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

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

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

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

68

Evaluation of Network-guided Random Forest for Disease Gene Discovery

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

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

69

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

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

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

70

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

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

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

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

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

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

71

Multi-ethnic Polygenic Risk Scores for Venous Thromboembolism

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

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

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

72

GWAS of Longitudinal Trajectories at Biobank Scale

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

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

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