 2%-56%, $p = 0.04$) of its effect. Children with high sun sensitivity and sunburn had more nevi, regardless of sun exposure.

A program of increasing sun protection in early childhood as a strategy for reducing nevi, when applied to the general population, may not equally benefit everyone.

131

No such thing as a free lunch: Assessing consistency of genotype imputation

Tae-Hwi Schwantes-An (1) Heejong Sung (1) Cristina M. Justice (1) Alexander F. Wilson (1)
(1) National Human Genome Research Institute

Imputation uses correlations between genotyped single nucleotide polymorphisms (SNPs) and reference panels such as HapMap or 1000Genomes to infer the genotypes of missing or untyped variants. The accuracy of imputation has been investigated in several studies and common variants are generally thought to be more accurately imputed than rare variants. Although considerable effort has been directed towards evaluating the accuracy of imputation, the consistency of imputation across replications has not been widely studied. In this study, we performed a simulation study to assess the consistency of imputation. Based on the 1000Genome European reference, 2,000 unrelated individuals were simulated for a 2Mb region on chr22 and their genotypes were masked and taken to be missing to match genotyping coverage of Illumina Omni 2.5 array. A total of 200 replications were performed, and ShapeIt and Impute2 were used to impute the masked missing variants. We calculated the assigned (imputed) non-reference allele counts (NRC) per SNP and per individual in order to measure the accuracy and consistency of imputation across the 200 replications. We found that imputation produces highly consistent NRC per SNP. However, the NRC is less consistent per individual. That is, although the total number of imputed alleles is fairly constant, the imputed allele (s) are found in different individuals across replications. Furthermore, the inconsistency we observed is not limited to the rare SNPs (minor allele frequency (MAF) < 0.01) but occurs across all MAF ranges. This result suggests that although imputation produces inconsistent genotypes at individual level across replications, and the inconsistency cannot be identified with current filtering schemes.

132

Analyzing case-parent trio data with the R package trio

Holger Schwender (1) Qing Li (2) Christoph Neumann (3) Margaret A. Taub (4) Samuel G. Younkin (5) Philipp Berger (1) Robert B. Scharpf (6) Terri H. Beaty (7) Ingo Ruczinski (4)
(1) Mathematical Institute, Heinrich Heine University, Duesseldorf, Germany (2) Inherited Disease Research Branch, National Human Genome Research Institute, Baltimore, MD, USA (3) Faculty of Statistics, TU Dortmund University,

Dortmund, Germany (4) Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA (5) Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA (6) Department of Oncology, Johns Hopkins University, Baltimore, MD, USA (7) Department of Epidemiology, Johns Hopkins University, Baltimore, MD, USA

Case-parent trio designs are frequently employed in genome-wide association studies to detect SNPs associated with disease. The most popular statistical tests in this study design are transmission/disequilibrium tests (TDTs), considering either the alleles or the genotypes as units in the analysis. While the genotypic TDT can be extended to the test for both gene-environment interactions and interactions between pairs of SNPs, higher-order interactions require procedures such as trio logic regression and trioFS that more cleverly search for association with disease without evaluating all possible interactions.
All these procedures are implemented in an R package called trio as user-friendly functions enabling efficient and fast association analysis. Besides these functions, trio also contains functionalities for, e.g., simulating trio data, estimating sample size and power, dealing with ped and vcf files, and estimating measures of linkage disequilibrium as well as LD-blocks. In this presentation, we exemplify the usage of the functions in trio, which is freely available at http://www.bioconductor.org, by reanalyzing data from the International Cleft Consortium comprising genotypes from about 2,000 children with different types of clefts.

133

Genetic heterogeneity results in variable heritability estimates in the presence of population substructure - epistasis as a cause for missing heritability

Ronnie Sebro (1)
(1) University of Pennsylvania

Genetic heterogeneity is a phenomenon where the phenotypic effect of the allele/s at one genetic locus is altered based on the allele/s at another genetic locus and is a form of epistasis. We assess the impact of genetic heterogeneity in the presence of population substructure on heritability estimates. Consider locus A having two alleles A or a, and locus B having two possible alleles B or b and these loci in complete linkage equilibrium. Assume the study population is comprised of two subpopulations where there is random mating and Hardy-Weinberg Equilibrium (HWE) within subpopulation but no intermixing between subpopulations.
100,000 random samples of 10,000 subjects were evaluated, assuming that genotype AABB has a phenotypic value of 0; genotypes AABb and AAbb have a value of 1 and all other genotypes have values of 0.5. The phenotypic variance was calculated: first assuming population substructure and then assuming HWE. The environmental variance was fixed equal to the phenotypic variance calculated assuming HWE. The heritability, H^2 is the ratio of the genetic variance, Var (G) to the sum of the genetic and environmental variance, Var (E), H^2 = Var (G)/ (Var (G) + Var (E)).
Wright's FST estimates varied from 0 to 0.99 with mean (median) of 0.14 (0.065) for the first locus and varied from 0 to 0.982 with mean (median) of 0.14 (0.064) for the second locus. Heritability estimates varied from 37.23% lower than to 96.3% higher than that calculated assuming HWE with mean (median) heritability estimates being 7.7% (2.7%) higher.
Population substructure results in variable heritability estimates, which are mostly inflated, but sometimes decreased relative to those calculated assuming HWE.

134

Polygene - by - Prenatal Environment Interaction in Autism Spectrum Disorder using Copy Number Variant Burden

Brooke Sheppard (1,2,3) Kelly S. Benke (4,2) Julie Daniels (5) Lisa A. Croen (6) Diana Schendel (7,8,9) Christine Ladd-Acosta (1,2) M. Daniele Fallin (2,4)
(1) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health (2) Wendy Klag Center for Autism and Developmental Disabilities, Johns Hopkins Bloomberg School of Public Health (3) Genetic Epidemiology Research Branch, National Institute of Mental Health (4) Department of Mental Health, Johns Hopkins Bloomberg School of Public Health (5) Department of Epidemiology, University of North Carolina (6) Autism Research Program, Division of Research, Kaiser Permanente Northern California (7) Section for Epidemiology, Department of Public Health, Aarhus University (8) Department of Economics and Business, National Centre for Register-Based Research, Aarhus University (9) Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH

Increased copy number variation (CNV) burden has previously been associated with Autism Spectrum Disorder (ASD). The prenatal period is also of increasing importance for understanding the environmental risk factors contributing to ASD etiology. However, evidence for specific risk factors has been inconclusive. A joint approach considering CNV burden and candidate prenatal risk factors simultaneously may elucidate genetic risk in ASD. The purpose of this work is to evaluate gene-environment interaction associated with ASD using estimates of CNV burden and prenatal environmental exposure data from the Study to Explore Early Development (SEED). SEED is a population-based case-control study of children aged 2–5 years with ASD and a control group drawn from the general population. A Hidden Markov Model approach (PennCNV) was used to call CNVs from Illumina genotype array data. Prenatal exposure to environmental risk factors such as smoking, alcohol, selective serotonin reuptake inhibitors, beta-2 adrenergic receptors, and maternal infection were assessed via self-report. Measures of CNV burden were assessed on both a genome-wide scale and restricted to ASD candidate regions. We will present results from testing interactions between CNV burden and five prenatal environmental risk factors in relation to ASD case status. This will be

the first effort to report genome-wide CNV burden by environment interaction in a population-based ASD case-control sample. These results may contribute to our understanding of how CNV burden and prenatal environmental risk factors are related to ASD etiology.

135

Using parental phenotypes in case-parent studies

Min Shi (1) David M. Umbach (1) Clarice R. Weinberg (1)
(1) National Institute of Environmental Health Sciences

In studies of case-parent triads, information is often collected about history of the condition in the parents, but typically parental phenotypes are ignored. Including that information in analyses may increase power to detect genetic association for autosomal variants. Our proposed approach uses parental phenotypes to assess association independently of the usual case-parent-based association test, enabling cross-generational internal replication for findings based on offspring and their parents. Our model for parental phenotypes also resists bias due to population stratification. We combine the information from the two generations into a single coherent model that can exploit approximate equality of parental and offspring relative risks to improve power and can also test that equality. We call the resulting procedure the Parent-phenotype Informed Likelihood Ratio Test (PPI-LRT). When some parental genotypes are missing, one can use the expectation-maximization algorithm to fit the combined model. We also develop a second composite test (PPI-CT) based on a linear combination of the parent-phenotype-based test statistic and that from the traditional log-linear, transmission-based test. We evaluate the proposed methods through non-centrality parameter calculations and simulation studies and compare them to the previously proposed approaches, parenTDT and combTDT. We show that incorporation of parental phenotype data often improves statistical power. As illustration, we apply our method to a study of young-onset breast cancer and find that it improve precision for SNPs in *FGFR2* and that estimated relative risks based on triads are closely replicated using the parental data.

136

Genetic Association Analysis of Low Frequency Variants: Prospective vs. Retrospective Penalized Logistic Regression with a Quantitative Covariate

Ji-Hyung Shin (1) Shelley B Bull (1)
(1) Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital & Dalla Lana School of Public Health, University of Toronto

For many complex diseases, genetic variants of low frequency are thought to be important, but standard logistic regression can fail for low variant counts and/or low disease prevalence. As an alternative, we evaluated Firth's penalized logistic regression, by simulation, under two analytic approaches: (i) prospective model with disease outcome as the response variable and genotype as a predictor variable; and (ii) retrospective model with genotype as the response variable and disease outcome as a predictor variable. We generated a binary disease outcome under a cohort design with disease risk depending on a bi-allelic variant and a quantitative covariate, which may or may not be correlated with one another, and evaluated bias in the genetic effect estimate and properties of the likelihood ratio (LR) test for genetic association.

For highly sparse data (< 40 minor allele carriers in a sample of 2000), all penalized and standard approaches can perform poorly. Otherwise, both prospective and retrospective penalized LR tests maintain valid type 1 error better than standard LR tests and have similar power when the genotype and quantitative covariate are uncorrelated, while the prospective approach performs better than the retrospective for correlated covariates. Moreover, the prospective odds ratio estimate tends to be less biased than the retrospective estimate. We recommend prospective penalized logistic regression as a useful alternative for analysis of binary traits and low frequency variants in the presence of a quantitative covariate such as a population structure principal component, and suggest that the retrospective model be applied cautiously.

137

Phenotypic variance explained by ancestry in admixed African Americans

Daniel Shriner (1) Amy R. Bentley (1) Ayo P. Doumatey (1) Guanjie Chen (1) Jie Zhou (1) Adebowale Adeyemo (1) Charles N. Rotimi (1)
(1) National Human Genome Research Institute

We surveyed 26 quantitative traits and disease outcomes to understand the proportion of phenotypic variance explained by ancestry in admixed African Americans. After inferring local ancestry as the number of African-ancestry chromosomes at hundreds of thousands of genotyped loci across all autosomes, we estimated the variance explained by local ancestry in two large independent samples of unrelated African Americans. We found that local ancestry at major and polygenic effect genes can explain up to 20% and 8% of phenotypic variance, respectively. These findings provide evidence that most but not all additive genetic variance is explained by genetic markers undifferentiated by ancestry. These results also inform the proportion of health disparities due to genetic risk factors and the magnitude of error in association studies not controlling for local ancestry.

138

The Renaissance of Linkage Analysis and Effects of Extreme High Density Genotype Data on Linkage Algorithms

Claire L. Simpson (1) Anthony Musolf (1) Joan E. Bailey-Wilson (1)

(1) Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore MD 21224

The inflation of multipoint linkage statistics in the presence of linkage disequilibrium (LD) between markers when there are missing founder genotypes is well established. The current era of whole exome (WES) and whole genome sequencing (WGS), and new genotyping arrays with exome content have brought renewed interest in family-based studies and there is renewed interest in linkage analysis methods. However, most methods and programs were developed before the genome-wide association study (GWAS) era and although some programs do incorporate methods to deal with intermarker LD, it is unknown whether these algorithms are adjusting adequately for the extremely high density of genotype data now available.
In studies of WES and exome array genotyping, we have observed highly inflated linkage statistics in both singlepoint and multipoint analyses, even when allele frequencies were estimated from the data, there were no missing founder genotypes and using programs that adjust for intermarker LD (MERLIN and MCLinkLD). Although it is possible to prune SNP data based on LD in programs such as PLINK, such approaches are not always ideal especially when dealing with multi-allelic variants and may not choose the most informative markers for linkage. In addition, the pruning process is only partly automatable and requires hands-on analysis of residual LD to ensure all intermarker LD has been removed. Here we present analyses of simulated data from Genetic Analysis Workshops 17 and 18, to demonstrate the problem. We also will present methods for dealing with the problem in the presence of multi-allelic markers and for prioritizing the most polymorphic and therefore likely most informative markers.

139

Extending Concepts of Gene-Environment Interaction Testing to a DNA Methylation Framework

Emily Slade (1) Peter Kraft (1)
(1) Harvard T.H. Chan School of Public Health

This paper was motivated by the successes of leveraging gene-environment interaction to increase the power of genetic association tests as well as considering the role played by DNA methylation in complex disease susceptibility. We extend the joint test for discrete genetic and environmental exposure data to the case of continuous methylation and environmental data to create a joint test that is powerful to detect methylation association even in the presence of effect modification by an environmental exposure. Through simulation, we show that this test has optimal or nearly optimal power as compared to a marginal test for methylation main effect and a standard test for methylation-environment interaction. We also show that this holds in the presence of methylation-environment correlation and measurement error in the environmental exposure. Thus, our joint test is a good option for detecting methylation association when little is known about the potential presence of effect modification by an exposure a priori due to its flexibility to be powerful across a wide range of true methylation-environment models.

140

Predicting maximally informative future experiments from existing repositories of gene expression data

Claire Smith (1) Brian Greco (2) Jacob O'Bott (3) Nathan Tintle (4)
(1) Yale University, New Haven (2) University of Texas, Austin (3) University of Maryland, Baltimore (4) Dordt College, Sioux Center

When considering large databases of genome-wide gene expression data, we have noted that many experiments show similar patterns of gene expression, such that many newly added experiments tend to be limited in their ability to provide substantially new insights into genome-wide metabolic and regulatory behavior. We have recently developed a method to predict experimental conditions which will be substantially different than existing gene expression data, and, thus, have the potential to provide maximal amounts of information about the regulatory and metabolic behaviors of hypothetically annotated genes and poorly understood pathways. In our method, we first utilize a metric to determine which genes have consistently low expression values across all experimental conditions, suggesting the genes are rarely, if ever, activated in the current set of experiments. We then utilize integrated regulatory and metabolic models to predict what type of experiments will activate these genes. We will present results from application of our method to large repositories of gene expression for E. coli and S. oneidensis.

141

A generalized joint location-scale association test for uncertain genotypes and related individuals

David Soave (1,2) Lei Sun (1)
(1) University of Toronto, Toronto (2) The Hospital for Sick Children, Toronto

In genetic association studies, it has been pointed out that a number of biologically meaningful scenarios, including GxG and GxE interactions, can lead to variance (scale) heterogeneity of a quantitative trait across genotype groups, and corresponding scale tests have been developed. Recently, a joint location-scale testing framework was proposed for association analysis of genotyped SNPs in a population sample, combining evidence from both the traditional mean (location) test and the newer scale test to achieve better power (Soave et al., 2015 AJHG *in press*).
Genotype uncertainty, however, is inherent in both imputed and sequenced data, and these DNA data are also commonly obtained on samples that include related individuals from families or cryptic relationships. A number of modified location tests have been proposed for these more complex data,

but the lack of a generalized scale test remains the bottleneck in broader application of the joint location-scale association test.

Among many testing procedures for variance heterogeneity, Levene's test is noted for its simplicity and robustness to departures from normality. Extending the work of Glejser (1969) and Iachine et al. (2010), we propose an easy-to-implement extension to Levene's method that formulates the variance test as a generalized least squares regression problem. Unlike Levene's original test, the regression framework is flexible to directly incorporate the probabilistic data associated with genotype uncertainty and the correlation structure associated with related individuals. Performance of the proposed method, including its validity and enhanced power, is demonstrated through simulations and applications.

142

S.A.G.E. Suite: a collection of software programs that enables Statistical Analysis for Genetic Epidemiology

Yeunjoo E. Song (1) Sungho Won (2) Sunah Song (3) Robert C. Elston (1)

(1) Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA (2) Department of Public Health Science, Graduate School of Public Health, Seoul National University, Seoul, Korea (3) Department of Electrical Engineering and Computer Science, Case School of Engineering, Case Western Reserve University, Cleveland, OH 44106, USA

The software package S.A.G.E. (Statistical Analysis for Genetic Epidemiology) is a well-known collection of C++ programs that provides a wide range of genetic analyses, such as familial correlation, segregation, linkage, and association analyses. From the first version released in 1987 to the most current version (v6.3) released in 2012, it has been extensively used for the analysis of family, pedigree and individual data by numerous researchers in the study of genetic epidemiology, especially with a point-and-click graphical user interface since v5.0. Here, we introduce the S.A.G.E. Suite, a new upgrade from the existing S.A.G.E. package. It contains S.A.G.E. v6.4, the latest version of the traditional package, with bug fixes and updates from the previous version. In addition, the new comprehensive S.A.G.E. Suite also includes two new packages, named qSAGE and rSAGE. qSAGE is a sequence-data enabled version of the S.A.G.E. package: it is designed to work directly with variation data in a VCF (Variant Calling Format) file from large sequencing projects such as for whole-exome and whole-genome studies. rSAGE is an R interface to the S.A.G.E. package: it allows users to run S.A.G.E. programs within the R environment and provides additional graphical outputs. With these new additions, the S.A.G.E. Suite provides a more complete selection of tools for various genetic analyses with both old and new data types, and with both old and new user interfaces. Funding: National Research Foundation of Korea Grant by the Korean Government (NRF-2014S1A2A2028559)

143

Rheumatoid susceptibility SNPs and their association with disease severity at presentation and methotrexate response

Jenna L. Strathdee (1) John C. Taylor (1) YEAR Consortium Tim Bongartz (2) James I. Robinson (3) Paul Emery (3) Ann W.Morgan (3) Jennifer H. Barrett (1)

(1) Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK (2) Division of Rheumatology, Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA (3) Leeds Musculoskeletal Biomedical Research Unit, Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK

Many genetic variants have now been identified that are associated with susceptibility to rheumatoid arthritis (RA), but little is known about genetic predictors of disease severity or treatment response. The aim is to investigate whether a person's genetic predisposition to developing RA is associated with disease severity at presentation or response to the usual first-line treatment (methotrexate (MTX)). All subjects ($n = 342$) were from the Yorkshire Early Arthritis Register and were treated with MTX. A weighted polygenic risk score (PRS) was calculated combining information on 101 SNPs associated with RA using estimates from a meta-analysis of RA susceptibility by Okada et al., Nature 506: 376–381, 2014. For each subject, the number of risk alleles was multiplied by the log of the odds ratio for that SNP from the meta-analysis; the PRS is the sum of these terms over all SNPs. The PRS was used as a continuous predictor in linear regression analysis of the initial measures: log of C-reactive protein (CRP), disease activity score 28 (DAS28), swollen joint count 28 (SJC28) and tender joint count 28 (TJC28), and 6-month change in each of those measures, adjusted for age and sex. PRS was also used as a predictor for family history and presence of rheumatoid factor and analysed by logistic regression, adjusted for age and sex. At a nominal 5% significance level, a higher PRS was associated with a lower baseline SJC28 ($p = 0.03$) and TJC28 ($p = 0.03$) and higher odds of having family history ($p < 0.001$) and rheumatoid factor ($p < 0.001$), but not with any other baseline or response measures. This analysis is now being expanded to better capture the association with *HLA*, which is currently represented by only one SNP.

144

Combining text mining and epistasis analyses identifies new atopy genes

Pierre-Emmanuel Sugier (1,2) Myriam Brossard (1,3) Amaury Vaysse (1,4) Chloé Sarnowski (1,3) Marie-Hélène Dizier (1,4) Marc Lathrop (5) Catherine Laprise (6) Florence Demenais (1,4) Emmanuelle Bouzigon (1,4)

(1) INSERM, UMR-946, Paris, France (2) Université Pierre et Marie Curie, Paris, France (3) Université Paris Sud, Paris, France (4) Université Paris Diderot, Paris, France (5) McGill University and Genome Québec Innovation Centre,

Montréal, Canada (6) Université du Québec à Chicoutimi (UQAC), Chicoutimi, Canada

A previous GWAS conducted for atopy in the French EGEA dataset with validation in the French-Canadian SLSJ dataset identified a single genome-wide significant signal (Pmeta = 9×10^{-9}) within adhesion G protein-coupled receptor V1 gene (*ADGRV1*).
To investigate whether the integration of prior knowledge based on text mining to prioritize genes for SNPxSNP interaction analysis may contribute to uncover new atopy genes, we conducted epistasis analysis between *ADGRV1* and a set of genes selected using two strategies: 1) genes harbouring at least one SNP at $PGWAS \leq 10^{-4}$; 2) genes selected in (1) and showing significant relationship through text mining using GRAIL (Gene Relationships Among Implicated Loci) applied to PubMed abstracts.
Based on EGEA GWAS results, 32 genes were selected for cross-gene SNPxSNP interaction analyses. The first approach led to 81,730 SNPxSNP interaction tests performed in EGEA and among which 403 reached the threshold of 0.005 and were taken forward for replication in SLSJ. Meta-analysis of the outcomes from these two datasets showed three SNP pairs with PmetaINT $\leq 10^{-4}$. The GRAIL method applied to the 32 selected genes identified a relationship between *ADGRV1* and three genes (*DNAH5*, *CHD7* and *ATP8B1*), reducing the number of interaction tests by 9-fold. The same three SNP pairs as previously identified had PmetaINT $\leq 10^{-4}$. One of these pairs at *ADGRV1* and *CHD7* (PmetaINT = 2×10^{-5}) reached the multiple-testing corrected threshold when using the text-mining based filtering.
This study highlights that integrating text mining and epistasis analysis facilitates the identification of new susceptibility genes.
Grants: ANR-11-BSV1-027, ANR-USPC-2013

145

Multivariate association test for rare variants controlling for cryptic and family relatedness

Jianping Sun (1) Karim Oualkacha (2) Celia M. T. Greenwood (3,1)
(1) McGill University (2) Université du Québec à Montréal (3) Lady Davis Institute

Since rare causal genetic variants are often enriched in families containing multiple affected individuals, family-based study designs are back in vogue for studying rare genetic variants. Recently, a few new methods have been developed to test for (inter-family or population) association between genetic variation in a small genomic region and a quantitative phenotype, while adjusting for the dependence between related individuals. Despite the sophistication of these methods, statistical power is often limited due to the rarity of the causal alleles and the genetic architecture at the locus. However, when there is the potential for a shared genetic underpinning of multiple quantitative traits, a simultaneous analysis of multiple traits may increase power to detect associations.
We have developed an approach for multivariate-phenotype rare-variant analysis in families by combining univariate test statistics for each trait. However, it is usually infeasible to derive the distribution of the optimal combination due to the complex unknown covariance structure among these univariate test statistics. Hence, we propose two different approaches by using either a copula model or a perturbation method to approximate the distribution of the combined test statistic. Simulations show that these two approaches are approximately valid with small deviations from the expected type I error when the approximated distributions close to the true ones, and more powerful than univariate tests when there are pleiotropic effects or correlated multiple traits. In this talk, we will present this work together with some simulations describing the performance of these two options.

146

Correlation structure of the genome

Heejong Sung (1) Tae-Hwi Schwantes-An (1) Alexa J. M. Sorant (1) Jeremy A. Sabourin (1) Cristina M. Justice (1) Alexander F. Wilson (1)
(1) Genometrics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, NIH, Baltimore, MD

As the number and density of genetic markers increases, it is becoming increasingly difficult to adjust for all the correlations between markers in statistical genetic analyses. An alternative approach is to identify regions of the genome that are "independent" and adjust for correlations within region. Recombination hotspots can be used to identify "independent" regions, although the degree of independence between these regions is unknown. In this study, we estimate the correlations between these adjacent regions and compare them to correlations between regions selected at random. Following Schwantes-An et al. (2014), the genome (HapMap 3 CEU unrelated data) is classified into hot or cold spots (tiles), regions including markers having greater or less than a predefined threshold on recombination rates (5, 10, 15, or 20 cM/Mb) between consecutive markers. Dependence between two tiles is measured by the average pairwise squared Pearson correlations between markers of one tile and markers in the other tile, denoted as AveCor. The mean AveCor between consecutive cold tiles across the genome decreases as the recombination rate threshold increases. For each threshold, this mean is statistically significantly different from the mean over a random 10,000 pairs of cold tiles from two different chromosomes. For most chromosomes, the mean AveCor of a random 10,000 pairs of cold tiles within the chromosome is not significantly different from that of pairs on different chromosomes. However, chromosome 14 is an exception, with higher correlation across the chromosome for all thresholds used. By defining regions based on recombination rate, one can investigate

correlation structures of the genome as well as having independent regions.

147

An Empirical Comparison of Interaction and Stratified Models to GxE Interactions Analysis: Smoking and Systolic Blood Pressure in the CHARGE Gene-Lifestyle Interactions Working Group

Yun Ju Sung (1) Thomas W. Winkler (2) Alisa K. Manning (3) Sharon Kardia (4) Xiaofeng Zhu (5) Kenneth Rice (6) Ingrid B. Borecki (7) Dabeeru C. Rao (1) James W. J. Gauderman (8) Adrienne L. Cupples (9)
(1) Division of Biostatistics, Washington University, St. Louis, MO 63110, USA (2) Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, 93051, Germany (3) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA (4) Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA (5) Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA (6) Department of Biostatistics, University of Washington, Seattle, WA 98195, USA (7) Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA (8) Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90089, USA (9) Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

Studying gene-environment (GxE) interactions is important, as they extend our knowledge of the genetic architecture of complex traits and may help to identify novel variants not detected via main-effects analysis alone. For quantitative traits, two models are widely used for studying GxE interactions. The 'interaction' model includes both SNP main and GxE effects in a single model. The 'stratified' model combines results from SNP main-effect analyses carried out separately within the exposed and unexposed groups. Although there have been several investigations including theoretical and simulation-based studies, an empirical comparison using real data is lacking. Here, we present our extensive comparison using over 10 million variants based on 1000 Genomes imputed data from 20 cohorts of European ancestry, comprising a total of 79,731 individuals. Our cohorts have sample sizes ranging from 456 to 22,983 and include both family-based and population-based samples. Within each model, we performed the 1 degree of freedom (DF) test of the interaction effect and the 2 DF joint test of the main and interaction effects. We compared results from two models at two levels, one using cohort-specific GWAS results and next using the results from meta-analysis. For most cohort-specific GWAS results, both approaches provided similar p-values (correlation over 0.96 in 14 cohorts). For meta-analysis results, however, agreement between the two models was much reduced despite increased sample sizes (correlation of 0.90). Disagreement was more pronounced at low frequency variants (with MAF < 5%). We also found that the two models may provide different results with inclusion of family-based cohorts, indicating a need for novel method development.

148

Evaluation of methodology for the analysis of "time-to-event" data in pharmacogenetic studies

Hamzah Syed (1) Andrea Jorgensen (1) Andrew P. Morris (1)
(1) University of Liverpool

Methods for the analysis of genome-wide association studies have focused on binary and quantitative traits. However, in pharmacogenetic studies, the outcome is often "time to event" after treatment intervention, for example remission, or occurrence of an adverse event. One approach is to dichotomise survival outcomes at a fixed end-point, and to treat the resulting phenotype as binary in a logistic regression framework, but this would be expected to result in a loss of power to detect association. We undertook a simulation study to compare the power of the Cox proportional hazards and logistic regression models for the analysis of time to event data in genetic association studies. We aimed to highlight scenarios with the greatest difference in power between these models. We considered a range of study designs including multiple treatments with differential effects on outcome and adverse events, and SNP-treatment interactions. The models were tested further by application to the Standard and New Anti-epileptic Drugs (SANAD) Study to test association of variants at the *ABCB1* locus in patients with epilepsy with three outcomes: time to first seizure, 12 month remission and withdrawal. The Cox proportional hazards model was demonstrated to be uniformly more powerful than the logistic regression analysis of dichotomised outcomes across simulation scenarios and generated stronger signals of association in the SANAD study. However, the difference in power between methods was highly dependent on the rate of censoring and number of events occurring in the study period. The findings of our study have important implications for the development of analytical protocols in the analysis of time to event data in pharmacogenetic studies.

149

Integrating biological knowledge and omics data using network and module guided random forests

Silke Szymczak (1) Michael Krawczak (1)
(1) Institute of Medical Informatics and Statistics, University of Kiel, Kiel, Germany

High-throughput technologies including microarrays and next generation sequencing allow comprehensive characterization of patients on many different molecular levels. However, building mathematical prediction models based on those omics data is challenging. A promising solution is to integrate external knowledge about structural and functional relationships between molecules into the

training process. Recently, several approaches have been proposed that use either modules of functionally related variables or directly employ the network structure to guide variable selection and tree building in random forests (RFs). We compare these novel methods including synthetic feature RF (Pan et al., 2014), module guided RF (Chen & Zhang, 2013), mutual information RF (Pan et al., 2013) and network constrained RF (Andel et al., 2015) with standard RF implementations. Evaluations are based on publicly available data sets measuring mRNA and miRNA expression. External biological information will be provided as a network based on information about protein-protein interactions and relationships between miRNAs and their target genes. For each phenotype results of two independent studies are used to assess training and test error as well as consistency of variable selection.

150

Meta-Analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs

Zheng-Zheng Tang (1) Dan-Yu Lin (2)
(1) Vanderbilt University, Nashville (2) University of North Carolina, Chapel Hill

There is heightened interest in using next-generation sequencing technologies to identify rare variants that influence complex human diseases and traits. Meta-analysis is essential to this endeavor because large sample sizes are required to detect associations with rare variants. In this work, we provide a comprehensive overview of statistical methods for meta-analysis of sequencing studies to discover rare-variant associations. Specifically, we discuss the calculation of relevant summary statistics from participating studies, the construction of gene-level association tests, the choice of transformation for quantitative traits, the use of fixed-effects versus random-effects models, and the removal of shadow association signals through conditional analysis. We also show that meta-analysis based on properly calculated summary statistics is as powerful as joint analysis of individual-participant data. In addition, we demonstrate the performance of different meta-analysis methods using both simulated and empirical data. We then compare four major software packages for meta-analysis of rare-variant associations – MASS, RAREMETAL, MetaSKAT, and seqMeta – in terms of the underlying statistical methodology, analysis pipeline, and software interface. Finally, we present a software interface, PreMeta, that integrates the four meta-analysis packages and allows a consortium to combine otherwise incompatible summary statistics.

151

Data integration in cancer genomics: non-coding mutations

Simon Tavaré (1)
(1) Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

We have been involved with the International Cancer Genome Consortium (ICGC) project on Oesophageal Adenocarcinoma. The project is producing whole-genome sequence data on 500 tumour-normal pairs, and has begun expression and methylation profiling of the same samples. To understand how the different regulatory landscapes are generated in cancer cells we also need to take into account changes in the DNA sequence in non-coding regions. Analysing the non-coding somatic mutations is a rather new field with little consensus with regards to the best approaches. In the talk I will describe our first steps towards assessing the significance of recurrence (among samples) and clustering of various regions of interest, and our attempts to predict functional consequences of such mutations.

152

Singular value decomposition in permutations for family-based association tests for quantitative traits

Bamidele O. Tayo (1) Richard S. Cooper (1)
(1) Department of Public Health Sciences, Loyola University Chicago Stritch School of Medicine, Maywood, IL

In family-based association testing for quantitative traits, some procedures such as the PLINK QFAM, use permutation procedure to correct for family structure when simple linear regression of traits on genotype that ignores family structure is performed. Other procedures such as those of QTDT that are based on variance components and are thus more appropriate for testing family-based associations also use permutations to derive empirical significance measures that correct for bias and multiple testing. The choice of permutation methods however varies. Objective: The aim of our study was to apply singular value decomposition (SVD) to permutations in family-based association tests for quantitative traits. Our approach is based on permuting trait values from which dependence among relatives has been eliminated. First, residuals are extracted from covariate-adjusted polygenic model for the study trait and then computation of SVD of the matrix of coefficients of relationship for each family is done. Eigenvectors and singular values from the SVD are then used to transform values of the trait to new values which are independent among related individuals. The new trait values are permuted across families in a way that satisfies required exchangeability for permutations and then before testing for association in each permuted dataset, phenotypic dependence among relates is re-established. We present results of our application of this approach to family-based association tests for anthropometric traits and blood pressure using the QTDT software.

153

Mixed Model Association Mapping in Admixed Populations

Timothy A. Thornton (1)

(1) Department of Biostatistics, University of Washington, Seattle, Washington, USA

Genetic association studies in recently admixed populations, such as African Americans and Hispanics, offer exciting opportunities for the identification of genetic variants that underlie phenotypic diversity. At the same time, heterogeneous genetic background and dependencies among sample individuals pose special challenges for complex trait mapping in admixed populations. Linear mixed models (LMMs) have garnered significant attention as a powerful approach for genetic association testing in the presence of sample structure, including population stratification, family structure and/or cryptic relatedness. Existing implementations of LMMs, however, may not appropriately account for the diverse genomes of admixed individuals. In this talk, we propose MMAAPS, a LMM method that appropriately accounts for sample structure in samples from admixed populations by (1) using individual-specific allele frequencies at SNPs that are calculated on the basis of ancestry derived from whole-genome analysis, and (2) partitioning recent and more distant genetic relatedness into two separate components. In simulation studies we demonstrate that MMAAPS provides improved type-I error rates and power over widely used LMM methods, such as EMMAX and GEMMA. The utility of MMAAPS is further demonstrated with applications to the Hispanic Community Health Study / Study of Latinos for genetic association mapping of hematology phenotypes.

154

Illustrating, quantifying and correcting for bias in post-hoc analysis of gene-based rare variant tests of association

Nathan Tintle (1) Kelsey Grinde (2) Alden Green (3) Michael O'Connell (4) Jaron Arbet (4) Alessandra Valcarcel (5) Jason Westra (1)
(1) Dordt College (2) University of Washington (3) Harvard University (4) University of Minnesota (5) University of Pennsylvania

To date, gene-based rare variant testing approaches have focused on aggregating information across sets of variants to maximize statistical power in identifying genes showing significant association with diseases. Beyond identifying genes that are associated with diseases, the identification of causal variant (s) in those genes and estimation of their effect is crucial for planning replication studies and characterizing the genetic architecture of the locus. However, we illustrate that straightforward single-marker association statistics suffer from bias introduced by conditioning on gene-based test significance, a phenomenon often referred to as "winner's curse." We illustrate the ramifications of this bias on power and type I error of single-marker association tests, outline parameters of genetic architecture that affect this bias, and propose a bootstrap resampling method to correct for this bias.

155

A non-parametric method for joint association analysis of sequencing and Imaging data

Xiaoran Tong (1) Qing Lu (1)
(1) Department of Epidemiology and Biostatistics, Michigan State University

The rapid development of whole genome sequence (WGS) technology coupled with magnetic resonance image (MRI) data mandates the development of analytical methods that are capable of utilizing both WGS and MRI data to identify predictive biomarkers associated with neurodegenerative diseases, such as Alzheimer's disease. The rich WGS/MRI data, however, brings the issue of "the curse of dimensionality" due to the vast number of sequencing variants and brain surface vertexes. In this work, we tackled the dimensionality issue of MRI data through a stacked denoising autoencoder (SDA) constructed using the deep learning algorithm, which reduces the dimensionality and maintains the majority of the information. For the WGS data, we use a weighted identity-by-state (IBS) kernel to aggregate information over multiple sequencing variants in a genetic region. A weighted U statistic is then used to evaluate the joint association of both imaging and sequencing data with the phenotype of interest. We show that our method maintains the correct type 1 error rate, while achieving high statistical power in comparison to methods using either sequencing or image data alone. To illustrate our approach, we apply the proposed method to the sequencing and image data from the Alzheimer's Disease Neuroimaging Initiative.

156

BMI as an effect modifier of a novel triglyceride-associated epigenetic mark

Vinh Truong (1) Nora Zwingerman (1) Irfahan Kassam (1) Dylan Aïssi (2) Jessica Dennis (1) Michael Wilson (3) Phillip Wells (4) Pierre-Emmanuel Morange (5) France Gagnon (1)
(1) Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada (2) INSERM, UMR's 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, Paris, France (3) Genetics and Genome Biology Program, SickKids Research Institute, Toronto, Canada (4) Department of Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Canada (5) INSERM, UMR's 1062, Nutrition Obesity and Risk of Thrombosis, Aix-Marseille University, Marseille, France

In a candidate gene study using biobanked blood DNA, we reported association of DNA methylation marks (DNAm) in the *CPT1A* gene with triglyceride (TG) levels. The *CPT1A* gene encodes for a protein expressed in liver that is essential for fatty acid oxidation, which supports the role of DNAm in TG regulation.
To further explore the role of epigenetic marks in the inter-individual variation of TG levels, we conducted a genome-wide investigation of blood DNAm in 214 individuals from

five large pedigrees ascertained on single probands with venous thrombosis (VT). Independent replication was tested in 350 unrelated VT cases. Models were adjusted for age, sex, cell type proportions; and for relatedness in the pedigree study. SNPs ± 1Mb of the top DNAm were tested for association. Sensitivity analyses for potential confounders and effect modifiers were conducted.
We replicated the recently reported DNAm-TG association in *ABCG1*, a lipid-related gene, and discovered a novel genome-wide significant association in the *PHGDH* gene, a glucose metabolism gene. The effect size at the *PHGDH* locus was stronger in the pedigree study ($\beta = -0.21$, $p = 2.3 \times 10^{-7}$) than in the unrelated cases study ($\beta = -0.08$, $p = 0.048$). A stratified analysis in the cases study revealed BMI as an effect modifier, with an effect only detected in individuals with BMI≥25 ($\beta = -0.17$, $p = 0.0098$ vs. normal BMI $\beta = -0.00$, $p = 0.99$). Five SNPs were significantly associated with the novel DNAm in both study samples.
The *PHGDH* locus is 34kb from *HMGCS2*, a gene with similar functions as *CPT1A*. HMGCS2 belongs to the HMG-CoA synthase family known for its pivotal role in cholesterol synthesis and ketogenesis. These findings may shed light on the molecular mechanisms underlying the obesity-related lipid regulation

157

A new 'front' in rule-based data mining for complex, heterogeneous, and noisy association analyses

Ryan J. Urbanowicz (1) Jason H. Moore (1)
(1) University of Pennsylvania

Biological and statistical phenomena such as epistasis and genetic heterogeneity can mask the relationship between genotypic and environmental risk factors/markers and phenotypes of interest. Most data mining or modeling approaches labor under restrictive assumptions such as the number of predictive variables, the application of a specific genetic model, linearity, or homogeneity in order to function effectively. Previously, we developed a rule-based machine learning algorithm called ExSTraCS for assumption-free classification, prediction, and knowledge discovery designed to be particularly advantageous in detecting, modeling and characterizing complex, noisy, multivariate, epistatic, and heterogeneous patterns of association. ExSTraCS flexibility comes from learning human interpretable rules that are evolved to individually capture subspaces of the overall pattern and collectively applied to form the predictive 'model'. One major challenge is to be able to compare and rank the 'value' of these evolved rules in a way that emphasizes both the accuracy and the correct coverage of the dataset in order to reduce overfitting and promote solution interpretability in noisy or heterogeneous problems. In the present study we introduce a Pareto-front-inspired methodology for the calculation of rule-fitness within ExSTraCS that provides a reliable, multi-objective global value metric. We find that this methodology significantly improves performance, interpretability, and allows for dramatic and simple rule compaction, across a spectrum of complex noisy simulation studies concurrently modeling epistatic and heterogeneous patterns with assorted heritabilities and sample sizes.

158

Evolving ancestry: The shift in individual ancestry composition over time

Digna R. Velez Edwards (1) Tracy L. MacGregor (1) Todd L. Edwards (1)
(1) Vanderbilt University

In observational studies and in clinical encounters, researchers and care providers frequently assign a single ancestry for each individual. We evaluated the genetic ancestry of individuals in a large cohort spanning over 100 birth years by evaluating ancestry informative markers (AIMs) and comparing proportions of racial ancestry with categorical race, as recorded in the electronic health record (EHR). Individuals included in this study are members of the BioVU repository, a cohort of DNA samples linked to deidentified EHRs. AIMs were extracted from the Illumina Infinium Human Exome Beadchip on 35,842 individuals. The data were evaluated using STRUCTURE software with the subjects from the 1000 Genomes Project, and the three ancestral clusters corresponding with African, European, and East Asian ancestry were further evaluated. We observed that individuals identified as White in the EHR have become more genetically admixed over time. The mean fractions of European ancestry have decreased for all cohorts by decade of birth from over 98% through the 1980's to 95.7% and 90.0% for the cohorts born in the 2000's and 2010's, respectively. In addition, the number of individuals with less than 70% attributable to any single ancestry has been increasing over time. These data indicate that younger cohorts of individuals exhibit higher levels of heterogeneity than older cohorts, particularly within those identified as White. We also observed increasing heterozygosity over time, with more rapid increases in younger individuals, implying increasing rates of migration of alleles between racial groups over time. Investigators studying traits in younger cohorts should consider these increasing levels of admixture when developing study designs.

159

The causal effect of adiposity on vascular dysfunction in healthy adolescents

Kaitlin H. Wade (1) Tauseef Khan (2) John E. Deanfield (2) Alun D. Hughes (3) Nish Chaturvedi (2) Abigail Fraser (1) Debbie A. Lawlor (1) George Davey Smith (1) Nicholas J. Timpson (1)
(1) Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, UK School of Social and Community Medicine, Faculty of Medicine and Dentistry, University of Bristol, Bristol, BS8 2BN, UK (2) Vascular Physiology Unit, Institute of Cardiovascular Science, University College London, London W1T 7HA, UK (3) Cardiometabolic

Phenotyping Group, Institute of Cardiovascular Science, University College London, London, W1T 7HA, UK

Adiposity is a known risk factor for vascular dysfunction in adults, with evidence that this effect may be present in early life. However, cross-sectional and case-control studies suffer from reverse causation, confounding and other sources of bias. In a large UK birth cohort, we used an allelic score comprising 97 genetic variants as an instrument to test the causal role of body mass index (BMI) on a range of vascular phenotypes measured at 11 and 18 years. We found positive effects of BMI on pulse rate (0.70bpm; 95% CI: 0.25, 1.16; $P = 2.51 \times 10^{-03}$), and both brachial artery diameter (0.04cm; 95% CI: 0.02, 0.05; $P = 5.49 \times 10^{-09}$) and compliance (0.02mm2/mmHgx10-3: 95% CI: −0.003, 0.04; P = 0.09) at age 11, and on left atrial size (0.08cm; 95% CI: 0.05, 0.11; $P = 1.03 \times 10^{-08}$), left ventricular diameter during diastole (0.04cm; 95% CI: 0.01, 0.07; P = 0.004), left ventricular mass (1.00g/m2.7; 95% CI: 0.64, 1.36; $P = 5.10 \times 10^{-08}$) and e', the peak velocity of the lateral mitral annulus in early diastole, (0.01cm/s; 95% CI: 0.0004, 0.03; P = 0.04) at age 18. We found inverse associations (not in the hypothesised causal direction) of BMI on pulse wave velocity (PWV) at 11 and 18 (-0.01m/s; 95% CI: −0.01, −0.002; P = 0.01 at age 11 and −0.01m/s; 95% CI: −0.02, −0.002; P = 0.01 at age 18). Although our results suggest that increased BMI is causally related to adverse cardiovascular health in childhood and adolescence, we found a paradoxical benefit of greater BMI on brachial artery compliance and PWV, supporting a complex aetiology in cardiovascular health. Replication of these findings using Mendelian randomization may help to better understand how adiposity influences cardiovascular health across the lifecourse.

160

A Near-Optimal Test of Association that Combines Case-Control and Affected Pedigree Designs

Meng Wang (1) William C. L. Stewart (1,2)
(1) Battelle Center for Mathematical Medicine at The Research Institute, Nationwide Children's Hospital, Columbus, OH, USA (2) Department of Statistics and Pediatrics, Ohio State University, Columbus, OH, USA

We present a new test of association that increases power by making more efficient use of heterogeneous data. Our proposed test–POPFAM+ uses near-optimal weights in a meta-analysis approach to accumulate evidence for association across case-control and family-based designs. Specifically, POPFAM+ identifies alleles that are both over-represented in cases, and preferentially transmitted from heterozygous parents to their affected offspring. Furthermore, POPFAM+ can accommodate any number of association tests (e.g. WQLS, PDT, GDT, etc.), provided that each test is normally distributed with known mean and variance under the null hypothesis. From the analysis of simulated data, we show that type I error is controlled, and that POPFAM+ is more powerful than GDT (generalized disequilibrium test) and WQLS (quasi-likelihood score test). Therefore, in addition to detecting common risk variants, POPFAM+ can also detect rare variants, especially if the minor allele frequency is greater than 2.5%. To demonstrate the utility of our approach, we applied POPFAM+ to an existing epilepsy data set comprised of affected families, independent cases, controls, and publicly available reference samples. Our secondary analysis of these heterogeneous data narrowed the region of interest significantly, which shows that POPFAM+ has the potential to dramatically influence the discovery of disease genes.

161

Empirical error rate estimation approach in next generation short reads sequencing

Jian Wang (1) Xuan Zhu (1) Bo Peng (1) Sanjay Shete (1)
(1) UT MD Anderson Cancer Center

During the past decade, next generation sequencing has developed rapidly, and has been harnessed by investigators to address a diverse range of biological problems such as quantification of gene expression and polymorphism and mutation discovery etc. The error rates are often high for next generation sequencing which affects the downstream genomic analysis. Recently, Wang et al. (2012) proposed a shadow regression approach to estimate the error rates based on the assumption of linear relationship between the counts of reads sequenced and the counts of reads containing errors (denoted as shadows). However, the linear assumption about the read-shadow relationship may not be valid. Therefore, it is necessary to estimate the error rates in a more accurate way. We proposed the empirical error rate estimation approach, which employs the cubic and robust smoothing splines to model the read-shadow relationship. We performed simulation studies using a frequency-based approach to generate the read and shadow counts directly which can mimic the real sequence counts data structure. We investigated the performance of the proposed approach and compare it to the shadow linear regression. From the simulation results, we observed that the proposed approach provided more accurate error rate estimations than the shadow linear regression for all the scenarios. We applied the proposed approach to assess the error rates for the sequence data from MicroArray Quality Control project, mutation screening study, Encyclopedia of DNA Elements project and bacteriophage PhiX DNA samples.

162

Joint analysis of multiple traits in rare variant association studies

Zhenchuan Wang (1) Xuexia Wang (2) Qiuying Sha (1) Shuanglin Zhang (1)
(1) Michigan Technological University (2) University of Wisconsin-Milwaukee

The joint analysis of multiple traits has recently become popular since it can increase statistical power to detect genetic variants and there is increasing evidence showing that

pleiotropy is a widespread phenomenon in complex diseases. Currently, most of existing methods for the joint analysis of multiple traits are to test association between one common variant and multiple traits. However, the variant-by-variant methods for common variant association studies may not be optimal for rare variant association studies due to the allelic heterogeneity as well as the extreme rarity of individual variants. Current statistical methods for rare variant association studies are for one single trait only. In this paper, we propose an Adaptive Weighting Reverse Regression (AWRR) method to test association between multiple traits and variants (can be both common and rare) in a genomic region. AWRR is robust to the directions of effects of causal variants and is also robust to the directions of association of traits. Using extensive simulation studies, we compare the performance of AWRR with canonical correlation analysis (CCA) and other two methods. Our results show that, in all of the simulation scenarios, AWRR is consistently more powerful than CCA. In most scenarios, AWRR is more powerful than the other two methods.

163

Kernel machine association testing for longitudinally-measured quantitative phenotypes

Zhong Wang (1) Ke Xu (2) Zuoheng Wang (2)
(1) Cornell University (2) Yale University

Recent developments in high-throughput sequencing technologies have made it possible to search for both rare and common genetic variants associated with complex diseases. Many phenotypes in health studies are measured at multiple time points. The rich information on repeated measurements on each subject not only provides a more accurate assessment of disease condition, but also allows us to explore the genetic influence on disease onset and progression. However, most association tests mainly focus on a single time point. To address this limitation, we propose LSKAT (Longitudinal Sequence Kernel Association Test), a region-based variants association test for longitudinal data, which extends the SKAT method for a single measurement to repeated measurements. LSKAT uses several variance components to account for the within-subject correlation structure of the longitudinal data, and the contributions from all genetic variants (common and rare) in a region. Additionally, we propose another test LMSKAT (Longitudinal Multi-Kernel Association Test) which allows for the time-varing genetic effects by using multiple kernels to detect genes affecting the temporal trends of the trait. In simulation studies, we evaluate the performance of LSKAT and LMSKAT, and demonstrate that they have improved power, by making full use of multiple measurements, as comparing to previously proposed tests on a single measurement or average measurements for each subject. We apply LSKAT and LMSKAT to testing with body mass index in Framingham heart study.

164

A fast and effective W-test for SNP-SNP interaction identification in GWAS with application on Bipolar disorder

Maggie H. Wang (1) Rui Sun (1) Inchi Hu (2) Pak Sham (3) Benny C. Y. Zee (1)
(1) The Chinese University of Hong Kong (CUHK) (2) Hong Kong University of Science and Technology (HKUST) (3) The University of Hong Kong (HKU)

Genetic association study has the objective of identification of disease susceptible loci from genome-wide data. One of the challenges is identifying epistasis effect underlies complex diseases. Many interaction methods were suggested, seeking a balance among the complexity of methods, efficiency of computation, and interpretability of results. In this talk, we will introduce a fast and effective W-test, which has an odds ratio interpretation comparing the distributional difference between cases and controls. We will demonstrate its superior performance among alternative methods in simulated data sets under different genetic architectures and varying sample size. The W-test's power is especially strong in moderate minor allele frequency environment, and robust to smaller sample sizes. The method is applied on two real GWAS bipolar disorder data sets from the United States and Europe. Besides finding genes with important bipolar association that are previously identified through biological experiments, the W-test can replicate gene-gene interaction from the two independent GWAS data sets. The validated gene pairs reside in key neuro-function pathways, which cannot be found through main effect. They contribute to the completion of the genetic heritability picture of bipolar disorder, and provide valuable pharmaceutical targets for treatment of the disease.

165

Gene and pathogenic variant discovery for Mendelian and Complex Familial Traits

Gao T. Wang (1) Hang Dai (1) Bo Peng (2) Regie Lyn Pastor Santos-Cortez (1) Suzanne M. Leal (1)
(1) Baylor College of Medicine (2) University of Texas M.D. Anderson Cancer Center

We provide methods and implementation of best practices to identify rare variants involved in the etiology of Mendelian and familial complex traits using next-generation sequence (NGS) data. We demonstrate through case examples, bioinformatics protocols to analyze exome and genome sequence data to elucidate pathogenic variants that are either de novo, underlie Mendelian phenotypes or complex traits with familial aggregation. We feature four types of commonly adopted study designs analyzing either genomes or exomes from (1) a single affected individual (2) multiple family members, (3) multiple families and (4) multiple unrelated individuals with a family history of disease. For each design we illustrate the procedures to integrate data from different sources, performing variant annotations, and selecting potentially pathogenic variants from single or multiple exomes based on several parameters, including but not limited to mode of inheritance, variant sharing among pedigree members, population minor allele frequency, functional annotation and prediction,

linkage mapping data and variants/genes previously implicated in disease etiology. Case studies were novel pathogenic variants have been discovered for nonsyndromic hearing impairment, thoracic aortic aneurysms and dissections, autism, Moyamoya disease, otitis media and rare autosomal recessive traits such as achromatopsia and trichothiodystrophy are used to illustrate the protocols and best practices. We have also developed an easy-to-use bioinformatics software, Variant Mendelian Tools (VMT), to implements the protocols which makes it possible for any researcher with NGS data to efficiently hunt down and identify pathogenic variants. VMT is designed to be flexible to accommodate regular updates from annotation databases and incorporation of new information from a variety of sources including public databases and in-house data, e.g. linkage regions. The analysis protocols we developed are distributed under the VMT platform which can be readily adapted and shared for a variety of projects, owing to the compact, human-readable syntax that VMT adopts. Our work is highly beneficial to clinicians and researchers who aim to identify pathogenic variants from NGS data but have minimal knowledge and experience in the use of Linux and programming languages, and/or in annotations and variant discovery using family data.

166

Genome-Wide Survey in African Americans Demonstrates Widespread Epistasis of Fitness in the Human Genome

Heming Wang (1) Yoonha Choi (2) Bamidele Tayo (3) Xuefeng Wang (4) Xiang Zhang (5) Uli Broeckel (6) Craig L. Hanis (7) Sharon L. R. Kardia (8) Susan Redline (9) Richard S. Cooper (3) Hua Tang (2) Xiaofeng Zhu (1)
(1) Department of Epidemiology and Biostatistics, Case Western Reserve University (2) Department of Genetics, Stanford University (3) Department of Public Health Science, Loyola University Medical Center (4) Departments of Preventive Medicine, Biomedical Informatics, and Applied Mathematics and Statistics, Stony Brook University (5) Department of Electrical Engineering and Computer Science, Case Western Reserve University (6) Human and Molecular Genetics Center, Medical College of Wisconsin (7) Department of Epidemiology, Human Genetics and Environmental Sciences, University of Texas Health Science Center at Houston (8) Department of Epidemiology, University of Michigan (9) Department of Medicine, Harvard Medical School

The role played by epistasis between alleles at unlinked loci in shaping population fitness has been debated for many years and the existing evidence has been mainly accumulated from model organisms. In model organisms, fitness epistasis can be systematically inferred by detecting non-independence of genotypic values between loci in a population and confirmed through examining the number of offspring produced in two-locus genotype groups. No systematic study has been conducted to detect epistasis of fitness in humans owing to the experimental constraints. In the African-American population gene flow from European into African ancestries creates statistical properties in the genome similar to what is observed in recombinant inbred lines. We demonstrate theoretically that fitness epistasis can create correlation of local ancestry between unlinked loci that can be examined. We inferred local ancestry across the genome in 16,252 unrelated African Americans and systematically examined the pairwise correlations between the genomic regions on different chromosomes. Our analysis revealed 37 pairs of genomic regions showing local ancestry correlation (p-value $< 6 \times 10^{-8}$) that can be potentially attributed to fitness epistasis. These genomic regions also harbored multiple genes with strong evidence of selection, including the Duffy locus and the glycosylphosphatidylinositol-anchored serine protease (*PRSS21*) that is associated with sperm-dysfunction. To our knowledge, this study is the first to systematically examine evidence of fitness epistasis across the human genome. Our results demonstrate that fitness epistasis is widespread in humans and may have an important impact on current efforts to map susceptibility genes.

167

mFARVAT: FAmily-based Rare Variant Association Test for multivariate phenotypes

Longfei Wang (1) Dandi Qiao (2,3) Michael Cho (2,4) Edwin K. Silverman (2,4) Sungho Won (1,5)
(1) Interdisciplinary Program in bioinformatics, Seoul National University, Seoul, 151-742, Korea (2) Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA (3) Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA (4) Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA (5) Graduate School of Public Health, Seoul National University, Seoul, 151-742, Korea

Family-based designs have been repeatedly shown to be powerful in detecting the significant rare variants associated with human disease. Furthermore, human diseases are often outcomes of multiple phenotypes, and thus we expect multivariate, family-based analyses to be efficient in detecting associations with rare variants. However, few statistical methods implementing this strategy have been developed. In this report, we describe one such implementation: the multivariate family-based rare variant association tool (mFARVAT).
mFARVAT is a quasi-likelihood-based score test for the rare variant association analysis with multiple phenotypes, and tests homogeneous and heterogeneous effects of each variant on multiple phenotypes. Simulation results show that the method is generally robust and efficient for various disease models, and we identify some promising candidate genes associated with chronic obstructive pulmonary disease.
Availability and Implementation: The software is freely available athttp://healthstat.snu.ac.kr/software/mfarvat/, implemented in C++ and supported on Linux and MS Windows.

168

Investigating the Association of Rare Genetic Variants with Blood Pressure traits

Helen R. Warren (1,2) Praveen Surendran (3) Fotios Drenos (4,5) Robin Young (3) James P. Cook (6,7) Alisa K Manning (8,9,10) Niels Grarup (11) Xueling Sim (12,13) Danish Saleheen (3,14,15) Folkert W. Asselbergs (16,17,18) Cecilia M. Lindgren (9,19,20) John Danesh (3,21) Louise V. Wain (6) Adam S. Butterworth (3) Joanna M. M. Howson (3) Patricia B. Munroe (1) CHARGE+ Consortium, T2D-GENES Consortium, GoT2DGenes Consortium, ExomeBP Consortium, CHD Exome+ Consortium

(1) Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK (2) NIHR Barts Cardiovascular Biomedical Research Unit, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK (3) Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, UK (4) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, UK (5) Centre for Cardiovascular Genetics, Institute of Cardiovascular Science, Rayne Building University College London, UK (6) Department of Health Sciences, University of Leicester, UK (7) Department of Biostatistics, University of Liverpool, UK (8) Department of Genetics, Harvard Medical School, Boston, USA (9) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA (10) Department of Molecular Biology, Massachusetts General Hospital, Boston, USA (11) The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark (12) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, USA (13) Saw Swee Hock School of Public Health, National University of Singapore (14) Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, USA (15) Center for Non-Communicable Diseases, Karachi, Pakistan (16) Department of Cardiology, University Medical Center Utrecht, Netherlands (17) Durrer Center for Cardiogenetic Research, ICIN-Netherlands Heart Institute, Utrecht, the Netherlands (18) Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, UK (19) The Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK (20) Wellcome Trust Center for Human Genetics, University of Oxford, UK (21) Wellcome Trust Sanger Institute, UK

GWAS analyses for blood pressure (BP) have primarily identified common variants with small effects, explaining < 5% of the total trait variance. The missing heritability may in part be explained by rare variants.

We have investigated associations of rare variants in the largest genetic study of BP to date. A total of ~350,000 individuals were genotyped at ~250,000 rare, low-frequency and common variants using the Exome Chip. We analysed medication-adjusted systolic BP, diastolic BP and pulse pressure as quantitative traits and hypertension as a binary trait. Single-variant analyses were performed in our discovery data (N~195,000) and 81 variants were selected for replication in independent samples (N~155,000). Loci were declared novel if they were Bonferroni significant in the replication data, or if the combined meta-analysis of discovery and replication data was genome-wide significant. We identified 30 novel loci associated with BP, and confirmed previously reported BP associations. For the first time, rare missense variants with large effects on BP were identified in the general population. Conditional tests using RareMetalWorker (RMW) found multiple independent signals within 5 novel loci, and new independent associations at 5 known BP loci, including a rare nonsense variant in *ENPEP*. Gene-based tests in RMW across different MAF thresholds identified 2 further novel genes, and found one gene, *ENPEP* from a known BP region, with evidence of multiple rare variant associations.

Pathway analyses showed an enrichment of BP genes at loci previously associated with cardiac abnormalities. Our results highlight novel rare variants in BP regulation, and combined with expression and functional data, pinpoint potential causal genes.

169

Robust Association Testing for Quantitative Traits and Rare Variants

Peng Wei (1) Ying Cao (1) Zhiyuan Xu (2) Yiwei Zhang (2) Jacy Crosby (1) Eric Boerwinkle (1) Wei Pan (2)

(1) Human Genetics Center, University of Texas School of Public Health (2) Division of Biostatistics, University of Minnesota

With the advance of sequencing technologies, it has become a routine practice to test for association between a quantitative trait and a set of rare variants (RVs). However, to our knowledge, there is no study yet on the robustness of association testing to the non-Normality of the observed traits, e.g., due to skewness, which is expected to be ubiquitous for quantitative traits. By extensive simulations, we demonstrate that SKAT and SKAT-O are not robust to heavily-tailed or right-skewed trait distributions with inflated Type I error rates; in contrast, our recently proposed SPU tests and their adaptive version called aSPU test, are much more robust. We further propose a robust version of the SPU and aSPU tests, denoted as SPUr and aSPUr. We demonstrate that in most situations the aSPU test achieves similar power to that of SKAT or SKAT-O when testing on a smaller number of RVs, while, for a larger number of RVs, the aSPU test is often more powerful than others, owing to its high data-adaptivity. We also compare different tests by conducting association testing of triglyceride using the NHLBI ESP exome sequencing data and brain imaging phenotypes using the ADNI whole genome sequencing data. In the former, the QQ plots for SKAT, SKAT-O and T1 tests of the untransformed triglyceride were severely

inflated, while, in the latter, SKAT could have severely inflated QQ plots even with Log or inverse-normal transformed imaging phenotypes. In both cases, aSPU and aSPUr controlled the lambda well. Due to its relatively high robustness to outliers and high power of the aSPU test, we recommend its use complementary to SKAT and SKAT-O. If there is evidence of inflated Type I error rate from the aSPU test, we would recommend the use of the more robust aSPUr test.

170

A stochastic search algorithm for finding multi-SNP effects using nuclear families

Clarice R. Weinberg (1) Min Shi (1) Alison Wise (3) David M. Umbach (1) Leping Li (1)
(1) National Institute of Environmental Health Sciences (NIEHS)

Given that biologic systems typically involve failure-resistant redundancy, phenotypes such as birth defects may occur only through the joint effects of exposures and several genetic variants. Such joint effects tend to produce a weak signal in typical GWAS analyses that assess only single-SNP associations. We describe an approach that uses case-parent triads and applies an "evolutionary" algorithm to stochastically search the large sample space that includes all sets of size S of potentially-interacting SNPs. We assess the performance of our algorithm using simulated but realistic data from the dbGaP GWAS data on families with the birth defect oral cleft. We simulate specific multi-SNP causal effects and then try to recover those causative complexes de novo. Initial simulations using our method show promising results in scenarios that involve two sets of four interacting SNPs, against a background of random cases. We are applying the refined method, using the original cleft data to explore a large set of candidate SNPs for epistatic effects related to risk of oral cleft.

171

Integrating Genotype, RNA Sequencing, and DNA Methylation Data to Investigate the Role of X Chromosome Inactivation in Ovarian Cancer

Stacey J. Winham (1) Nicholas B. Larson (1) Sebastian M. Armasu (1) Zachary C. Fogarty (1) Melissa C. Larson (1) Brooke L. Fridley (2) Ellen L. Goode (1)
(1) Mayo Clinic (2) University of Kansas

X chromosome inactivation (XCI) randomly silences transcription of one of the two homologous copies of the female X chromosomes to equalize levels of gene expression with males. Studies suggest that XCI is skewed (non-random) in lymphocytes from epithelial ovarian cancer (EOC) patients, but its role in tumors is unknown. We integrated germline genotype with tumor RNA sequence and DNA methylation data to examine the role of XCI in 113 EOC patients. By combining RNA-Seq and imputed genotype data, we measured allele-specific expression (ASE) for 481 X chromosome genes and identified the active alleles for each tumor. XCI is regulated by DNA methylation, so we also leveraged tumor methylation data of the X chromosome promoters to guide inference for genes that did not have sufficient ASE reads. X chromosome ASE and methylation patterns were unique, showing a higher degree of ASE imbalance and 50% methylated CpG sites compared to the autosomes. Because ASE imbalance was associated with promoter methylation ($P < 10^{-6}$), we predicted ASE imbalance based on promoter methylation using a mixed effects model. To assess global XCI patterns across 718 X chromosome genes, we performed separate clustering via non-negative matrix factorization on the promoter methylation, observed and predicted ASE imbalance data and consistently identified two XCI clusters. Methylation clusters were associated with clinical factors, including tumor histology ($P = 0.03$), stage ($P = 0.003$), surgical debulking ($P = 0.002$), and time to recurrence ($P = 0.01$). Tumors with less promoter methylation had shorter time to recurrence (HR = 1.84), after covariate adjustment. These results suggest that XCI may play a role in EOC, but future studies to examine somatic changes are needed.

172

A systematic evaluation of approaches for stratified genome-wide association meta-analyses to identify gene-strata interaction effects

Thomas W. Winkler (1) Zoltan Kutalik (2,3) Iris M. Heid (1)
(1) Department of Genetic Epidemiology, University of Regensburg, 93053 Regensburg, Germany (2) Institute of Social and Preventive Medicine, CHUV-UNIL, 1010 Lausanne, Switzerland (3) Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Genome-wide association meta-analyses (GWAMA) stratifying for a dichotomous non-genetic factor (S) have successfully identified a multitude of gene-strata (G x S) interaction effects, such as gene-sex interactions for obesity traits. However, there is still uncertainty about what screening approach is ideal for a given scenario and a systematic evaluation of approaches is lacking.

We here performed exhaustive simulations and analytical computations to evaluate type 1 error and power of different stratified GWAMA approaches to identify G x S interaction. We employed a stratum-difference test to infer G x S interaction from the stratified model and considered various study designs (i.e., balanced and unbalanced strata designs), various types of interactions (i.e., qualitative, pure and quantitative), as well as realistic sample size and effect size configurations.

For balanced strata designs, the genome-wide screen for stratum-difference is the best approach to detect qualitative interaction; interestingly, the approach of filtering for overall association followed by a test for stratum-difference is the best approach for pure and quantitative interaction. Filtering for joint association prior to an interaction test renders type 1 error control very difficult when both steps are conducted in the same data set.

For unbalanced stratum sizes our power computations yielded similar recommendations and demonstrated that it is generally more likely to identify interactions with stronger effect in the larger stratum.
In summary, our recommendations may guide future genome-wide G x S interaction screens. All considered approaches have been implemented in our R package EasyStrata.

173

Enabling improved low frequency variant imputation in multi-ethnic studies

Genevieve L. Wojcik (1) Christopher R. Gignoux (1) Christian Fuchsberger (2) Daniel Taliun (2) Ryan Welch (2) Alicia R. Martin (1) Henry R. Johnston (3) Suyash Shringarpure (1) Charit Pethiyagoda (4) Jared O'Connell (4) Luana McAuliffe (4) Zhaohui S. Qin (3) Kathleen C. Barnes (5) Goncalo Abecasis (2) Christopher S. Carlson (6) Hyun M. Kang (2) Michael Boehnke (2) Carlos D. Bustamante (1) Eimear E. Kenny (7)
(1) Department of Genetics, Stanford University School of Medicine, Stanford CA (2) Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor MI (3) Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta GA (4) Illumina, Inc, 5200 Illumina Way, San Diego CA (5) Department of Medicine, Johns Hopkins University School of Medicine, Baltimore MD (6) Fred Hutchinson Cancer Research Center, Seattle WA (7) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York NY

The emergence of large sequenced reference panels in the past year has facilitated a new focus on accurate imputation of low frequency variants (LFV; 0.005-0.05 MAF) with a new generation of genotyping arrays. Tag SNP selection poses several challenges as LFVs tend to be continentally or even population specific. To address these challenges, we have developed a novel algorithm to select tag SNPs considering both population-specific and trans-ethnic tagging performance, maximizing imputation accuracy rather than pairwise coverage. This is achieved via a leave-one-out internal validation approach that allows direct comparison of tag SNP performance for each iteration of potential scaffold sites. This method was recently applied to boost diverse population coverage for the Multi-ethnic Genotyping Array (MEGA), a collaboration between Illumina and multiple consortia (PAGE, CAAPA, T2D Genes), by leveraging the whole genome sequences available in the 1000 Genomes Project Phase 3 (TGP) release. We will present results that explore the effect of minimum r^2 and minor allele frequency (MAF) threshold for tag SNP selection using various methods of prioritization across the 26 TGP populations. By prioritizing tags that contribute information across multiple populations, this method provides enhanced imputation accuracy compared to naive selection. When applied to the 1.5M GWAS scaffold on MEGA, imputation accuracy was > 0.89 for SNPs with MAF $\geq$ 0.005 and $\geq$0.95 for MAF $\geq$ 0.05 in all 6 TGP continental populations. This unified framework for tag SNP selection and imputation evaluation will be useful for designing reference panels, large multi-ethnic epidemiological studies and biobanks, as well as future biological repositories.

174

Identifying shared genetic risk for head circumference and ASD using genetic linkage in combination with exome sequencing

Marc R. Woodbury-Smith (1) Andrew D. Paterson (2) Hilary Coon (3)
(1) McMaster University (2) University of Toronto (3) University of Utah

The aim of this study was to identify heritable genetic loci for head circumference in pedigrees segregating Autism Spectrum Disorder (ASD). Sixty-seven pedigrees (N = 612) having at least two family members with ASD were recruited in Utah as part of ongoing projects exploring the genetics of ASD. The pedigrees comprised 20 large extended pedigrees of 6–9 generations, along with smaller multigenerational pedigrees. Subjects were genotyped using the CIDR 6k SNP linkage panel and head circumference was available for at least one time point. Genome-wide linkage analysis of this discovery sample revealed suggestive evidence of linkage at 6p22.3 (LOD = 2.7) when covarying for age, sex and height. The signal became attenuated when height was removed from the analysis.
For replication, we used the Autism Genome Project (AGP) sample of 1,397 multiplex ASD pedigrees recruited from ten sites in North America and Europe, excluding any samples overlapping with our discovery sample. Genotypes were obtained using the Affymetrix 10K SNP arrays. The results of our genome-wide linkage analyses were consistent with suggestive evidence of linkage at 1q25.3 (LOD = 2.6), covarying for age and sex with or without height, with no evidence of linkage at 6p.
Samples from eight of the Utah extended pedigrees (60 relatives with autism spectrum disorder and 26 unaffected relatives) were also exome sequenced. Variants shared between individuals, and thereby possibly explaining the linked region, were annotated for both the complete sample, and for those pedigrees contributing most strongly to the linkage signal. These results, and in particular the implication of these two regions for our understanding of the genetics of brain growth and ASD, are presented and discussed

175

Detecting Gene-Gene Interactions for Cleft Lip with/without Cleft Palate in Targeted Sequencing Data

Yanzi Xiao (1) Terri H. Beaty (1) Margaret A. Taub (2) Ferdouse Begum (1) Jacqueline B. Hetmanski (1) Meg M. Parker

(1) Alan F. Scott (3) Ingo Ruczinski (2) Holger Schwender (4) Mary L. Marazita (5) Elizabeth J. Leslie (5) Dan Kobaldt (6) Jeff Murray (7)
(1) Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University (2) Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University (3) Institute of Genetic Medicine, School of Medicine, Johns Hopkins University (4) Mathematical Institute, Heinrich Heine University (5) Department of Oral Biology, School of Dental Medicine, University of Pittsburgh (6) The Genome Institute, Washington University School of Medicine (7) Department of Pediatrics, Carver College of Medicine, University of Iowa

Non-syndromic cleft lip with or without cleft palate (NSCL/P) is the most common craniofacial birth defect in humans, affecting 1 in 700 live births. This malformation has a complex etiology where multiple genes and several environmental factors influence risk. At least a dozen different genes have been confirmed to be associated with risk of NSCL/P in previous studies. Several authors have suggested gene-gene (GxG) interaction may be important in the etiology of this complex and heterogeneous malformation.
We aimed to detect gene-gene interactions for cleft lip with/without cleft palate in targeted sequencing data.
We used targeted sequence data on 13 regions identified by previous studies spanning 6.3 MB of the genome in a study of 1,498 case-parent trios. We used R-package Trio to perform a likelihood ratio test (LRT) to test for GxG interaction in both a 1 df test and a 4 df test. To adjust for multiple testing, permutation test was performed to generate empiric p-values. The most significant 4df LRT was seen with rs6029315 in *MAFB* and rs6681255 in *IRF6* ($p = 3.8 \times 10^{-8}$) in the European group, which remained significant ($p = 0.02$) after correcting for multiple comparison via permutation tests. Only 2% of replicates generated under the null hypothesis exceeded this observed test statistic. However, we found no pairwise interaction yielding an empirical $p < 0.05$ in the Asian trio group.
Our results suggest that there is statistical GxG interaction between *IRF6* and *MAFB* in the European population. Because *IRF6* is the only gene that has shown consistency across different types of genetic studies, evidence of statistical interaction between markers in/near the genes *IRF6* and *MAFB* is especially interesting.

176

Modified Screening and Ranking Algorithm (modSaRa) for Copy Number Variation Detection

Feifei F. X. Xiao (1) Xiaoyi X. M. Min (2) Heping H. Z. Zhang (2)
(1) University of South Carolina (2) Yale School of Public Health

Copy number variation (CNV) is an important type of structural variation and is associated with human complex disorders, such as autism, growth retardation and HIV progression. Duplication or deletion on any of the two copies of genomic segments results in CNV of this region. To detect CNVs in human genome, several change-point based techniques have been proposed. However, these methods usually present high computational complexity, given that the data points are repeatedly used in the process of determining change-points along the same sequence. Moreover, some practical issues arise when these methods are applied to real data, such as handling the heavy-tailed distribution and identifying the biologically meaningful copy number states. In this study, we propose a modified screening and ranking algorithm (modSaRa). This algorithm is suitable for high-throughput genetic data due to its low computational complexity. The aforementioned issues in CNV detection are also addressed. First, modSaRa is robust to the violation of the normal assumption. Second, a novel approach using a normal mixture model coupled with a modified BIC criterion is proposed for filtering false positives and further clustering the potential CNV segments to copy number states. We compare modSaRa with an alternative method, circular binary segmentation (CBS). In both simulated and real data studies, modSaRa outperforms CBS in detecting CNVs with higher sensitivity and specificity.

177

Exome sequencing to identify the genetic bases for lysosomal storage diseases of unknown etiology

Jinchuan Xing (1) Nan Wang (1) Erika Gedvilaite (1) Yeting Zhang (1) Dibyendu Kumar (2) Robert Donnelly (3) David Sleat (4,5) Peter Lobel (4,5)
(1) Department of Genetics, Rutgers University (2) Waksman Institute of Microbiology, Rutgers University (3) Molecular Resource Facility, Rutgers - New Jersey Medical School (4) Center for Advanced Biotechnology and Medicine, Rutgers University (5) Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers University

Lysosomes are membrane-bound, acidic eukaryotic cellular organelles. As an enzyme container, they play important roles in the degradation of macromolecules. Monogenic mutations resulting in the loss of enzyme activities in the lysosome may lead to severe health problem, such as neurodegeneration, early death, etc. These conditions are categorized as lysosome storage diseases (LSDs). In some cases, mutations that result in atypical clinical presentation or defects in previously undescribed lysosomal disease genes complicate the identification of the underlying genetic defect. Here, we performed whole exome sequencing on 14 suspected LSD cases, with the goal of finding the causal mutations in each case. From the raw sequence data, we identified DNA variants using three variant discovery pipelines: the Genome Analysis Toolkit, LifeScope and CLC Genomics Workbench. We then used the Variant Annotation Analysis Selection Tool (VAAST) to prioritize disease-causing mutations in 848 candidate LSD genes. As a probabilistic disease gene finder, VAAST integrates allele

frequency, amino acid substitution severity and conservation information into a composite likelihood framework. A number of candidate variants have been identified and validated, and we are performing downstream proteomic analyses to investigate the potential connection between our candidate variants and LSDs. Our results will shed light on the genetic basis of LSDs.

178

A Novel Multiple-SNP Approach for Fine-Mapping Studies

Jingxiong J. X. Xu (1,2) Hilmi H. O. Ozcelik (3,4) Maciej M. K. Kwiatkowski (5) Bharati B. B. Bapat (6,7) Eleftherios E. P. D. Diamandis (7) Alexandre A. R. Z. Zlotta (8,9) Laurent L. B. Briollais (1,2)
(1) Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada (2) Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada (3) Fred A. Litwin Centre for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada (4) Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, ON, Canada (5) Department of Urology, Cantonal Hospital Aarau, Aarau, Switzerland (6) Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada (7) Department of Pathology and Laboratory Medicine, University Health Network, Toronto, ON, Canada (8) Department of Surgery, Urology. Mount Sinai Hospital. Toronto, ON, Canada (9) Department of Surgical Oncology, Urology. Princess Margaret Hospital, University Health Network. Toronto, ON, Canada

The clustering of tightly linked genetic markers is commonly observed in fine-mapping studies where the goal is to refine the location of a previously identified locus. Fine-mapping studies usually involve genotyping additional markers in the region of interest or alternatively imputing them using a reference genetic panel. Fine mapping studies provide opportunities to identify the causal loci underlying complex human diseases but also yield many statistical challenges from the complex correlations and high LD between the markers.
As an illustrative example, the Kallikrein (*KLK*) region on chromosome 19 entails an important family of genes, clustered in a region that spans 261,558 bp, that display significant homology to each other at both the nucleotide and protein levels. The Prostate Specific Antigen (PSA) marker is the serum level of the KLK3 protein. To better understand the implication of SNPs in the *KLK* region and their association with prostate cancer (PCa), we performed a fine-mapping study of the entire *KLK* region. The study included 123 original SNPs and 925 imputed SNPs from a sample of 920 cases and 904 controls.
To model the joint effect of SNPs and account for their complex dependence structure, we proposed a novel approach based on Bayesian Graphical Model (BGM) and an efficient algorithm, MOSS, to perform the model search and SNP selection. BGM combines graphical model with the Bayesian framework, and allows prior information to be used for inference. Our simulation studies assessed the performances of BGM. The data analysis showed two genetic paths leading to PCa, one involving PSA as intermediate phenotype and implicating the region between *KLK12* & *KLK13* and the other one independent of PSA, implicating the *KLK4* region.

179

Novel Association Testing Based on Genetic Heterogeneity in GWAS

Zhiyuan Xu (1) Wei Pan (1)
(1) University of Minnesota

The commonly used association tests in GWAS all aim to detect single SNPs with allele frequency (AF) differences between cases and controls. Although many disease-associated SNPs have been identified, they can only account for a small proportion of the genetic variance for a common disease. Two of several possible reasons are small effect sizes and genetic (or locus) heterogeneity, which are related to each other and have been confirmed by completed GWAS. Under genetic heterogeneity it is assumed that a complex disease is not caused by a single SNP or single gene, but by multiple variants in multiple functionally related genes forming a pathway, motivating the application of pathway analyses. However, as for single SNP tests for AF differences, in addition to incomplete annotations of biological pathways, existing pathway methods do not directly take advantage of existing genetic heterogeneity to boost statistical power. Here we propose and apply a single SNP-based test to explicitly account for genetic heterogeneity: it is assumed that a patient population of a complex disease can be decomposed into multiple subpopulations, each with a possibly different set of causal SNPs; under this model, even if the AF of a causal SNP for the whole patient sample is the same as that of the controls, in which the conventional AF tests have no power, our proposed test retains power to detect the causal SNP. We apply the proposed test to the WTCCC data, demonstrating that it can detect novel SNPs and loci missed by the conventional AF tests. For example, when applied to the WTCCC Crohn's disease data, the proposed test detected 128 significant SNPs, while the conventional trend test identified only 60 significant SNPs, among which 46 were common. In addition to annotated pathways, the new test can be generalized to detect de novo (i.e. un-annotated) disease-associated pathways. Its relationships with (and advantages over) some existing tests will be discussed.

180

Detecting Association of Rare and Common Variants based on Cross-Validation Prediction Error

Xinlan Yang (1) Shuanglin Zhang (1) Qiuying Sha (1)
(1) Michigan Technological University

Despite the extensive discovery of disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants can explain additional disease risk or trait variability. Although sequencing

provides a supreme opportunity to investigate the roles of rare variants in complex diseases, detection of these variants in sequencing-based association studies presents substantial challenges. In this article, we proposed a novel statistical test based on cross-validation prediction error (PE) to test the association between variants in a genomic region and a complex trait of interest. PE is based on the ridge regression of trait values on the genotypes of the variants in a genomic region. We then proposed to use a weighted combination of variants with two different weighting schemes, the weight suggested by Madsen and Browning (WS-PE) and the optimal weighting scheme (TOW-PE) in the PE approach. The proposed methods, WS-PE and TOW-PE, aim to test the genetic effects of both rare and common variants. Both of them allow covariates and can control for population stratification. We used extensive simulation studies to evaluate the type I errors of these two methods and to compare the powers of the proposed methods with existing WS and TOW. Simulation studies showed that proposed methods have correct type I error rates. Simulation results also showed that WS-PE is consistently more powerful than WS, and TOW-PE is consistently more powerful than TOW.

181

Integrated analysis of germline, omic and disease data

Zhao Yang (1) Duncan C. Thomas (1) David V. Conti (1,2)
(1) Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA (2) Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, CA, USA

The availability of various omics data, such as metabolites, expression, and somatic profiles, facilitates potentially new insights into the underlying etiologic mechanism of disease. However, such data presents many analytic challenges including effect heterogeneity and high dimensionality. Recently proposed methods for omic data often ignore the underlying causal relationships of the various data types and focus mostly on data reduction by estimating underlying clusters. Here, we present a novel approach for the integrated analysis of germline, omic and disease data. Via a specific directed acyclic graph (DAG), we use a latent variable to relate information from germline genetic data to either a continuous or binary disease outcome. Within a measurement error framework, the omic data is viewed as a flawed measure of underlying latent clusters, categorized to simplify interpretation. We use an expectation-maximization (EM) algorithm to simultaneously estimate the unobserved latent clusters and model parameters, including genetic effects on the latent cluster and the impact of the cluster on omic patterns and on the disease outcome. Additionally, we incorporate penalized methods for variable selection in a high dimensional setting for both the genetic data and the omic data. Using simulations, we demonstrate the ability of our approach to accurately estimate underlying clusters and their corresponding genetic, omic and disease effects. Moreover, we demonstrate the feasibility of the variable selection to identify genetic and omic factors as both the means and correlational structures are varied. We discuss extensions to accommodate ascertainment and missing data.

182

Pharmacogenetics of Acute Coronary Syndrome

Peng Yin (1) Andrea Jorgensen (1) Andrew P. Morris (1) Richard Turner (2) Richard Fitzgerald (2) Rod Stables (3) Anita Hanson (2) Munir Pirmohamed (2)
(1) Department of Biostatistics, University of Liverpool, UK (2) Department of Molecular & Clinical Pharmacology, University of Liverpool, UK (3) Liverpool Heart and Chest Hospital, Liverpool, UK

To identify loci associated with response to cardiovascular drugs, we have undertaken a genome-wide association study in 1470 patients recruited to a UK prospective pharmacogenetic trial of acute coronary syndrome (PHACS). Approximately 8% of the patients had another cardiovascular event in one year after hospital discharge, including myocardial infarction (MI) or stroke, and in some cases resulting in death. We began by identifying clinical risk factors for drug response: age ($p = 3.2 \times 10^{-7}$), prior MI ($p = 0.0048$), diabetes ($p = 0.032$), ACE inhibitor use pre-admission ($p = 0.0072$), aspirin use at discharge ($p = 0.0066$) and gender ($p = 0.12$). Patients were genotyped using the Illumina OmniExpress array. After quality control, the genotype scaffold was imputed up to the 1000 Genomes Phase I reference panel (all ancestries, March 2012 release). We tested for association of SNPs with outcome under an additive dosage model in a logistic regression framework after adjusting for the clinical factors identified above and principal components to account for population structure. Variants mapping to the LRPPRC gene demonstrated strong evidence of association: lead SNP rs65544733, minor allele frequency 0.37, odds ratio (95% CI) 1.96 (1.49-2.63), $p = 2.9 \times 10^{-7}$. When stratifying the analysis by drug, the association with this variant was strongest in patients treated with statins ($p = 6.5 \times 10^{-8}$). Mutations in *LRPPRC* have previously been associated with Leigh syndrome, which causes lowered levels of cytochrome C oxidase, a key enzyme in aerobic metabolism. Our study highlights that variants mapping to *LRPPRC* are associated with response to cardiovascular drugs in CHD patients, in particular those treated with statins.

183

A quality control framework for exome sequencing studies to reduce bias from heterogeneous sequencing platforms

Yao Yu (1) Hao Hu (1) Chad Huff (1)
(1) Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center

Whole-exome sequencing data is now increasingly being available to the research community for secondary analyses, providing new opportunities for large-scale exome

association studies. However, the heterogeneous nature of data presents major barriers in exome mega-analysis efforts. Minor differences in sample preparation, target capture, and sequencing protocols often result in platform-specific biases which overwhelm subtle signals of disease association. To mitigate the problem, we have developed a quality control (QC) framework for whole-exome sequencing case-control studies, consisting of optimizing variant calling procedures, assessing population stratification based on well-behaved exonic markers, and applying multiple approaches to identify and filter variants with platform-specific biases. We apply our framework to a gene-based exome case-control study of individuals with European ancestry involving 314 TCGA skin melanoma cancer cases and 2,424 population controls from various sources. Compared to standard joint genotyping and QC metrics, utilizing our framework can reduce the genome-wide inflation factor (λ) from 2.25 to 1.34, while retaining power to replicate well-established melanoma-gene associations. The results demonstrate that our framework has the potential to greatly mitigate Type I error inflation resulting from heterogeneous sequencing platforms.

184

Evaluation of Copy Number Variation (CNV) detection methods in whole exome sequencing data

Peng Zhang (1) Hua Ling (1) Elizabeth Pugh (1) Kurt Hetrick (1) Dane Witmer (1) Nara Sobreira (2) David Valle (2) Kimberly Doheny (3)
(1) Center for Inherited Disease Research, Institute of Genetic Medicine, The Johns Hopkins School of Medicine (2) Institute of Genetic Medicine, Johns Hopkins University School of Medicine (3) Center for Inherited Disease Research, Institute of Genetic Medicine, The Johns Hopkins School of Medicine

As part of the Baylor Hopkins Center for Mendelian Genomics (CMG) (http://www.mendelian.org/), CIDR has performed whole exome sequencing (WES) for over 900 samples to discover the genetic basis for Mendelian conditions. We have worked on calling copy number variations (CNVs) from our WES and have successfully identified a causal deletion from two families. But more often, calling CNVs from WES is a challenge because of the sparseness of the target regions and the non-uniform distribution of reads across genome. Consequently, the concordance across different CNV calling methods is usually low, and it's often hard to evaluate the called CNVs from different programs with real sequencing data.
In this study, we plan to simulate data similar as those from our WES data including deletions and duplications, and then to evaluate the sensitivity and specificity of different CNV calling methods including ExomeDepth, EXCAVATOR, XHMM and CANOES. We will test how these methods perform with different parameter settings including CNV sizes and frequency. We then use the results to evaluate the CNV calls for our WES samples from CMG.

185

Two-Step Testing Approaches for Detecting Quantitative Gene-Environment Interactions in a Genome-Wide Association Study

Pingye P. Zhang (1) John J. L. Morrison (1) James W. J. Gauderman (2)
(1) Department of Preventive Medicine, University of Southern California, Los Angeles, USA (2) Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

A genome-wide association study (GWAS) typically is focused on detecting marginal genetic effects. However, many complex traits are likely to be the result of the interplay of genes and environmental factors, and many trait-related loci may be specific only to a subgroup of the population defined by some environmental factors. These SNPs may have a weak marginal effect and thus unlikely to be detected from a scan of marginal effect, but may be detectable in a gene-environment (G x E) interaction analysis. However, standard analysis of G x E interaction is known to have low power, particularly when one corrects for testing of multiple SNPs. There are two 2-step methods for conducting a genome-wide interaction scan (GWIS) aimed at improving efficiency by prioritizing SNPs most likely to be involved in a G x E interaction using a screening step. For example, Kooperberg and Leblanc (2008) suggested screening on genetic marginal effect. Alternatively, Paré et al. (2010) proposed screening on variance heterogeneity induced by interaction. In this paper, we show that the Paré et al. approach has an inflated false-positive rate in the presence of environmental marginal effect, and we propose an alternative that remains valid. We also propose a novel 2-step approach that combines the two screening approaches, and provide simulations demonstrating that the new method outperforms other GWIS approaches for some underlying models. Application of this method to a G x Hispanicity scan for childhood lung function reveals one interesting SNP that was not identified from marginal-effect scans.

186

The Principal Components Analysis Propensity Scores (PCAPS): A Practical Approach to Population Stratification in Genome-wide Association Studies

Huaqing Zhao (1) Nandita Mitra (2) Peter A. Kanetsky (3) Katherine L. Nathanson (2) Timothy R. Rebbeck (2)
(1) Temple University School of Medicine (2) University of Pennsylvania (3) Moffitt Cancer Center

Currently, the most widely used method to address population stratification (PS) in genome-wide associations studies (GWAS) is principal components analysis (PCA). The main challenge of PCA lies in choosing numbers of PCAs to include as covariates. The original approach is to use the 10 PCAs which is common but arbitrary. One alternative is to use the Tracy-Widom statistic to select significant PCAs. However,

this may result in a liberal selection of PCAs. We developed genomic propensity scores (GPS) and extended GPS (eGPS) to correct for PS. However, GPS and eGPS are not readily applicable for GWAS data. To overcome these limitations, we combine all these approaches to estimate principal components analysis propensity scores (PCAPS) based on PCAs selected by Tracy-Widom statistic to control PS in GWAS. The advantages of PCAPS are the uniform selection of PCAs and the potential ability of handling outliers. The proposed PCAPS is able to correct bias due to PS with outliers at SNP and subject levels since PCAPS values are calculated for each specific SNP and subject. In contrast, PCA values are calculated based on the profile of all SNPs for a given subject and the same values will be adjusted for each SNP. Our simulation GWAS studies have shown that our PCAPS method can adequately control bias due to PS in GWAS. In addition, PCAPS is able to reduce spirious associations compared with PCA. We illustrate our approach in a case-control GWAS of testicular germ cell tumors (TGCT). Our new method consistently yields narrower confidence intervals of odds ratio than PCA in the TGCT data. PCAPS is a data reduction tool without additional software. We believe the newly proposed PCAPS method will provide an innovative and practical way of correcting PS in GWAS studies.

187

A statistical Approach for Testing Gene by Microbiome Interactions

Ni Zhao (1) Michael C. Wu (1)
(1) Fred Hutchinson Cancer Research Center

Although GWAS and next generation sequencing studies have helped elucidate the genetic component of complex traits, a comprehensive understanding of such traits often requires evaluating other risk factors, such as the human microbiome. Despite the well-known importance of both microbiome and genotype underlying many traits, there has been little work on assessing their joint effect. A problem of particular interest is identifying genetic variants that modify the effect of microbiome composition and the discovery of genotypes for which the effect of microbiome composition is heterogeneous. Yet, how to test for statistical interactions between microbiome composition and genetic variants remains unaddressed. We propose a new strategy for assessing the interaction between individual variants and microbiome using a regression based framework. In particular, following standard approaches, we encode microbiome as a matrix of pairwise distances between individuals where the distance is defined based on a phylogenetically informed metric. The distance can then be transformed to a kernel matrix of pairwise similarities between individuals in the study. This enables us to exploit the kernel machine framework (e.g. SKAT) to model the relationship between a trait and microbiome, the main effect of a genetic variant, and their interaction. Modifications to the standard kernel machine framework are necessary to accommodate the dimensionality and complexity of microbiome composition data. Under this modified framework, we develop a score test for testing the interaction term. Simulations and real data applications show that our approach has reasonable power and correct type I error.

188

Bayesian analysis of polygenic effects

Jing Hua Zhao (1) Jian'an Luan (1)
(1) MRC Epidemiology Unit

An estimate of the total genetic effects on a phenotype of interest broadly indicates the scope of genomewide association studies (GWASs), which now play a significant role in identification and characterization of the underlying genetic variants. However, it can involve considerable uncertainty. Mixed models serve as a generic framework for this problem since they accommodate a variety of outcomes as a function of environmental factors together with polygenic effects. Indicators of the genetic influence from these models include estimated proportions of phenotypic variance attributable to genetic factors alone (heritability, h^2) or combined with environmental factors (R^2). The models differ from those seen in general statistics as the polygenic effects is represented by a random variable, "vertically" correlated among relatives as well as individuals in a general population due to identity-by-descent gene sharing. Here we focus on Bayesian mixed models and compare their performance with the frequentist counterparts. Through simulated and real data we showed that the two approaches can give comparable point estimates but a Bayesian approach is desirable with its ability to use prior information and produce posterior distributions. Implemented both in general and special software, our analysis highlighted technical issues which need to be tackled, especially with genomic relationship matrix.

189

LASSO-Based Approaches for Joint Tests of Genetic Main Effects and Gene-Environment Interactions

Jie Zheng (1) Dabeeru C. Rao (2) Gang Shi (3)
(1) Xi'an Jiaotong University (2) Washington University in St. Louis (3) Xidian University

The least absolute shrinkage and selection operator (LASSO) regression has the so-called shrinkage property and has attracted considerable interest in genetic studies. We generalized the single-marker two degrees-of-freedom (2 df) joint test of genetic main effects and gene-environment interactions, which is known to be more powerful than either of the marginal tests if an interaction exists, to a multiple-marker version using the LASSO regression. We considered two LASSO-based approaches: one analyzed the genetic main effects of multiple SNPs and their interactions with an environmental factor, and the other employed the group LASSO regression. SNPs with a nonzero regression coefficient were

considered significant. The 2 df linear regression was used as the benchmark test for comparisons. Bayesian information criterion (BIC), Mallow's Cp, and Stine's Sp were compared with respect to their ability to select constraint parameters of the two LASSO approaches. Based on simulation studies, we showed that BIC and Mallow's Cp tend to generate over-fitted models resulting in large type I errors for both the LASSO methods. Choosing constraint parameters based on Stine's Sp demonstrated acceptable empirical type I errors. Since Stine's Sp does not control type I errors directly, the two LASSO methods showed varying empirical false positive rates for different sample sizes. Group LASSO displayed lower type I errors and statistical power than did the 2 df regression test when sample sizes were relatively small and higher type I errors and statistical power when sample sizes were larger. For computational efficiency, the first approach was much faster than the second when using the least-angle regression solver.

190

A Comparison of Methods for Joint Association Analysis of Multiple Traits

Huanhuan Zhu (1) Shuanglin Zhang (1) Qiuying Sha (1)
(1) Michigan Technological University

Genome-wide association studies (GWAS) have identifies many variants that each affects multiple traits, which suggests that pleiotropic effects on human complex traits may be widespread. Therefore, statistical methods that can jointly analyze multiple traits in GWAS may have advantages over analyzing each trait individually. Some statistical methods have been developed to utilize such multivariate traits in genetic association studies, however the power comparisons of these methods are not well known. In this study, we compared some of the existing methods for association studies using multiple traits, which include O'Brien's method, Cross-Validation method, Optimal Weight method, TATES, PCH, CCA, MANOVA and MultiPhen. We used simulation studies to compare the powers of these methods under a variety of scenarios, including different numbers of traits, different minor allele frequencies, different mean models and variance models. Our results showed that there is no single method, which has consistently good performance among all the scenarios. Each method has its own advantages and disadvantages. Our goal of this study is to provide researchers with useful guidelines on selecting statistical methods in the application of real data with multiple traits.

191

Methylome scan of PAI-1 plasma levels identifies a locus with putative epigenetic mediation of genetic effect

Nora Zwingerman (1) Irfahan Kassam (1) Vinh Truong (1) Dylan Aïssi (2) Jessica Dennis (1) Michael Wilson (3) Phil Wells (4) Pierre-Emmanuel Morange (5) David-Alexandre Trégouët (2) France Gagnon (1)
(1) Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada (2) INSERM, UMR'S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, Paris, France (3) Genetics and Genome Biology Program, Peter Gilgan Centre for Research and Learning, Toronto, Canada (4) Department of Medicine, Faculty of Medicine, Ottawa, Canada (5) INSERM, UMR'S 1062, Nutrition Obesity and Risk of Thrombosis, Aix-Marseille, France

Plasminogen activator inhibitor type 1 (PAI-1) is the main inhibitor of fibrinolysis, and elevated levels are associated with increased thrombosis risk. PAI-1 levels are heritable, with estimates of ~50%. The SNPs identified by meta-GWAS explain less than 3% of PAI-1 variance; known non-genetic factors account for 36–47%. DNA methylation (DNAm) marks can regulate gene transcription and may account for the unexplained PAI-1 genetic variance.

To explore the role of DNAm on PAI-1 variance, we conducted a genome-wide study of blood DNAm in 202 individuals from 5 large pedigrees ascertained on single probands with venous thrombosis (VT). Linear mixed effect models were adjusted for relevant covariates, and accounted for cellular heterogeneity and family structure. Heritability estimates were computed using maximum-likelihood variance component approach. SNPs ±150kb of the top DNAm site were tested for association, followed by causal inference testing (CIT). Independent replication was tested in 350 VT cases.

Five genome-wide significant DNAm sites - accounting for 23% of PAI-1 variance - were identified in 1 gene and 1 CpG Island (FDR $q < 0.025$–0.072). The top DNAm site is located in *SMOC2* gene, and showed to be highly heritable (69%). A nearby SNP was found associated ($p < 10^{-5}$) with the *SMOC2* DNAm. CIT showed the SNP effect on PAI-1 levels to be fully mediated by the DNAm ($p < 0.025$). The SNP-DNAm effect was replicated ($p < 10^{-4}$), while the relationships with PAI-1 levels were not replicated. *SMOC2* is highly expressed during wound healing, process where PAI-1 levels are elevated to help maintain hemostasis. In summary, a methylome scan identified a DNAm site, in a biologically relevant gene, that could mediate a SNP-PAI-1 association.