Human Heredity

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

Published online: April 4, 2017

45th European Mathematical Genetics Meeting (EMGM) 2017

Tartu, Estonia, April 4–7, 2017

Abstracts

Guest Editor
Krista Fischer, Tartu

KARGER Basel · Freiburg · Paris · London · New York · Chennai · New Delhi · Bangkok · Beijing · Shanghai · Tokyo · Kuala Lumpur · Singapore · Sydney

Human Heredity

Invited 1

1
Multivariate Analyses of Summary Statistics from Genome-Wide Association Studies: The Role of Covariance Structures

Matti Pirinen (joint work with Christian Benner)

Institute for Molecular Medicine Finland (FIMM), Department of Mathematics and Statistics and Department of Public Health, University of Helsinki, Helsinki, Finland

Recently, various novel statistical methods have been developed to work directly on summary statistics from Genome-Wide Association Studies (GWAS). This is a promising approach to utilize the increasing GWAS sample sizes while avoiding privacy concerns and logistics of sharing individual-level genotype data. Examples include estimation of heritability and genetic correlations, gene-level tests, risk prediction, *z*-score imputation and fine-mapping of causal variants.

In addition to GWAS results, these approaches require estimates of genotype-genotype and/or phenotype-phenotype covariance structures to properly account for dependences between variables. Hence, key questions are from where to take these covariance estimates and how do they perform in practice. I consider these questions in the context of our recent multivariate methods metaCCA and FINEMAP.

Our theoretical and empirical results suggest that problems with external estimates of genotype-genotype correlations will increase with GWAS sample size. As a solution, we introduce a software tool LDstore for efficient estimation, storing and sharing of correlation information and outline how LDstore helps downstream analyses within a GWAS consortium.

Contributed 1: GWAS: Methodology and Applications

2
A Genome-Wide Two-Component Mixture Model Expectation-Maximisation Algorithm for Time to Event Data

Ben Francis[a], Peng Yin[a], James P. Cook[a], Andrea L. Jorgensen[a], Jane Hutton[b], Andrew P. Morris[a]

[a]Department of Biostatistics, University of Liverpool, Liverpool, [b]Department of Statistics, University of Warwick, Coventry, UK

Traditional survival analysis of time to event (TTE) data assumes that all individuals will experience the event of interest (EOI). In pharmacogenetics studies, there are patients who will not experience the therapeutic effect of a drug regardless of dosage or duration of prescription. Those who are unable to experience the EOI are deemed to be the part of the "cure fraction". Therefore, those with censored survival time in these particular pharmacogenetics studies can either be patients who will go on to experience the EOI after being lost to follow-up and those who are not able to experience the EOI.

Modelling TTE data consisting of those susceptible to the EOI and the cure fraction requires a "two-component" approach; enabling estimation of the effect of covariates on both susceptibility and the time to the occurrence of an EOI. One widely-used method incorporates an accelerated failure time model and expectation-maximisation algorithm ("the full model") but is too computationally intensive to be applied genome-wide with a minimum run time of 56 seconds per SNP for a dataset of one thousand patients. To circumvent this problem, we obtained survival and susceptibility residuals from a model including clinical covariates only, to then use as phenotypes in a multivariate "reverse regression" analysis ("the residual model") which results in a one-off run time of 56 seconds to calculate residuals and then 0.01 seconds per SNP for the same dataset.

To assess the performance of this approach, we performed detailed simulations incorporating a range of models of SNP effect on survival and susceptibility comparing the likelihood ratio tests results from the full and residual models. Under a null model of no association of a SNP with survival or susceptibility, the type I error rates of all analytical approaches were maintained. Over the range of association models considered, the multivariate reverse regression approach was no less powerful for detecting survival and susceptibility effects than the full two-component model.

In conclusion, we have developed a novel "approximate" computationally efficient approach to enable genome-wide analysis of

E-Mail karger@karger.com
www.karger.com/hhe

two-component TTE data. This approach is able to determine whether there are factors across the genome that influence time to treatment response and/or susceptibility to treatment response. The interaction of significant genetic and clinical or treatment factors can then be investigated in the full model.

3

Multi-Phenotype Genome-Wide Meta-Analysis of Lipid Levels and BMI in 64,736 Europeans Suggests Shared Genetic Architecture

Marika Kaakinen[a,b], *Vasiliki Lagou*[c,d], *Reedik Mägi*[e], *Krista Fischer*[e], *Andrew P. Morris*[f], *Inga Prokopenko*[b], *for the ENGAGE Consortium*

[a]Department of Medicine, Division of Experimental Medicine and Toxicology, Imperial College London, London, [b]Department of Genomics of Common Disease, Imperial College London, London, UK; [c]VIB Center for Brain & Disease Research, Leuven, [d]KU Leuven, Department of Microbiology and Immunology, Leuven, Belgium; [e]Estonian Genome Center, University of Tartu, Tartu, Estonia; [f]Institute of Translational Medicine, University of Liverpool, Liverpool, UK

Serum lipid levels and obesity share biochemical pathways, suggesting influence by common genetic factors. Genetic association analysis of multiple correlated phenotypes simultaneously allows for detection of such shared genetic effects with improved power. Within the ENGAGE consortium, we performed a multi-phenotype genome-wide association study (MP-GWAS) on three blood lipids (high-/low-density lipoprotein cholesterol and triglycerides [HDL-C/LDL-C/TG]) and body-mass index (BMI). We used GWAS data imputed to the 1000 Genomes reference panel (Phase 1) of up to 64,736 individuals from 22 European-ancestry studies. Each study performed the MP-GWAS by fitting a "reverse regression" model between each imputed single-nucleotide variant (SNV) and the linear combination of the three blood lipids and BMI, using the SCOPA software, i.e. $SNV_i = \beta_{1i}\times HDL\text{-}C + \beta_{2i}\times LDL\text{-}C + \beta_{3i}\times TG + \beta_{4i}\times BMI + \varepsilon_i$, where $i = 1, \ldots, n$, n is the maximum number of SNVs tested and $\varepsilon_i \sim N(0,\sigma^2)$. Study-specific variance-covariance matrices for each variant were combined in a meta-analysis within the META-SCOPA software. We considered only variants with MAF >1%, *P*-value for HWE >0.0001 and data present in at least half of the studies. Empowered by the joint analysis, we identified 14 novel common variant loci at genome-wide significance ($P < 5\times10^{-8}$) at/near *CCDC18*, *CANX*, *LINC00681*, *BC036431*, *CGNL1*, *SDC1*, *SLC8A1*, *EPHA6*, *SPATA4*, *MAGI2*, *CTSB*, *BC014119*, *SMCO4* and *CNTN5*. The latter nine of the loci showed effects at nominal significance ($P < 0.05$) both on lipids and BMI in the joint model, suggesting shared genetic architecture worth further investigation. We additionally detected 41 and 9 previously established lipid and BMI loci, respectively, of which only 32 and 4 would have been identified if the traits had been analysed with traditional single-phenotype methodology. Our analyses demonstrate the improved power of the MP-GWAS approach as compared to single-phenotype GWAS, and show its ability to detect multi-phenotype effects.

4

Using Penalised Regression to Predict Treatment Response in Rheumatoid Arthritis from SNP data

Svetlana Cherlin, *Heather J. Cordell*

Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, UK

In a typical genome wide association study (GWAS), several thousands to several millions of SNP markers are genotyped in a sample size of several hundreds to several thousands of individuals, thus leading to many more predictor variables than response variables, causing model overfitting. Overfitted models are likely to demonstrate poor predictive ability when applied to a new data. To overcome this problem, penalised regression methods have been proposed, aiming at shrinking the coefficients towards zero.

We explore prediction of treatment response in Rheumatoid Arthritis patients from SNP data using one specific penalised regression approach, namely the least absolute shrinkage and selection operator (lasso regression). One important property of the lasso penalty is that it allows the coefficients to be set to exactly zero, thus performing variable selection. We use 10-fold cross validation to assess predictive performance, with nested 10-fold cross validation used to specify the penalty parameter.

The approach is applied to data from the MATURA (Maximum Therapeutic Utility in Rheumatoid Arthritis) consortium. The data comprises treatment response measures and SNP data from approximately 1 K patients and 5 M SNPs, with none of the SNPs reaching genome-wide significance in a univariate regression analysis. We apply lasso regression to these data after reducing the number of SNP to approximately 56 K using LD-based clumping. The results illustrate poor predictive ability as assessed by examining the correlation coefficient and the calibration slope. In order to investigate these results, we apply lasso regression on a simulated data set where causal SNPs have genome-wide significant effects, and the results illustrate a good prediction performance.

We also examine the effect of sample size on the prediction in the case where the causal SNPs are not genome-wide significant. The simulation study shows that when the sample size comprises a few hundreds of individuals, SNP effects are heavily penalised resulting in a poor predictive performance. Increasing the sample size to a few thousands of individuals has the effect of a much smaller penalisation of the true effects, thus greatly improving the prediction. Overall our results suggest that lasso regression requires strong effects or large sample sizes in order to achieve good prediction.

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

5

Quantifying the Extent to Which Index Event Biases Influence Large Genetic Association Studies

Hanieh Yaghootkar[a], *Michael P. Bancks*[b], *Sam E. Jones*[a], *Aaron F. McDaid*[c,d], *Robin Beaumont*[a], *Louise Donnelly*[e], *Andrew R. Wood*[a], *Archie Campbell*[f], *Jessica Tyrrell*[a], *Lynne J. Hocking*[g], *Marcus A. Tuke*[a], *Katherine S. Ruth*[a], *Ewan R. Pearson*[e], *Anna Murray*[a], *Rachel M. Freathy*[a], *Patricia B. Munroe*[h,i], *Caroline Hayward*[j], *Colin Palmer*[e], *Michael N. Weedon*[a], *James S. Pankow*[b], *Timothy M. Frayling*[a,&], *Zoltán Kutalik*[c,d,&]

[a]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK; [b]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA; [c]Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, [d]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [e]Division of Cardiovascular & Diabetes Medicine, Medical Research Institute, University of Dundee, Dundee, Scotland; [f]Generation Scotland, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, [g]Institute of Medical Sciences, University of Aberdeen, Aberdeen, Scotland; [h]Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, [i]NIHR Barts Cardiovascular Biomedical Research Unit, Barts and The London School of Medicine, Queen Mary University of London, London, UK; [j]Generation Scotland, MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, Scotland
[&]These authors jointly supervised this work

As genetic association studies increase in size to 100,000 s of individuals, subtle biases may influence conclusions. One possible bias is "index event bias" (IEB) that appears due to the stratification by, or enrichment for, disease status when testing associations between genetic variants and a disease-associated trait. We aimed to test the extent to which IEB influences some known trait associations in a range of study designs and provide a statistical framework for assessing future associations. Analysing data from 113,203 non-diabetic UK Biobank participants, we observed three (near *TCF7L2, CDKN2AB* and *CDKAL1*) overestimated (BMI-decreasing) and one (near *MTNR1B*) underestimated (BMI-increasing) associations among 11 type 2 diabetes risk alleles (at $P < 0.05$). IEB became even stronger when we tested a type 2 diabetes genetic risk score composed of these 11 variants (–0.010 SDs BMI per allele, $P = 5\times10^{-4}$), which was confirmed in four additional independent studies. Similar results emerged when examining the effect of blood pressure increasing alleles on BMI in normotensive UK Biobank samples. Furthermore, we demonstrated that, under realistic scenarios, common disease alleles would become associated at $p < 5\times10^{-8}$ with disease-related traits through IEB alone, if disease prevalence in the sample differs appreciably from the background population prevalence. For example, some hypertension and type 2 diabetes alleles will be associated with BMI in sample sizes of >500,000 if the prevalence of those diseases differs by >10% from the background population. In conclusion, IEB may result in false positive or negative genetic associations in very large studies stratified or strongly enriched for/against disease cases.

Contributed 2: Families and Relatedness

6

Estimation of Realized Relatedness: Contiguity Matters

Elizabeth Thompson, Bowen Wang

Department of Statistics, University of Washington, Seattle, WA, USA

Relatedness between a pair of individuals is classically measured as twice the coefficient of kinship; it is a measure of the proportion of genome shared identical by descent (IBD) between the two individuals. Given a pedigree structure, kinship coefficients relative to pedigree founders can be computed. However, there are two issues: (1) the process of meiosis has high variance, so that realized proportions of genome shared may differ widely from pedigree expectations, and (2) in population samples the reference population of interest may be further back in time, and pedigrees are often unknown. Particularly, for example, in the managed populations of endangered species, populations are small, there are close but unknown relatives or inbreeding, and there may have been strong selection, whether intended or unintended.

In such cases, realized relatedness provides a far more informative picture of the structure of a population sample. Modern SNP data provide a basis for estimation of realized relatedness, but many current methods do not allow for inbreeding, do not adjust for allelic associations (linkage disequilibrium), and do not take into consideration that genomes descend in large segments from generation to generation. Not only do allelic associations themselves arise from this segmental descent of genome, but methods that model this contiguity of descent can provide much more efficient estimators of relatedness. We present models, methods and results of some new estimators of realized relatedness.

7

A Transmission Based Association Test for Multivariate Phenotypes Using Quasi Likelihood

Saurabh Ghosh, Hemant Kulkarni

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

The classical transmission disequilibrium test (TDT) [Spielman et al. 1993] based on the trio design is an alternative to the population based case-control design to detect genetic association as it protects against population stratification. Since the manifesta-

tion of most complex diseases are governed by multiple precursor traits, it has been argued that it may be a more prudent strategy to study a multivariate phenotype comprising these precursors. One of the statistical challenges in analyzing multivariate phenotypes is to incorporate both quantitative and qualitative variables in the vector of phenotypes. We modify the classical TDT for quantitative traits based on logistic regression [Waldman et al. 1999, Haldar and Ghosh 2015] to include multivariate phenotypes. We adopt a quasi likelihood approach [Wedderbur 1974] based on Generalized Linear Regression to develop a test of association for multivariate phenotypes. Since the Generalized Estimating Equation (GEE) approach [Gourieroux, Monfort, and Trognon 1984; Liang and Zeger 1986] used for solving the quasi likelihood equation is highly influenced by outliers, we use a modified Resistance Generalized Estimating Equation approach (RGEE) [Hall, Zeger, and Bandeen-Roche 1996, Preisser and Qaqish 1999] to down weight the outliers. We also explore a modified model that includes information on allelic transmission from both parents. We perform extensive simulations under a wide spectrum of genetic models and different correlation structures between the components of a multivariate phenotype. We compare our method with the FBAT test procedure [Lake et al. 2002] as well as separate univariate analyses of the component phenotypes and find that the proposed method that incorporates information on both parents is more powerful than the other approaches. We apply our method to analyze a multivariate phenotype related to alcoholism using data from the Collaborative Study on the Genetics of Alcoholism (COGA) project.

8

On Inferior Power of Recently Developed Family-Based Association Analysis Methods for Next-Generation Sequencing Studies of Rare-Variants

Tero Hiekkalinna[a,b], *Joseph Terwilliger*[a,c,d], *Markus Perola*[a,b,e]

[a]National Institute for Health and Welfare, [b]Institute for Molecular Medicine Finland FIMM, Helsinki, Finland; [c]Department of Psychiatry, Department of Genetics and Development, Gertrude H. Sergievsky Center, Columbia University, New York, NY, [d]Division of Medical Genetics, New York State Psychiatric Institute, New York, NY, USA; [e]University of Tartu, Estonian Genome Center, Tartu, Estonia

In this post-GWAs era, whole genome sequencing (WGS) techniques will be used to generating a vast amount of data which will raise enormous computational challenges for human genetics researchers. Scientists are now returning to family-based designs, noting that under every model considered, they provide higher power for gene mapping. Subsequently, several new methods for family-based association (FBA) analysis on WGS data have been recently published. However, these newly developed methods have not been compared to classical FBA methods, such as PSEUDOMARKER.

To this end, we have evaluated usability and compared the empirical type I error rates and power of several classical methods and recently published implementations of FBA tests (burden tests, kernel statistics) using datasets composed of mixtures of singletons and families. We compare the performance of these tests under three hypotheses: (a) No linkage and no association; (b) Complete linkage and no association; (c) Complete linkage and association. The first two represent potential null hypotheses in (a) joint tests of linkage and association; and (b) conditional tests of association given linkage. Power was estimated for all methods for both joint tests and conditional tests, over a range of effect sizes for the functional locus, and a range of LD between the functional locus and a marker.

The most notable results are that new methods are not always easy to use, sometimes even impossible to use and documentation is often insufficient. More alarmingly, some of them have enormous type I error rates. This can lead to an unacceptable rate of spurious conclusions about the presence of association, leading to fruitless follow up studies. Power is shown to be optimal for methods based on full likelihood analysis of complete pedigree data, such as our PSEUDOMARKER. Results from a simulation-based study will be presented.

We strongly suggest that new software for gene mapping should not be allowed to be published unless authors provide fully functioning software with all features which they claim it can perform and complete usage documentation with example data. This would be quality improving procedure to assure that high quality software would be only published and researchers world-wide would not waste their valuable time on useless software.

9

Statistical Methods for Multivariate Phenotypes for Admixture and Related Data

Mariza de Andrade[a], *Brandon J. Coombes*[b], *Julia M.P. Soler*[c]

[a]Mayo Clinic, Rochester, MN, [b]Mayo Clinic, Rochester, MN, USA; [c]University of São Paulo, São Paulo, SP, Brazil

Several models and statistical methods have been proposed to identify genetic variants with pleiotropic effects in humans, animals and plants. From the proposed models there is a variation about the number of phenotypes and the source of the data (related or unrelated). When analyzing several phenotypes using related data simultaneously there is a limitation due to the number of parameters to be estimated unless several restrictions are applied. In this paper we discuss several methods that take into account the familial relation either using the kinship matrix or using the residuals with the novelty to include the population admixture in the model. As an application we use the Baependi family study that consists of 80 families and 1,100 subjects, genotype data from Affymetrix 6.0, and the metabolic syndrome variables (as our multivariate phenotypes) from Brazil. These families are highly admixed with on average 70% European Ancestry, 13% African Ancestry, and 18% Native Americans. Our goal is to compare the first principal component of heritability using the metabolic syndrome variables, multivariate phenotype analysis [2, 3], both methods using the variance components approach taking into account the family structure, and compare with methods that used the residuals of the polygenic model as the phenotypes.

Contributed 3: Analysis of Whole-Genome Sequence Data: Methodology and Applications

10
The Rare Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer's Disease Whole Genome Sequence Data

Suzanne M. Leal[a], *Zongxiao He*[a], *Di Zhang*[a], *Alan E. Renton*[b], *Biao Li*[a], *Linhai Zhao*[a], *Gao T. Wang*[a], *Alison M. Goate*[b], *Richard Mayeux*[c]

[a]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, [b]Department of Neuroscience and Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY, [c]Department of Neurology, Taub Institute on Alzheimer's Disease and the Aging Brain, and Gertrude H. Sergievsky Center, Columbia University, New York, NY, USA

Whole genome and exome sequence data can be cost effectively generated for families to detect rare variant (RV) associations. Causal variants that aggregate in families usually have larger effect sizes than those found in sporadic cases, therefore family-based design can be a more powerful approach than population-based designs. Moreover, some family-based designs are robust to confounding due to population admixture/substructure. We developed a RV extension of the generalized disequilibrium test (GDT) to analyze sequence data obtained from nuclear and extended families. The GDT utilizes genotype differences of all discordant relative pairs to assess associations within a family and the RV extension combines the single-variant GDT statistic over a genomic region of interest. The RV-GDT has increased power by efficiently incorporating information beyond first-degree relatives and allows for the inclusion of covariates. Using simulated genetic data, it is demonstrated that the RV-GDT method has well-controlled type I error rates, even when applied to admixed populations and populations with substructure. It is more powerful than existing family-based RV association methods, particularly for the analysis of extended pedigrees and pedigrees with missing data. Whole genome sequence data from Alzheimer's disease (AD) families from the AD Sequencing Project were analyzed to illustrate the application of the RV-GDT and several suggestive associations were found including an association with *TNK1* for which a common variant association with AD was previously reported. Given the capability of the RV-GDT to adequately control for population admixture/substructure, analyze pedigrees with missing genotype data and its superior power over other family-based methods, it is an effective tool to elucidate the involvement of RVs in the etiology of complex traits.

11
Are Allele Counts from NGS-Data a Better Alternative than Called Genotypes for Testing Phenotype-Genotype Associations?

Rosa Gonzalez Silos, Justo Lorenzo Bermejo

Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

Next generation sequencing (NGS) is revolutionizing research and practice in human genetics. Technical advances have already translated into large collections of NGS-data and the need of new techniques to analyse them. Called genotypes are typically used to investigate the relationship between a phenotype of interest and a particular genetic variant. Genotypes are usually called using probabilistic methods, which rely on quality scores and allele counts after base-calling and alignment. Here we investigate an alternative approach which takes into account the uncertainly of called genotypes by directly using allele counts to test genetic association.

We simulated hypothetical NGS-association studies based on chromosome 21 genotype data from 1179 HapMap individuals, which was grouped into 18 classes according to the individual genotype, the reference, and the alternative alleles. At each genetic position, phenotypes were assigned independently of, and depending on, individual genotypes. Genotype quality scores and allele counts and were simulated using real NGS data from eight participants in the Personal Genome Project. First, quality scores were drawn according to observed frequencies for each genotype class. Then, reference and alternative allele counts were simulated according to a bivariate normal distribution which reflected observed real counts. After simulation of genotype quality scores and allele counts, genotypes were called with the Haplotype Caller.

A wide set of statistical techniques was used to explore the relationship "allele counts-phenotype" and "called genotype-phenotype". They included standard and robust linear, Poisson, negative binomial, logistic and ordinal regression models, ANOVA and Kruskal Wallis tests. Since allele counts show an excess of zeros, Hurdle and zero inflated regression were used too.

The different approaches for testing genetic association were compared regarding type-I-error rate and statistical power. In addition to overall comparisons, stratified analyses were conducted according to variant characteristics – including the type and frequency of the genetic variant. Details on simulations, methodology and results will be presented at the meeting.

12

On the Association Analysis of Genome-Sequencing Data: A Spatial Clustering Approach for Partitioning the Entire Genome into Non-Overlapping Windows

Heide Loehlein Fier[a,b], *Julian Hecker*[b], *Dmitry Prokopenko*[b], *Scott T. Weiss*[c], *Rudolph E. Tanzi*[d], *Christoph Lange*[a,c]

[a]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [b]Working Group of Genomic Mathematics, University of Bonn, Bonn, Germany; [c]Channing Laboratory, Brigham and Women's Hospital, Department of Medicine, Harvard Medical School, Cambridge, MA, [d]Genetics and Aging Research Unit, MassGeneral Institute for Neurodegenerative Disease, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

For the association analysis of whole-genome sequencing (WGS) studies, we propose an efficient and fast spatial-clustering algorithm. Compared to existing analysis approaches for WGS data, that define the tested regions either by sliding or consecutive windows of fixed sizes to the variant data, a meaningful grouping of nearby variants into consecutive regions has the advantage compared to sliding window approaches that the number of tested regions is likely to be smaller and compared to consecutive fixed-window approaches that nearby-variants are likely to be grouped together. Given existing biological evidence that disease-associated mutations tend to physically cluster in specific regions along the chromosome, the identification of meaningful groups of nearby located variants could thus lead to a potential power gain for association analysis.

Our algorithm defines consecutive genomic regions based on the physical positions of the variants, assuming an inhomogeneous Poisson process and groups together nearby variants. As parameters are estimated locally, the algorithm takes the differing variant density along the chromosome into account and provides locally optimal partitioning of variants into consecutive regions.

We discuss the theoretical advances of our algorithm compared to existing, window-based approaches and show the performance and advantage of our introduced algorithm by an application to Alzheimer's disease WGS data. Our analysis identifies a region in the ITGB gene that potentially harbors disease susceptibility loci for Alzheimer's disease. The region-based association signal of ITGB replicates in an independent data set and achieves formally genome-wide significance.

13

Systematic Characterization of VNTR Polymorphisms Using Deep Whole Genomes

Mart Kals[a,b], *Tarmo Puurand*[c], *Maris Teder-Laving*[a], *Tõnu Esko*[a,d]

[a]Estonian Genome Center, University of Tartu, Tartu, [b]Institute of Mathematics and Statistics, University of Tartu, Tartu, [c]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; [d]Broad Institute, Cambridge, USA

Human genome has been systematically characterized since early 2000s but the focus has mainly been on single nucleotide variation (SNV) while the structural variation (like copy number variation and repeative elements) have been less studied. We have used deep whole-genome sequences (30×, PCR-free technology) of 473 individuals from Estonian Biobank cohort to systematically identify VNTRs (variable number tandem repeats) and STRs (short tandem repeats). STRs motifs of periods 1 through 6 bp are called using lobSTR and motifs above 6 bp by k-mer methodology, where the BAM files are indexed for a list of sequence combination for oligomer. In both cases, Tandem Repeat Finder database containing all known tandem repeat sequences (200,000+) is used to obtain the genomic coordinates of structural variants. In summary, we identified >217 K VNTRs, of which >60 K had a repeating element longer than 24 bp.The number of repeats per VNTR ranged from 3 to 2000. We validated our approach through Sanger sequencing for several VNTR classes (repeat count accuracy above 85%). Next, we used comprehensively available set of omics profiles (including RNAseq, methylation, MS-LC and NMR metabolomics and a wide set of clinical labs, like blood cell counts) to systematically estimate the downstream effects of VNTR polymorphisms on biological processes and cellular pathways. We focused on protein coding genes and as an example, we identified 265 robust (FDR <0.05) cis-eQTLs in blood. Furthermore, we observed that several VNTRs are in LD with the GWAS hits, indicating similar underlying causal mechanism. In conclusion, we have developed a computational framework to call VNTR polymorphisms and through systematic analysis within a phenotypically rich sample, have demonstrated a wide range of phenotypic consequences.

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

Contributed 4: Mendelian Randomization

14
Identifying Novel Genes Whose Tissue-Specific Expression Level Causally Influence Complex Human Traits

Eleonora Porcu[a,b], *Alexandre Reymond*[a], *Zoltán Kutalik*[b,c]

[a]Center for Integrative Genomics, University of Lausanne, Lausanne, [b]Swiss Institute of Bioinformatics, Lausanne, [c]Institute of Social and Preventive Medicine, CHUV and University of Lausanne, Lausanne, Switzerland

In the last decade Genome-Wide association studies (GWAS) identified thousands variants associated with hundreds complex traits but in many cases the underlying biological reason for a trait association is unknown. Given that many GWAS loci fall outside coding regions and the important role of regulatory variations in shaping complex phenotypes, gene expression QTLs (eQTL) can provide an interpretation of the variants in a GWAS region and the complex trait.

Recently, a study (Zhu et al. Nature Genetics 48, 481–487 (2016)) applied Mendelian Randomization (MR) approach using a single instrumental variable to search for the most functionally significant genes at the loci associated in GWAS for complex phenotypes by integrating association summary statistics and eQTL data.

Here, we propose an advanced MR approach that uses multiple SNPs jointly as instruments and multiple gene expression traits simultaneously. Our method is not only more powerful than the one previously published, but verifies MR assumptions more rigorously. We applied the method to schizophrenia using summary data from the latest GWAS meta-analyses and various eQTLs datasets in different tissues and identified 53 genes causally implicated in schizophrenia. At the same time, we replicated 10 out of 17 signals reported in Zhu et al and flagged up the remaining due to possible violations of MR assumptions.

Among our results we find examples of causally associated genes – identified by our method – with a) overestimated effect in single-gene analysis; b) insignificant association via single-gene analysis; c) tissue-specific causal effects; d) no genome-wide significant SNP nearby in previous GWAS analysis. These results suggest that a multi-SNP and multiple-gene approach verifies MR assumptions rigorously and is a powerful test, which can reveal new insights into the tissue-specific transcriptomic regulation of complex traits.

In order to translate our findings into a functional understanding of disease processes, we applied tissue-specific network analysis using ~400 tissue-specific gene regulatory circuits. We show how schizophrenia-associated genes disturb regulatory modules in tissues that are specific to schizophrenia, giving new insights on biological mechanisms of the disease. Furthermore, our results demonstrate the importance of integrating data from various tissues when trying to interpret GWAS results using gene expression as an intermediate phenotype.

15
A Framework for the Investigation of Pleiotropy in Two-Sample Summary Data Mendelian Randomization

Jack Bowden[a], *M. Fabiola Del Greco*[b], *Cosetta Minelli*[c], *George Davey Smith*[a], *Nuala A. Sheehan*[d], *John R. Thompson*[d]

[a]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; [b]Center for Biomedicine, EURAC research, Bolzano, Italy; [c]Population Health and Occupational Disease, NHLI, Imperial College, London, [d]Department of Health Sciences, University of Leicester, Leicester, UK

Mendelian randomization (MR) uses genetic data to probe questions of causality in epidemiological research, by invoking the Instrumental Variable (IV) assumptions. In recent years it has become commonplace to attempt MR analyses by synthesising summary data estimates of genetic association gleaned from large and independent study populations. This is referred to as two-sample summary data MR. Unfortunately, due to the sheer number of variants that can be easily included into summary data MR analyses, it is increasingly likely that some do not meet the IV assumptions due to pleiotropy. There is a pressing need to develop methods that can both detect and correct for pleiotropy, in order to preserve the validity of the MR approach in this context. In this paper we aim to clarify how established methods of meta-regression and random effects modelling from mainstream meta-analysis are being adapted to perform this task. Specifically, we focus on two contrasting approaches: the Inverse Variance Weighted (IVW) method which assumes in its simplest form that all genetic variants are valid IVs, and the method of MR-Egger regression that allows all variants to violate the IV assumptions, albeit in a specific way. We investigate the ability of two popular random effects models to provide robustness to pleiotropy under the IVW approach, and propose a statistic to quantify the relative goodness-of-fit of the IVW approach over MR-Egger regression.

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

16

Bayesian Association Scan Reveals Novel Loci Associated with Human Lifespan and Pinpoints Causally Implicated Transcriptome Biomarkers

Aaron F. McDaid[a,b], *Peter K. Joshi*[c], *Eleonora Porcu*[b,d], *Andrea Komljenovic*[b,e], *Hao Li*[f], *Vincenzo Sorrentino*[f], *Maria Litovchenko*[b,g], *Roel P.J. Bevers*[b,g], *Sina Rüeger*[a,b], *CHARGE Consortium, Alexandre Reymond*[d], *Murielle Bochud*[a], *Bart Deplancke*[b,g], *Robert W. Williams*[h], *Marc Robinson-Rechavi*[b,e], *Fred Paccaud*[a], *Valentin Rousson*[a], *Johan Auwerx*[f], *James F. Wilson*[c,i], *Zoltán Kutalik*[a,b]

[a]Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, [b]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [c]Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland; [d]Center for Integrative Genomics, University of Lausanne, Lausanne, [e]Department of Ecology and Evolution, University of Lausanne, Lausanne, [f]Laboratory of Integrative and Systems Physiology, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, [g]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics, Lausanne, Switzerland; [h]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, USA; [i]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland

It is believed that lifespan is highly polygenic, probably influenced by many genetic variants via predisposition to certain diseases or altering behaviour. Many genetic association studies have focussed on extreme case-control longevity (>90 y) analysis and revealed only very few robustly associated SNPs (mainly at the APOE and FOXO3 loci). Others used self-reported parental age of death in general population cohorts, but this approach only added one new locus (near CHRNA3/5).

Since most life-shortening events are disease-related, we developed a Mendelian randomization-based approach combined with 207 previous disease GWAS results to derive independent prior effects on longevity for all SNPs genome-wide. 11 of these 207 traits were shown to be predictive of lifespan, and therefore the prior effect of a SNP on lifespan could be estimated by summing up the products of the estimated effect of the SNP on the 11 traits and the causal (MR) estimates of the effect of each trait on lifespan. The 11 traits include cholesterol levels, diabetes, coronary artery disease, and smoking, alongside educational level, height and schizophrenia.

We found that these priors strongly correlate (r = 0.17) with SNP effects on parental age of death in the UK Biobank (n = 116,279). Combining these priors with the traditional lifespan association summary statistics from the UK Biobank, we derived Bayes Factors for each SNP, and we found 16 independent SNPs with Bayes factors significantly larger (at 5% false discovery rate) than expected under the null hypothesis of no genetic effect on lifespan.

Convincingly, all but two of these 16 have already been implicated in human diseases; nine of them replicate ($P < 0.05/16$) in at least one independent longevity study; all but three show depletion of the life-shortening alleles in older individuals in the UK Biobank. Further analysis revealed that brain expression levels of nearby genes (RBM6, SULT1A1, CHRNA5) might be causally implicated in longevity. Gene-expression and caloric restriction experiments in mice confirm the conserved role for RBM6 and SULT1A1 in lifespan determination. We believe that our study greatly contributes to the discovery and understanding of the biological mechanisms of longevity genetics.

17

Identification of Genetic Causal Pathways: Mendelian Randomization and Statistical Colocalization

Hui Guo[a], *Carlo Berzuini*[a], *Jeanine Houwing-Duistermaat*[b,c], *Markus Perola*[d]

[a]Centre for Biostatistics, School of Health Sciences, The University of Manchester, Manchester, [b]School of Mathematics, University of Leeds, Leeds, UK; [c]Leiden University Medical Center, Leiden, The Netherlands, [d]Diabetes and Obesity Research Program, University of Helsinki, Helsinki, Finland

Genome-wide association studies have found many genetic variants, in particular, single nucleotide polymorphisms (SNPs), associated with certain clinical outcomes of importance (*e.g.* cardiovascular disease). Some of these SNPs are also associated with modifiable exposures (*e.g.* gene expression). This overlap can be leveraged to learn about the biological mechanisms underlying these outcomes. Interventions can then be tailored on the basis of well-targeted causal pathways. Existing state-of-the-art methods including Mendelian randomization (MR) and statistical colocalization make important steps towards genetic causal inference.

Both MR and colocalization exploit summary statistics from two association studies: SNP-exposure and SNP-outcome. However, they are different in several ways. First, MR is tailored to identify a causal effect of the exposure on the outcome. Thus, SNPs used as instruments need not be causal for either. Colocalization is designed to identify common SNPs which are causal for both. Second, MR uses exposure associated SNPs as instruments, while colocalization often starts with outcome associated regions and identifies an exposure that explains the effect of the SNPs on the outcome. Third, both approaches can be applied to the scenario which comprises three essential components: SNPs, exposure and outcome. However, colocalization can also be used to test for common causal signals of multiple variables that occur in no particular temporal order. Moreover, colocalization cannot detect the direction of a causal relationship, while MR is a useful tool for such purposes.

Taken together, both MR and colocalization have their own limitations. Therefore there is a need to take forward strengths of the two methods in a unifying approach. We will focus on MR Egger regression and Bayesian colocalization, and possible ways of combining them appropriately. We will make the use of data collected from the Dietary, Lifestyle and Genetic determinants of

Obesity and Metabolic syndrome (DILGOM) study which comprises genotype information, gene expression level and concentration measures of 137 metabolites from a Finnish cohort. This study enables integrative analysis on potential causal pathways from genotypes to metabolites through gene expression level, which will ultimately help further study aetiology of metabolic disorders.

Contributed 5: Post-GWAS, Polygenic Prediction

18
Tree-Based Transcriptome Wide Association Studies in Autoimmune Diseases

Nastasiya Grinberg[a], *Chris Wallace*[a,b]

[a]Department of Medicine and [b]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Genome-wide association studies (GWAS) have been hugely successful over the last decade, identifying over 25 k variants associated with more than 1500 traits (phenotypes). However, interpretability remains a criticism of GWAS – most disease-associated variants situated in regulatory regions have not yet been linked to the genes they regulate. One approach to linking GWAS variants to traits through genes is by using an eQTL dataset to learn imputation rules with which to impute gene expression in GWAS samples, enabling gene-based instead of variant-based, disease-control comparisons (transcriptome wide association studies; TWAS).

We aimed to improve the predictive accuracy of the first part of this procedure, constructing predictive models for gene expression, in order to increase power of the TWAS overall. Currently cis-based linear models, such as various regularised regressions are used. We propose to use a powerful machine learning method of random forests (RF) for eQTL mining. One of the advantages of RF is its inherent non-linear nature which enables it to take complex interactions between SNPs into account. Furthermore, we propose to also relax the cis restriction by mining for and including trans-eQTLs in our random forest models. We demonstrate that RF generally outperforms linear methods: RF beats lasso, ridge and elastic net regressions with average advantages of 5%, 6% and 12%, respectively, in most of ~15 K regressions we performed. Moreover, RF beats lasso, the strongest of the three alternative methods, by more than 10% in ~2 K cases. Finally, we show that a multi-task approach to eQTL-mining, taking advantage of correlation between gene expression in different cell types, can further improve prediction in specific situations.

We applied our tree-TWAS to eQTL data from a variety of immune cells (monocytes, B cells and T cells) and used it to impute transcription into GWAS subjects with autoimmune diseases and healthy controls. We highlight several case studies illustrating the ability of the two-step TWAS procedure to discover disease association signals not detectable at the GWAS level and leading to identification of putative implicated genes (supported by literature). We also highlight an important caveat of these methods: that a second stage of testing is required to eliminate cases where eQTL and disease causal variants are in LD, but not colocalised.

19
Improved Summary Statistics Imputation for Studies with Variable Missingness Pattern or Multi-Ethnic Samples

Sina Rüeger[a,b], *Aaron F. McDaid*[a,b], *Zoltán Kutalik*[a,b]

[a]Institute of Social and Preventive Medicine, University Hospital (CHUV), Lausanne, [b]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Genome-wide association studies (GWASs) provide an estimated effect size of each genotyped variant on complex traits. Despite dropping sequencing costs, the most common technology for genomic data collection remains to be genotyping, which however only covers a fraction of the full genome.

Many methods that use GWAS summary statistics for follow-up analyses, such as calculating genetic correlation (via LD score regression) or causal inference (Mendelian randomisation), need summary results for the full genome, or at least a set of overlapping SNPs. These challenges require imputation, where typed variants act as tag SNPs to infer un-typed variants. Although genotype imputation provides a robust solution, with the frequent emergence of new reference panels it requires time and effort from investigators. On the contrary, summary statistics imputation needs no new analysis from participating studies. Summary statistics imputation requires two types of inputs: summary statistics of typed SNPs and the pair-wise LD between all variants involved. The latter is estimated from sequencing data of an external reference panel. We developed two extensions to summary statistics imputation that improve the estimation of the latter quantity. First, we improved the estimation of the correlation matrix when summary statistics come from a GWAS with different population composition than the reference panel. Second, we addressed the problem that summary statistics for different SNPs may be derived from different sample sizes.

If the summary statistics is resulting from a GWAS with possibly mixed populations, choosing an appropriate reference panel may prove to be difficult. Instead of estimating the correlation matrix using the full reference panel (e.g. the European sample), we propose to derive a re-weighted LD matrix to better match the LD structure of the GWAS population. We tested our method in a simulation framework. Compared to the default – using the full European panel – we yield in a smaller mean squared error for the majority of the cases. We also outperform the existing method Adapt-Mix.

The sample of individuals used to compute a Z-statistic might vary from SNP to SNP, for example when meta-analysing different genotype chips. As a result of this variation, the correlation between Z-statistics will change. We show in realistic simulations that accounting for sample size variation can reduce the mean squared error of the corresponding standardized effect estimate by

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

>50% when the true effect is relatively strong (>0.1% explained variance). This method requires only the available sample size for each SNP.

20

Genomic Prediction of Type 2 Diabetes and Coronary Heart Disease Following Diabetes

A.S. Havulinna[a,b], *G. Abraham*[c], *K. Kristiansson*[b], *M. Perola*[b], *V. Salomaa*[b], *M. Inouye*[c], *S. Ripatti*[a]

[a]FIMM – Institute for Molecular Medicine Finland, Helsinki, [b]THL – National institute for Health and Welfare, Helsinki, Finland; [c]Centre for Systems Genomics, School of BioSciences, The University of Melbourne, Melbourne, Australia

The heritability of type 2 diabetes is high (>50%), but common variants capture only ~10% of the heritability of type 2 diabetes. These genome-wide significant SNPs have been used to form genetic risk scores with tens of SNPs, but this doesn't solve the missing heritability problem. We have recently shown the benefit of using thousands of LD-pruned SNPs for constructing a polygenic genomic risk score (GRS) for coronary heart disease (CHD), with more accurate risk prediction and stratification. Using an accurate GRS, risk prevention can be better targeted for those high-risk individuals, who are not detected using traditional risk factors only. Recently, Fischer et al. showed that a similar approach stratified the risk of diabetes in Estonian samples. In this project, we test a highly polygenic prediction in two settings: 1) by constructing and evaluating the performance of a polygenic GRS to predict first-ever diagnosis of type 2 diabetes in large-scale Finnish population-based cohorts, and 2) evaluate the behaviour of both the CHD GRS and the diabetes GRS on CVD outcomes after the diabetes diagnosis.

We derive the GRSs using the results from DIAGRAM meta-analysis and test the performance of the GRSs independently in the FINRISK and H2000 studies. We use C-statistics, NRI, IDI and model calibration to assess the performance. Similar to Goldstein et al., we evaluate the effect of using different LD-pruning thresholds, p-value thresholds and imputation quality thresholds in the GRS performance. We also test whether the use of Finnish LD structure instead of LD from non-Finnish data makes a difference.

Our preliminary data from FINRISK show that LD pruning has large effect on prediction of incident diabetes. When taking the relative weights from each SNP from the external DIAGRAM data, the hazard ratio for incident diabetes is 1.30 per GRS SD (95% CI: 1.21–1.39, $p = 5.1\times10^{-14}$). We are now extending the evaluation of the scores for CVD risk assessment after the diabetes diagnosis and validating the results in a non-Finnish study.

21

POLARIS: Polygenic LD-Adjusted Risk Score Set-Based Method

Emily Baker[a], *Karl Michael Schmidt*[b], *Peter Holmans*[a], *Michael O'Donovan*[a], *Valentina Escott-Price*[a]

[a]Medical Research Council Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, [b]School of Mathematics, Cardiff University, Cardiff, Wales, UK

Set-based analysis can gain power over single SNP analysis to detect disease associations by using the combined effect of SNPs within a set. Polygenic risk scores are a widely used method to summarise the additive trait variance captured by a set of SNPs, and can increase the power of set-based analyses by leveraging large public GWAS datasets. We propose the application of polygenic risk scores as a set-based method with an additional component of adjustment for linkage disequilibrium; this informs the analysis with previously reported effect sizes of a SNP's association to disease, and accounts for linkage disequilibrium between SNPs. This method is termed POLARIS: POlygenic Linkage disequilibrium-Adjusted RIsk Score. The POLARIS method identifies the linkage disequilibrium structure of SNPs in the test set using spectral decomposition of the SNP correlation matrix and adjusts the effect sizes from the discovery set used in the risk score. POLARIS scores are calculated per subject per set and the overall association of the set is determined using logistic regression on the adjusted polygenic risk score and additional population covariates. We used simulations to compare the power of POLARIS to MAGMA's set based approach. We then applied these approaches to real data to check the consistency with simulations. We observed that POLARIS has greater power than MAGMA in the majority of simulated scenarios, which reflect real genetic data. We also designed a software tool for whole genome, gene and pathway analysis, which reads data of standard format and will operate on any computing platform.

Contributed 6:
GWAS: Methodology and Interpretation

22
Genetic Classification of Individuals from a Large Schizophrenia Dataset Using Artificial Neural Networks

Carlos Pinto, Michael Gill, Elizabeth Heron

Data Analysis Unit, Neuropsychiatric Genetics Research Group, University of Dublin, Dublin, Republic of Ireland

Genome wide association studies (GWAS), that capture common genetic variation in the form of hundreds of thousands of genotyped single nucleotide polymorphisms (SNPs) are now commonplace. Such studies are used to identify SNPs associated with a trait. They can also be used for identifying affected and unaffected individuals on the basis of their genotypes. This holds out the possibility of applications in the clinical and public health areas even in the absence of an understanding of the genetic mechanisms involved.

Recent approaches to the problem have focused on the polygene risk score (PRS) method. A polygene score for an individual is created by summing the number of associated alleles an individual possesses, each weighted by the strength of that association. Affected individuals tend to lie towards the higher end of the distribution of scores.

The method has recently been applied to a large schizophrenia dataset generated by the Psychiatric Genomics Consortium (PGC). This dataset consists of approximately 34000 cases and 46000 controls sourced from 49 non-overlapping case-control subsets of varying ancestry. Results indicate that while the PRS can yield useful results, it has insufficient discriminatory power to allow for individual prediction at a useful level of accuracy.

We have investigated an alternative approach to classification, using an artificial neural network (ANN). An ANN consists of a number of nodes, each of which processes information and passes it to other nodes. There is an "input layer" which receives its inputs from the user, an "output layer" which delivers the outputs and one or more intermediate "hidden" layers. In training, inputs and outputs are specified and the ANN adjusts the output of each node relative to its input accordingly. In this way the ANN "learns" to recognise patterns.

In our application the inputs are the genotypes of an individual. The output is the case or control status of the individual. The network is trained on a subset of the data. After training, the network is tested on an independent subset to assess its accuracy in predicting case/control status.

In this talk we will describe our method and compare the performance of the ANN with the PRS in classifying cases and controls from the Psychiatric Genomics Consortium schizophrenia datasets.

23
Use of National Health Service Electronic Dental Treatment Records in Dental Public Health Genetics Research

M.L. Bermingham[a], *A. Campbell*[a], *D.J. Porteous*[a,b], *A.W.G. Walls*[c]

[a]Centre for Genomics and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, [b]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, [c]Edinburgh Dental Institute, University of Edinburgh, Edinburgh, Scotland

Electronic health records provides unprecedented opportunity for its re-use for many tasks, including public health genetics research. However, electronic health records data from clinical settings, such as dental practices can be inaccurate or of insufficient granularity to be of use. In this study we wish to determine the utility of National Health Service (NHS) electronic dental treatment records in dental public health genetics research. The objective of this study is to estimate the heritability of periodontal disease using NHS electronic dental treatment records linked to health and non-health data within the Generation Scotland: Scottish Family Health Study (GS:SFHS). We linked 852,355 NHS Scotland electronic dental treatment records *from April 2000 to July 2015* to 20,626 participants within GS:SFHS with pedigree, genomic, sociodemographic and clinical data. We then conducted a proof-of-principle genetic analysis using periodontal (gum) disease treatment records. The dataset analysed consisted of 160,508 dental treatment records from 13,717 study participants; 3,387 of which were periodontal treatment records (from 2,192 study participants). We adjusted for the effects of previous treatment record, interval since last treatment, age, sex, treatment year, and treatment month, Scottish Index of Multiple Deprivation (SIMD), alcohol consumption, diabetes diagnosis, and smoking status in a linear model in the statistical software ASReml. We then calculated the mean risk of periodontal disease for each study participant based on residuals extracted from the aforementioned model. Genome-wide complex trait analysis (GCTA; with correction for population stratification) was used to estimate the pedigree and genomic based heritability of periodontal disease. We estimated the familial heritability of periodontal (gum) disease at 10.42% (95% confidence interval 5.97–14.88%). The genomic component did not contribute significantly to the heritability estimate. We have demonstrated the usefulness of electronic dental treatment records in dental public health genetics research. This study has also, to the best of our knowledge provided the first population based estimates of the genetic parameters for periodontal disease; confirming its familial nature. The invaluable and unique NHS Scotland electronic dental treatment records resource will allow the acceleration of dental public health genetics research in Scotland and the exploration of research questions that could not be considered previously.

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

24

Improved Statistical Methods and Bioinformatic Tools to Integrate Multi-Omic Data in Disease Association Studies

Juan R. Gonzalez

Bioinformatics Research Group in Genetic Epidemiology (BRGE), Barcelona Institute for Global Health (ISGlobal), Spain

Reduction in the cost of genomic assays has generated large amounts of biomedical-related data. As a result, current studies perform multiple experiments in the same subjects. For instance, epidemiological studies include SNP array, gene expression and methylation data among others. Knowing existing methodologies for analyzing and integrating this type of data in disease association studies is crucial. In this talk, different issues related to how to tackle this problem will be presented. We will start by introducing MultiDataSet, a new R class based on Bioconductor standards, designed to encapsulate and properly manage multiple data sets. This tool will allow us to deal with the usual difficulties of dealing with multiple and non-complete data while offering a simple and general way of subsetting features and selecting samples. Then, statistical methods to integrate different types of omic data will be presented. These will include multivariate methods such as generalized canonical correlation (GCCA) or multiple coinertia analysis (MCIA). The final topic will cover how to enhance GCCA to consider real problems where there are missing individuals of a given omic dataset. This is a really common scenario in most studies since researchers normally combine omic data of samples belonging to the same cohort but obtained from different projects that normally provide data of different sample sizes. The proposed methodology will be illustrated by using data from the TCGA project where information about different omic data is public available for a large number of cancer samples. We will also show how the statistical power is dramatically reduced when analyzing more than one omic data at the same time since current multivariate methods only allow the analysis of complete cases.

Acknowledgments: This work has been partly funded by the Spanish Ministry of Economy and Competitiveness (MTM2015-68140-R).

25

IMHOTEP – A Composite Score Integrating Popular Tools for Predicting the Functional Consequences of Non-Synonymous Sequence Variants

Amke Caliebe[a], *Carolin Knecht*[a], *Matthew Mort*[b], *Olaf Junge*[a], *David N. Cooper*[b], *Michael Krawczak*[a]

[a]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [b]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

The *in silico* prediction of the functional consequences of mutations is an important goal of human pathogenetics. However, bioinformatic tools that classify mutations according to their functionality employ different algorithms so that predictions may vary markedly between tools. We therefore integrated nine popular prediction tools (PolyPhen-2, SNPs&GO, MutPred, SIFT, MutationTaster2, Mutation Assessor and FATHMM as well as conservation based Grantham Score and PhyloP) into a single predictor. The optimal combination of these tools was selected by means of a wide range of statistical modelling techniques, drawing upon 10 029 disease-causing single nucleotide variants (SNVs) from Human Gene Mutation Database and 10 002 putatively 'benign' non-synonymous SNVs from UCSC. Predictive performance was found to be markedly improved by model-based integration, whilst maximum predictive capability was obtained with either random forest, decision tree or logistic regression analysis. A combination of PolyPhen-2, SNPs&GO, MutPred, MutationTaster2 and FATHMM was found to perform as well as all tools combined. Comparison of our approach with other integrative approaches such as Condel, CoVEC, CAROL, CADD, MetaSVM and MetaLR using an independent validation dataset, revealed the superiority of our newly proposed integrative approach. An online implementation of this approach, IMHOTEP ('Integrating Molecular Heuristics and Other Tools for Effect Prediction'), is provided at http://www.uni-kiel.de/medinfo/cgi-bin/predictor/.

Invited 2: Population Genetics

26

Personal Ancestry Inference at the Finest Scale Can Detect the County of Origin of the Sailors that Sailed with Cook to the Society Islands

Daniel J. Lawson

Department of Mathematics, University of Bristol, Bristol, UK

Chromosome Painting has revealed genetic differences within the UK at a very fine scale (Leslie *et al.* 2015), with structured genetic variation within a single county in some cases (such as Cornwall & South Wales). However, in that work, it was not possible to genetically distinguish much of England, which appeared as a single homogeneous group. Here, we describe an extension to the

FineSTRUCTURE (Lawson *et al.* 2012) clustering that can further distinguish ancestry even within England; for example, identifying regions such as Norfolk, the Midlands and the South as genetically distinct. The approach works by using the known county locations to craft genetic features to use in unsupervised clustering. Specifically, we group individuals by their geographic sampling location into reference donor populations. This forms an ancestry profile – which can be viewed as a careful choice of feature vector – that still allows unsupervised genetic clustering for all individuals.

Further, we describe how this approach allows individuals to be described as an admixture of the inferred geographical clusters. This allows ancestral information to be recovered for individuals who are not purely represented by a single geographical location. This also allows us to characterise the genetic relationship between the inferred clusters, several of which represent drift that is most strongly represented by a particular geographical region (including Cornwall, Wales, Scotland and the North of England) and others of which represent characteristic admixture proportions between these ancestral drifted populations.

Beyond improving resolution, this approach facilitates personal genomics because individuals can be represented in terms of the fixed reference panel. We demonstrate the utility of the approach by describing the ancestry of the UK10K participants in terms of the new, high resolution POBI clusters. Previously, a similar analysis (UK10K Consortium 2015) without geographical information inferred little population structure in the UK from these samples, but now we have a rich representation of their population structure, including an assessment of admixture from outside the UK. Even more excitingly, we can describe the county of origin of the ancestry found in present-day Society islanders, which dates to Captain Cooks repeated visits to the island. An analysis of which sailors left DNA reveals surprising insights into the European ancestry introgression process in the Polynesian peoples.

27

Reconstructing Past History from Whole-Genomes: An ABC Approach Handling Recombining Data

Flora Jay[a,b], *Simon Boitard*[c], *Frédéric Austerlitz*[a]

[a]Laboratoire EcoAnthropologie et Ethnobiologie, CNRS/MNHN/Université Paris Diderot, Paris, [b]Laboratoire de Recherche en Informatique, CNRS/Université Paris-Sud/Université Paris-Saclay, Orsay, [c]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosan, France

In population genetics, a key interest is to reconstruct the demographic history of a population using its genetic data. This history can be characterized by multiple events such as migration of individuals, admixture with another population, or changes in population size. With the availability of large-scale genomic data numerous methods have arisen for untangling complicated histories or retrieving a detailed picture of a population at different time periods.

Although genomes are known to be extremely informative about demography, there are many ways to extract this information. We present an approach designed for inferring past population sizes for an intermediate number of fully sequenced genomes. It relies on Approximate Bayesian Computation (ABC), a simulation-based statistical framework for generic model comparison and parameter inference. We demonstrated how the specificities of DNA sequencing data (namely haplotypic information, long range genetic correlation and genotyping errors) can be handled using ABC and fast genetic simulators, and further infer histories of successive bottleneck and expansions in human populations.

Contributed 7: Population Genetics: Methodology and Applications

28

Estimating the Onsets of Selection on New Mutations and Standing Variation in Human Populations

Shigeki Nakagome[a,b], *Richard R. Hudson*[a,c], *Anna Di Rienzo*[a]

[a]Department of Human Genetics, University of Chicago, Chicago, IL, USA; [b]School of Medicine, Faculty of Health Sciences, Trinity College Dublin, the University of Dublin, Dublin, Ireland; [c]Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA

Genetic variation harbours signatures of natural selection driven by pressures that are often unknown. Estimating the ages of selection signals may allow reconstructing the history of environmental changes that shaped human phenotypes and diseases. Natural selection on new mutations, often referred to as a selective sweep, is expected to reduce levels of genetic variation linked to an advantageous allele. However, a signature of selection on pre-existing alleles is complicated and strongly depends on when selection has happened (t) and what the frequency of standing variation (f_t) is at the time of selection. The difficulty of choosing an evolutionary model comes from a lack of informative statistics that can capture subtle differences in a pattern of genetic variation between standing variation and neutral models. Here, we apply kernel Approximate Bayesian Computation (kernel ABC) to use a large amount of information on the haplotype structure and the full site frequency spectrum. Our aims are to discriminate between neutrality and different selection models, *i.e.*, selection on new mutation or on standing variation, and to estimate the age of selection for alleles with selection signals. Assuming that an advantageous allele is present at intermediate frequency in a population (f_{obs} = 50%), we evaluate the performance of model selection using Bayes factors to reject the neutral model and of age estimation under positive selection on a new mutation or a standing neutral allele. We show that our method has sufficient power to distinguish positive selection from the neutral model even when f_t is as low as 10%, and t is relatively old (1,200 generations ago). Although kernel ABC can accurately estimate the onset of selection on a new mutation, the estimation is prone to bias in priors on t and f_t under the standing variation model possibly due to a wide range of the parameter space on f_t. Then, we extended our approach to incorporating ancient DNA data that can give more information on t and

f_t in the past. Our simulation study shows a significant reduction of the bias by confining the parameter space based on the ancient DNA observations. These results suggest that the kernel ABC with the high-dimensional data provides a useful and a flexible framework to understand the evolutionary dynamics of alleles, which are associated with adaptive traits.

29

Revisiting the Male Genetic Landscape of China

Michael Nothnagel[a], *Sascha Willuweit*[b], *Lutz Roewer*[b], *on behalf of the China Y-STR study group*

[a]Cologne Center for Genomics (CCG), University of Cologne, Cologne, [b]Institute of Legal Medicine and Forensic Sciences, Charité – Universitätsmedizin Berlin, Berlin, Germany

Numerous studies have investigated China's genetic diversity. Early studies reported a genetic distinction between Northern and Southern Han Chinese, while others showed a picture of more subtle differences. Here, we investigated the distribution of Y chromosome variation across 28 administrative regions as well as 19 recognized Chinese nationalities in continental China to assess the impact of recent demographic processes. To this end, we analyzed 37,994 Y chromosomal 17-marker haplotype profiles from the YRHD database with respect to forensic diversity measures and genetic distance between groups defined by administrative boundaries and ethnic origin, representing the largest genetic study on China to date. We observed high diversity throughout across all investigated Chinese provinces and ethnicities. Kazakhs and Tibetans showed the strongest significant genetic differentiation from the Han and other groups. However, differences between provinces were, except for those located on the Tibetan plateau, less pronounced. This discrepancy is explicable by the sizeable presence of Han speakers, who showed high genetic homogeneity all across China, in nearly all studied provinces. We also observed a subtle genetic North-South gradient in the Han, confirming previous reports of a clinal distribution of Y chromosome variation and being in notable concordance with the previously observed spatial distribution of autosomal variation. Our findings shed light on the demographic changes in China accrued by a fast-growing and increasingly mobile population.

30

Investigating Fine-Scale Population Structure in UK BioBank

James P. Cook, *Andrew P. Morris*

Department of Biostatistics, University of Liverpool, Liverpool, UK

The United Kingdom (UK) is a genetically diverse population with strong genetic differences between regions, which may adversely affect genome-wide association studies (GWAS) of complex traits if not fully accounted for in the analysis. Principal components, calculated from a genetic relatedness matrix, are routinely included in regression models to account for population structure in GWAS.

UK BioBank provides an opportunity to examine fine-scale UK population structure in unprecedented detail, with a first release of genetic data consisting of ~150,000 genotyped individuals recruited from 23 centres. Easting and Northing coordinates were collected for every participant at recruitment and birth, and genotypes were imputed up to a combined 1000 Genomes and UK10K reference panel.

We have performed univariate GWAS analysis of imputed SNPs using the Easting and Northing coordinates at birth as continuous phenotypes, with and without adjustment for principal components. In analyses adjusted for 6 principal components, variants mapping in *TLR1* showed the strongest signal of association genome-wide with both Northings (rs4833095, previously associated with asthma and hay fever, $p = 3.9\times10^{-126}$) and Eastings (rs4543123, previously associated with alcohol tolerance, $p = 1.4\times10^{-15}$), representing a North-West to South-East cline in allele frequencies across the UK. However, after adjusting for 15 principal components, the strongest *TLR1* association with Northings was reduced to $p = 2.3\times10^{-21}$, whilst no genome-wide significant associations ($p < 5\times10^{-8}$) with Eastings remained. The use of linear mixed models to adjust for fine-scale structure (implemented using Bolt-LMM) further reduced the number of genome-wide significantly associated regions in the Northings analysis from 10 (adjusting with 15 principal components) to 7, and *TLR1* remained the strongest association signal.

We have also performed GWAS of blood pressure and asthma traits. We demonstrate that further adjustment with additional principal components is not essential because these traits are not confounded by population structure. Finally, we analysed rs4833095 with a self-reported asthma phenotype in a case-control analysis with 14,926 cases and 117,169 controls, revealing that although the difference in association strength was small, the inclusion of additional principal components resulted in the variant reaching genome-wide significance ($p = 6.7\times10^{-9}$ with 15 PCs, $p = 9.5\times10^{-8}$ with 6 PCs).

Our study has important implications for accounting for population structure in large scale GWAS performed in UK BioBank, and highlights the need for caution in interpreting association results from regression models incorporating principal components.

31

Fine-Scale Human Genetic Structure in France

Aude Saint-Pierre[a–c], *Céline Bellenguez*[d–f], *Luc Letenneur*[g,h], *Claudine Berr*[i,j], *Carole Dufouil*[g,h], *Philippe Amouyel*[d–f,k], *Emmanuelle Génin*[a–c]

[a]Université de Bretagne Occidentale, Brest, [b]Inserm UMR1078, Brest, [c]Centre Hospitalier Régional Universitaire de Brest, Brest, [d]Inserm, U744, Lille, [e]Université Lille 2, Lille, [f]Institut Pasteur de Lille, Lille, [g]U897, Inserm, Bordeaux, [h]Université Bordeaux 2, Bordeaux, [i]U1061, Inserm, Montpellier, [j]Université Montpellier, Montpellier, [k]Centre Hospitalier Régional Universitaire, Lille, France

Characterizing geographical population structure is critical to genetic studies of disease as it is an important cause of false positive results in genome wide association studies (GWASs). The genetic structure of several countries in Europe has been carefully studied but there is a lack of descriptive study of the French population. Indeed, apart from the work of Karakachoff et al (2015) that focused on the western part of France and detected interesting stratification, no study so far provided a comprehensive look at the French genetic landscape. Here we describe the genetic structure of the French population at a fine-scale using genetic data from the 3 Cities (3C) Study, a population cohort of French elderly individuals that served as controls in several GWAS conducted with French patients. From this cohort, we had access to 4,433 genotyped individuals sampled in three regions of France but born all over France.

We selected a subset of 770 individuals to cover evenly the different regions of France and applied methods that utilize haplotype information for detecting fine-scale population structure. The 770 individuals were partitioned into homogeneous clusters using CHROMOPAINTER and fineSTRUCTURE analysis (Lawson et al. 2012). Six clusters were identified that correlate well with the geographic origin of individuals. The coarsest level of genetic differentiation separates the samples from southwestern French from all the others. Subsequent splits reveal more subtle differentiation except for samples from western France which showed a relatively high degree of homogeneity.

For each cluster we used CHROMOPAINTER to estimate an "ancestry profile" which characterises the ancestry of the cluster as a mixture of the reference sample. Using the subsample of 770 individuals as a reference sample to assign the remaining 3C individuals, we found that the cluster assignment was coherent with the places of birth of individuals. The same procedure was applied using the five European samples from the 1000 Genomes Project as a reference sample. Contribution from European populations shows a cline roughly north-south, in ancestry profiles. Spain (IBS) is the largest contributor of the southwest and south clusters while the highest contribution of Great-Britain (GBR) population is observed in Brittany.

In conclusion, we provide evidence that there exist some levels of genetic stratification in France. The French population could roughly be divided into 6 genetic clusters that correlate well with geography. The knowledge of this stratification pattern will be useful to design robust and powerful association studies.

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638

Human Heredity

First Session

P1

Detecting Cancer Related Mutations in Cell-Free DNA of Lung Cancer Patients

Madli Tamm[a], *Kersti Oselin*[b], *Paula Ann Kivistik*[a], *Mart Kals*[a], *Katrin Keerma*[a], *Retlav Roosipuu*[c], *Tiina Leismann*[b], *Hanno Roomere*[d], *Andres Metspalu*[a,e], *Neeme Tõnisson*[a,d]

[a]Estonian Genome Center, University of Tartu, Tartu, [b]North Estonia Medical Centre, Tallinn, [c]Department of Pathology, Tartu University Hospital, Tartu, [d]Department of Genetics, Tartu University Hospital, Tartu, [e]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

According to WHO, lung cancer is the leading cause of cancer deaths worldwide. Besides invasive tumor biopsy, molecular diagnostic analysis of tumor cell-free DNA is becoming increasingly popular as a method enabling to capture the whole tumour heterogeneity in almost real-time setting. This may have applications in both early diagnostics, as well as therapeutic monitoring of alterations predictive for drug response.

Our primary objective was to set up a straightforward platform for the analysis of mutations relevant in targeted drug therapy. Our study cohort comprises of 70 lung adenocarcinoma patients who have donated blood plasma samples prior to initiation of chemo- or targeted therapy. The cohort will be longitudinally monitored and repeated blood samples collected upon progression. FFPE tumor samples have been available for approximately 50% of the study subjects. We have currently set up an allele-specific fragment analysis (FLA) workflow for EGFR gene common mutations and an amplicon-based multiplex next-generation sequencing (NGS) in 5 lung cancer-related genes (EGFR, BRAF, HER2, KRAS, PIK3CA). Test results by two major certified diagnostic laboratories were considered to be the gold standard and compared to the results of FLA and NGS.

After filtering noise and germline events our results showed that both approaches detect <1% mutant allele content. The methods were compared in both cfDNA and FFPE analysis, as well as with the available clinical DNA data. The overall concordance between diagnostic laboratory results and FLA, NGS manual analysis and NGS software based analysis results was 93%, 91% and 89%, respectively. The sensitivity of analysing FFPE DNA material was higher than analysing cfDNA but cfDNA had better specificity over FFPE DNA material.

Our further aim is to collect longitudinal samples and perform a larger screen of drug susceptibility and resistance mutations in our study cohort, involving detection of point mutations, copy number analyses and DNA methylation changes. We strongly believe that cfDNA analysis will become a clinical routine in various cancers in the near future.

P2

An Investigation into the Nutrigenomics of Pancreatic Cancer Using Data from the EPIC Study

Anna Urich[a], *Marika Kaakinen*[b], *Longda Jiang*[a], *Marc Gunter*[c], *Inga Prokopenko*[a]

[a]Department of Genomics of Common Disease, Imperial College London, London, [b]Department of Medicine, Division of Experimental Medicine and Toxicology, Imperial College London, London, UK; [c]Section of Nutrition and Metabolism, International Agency for Research on Cancer, 69372, Lyon, France

Pancreatic cancer (PanC) is one of the deadliest cancers with a 7% five-year survival rate. Early diagnosis poses a challenge to clinicians as symptoms are non-specific and arise at an advanced stage. Epidemiologic, *in vitro* and animal studies have suggested the role of diet in cancer risk and prognosis. We have conducted a genome – wide association study (GWAS) and genome–wide gene–environment (GxE) interaction analyses to dissect the genetic effects influencing PanC risk and those affecting PanC in relation to the nutrient consumed.

We used epidemiologic and genotype data from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort study from 518 cases and 5297 controls. Genotype data was imputed to the 1000 Genomes Project Phase 3 reference panel (June 2014) and yielded a total of 5,535,845 common SNPs (MAF >0.05) for analyses. We assumed log-additive genetic model and performed GWAS using logistic regression implemented in SNPTEST. The GxE interaction tests were conducted using QUICKTEST with logistic regression to test for interaction between eight nutritional variables and allelic effects of DNA variants in EPIC. The interaction model for each nutritional variable was defined as follows:

$$log(P(D)/1–P(D)) = \beta_0 + \beta_G \times G + \beta_E \times E + \beta_{GE} \times G \times E,$$

where D is the cancer status, G is the genotype and E is the environmental exposure (nutrient intake).

We confirmed allelic effects at all previously reported PanC loci and detected novel genome-wide significant association (P =

5.0×10^{-8}) with PanC at four novel loci (*GSTT2B/GSTT2, HLA–DRB1, HLA–DQB1, KIR3DL2*). We assumed two independent GxE GWA analyses of fat and carbohydrates, since other six nutrients represented fractions of these two and were highly correlated with them ($r > 0.5$). The GxE interaction analyses with sugar intake yielded one genome-wide significant association, corrected additionally for two GWA analyses, at *LINC01365* ($P = 1.59\times10^{-8}$, OR [95% CI] = 1.0013 [1.00088–1.0018]) and one locus at *ST6GALNAC3* just above the genome-wide significance threshold ($P = 5.16\times10^{-8}$, OR [95% CI] = 1.0007 [1.0004–1.0009]).

Novel loci highlighted *GSTT2B/GSTT2* genes, members of the *GSTT* gene family previously implicated in cigarette-smoke induced adenocarcinoma of the exocrine pancreas with a function to protect cells from a range of endogenous and exogenous chemicals and oxidative stress. Additionally, *ST6GALNAC3* codes for an enzyme related to the O-linked glycosylation pathway involved in the aberrant glycosylation of mucins observed in pancreatic tumours. In order to rule out spurious associations due to biases or chance alone, replication of novel loci in wider PanC GWAS as well as in the GxE interaction analysis of sugar intake and PanC disease risk is essential.

P3

Genetic Effects on Chromatin Accessibility Foreshadow Gene Expression Changes in Macrophage Immune Response

Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew Knights, Gordon Dougan, Daniel Gaffney

Wellcome Trust Sanger Institute, Hinxton, UK

Gene expression quantitative trait loci (eQTL) mapping studies can reveal the functions of common genetic variants. However, many genetic effects may go unobserved when cells or tissues are sampled in a single state. This is particularly true for immune cells, whose cellular function and gene expression can be substantially altered by external cues. We differentiated macrophages from induced pluripotent stem cells in 86 unrelated, healthy individuals derived by the Human Induced Pluripotent Stem Cells Initiative (HIPSCI), and profiled gene expression and chromatin accessibility in four experimental conditions: naïve, interferon-gamma (IFNγ) treatment, *Salmonella* infection and IFNγ treatment followed by *Salmonella* infection. We detected gene expression QTLs (eQTLs) for 5,383 genes, and chromatin accessibility QTLs (caQTLs) for 32,918 accessible regions, including hundreds of long-range interactions. We show that profiling even a small number of additional cellular states substantially increases the number of eQTLs that we can confidently colocalise with a known disease association, with approximately 30% new disease-eQTL pairs discovered in each additional state. Furthermore, we show that approximately 50% of stimulus-specific effects on gene expression manifest in naïve cells where they alter chromatin accessibility alone. Our results suggest that many disease-associated genetic variants lie in regulatory elements in a 'primed' state waiting for an appropriate environmental signal before regulating gene expression.

P4

New Model for Association Analysis of Multivariate Traits in Family-Based Samples

G.R. Svishcheva[a,b], *N.M. Belonogova*[a], *T.I. Axenovich*[a,c]

[a]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, [b]Vavilov Institute of General Genetics, the Russian Academy of Sciences, Moscow, [c]Novosibirsk State University, Novosibirsk, Russia

Recently the multiPhen method using the reverse multiple linear regression was proposed in contrast to the classical linear regression usually applied in GWAS [O'Reilly et al., 2012]. This regression technique considers the phenotypic data as predictors, while the genotypic data as independent variables. Thus, the requirements of normality of distribution are imposed to the genotypic but not phenotypic data. Since genotypes belong to the categorical data, the model predicts not the genotypes themselves, but their probabilities, using the logistic regression approach. MultiPhen demonstrates sufficiently high statistical power; however, the model is not applicable to the analysis of family-based samples because the genotypes of different pedigree members are not independent.

We propose a new model of linear regression for multivariate traits involved in a common biological process or the complex etiology of diseases, in the family-based samples.

For a family-based sample of n individuals, let Y denote a $(n\times t)$ matrix of known values of t traits adjusted on covariate effects and standardized, g denote a $(n\times1)$ vector of known genotypes of a genetic variant of interest. The genotypes are coded as the number of minor alleles (0, 1 or 2).

First of all, we introduced special matrix transformations which convert real genotypes and phenotypes into family independent values:

$$g^* = Q_g g \text{ and } Y_i^* = Q_{Yi} Y_i \text{ and for } i = 1, t.$$

Here

$$Q_g = \left(\frac{1}{t}\sum_{i=1}^{t}\Lambda_i^{-1/2}\right)T \text{ and } Q_{yi} = \Lambda_i^{-1/2}T,$$

where Λ_i is a diagonal matrix of eigenvalues of matrix of correlations between individuals V_i, and T is a matrix of eigenvectors of V_i. The correlation matrix V_i is given as $h_i^2R+(1-h_i^2)I$, where h_i^2 is the heritability of the i-th trait, and R and I are $(n\times n)$ relationship and identity matrices, respectively.

After such transformations, genotypic data do not belong to the categorical type. It means that the logistic regression model used in multiPhen for independent samples cannot be used here.

We compared distributions of genotypic data before and after the transformations, using the GAW19 data [Blangero et al., 2015]. Results showed that the transformed genotypes have the distributions being very close to the normal distribution. Therefore, it is possible to use a classical multiple regression model to describe reverse regression:

$$g^* = e\mu + Y^*\beta + E.$$

Here β is a ($t\times1$) fixed vector of regression coefficients defining reverse effects of the genetic variant on the multivariate traits, e is a ($n\times1$) vector of units, m is intercept, and E is a ($n\times1$) vector of random effects distributed as $N(0, \sigma^2 I)$.

formulation of the imputation model. Future work will include expanding the analyses to other proportions of missingness, wider range of phenotypes and correlations between them as well as to wider set of genetic variants to better mimic the GWAS setting, by using simulated phenotype and genotype data.

P5

Modern Approaches to Address Missing Data in Multi-Phenotype Genome-Wide Association Studies

Mila D. Anasanti[a], *Marika Kaakinen*[a,b], *Marjo-Riitta Jarvelin*[c], *Inga Prokopenko*[a]

[a]Department of Genomics and Common Disease, Imperial College London, London, [b]Department of Medicine, Division of Experimental Medicine and Toxicology, Imperial College London, London, [c]Department of Epidemiology and Biostatistics, Imperial College London, London, UK

Multi-phenotype genome wide association studies (MP-GWAS), as joint analysis of correlated traits, play an important role to increase the power for locus discovery. However, as the number of phenotypes increase, the chance of missing data for each individual is also higher compared to single-phenotype analysis. Most studies conducted on multiple phenotypes simply ignore the data missingness issue and use the complete case (CC) analysis – the default in many statistical approaches. An appropriate handling of missing phenotype data increases power for the analysis and produces less biased parameter estimates. We investigated the properties of the modern approaches, namely multiple imputation (MI) and expectation-maximum (EM) algorithm within the MP-GWAS framework, and compared them with CC analysis using simulation studies. We used the Northern Finland Birth Cohort (NFBC1966) study with complete data on 147 variables and 5063 individuals. We chose three correlated variables: triglycerides (TG), low-density lipoprotein cholesterol (LDL-C) and high-density lipoprotein cholesterol (HDL-C) with the following correlations between them: $r_{TG–HDL\text{-}C} = -0.32$, $r_{TG–LDL\text{-}C} = 0.31$ and $r_{HDL\text{-}C–LDL\text{-}C} = -0.19$. To start with, we chose one variant, the rs174564 from the *FADS1* region, which has repeatedly been associated with the selected three phenotypes. We simulated 5% of missingness under the three mechanisms of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The resulting betas and standard errors from the MP-GWAS after applying the MI and EM were compared to the true values from full data analysis as well as to those from the CC analysis. These preliminary analyses show that MI performes at best out of the all tested approaches, even under the scenario of MNAR, although MI assumes at least MAR. The results from the CC analysis and EM algorithm were biased under MNAR, showing large differences between the effect estimates and standard errors from the full and imputed data analyses (EM: $\Delta\beta_{TG} = 0.02316$, $\Delta\beta_{LDL\text{-}C} = 0.00584$, $\Delta\beta_{HDL\text{-}C} = 0.02679$, $\Delta SE_{TG,LDL\text{-}C,HDL\text{-}C} > 0.001$; CC: $\Delta\beta_{TG} = 0.02408$, $\Delta\beta_{LDL\text{-}C} = 0.0147$, $\Delta\beta_{HDL\text{-}C} = 0.02583$, $\Delta SE_{TG,LDL\text{-}C,HDL\text{-}C} > 0.001$), while the estimates from MI showed only minor deviances from the true values ($\Delta\beta_{TG} = 0.00033$, $\Delta\beta_{LDL\text{-}C} = 0.00014$, $\Delta\beta_{HDL\text{-}C} = 0.00157$, $\Delta SE_{TG,LDL\text{-}C,HDL\text{-}C} > 0.001$). However, the required time for imputation is longer for MI due to computational burden as well as pre-analysis, including a careful

P6

Evaluation of Methods for Genetic Association Test When the Quantitative Trait is Subject to Detection Limit

Siying Huang[a,b], *Vinh Truong*[a], *France Gagnon*[a]

[a]University of Toronto, Dalla Lana School of Public Health, Toronto, Canada; [b]Paris Cardiovascular research Centre (PARCC), Institut national de la santé et de la recherche médicale – U970 Team 3, Paris, France

Background: When the trait distribution is constrained by the detection limit of bioassay, the estimate of genetic effect in association test can be inaccurate. We aim to evaluate the bias introduced by detection limit in the trait and to evaluate performance of several existing methods on detection limit problem in the literature in the context of genetic association studies.

Methods: We conducted a Monte-Carlo simulation study to evaluate the performance of existing methods: deletion, substitution (d, d/2, d/$\sqrt{2}$, 0) and single imputation by robust regression on order statistics method (ROS), multiple imputation using ROS, Kaplan-Meier (K-M), and maximum likelihood estimate (MLE) methods for latent distribution estimate and MLE regression. We simulated 1000 samples of independent observations (n = 300, n = 1000) under additive, dominant and recessive models with minor allele frequency (MAF) at 0.05, 0.2 and 0.5. The detection limit thresholds were set at 10, 15, 25, 50 and 65 percentile of the trait distribution. We evaluated bias, accuracy, coverage, power and type 1 error rate across these methods.

Results: As the proportion of below detection level increased, bias, accuracy, coverage and power tended to decrease in a fashion that depends on MAF and genetic models. Among all methods MLE regression showed to be robust to the detection limit problem disregard percentage except under recessive model and MAF = 0.05. Single imputation and multiple imputation produced less than 5% bias and had 95% coverage when proportion below detection limit was less than 25%. The traditional substitution methods produced most unreliable results. Type 1 error rate was contained within 5% in most of scenarios and exhibited increasing trend in substitution methods and decreasing trend in multiple imputation methods as more samples are below detection limit.

P7

SurvivalGWAS_RV: Software to Test Rare Variant Association with "Time-to-Event" Outcomes

Hamzah Syed, Andrea L. Jorgensen, Andrew P. Morris

Department of Biostatistics, University of Liverpool, Liverpool, UK

Methodology and software for the analysis of common variants within genome-wide association studies (GWAS) have been extensively used and developed for a range of different outcomes, including "time to event" data. These approaches have identified many loci for a variety of complex traits and diseases, but together they account for only a small proportion of the genetic variance. It is considered that rare genetic variants, typically defined as having minor allele frequency <5%, may account for some of the "missing heritability" of human traits. Rare variants are most often analysed within "functional units" using burden or dispersion tests, each with their own benefits and limitations dependent on the underlying genetic architecture of the trait. Software implementing these methods, such as PLINK, EPACTS and GRANVIL, are well developed for binary and quantitative traits, and are capable of handling the scale and complexity of whole genome sequence data, integrating these rare variant methods within a generalised linear regression framework. However, the methodology has not been widely adapted for time to event phenotypes, which have become increasingly important in pharmacogenetic research, where the outcome of interest could be time to death, disease remission or occurrence of an adverse drug reaction. To address this need, we have developed the SurvivalGWAS_RV software implemented using C# and run on Linux, Windows & Mac OS X operating systems. SurvivalGWAS_RV is able to handle large scale genome-wide data, directly assayed or imputed. SurvivalGWAS_RV currently supports analysis using the burden test, Madsen-Browning weighted burden test and sequence kernel association test (SKAT) within a Cox proportional hazards or Weibull regression model. The software also allows for multiple covariates and inclusion of SNP-covariate interaction effects. In conclusion, we introduce a new console application analysis tool for rare genetic variants with time to event outcomes. With its particular relevance to pharmacogenetic studies, SurvivalGWAS_RV will enable the discovery of novel genes associated with patient response to treatment for a range of complex human diseases, ultimately allowing personalisation of therapeutic intervention.

P8

Application of Advanced Mendelian Randomization Methods, to Investigate Whether Low Education Causes Coronary Heart Disease

Taavi Tillmann[a†], *Julien Vaucher*[b†], *Aysu Okbay*[c], *Anne Peasey*[a], *Krista Fischer*[d], *Giovanni Veronesi*[e], *Jack Bowden*[f], *George Davey Smith*[f], *Martin Bobak*[a], *Michael V. Holmes*[g, h]

[a]Department of Epidemiology & Public Health, University College London, London, UK; [b]Department of Internal Medicine, Lausanne University Hospital, Lausanne, Switzerland; [c]Department of Applied Economics, Erasmus University, Erasmus, Netherlands, [d]Estonian Genome Center, University of Tartu, Estonia; [e]Research Center in Epidemiology and Preventive Medicine, University of Insubria, Varese, Italy; [f]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, [g]Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, [h]Medical Research Council Population Health Research Unit at the University of Oxford, Oxford, UK

[†]contributed equally

Trials give superior causal inference than observational studies, due to the random distribution of known and unknown confounders and the removal of reverse causation. Unfortunately, many associations cannot be studied by trials. Observational studies have consistently shown associations between lower education attainment and increased incidence of coronary heart disease (CHD). However, it is not known whether this association is causal, i.e. would increasing general education prevent heart disease?

Mendelian randomization (MR) uses genetic variants associated with a modifiable trait (e.g. education), to infer whether environmental interventions to the same trait would alter the subsequent risk of disease (e.g. CHD). If the genetic variants are randomly allocated before birth, this makes the genetic part of the trait independent from any measured and unmeasured confounders that can otherwise bias analyses. The non-modifiable nature of genes also minimizes the possibility of reverse causation. This allows causal effects to be determined with substantially less bias than in observational epidemiology, with results better approximating those from trials. Furthermore, conducting MR analyses in the opposite direction can directly test hypotheses of reverse causation, and advanced methods allow the interrogation of pleiotropy (which could invalidate the analysis) as well as mediation.

To date, there have been no sufficiently powered trials or MR studies investigating whether socioeconomic risk factors could have a causal role in the development of disease. By using public GWAS meta-analysis data from two large consortia (SSGAC, CARDIoGRAMplusC4D), we used 162 SNPs associated with education, and tested their association with CHD risk. We found that 3.6 years of additional education (i.e. 1-SD increase) would lower the risk of coronary heart disease by a third (odds ratio = 0.67, 95% confidence interval [CI], 0.59 to 0.77, p = 0.01). Sensitivity analyses (using MR-Egger and median-MR) found no support for the hypothesis that these findings were driven by pleiotropy. From the opposite direction, genetic liability for CHD was not associated with adverse educational attainment. Exploratory analyses sug-

gested that the causal association between education and CHD was partly driven by improvements in blood lipid profiles and reductions in smoking and BMI.

Our findings support the hypothesis that increasing time spent in education would substantially reduce the risk of coronary heart disease, and improve population health. These findings offer support for policy interventions that increase education in order to improve population health. We demonstrate the feasibility of applying such Mendelian randomization methods in order to address causal questions about socioeconomic factors.

P9

Introducing Linear Slichter Regression as a Method of Detecting and Correcting for Pleiotropic Bias in Mendelian Randomization Analyses

Wes Spiller, Jack Bowden, George Davey Smith

MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

Mendelian randomization is a popular approach to examining causal relationships in epidemiology, however, it remains controversial as causal effect estimates may exhibit pleiotropic bias. Methods such as MR-Egger regression have proven effective in a two-sample summary MR setting, but require the use of many genetic instruments. Such methods are often inappropriate for individual level data, where it is standard practice to combine variants into allelic scores to overcome weak instrument bias. Recent work (Slichter 2014, Cho et al, 2015) has highlighted the potential use of gene-environment interactions in detecting pleiotropic bias, by identifying target population subgroups where treatment assignment is independent of instrument status. For such groups one would expect the instrument and outcome to be independent, whilst an observed association serves as evidence of pleiotropy. We present linear Slichter regression (LSR) as a formal statistical method to identify and correct for pleiotropic bias using gene-environment interactions. This is achieved within a linear regression framework, regressing the instrument-outcome associations upon the instrument-treatment associations for each level of the interaction covariate. This yields a corrected causal effect estimate, and an estimate of average pleiotropic effect. The technique can be applied with a single instrument. Moreover, the instrument and treatment need not be strictly independent for any covariate group, as long as there is some variation in the strength of dependency between groups. We illustrate the effectiveness of LSR using simulations and data from UK Biobank to assess the role of alcohol consumption upon systolic blood pressure.

P10

A Pre-Existing Isolation by Distance Gradient in West Eurasia May Partly Account for the Observed "Steppe" Component in Europe

Luca Pagani[a], *Lehti Saag*[a], *Anto Aasa*[b], *Flora Jay*[c], *Mait Metspalu*[a]

[a]Estonian Biocentre, Riia 23b, 51010 Tartu, [b]Department of Geography, University of Tartu, 51010 Tartu, Estonia; [c]LRI, Paris-Sud University, CNRS UMR 8623, Orsay, France

It has been proposed that modern European populations can be modelled, by and large, as a three-way mixture of Hunter-Gatherer, Anatolian Neolithic and Steppe components that took place after 6kya (Haak et al. 2015, Allentoft et al. 2015). Particularly the pre-existing Hunter-Gatherer are thought to have admixed with incoming Early Neolithic people from Anatolian and, subsequently, with people carrying a "Steppe" component from the East. These people were likely bearing the so called Yamnaya and/or Corded-Ware cultures, and their initial impact of the European gene pool was estimated to be as high as 75% (Haak et al. 2015).

However ancient DNA samples from East European and Caucasian Hunter-Gatherers as well as from Early Iranian Neolithic, dating from before the Yamnaya expansion, already show signs of this so called "Steppe" component (Lazaridis et al. 2016). Such an observation is compatible with the presence of a pre-existing genetic gradient ranging from Caucasus/Iran all the way to Europe, which likely formed through isolation by distance over thousands of years.

Here we show that such a gradient, defined as decrease of "steppe" component with distance from Iran, can be inferred from ancient samples pre-dating the Yamnaya expansion ($r^2 = 0.93$).

When analysed in the light of this gradient, later ancient and modern samples from Europe still display an excess of Steppe component, however this excess is less pronounced than previously estimated. Additionally we found that, of the analysed samples, modern South Asians show the highest excess of "steppe" component, pointing to the documented, recent links between the Caucasus/Iran populations and the South Asian peninsula.

P11

Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets

Silke Szymczak

Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany

Machine learning methods and in particular random forests are promising approaches for prediction based on high dimensional omics data sets. They provide variable importance measures to rank predictors according to their predictive power. If building a prediction model is the main goal of a study, often a minimal set of variables with good prediction performance is selected. However, if interpretation is more important, approaches that aim to identify all relevant variables should be preferred.

We evaluated several of these variable selection procedures based on simulated data as well as publicly available experimental methylation and gene expression data. Our comparison included the Boruta algorithm, the vita method, recurrent relative variable importance (r2VIM), a permutation approach (PERM) and its parametric variant (Altmann) as well as recursive feature elimination (RFE).

In the simulation studies, Boruta was the most powerful approach, followed closely by the vita method. Both approaches demonstrated similar stability in variable selection, while vita was the most robust approach under a pure null model without any predictor variables related to the outcome. In the analysis of the different experimental data sets vita demonstrated slightly better stability in variable selection and was less computationally intensive than Boruta.

In conclusion, we recommend the vita approach for the analysis of high dimensional data sets. In case of more traditional low dimensional data, vita cannot be applied, but Boruta is a good alternative.

P12

Pathway-Induced Allelic Spectra of Diseases

George Kanoungi, Peter Nürnberg, Michael Nothnagel

Cologne Center for Genomics, University of Cologne, Cologne, Germany

The success of identifying causal genetic variation and understanding its contribution to disease aetiology crucially depends on the genetic architecture of those diseases. Many disease models were proposed to comprehend the impact of the genetic architecture on the epidemiological parameters of diseases. While the additivity of allelic effects is a frequent assumption in modelling disease aetiology, the growing awareness of the important role of gene networks and pathways in complex disease etiology puts this assumption into question. We therefore introduced non-additive disease models based on three different pathway motifs under either a dominant or a recessive genotype model and, using extensive large-scale population forward-in-time simulations, studied their effects on possible spectrum of disease epidemiological parameter values. We first successfully validated our simulations by replicating previously published results. When employing the pathway-based models and comparing them to the corresponding results from an additive model, we found that a serial pathway structure effectively put an upper limit to the sibling recurrence risk one magnitude smaller than those observed under an additive model for diseases with a prevalence below 0.05. On the other hand, pathway motifs that provided some tolerance against deleterious mutations yielded results comparable to the additive model, although odds ratio values were on average higher for some of these pathway models compared to the additive one for diseases with otherwise similar epidemiological parameters. We also applied a generalized pathway model, based on a mixture of three different motifs, to a repeatedly simulated dataset replicating the pathway structure implicated in the MODY disorder. In this application, the obtained allelic spectra of risk variants from the simulations were comparable to those from previously published MODY studies. In the majority of all replications, we could identify the MODY risk gene most likely affected by rare deleterious mutations. Furthermore, we identified potential candidates for disease etiology that may deserve further study in the future. In summary, consideration of pathways in simulation studies of complex diseases can improve their sensitivity towards several kinds of relations that finally rule gene's activities in the cell. Consequently that could contribute to a deeper understanding of the etiology of these diseases.

P13

Comparing Distributions of Polygenic Risk Scores of Type 2 Diabetes and Coronary Heart Diseases within Different Populations

Sulev Reisberg[a–c], *Tatjana Iljašenko*[a], *Kristi Läll*[d,e], *Krista Fischer*[e], *Jaak Vilo*[a–c]

[a]University of Tartu, Institute of Computer Science, Tartu, [b]Software Technology and Applications Competence Centre, Tartu, [c]Quretec Ltd, Tartu, [d]University of Tartu, Institute of Mathematics and Statistics, Tartu, [e]Estonian Genome Centre, University of Tartu, Tartu, Estonia

Polygenic risk scores are gaining more and more attention for estimating genetic risks for liabilities, especially for noncommunicable diseases. They are now calculated using thousands of DNA markers. In this paper, we compare the score distributions of two previously published very large risk score models within different populations. We show that the risk score model, built upon the data of one population, cannot be applied to another population without taking into account the target population's structure. We also show that if an individual is classified to the wrong population, his/her disease risk can be systematically incorrectly estimated.

Second Session

P14

Implication of Multivariate Analysis to Correlate Variables in the Selection of Specific Characters of Interest

D. Almorza[a], *M.V. Kandus*[b], *J.C. Salerno*[b]

[a]University of Cádiz, Spain; [b]INTA, Genetic Institute (IGEAF) Castelar, Argentina

Principal components analysis increases the process of selection of important characters, according to the objective of the selection that is sought to improve. In this way, the behavior of genotypes can be significantly differentiated by measuring the available variables of the phenotypes in the study material considered (weigh of ears, pin diameter, ear length and number of grain rows).

In this work we used contrasting stable materials to be able to differentiate them in the principal components analysis. Infostat was the statistic software used in this work. Tests were performed to evaluate inbred maize lines, using a completely randomized design with three replicates, considering the variables indicated as responsible for performance.

The results showed significant correlations among the variables and also allowed to separate the different groups of lines with high and low values for the characters studied, facilitating the selection process and shortening the time of obtaining elites to improve the final product.

To summarize the relation among the variables and the behavior of the maize lines evaluated we used biplot graphic representation. Maize lines evaluated included twelve flint lines (9, 61, 70, 73, 76, 91, 96, 101, 104, 123, LP521 and LP109) and three dent lines (B73, B14 and MO17).

P15

Focused Multidimensional Scaling: An Interactive, User-Friendly Tool for Visualization of Distance Matrices

Lea Urpa, Simon Anders

Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Data visualization, particularly in exploratory data analysis, is key to discovering new relationships within data and generating novel hypothesis from data. Current static tools for visualization of data from distances matrices via multidimensional scaling (MDS) are widely used in genetics research but necessarily lose many important details, as it is impossible to faithfully plot a matrix of pairwise dissimilarity measures on a 2D plane. We introduce a new approach named focused multidimensional scaling (focused MDS), where the user chooses a particular data point (the focus) around which all other points are iteratively plotted in their exact distance to the focus. Non-focus points are plotted in relation to one another as exactly as possible, accomplished by iteratively plotting points in polar coordinates with fixed radius and optimized angle in relation to all previously plotted points. The user may interactively choose another focus point, and the graph will rearrange to reflect the distances to the new focus. Focused MDS can quickly and easily visualize data from gene expression correlations, drug resistance and sensitivity testing correlations, or any other type of data that can be summarized by a distance matrix or be displayed in a heat map. Our interactive tool for focused MDS illuminates unintuitive relationships in datasets and assists in generating novel hypotheses from omics data.

P16

The Shortest Path to Significance

Julian Hecker[a], *Dmitry Prokopenko*[b], *Heide Löhlein Fier*[a,b], *Christoph Lange*[a,c]

[a]Harvard T.H. Chan School of Public Health, Boston, MA, USA; [b]Institute of Genomic Mathematics, Bonn, Germany; [c]Brigham and Women's Hospital, Boston, MA, USA

During the last years, with the development of sequencing technologies dense high-throughput whole-genome sequencing data has become increasingly available. In addition, sophisticated phasing algorithms as well as efficient sequencing techniques will provide the possibility to obtain the phase information for studies with large sample sizes in the future. As discovered already during the last decade, the phased haplotype information can give more detailed insights into the architecture of diseases and can be used to construct more powerful association tests. Motivated by the haplotype-block structure of the genome, the aim was to capture and incorporate the correlation of SNPs in regions with small recombination rates. The aforementioned approach can lead to scenarios where the degrees of freedom of the analysis are reduced. However, the application of haplotype-based association tests was complicated by the phase uncertainty. Now, we focus again on the methodology of haplotype-based tests, since exact phased haplotype data becomes more available. The standard approach for a whole-genome sequencing association study is to perform single-variant tests for all common genetic variants. While ignoring the additional information provided by the phase and the correlation structure, this approach requires the correction for a large number of statistical tests. Here, we describe an efficient algorithm for phased haplotype data to construct a testing strategy which minimizes the number of performed statistical tests, including all common genetic variants and using only stable haplotype testing configurations.

P17

Not all K-Mers are Equal – Some are Interesting, Some are Boring

Lauris Kaplinski, Maido Remm

Department of Bioinformatics, Institute of Molecular and Cell Biology, Tartu University, Tartu, Estonia

Alignment-free k-mer based methods have proven themselves to be useful tools for fast and accurate analysis of sequencing data. In current applications tag k-mers are used to find already known biological features. We present a novel methodology that allows one to scan population sequencing reads for "interesting" k-mers without any previous information about specific sequences of interest.

This is achieved by calculating the distribution of each k-mer in population sequencing samples and finding the frequencies of individuals with specific number of copies of given k-mer. K-mers that have identical counts in all individuals, or whose distribution can be explained by simple Mendelian inheritance of few alleles,

are probably not associated with any interesting sequence features. On the other hand those, whose distribution cannot be explained with simple inheritance models, can be studied further.

Our approach allows one to quickly discard most of the sequencing reads that do not contain any new information and study in depth only those that are "abnormal". This has a potential to both reduce the computation time in different analyses and enhance signal-noise ratio of acquired information.

P18

Bayesian Classification of Vaccine-Specific B-Cells from Repertoire Sequencing Data

Anna Fowler[a], *Gerton Lunter*[b]

[a]Department of Biostatistics, University of Liverpool, Liverpool, [b]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

B cell receptors (BCRs) are a component of the adaptive immune system that recognise and bind antigens. In order to have a healthy immune system, a diverse set of BCRs, capable of recognising many different antigens, is required. This BCR diversity is generated through a complex process of somatic recombination and hyper-mutation, thought to be capable of generating over 10^{13} unique BCRs. This is far greater than the 10^6 unique B cells estimated to be present in a single individual, resulting in very little overlap between any two BCR repertoires. NGS has allowed us to capture these somatic differences at the resolution of individual B cells, through targeted mRNA sequencing of the variable region of the BCRs.

In response to vaccination, B cells that bind the vaccine antigen will proliferate and hyper-mutate. Identifying sequences from vaccine-specific B cells has the potential to provide a cheap and accurate correlate of immune protection and add to our understanding of immunology.

We develop a Bayesian classification approach to infer those sequences that are vaccine-specific in BCR repertoires collected pre- and post-vaccination. The model is built up hierarchically. The presence of a sequence in an individual is dependent on the sequence classification, and then the abundance of a sequence in any single sample is dependent on both presence in the individual and sequence classification. This allows us to identify sequences from vaccine-specific B-cells as those that are likely to be seen in multiple individuals and/or in high abundance post vaccination.

We apply this algorithm to data from 5 subjects vaccinated against Hepatitis B and sampled at days 0, 7, 14, 21 and 28, relative to vaccination, using samples from total B cells. B cells from these samples were also sorted experimentally for Hepatitis B surface antigen vaccine-specificity. Using the sequences from cells enriched for vaccine specificity, we estimate that our algorithm has 70% sensitivity for detecting antigen-specific sequences. Additionally, we find evidence for convergent evolution in the CDR3, with sequences identified as vaccine specific present in greater numbers of samples at later time points.

P19

Two-Stage Variant Calling Algorithm for Next-Generation Sequencing Experiments

Sarunas Germanas

Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania

Next-generation sequencing (NGS) is often used to identify genetic variants. This sequencing technique suffers from large sequencing errors and demands sophisticated mathematical methods to control this problem. Therefore probability of variant detection depends on the variant caller. There are several variant calling procedures which assume constant variant calling threshold across genetic positions of target region. These assumptions may lead to smaller sensitivity and specificity of a variant caller. Here we propose a novel variant calling approach. It consists of two stages – clusterization of target region regarding the estimated genotype probabilities computed using another variant calling method (for example method used by MAQ package) and adaptation of variant calling threshold for every subregion. For clusterization of the target region we use change point detection methods (for example, non-parametric likelihood test). We apply offered and known methods to 1000 genomes data with known mutated and not mutated positions.

P20

An Evaluation of De Novo Mutation Rate in Lithuanian Exome

L. Pranckėnienė, A. Jakaitienė, V. Kučinskas

Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

De novo mutations are the ultimate source of genetic variation and one of the driving forces of evolution. Early studies of mutation rates in humans focused on relatively small and specific loci or on the *de novo* incidence of disease. Family-based whole-genome sequencing has begun to identify *de novo* mutations that provide more direct measures of mutation rates. The objective of our research is an evaluation of *de novo* mutation rate in Lithuanian whole exome's data.

Sequencing of Lithuanian population 48 parent-offspring trios exomes was performed using SOLiD 5500 sequencing system. Lifescope was used to retrieve sequencing data. *De novo* variants (DNV) called by two alternatives programs: VarScan and VarSeq. Called *de novo* variants were filtered by applying the following filters by SnpSift software: 1) genotype quality of the individual (≥50); 2) number of reads at each site (>20). In addition, all called and filtered possible *de novo* single nucleotide variants were manually reviewed by the Integrative Genomics Viewer. We used R package to estimate the rate of *de novo* mutations.

We estimated 189 (VarScan) and 121 (VarSeq) *de novo* single nucleotide mutations respectively. The probability of calling position as a *de novo* mutation in a trio we calculate independently for each family. Assuming independence, we calculate probability of

de novo mutation per family for each called position as the product of the probability of the child being heterozygous and probabilities of the parents being homozygous reference. We calculate the latter with respect to sequencing depth as well. We obtained that mutation rate is somewhat higher for Lithuanian exome data as compared to 1.5×10^{-8} estimate published in other research. This is the first attempt to estimate *de novo* mutation rate for family trios from Lithuanian population. The verification of the estimate using context-specific mutation data is foreseen.

This work supported by the LITGEN project (VP1-3.1-ŠMM-07-K-01-013), funded by the European Social Fund under the Global Grant Measure.

P21

How to Phase Genomic Data with Indels? A Comparative Study

Shabbeer Hassan, Javier Nunez-Fontarnau, Himanshu Chheda, Pyry Helkkula, Paavo Häppölä, Ida Surakka, Priit Palta, Aarno Palotie, Samuli Ripatti, for the SISu project group

Institute for Molecular Medicine (FIMM), Finland

The estimation of haplotypes (group of genes that are inherited together from a single parent) from SNP genotypes, is commonly referred to as 'phasing'. Determination of haplotype phase is an important methodological issue as we are in the era of large-scale sequencing. Many of its applications, such as imputing low- frequency variants and characterizing the relationship between genetic variation and disease susceptibility, are highly relevant to sequenced data. Haplotype phase can be generated through laboratory-based experimental methods, or it can be estimated using computational approaches. As experimental determination, could be very expensive, computational phasing along with imputation for genotyped chip data is generally the preferred solution. However, a major shortcoming of currently available phasing software are their inability to handle short insertions/deletions (indels) and longer copy number variants (CNVs). As indels and CNVs have been found to play an important role in both Mendelian and complex diseases, phasing of such indels is an important feature.

The aim of this work is to use and compare currently available phasing software (SHAPEIT2 and EAGLE2) by calculating various statistical metrics (switch error rates & imputation concordance) to evaluate solutions (indel omission, indel trimming, native EAGLE2 indel handling) for this burgeoning computational problem in genomics. Results indicate that phasing with trimmed indels yields the lowest switch error rates (0.6–0.9%) among all the phasing solutions. Native EAGLE2 indel handling produces higher switch error rates (1–1.5%) when compared to indel trimming. Imputation concordances were comparable across all phasing solutions (96–98%) and indel trimming does not induce any imputation errors when compared to other phasing solutions. Indel trimming is the first viable computational method which can allow for phasing of indels and other structural variants in any genomic data.

P22

Improved Imputation Accuracy of Rare and Low-Frequency Genetic Variants Using Population-Specific High-Coverage Whole-Genome Sequencing Data Based Imputation Reference Panel

Mario Mitt[a,b], *Mart Kals*[a,c], *Kalle Pärn*[a,d], *Stacey B. Gabriel*[e], *Eric S. Lander*[e], *Aarno Palotie*[d,e], *Samuli Ripatti*[d], *Andrew P. Morris*[a,f], *Andres Metspalu*[a,b], *Tõnu Esko*[a,e], *Reedik Mägi*[a], *Priit Palta*[a,d]

[a]Estonian Genome Center, University of Tartu, Tartu, [b]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, [c]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [d]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; [e]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; [f]Department of Biostatistics, University of Liverpool, Liverpool, UK

Genetic imputation has shown to be a cost-efficient way to improve the power and resolution of genome-wide association studies. Current publicly accessible imputation reference panels accurately estimate genotypes for common variants with minor allele frequency (MAF) ≥5% and low-frequency variants (0.5≤ MAF <5%) across diverse populations, but the imputation of rare variation (MAF <0.5%) is still rather limited.

We used high-coverage (30×) whole-genome sequencing data of 2,244 Estonian individuals from the Estonian Biobank to create a population-specific imputation reference panel. The resulting EGCUT panel together with two publicly accessible imputation reference panels (1000G and HRC) were utilised to impute 6,394 microarray-genotyped Estonians using IMPUTE2 software. We compared the results while dividing the dataset into different categories based on imputed genotype quality estimate and MAF.

The evaluation of overall imputation quality for each panel based on IMPUTE2 *INFO* metric showed that although EGCUT consists of substantially fewer haplotypes compared to HRC or considerably fewer SNVs compared to 1000G panel, it results in notably more low-frequency and rare SNVs imputed with high quality. To assess the accuracy of imputed genotypes directly, we utilised GATKs *GenotypeConcordance* module. While using 505 whole-exome sequences available for a subset of imputed EGCUT individuals as the truth set, both *sensitivity* and *discordancy rate* metrics showed that the population-specific panel yielded better accuracy throughout allelic spectrum.

We observe that, although publicly accessible imputation reference panels do cover a proportion of low-frequency and rare variation in different populations, the vast majority of rare variants present in these reference datasets are imputed with relatively low confidence and therefore cannot be used in downstream analyses. Furthermore, as there is a considerable proportion of population-specific variation that cannot be imputed with currently publicly available imputation reference panels, imputation of low-frequency and rare variants is considerably more accurate with a population-specific reference panel or if one is used in combination with a publicly available reference such as the 1000G panel.

P23

Personalised Bayesian Dose Adjustment of Vancomycin in Neonates

Tõnis Tasa[a–c]*, *Tuuli Metsvaht*[d], *Riste Kalamees*[c], *Lili Milani*[b], *Jaak Vilo*[a], *Irja Lutsar*[c]

[a]Institute of Computer Science, University of Tartu, J. Liivi 2, 50409, Tartu, [b]Estonian Genome Center, University of Tartu, Riia 23B, 51010 Tartu, [c]Department of Microbiology, University of Tartu, Ravila 19, 50411, Tartu, [d]Clinic of Anaesthesiology and Intensive Care, Tartu University Hospital, L. Puusepa 8, 51014, Tartu, Estonia

Our main aim was to study the precision of individualised dosing and to benchmark it against retrospectively observed target attainment levels. This required the development and validation of a computerised dose adjustment application, DosOpt, to guide the dose selection.

Model fitting in DosOpt uses Bayesian methods for deriving individual pharmacokinetic estimates from population priors and patient therapeutic drug monitoring measurements. These are used in simulating time-concentration curves for target constrained dose optimisation. We validated DosOpt using retrospective clinical data. Prediction accuracy and optimised time-concentration profiles with adjusted doses were benchmarked against attainment of target concentrations observed in clinical practice.

DosOpt is freely available at www.biit.cs.ut.ee/dosopt. We collected data for validation from 121 patients with at least one vancomycin concentration measurement. Three individual concentrations instead of population model estimates decreased the mean absolute percentage error from 61.2% to 22.8% and increased the probability of target attainment for trough concentrations within targeted range of 10–15 mg/L (median, 95% CI) from 16% (11%–24%) to 43% (21%–47%).

DosOpt only requires a small number of patient concentration measurements to significantly improve target attainment in recommended concentration ranges above retrospectively observed levels.

P24

A Comparison of 5 Software Implementations Conducting Mediation Analysis

Liis Starkopf[a], *Thomas Alexander Gerds*[a], *Theis Lange*[a,b]

[a]Section of Biostatistics, Department of Public Health, University of Copenhagen, [b]Center for Statistical Science, Peking University, Peking, P.R. China

Despite the existing well-developed statistical methods for mediation analysis, applied researchers are still faced with challenges when implementing mediation analysis. This talk provides practical advice on conducting counterfactual based mediation analysis in major statistical software packages. In epidemiology and many other scientific disciplines, mediation analysis is an important tool for understanding the causal mechanisms. Specifically, mediation analysis allows to disentangle the indirect effect of an exposure on an outcome through a given intermediate variable, the mediator, from the remaining direct effect through other non-specified mediators. Developments in causal inference have greatly extended the theoretical framework and have led to a number of a distinct estimation strategies of direct and indirect effects. Software implementations of mediation analysis are now available for standard statistical software packages allowing the researcher to perform mediation analysis with most data types. This talk introduces, compares and contrasts five estimation approaches for mediation analysis including software implementations in R, SAS, SPSS and STATA. The main contribution of this comparison is to give practical advice for applied researchers who wish to conduct counterfactual based mediation analysis. We provide mathematical details and a discussion of the scope and limitations of each estimation method. Finally, we illustrate the approaches with an empirical example and assess their performance in small samples through a simulation study.

P25

Estimation of SNP Heritabilities Using 30,000 Finns with 45+ Years of Health Registry Data

Sanni Ruotsalainen[a], *Heidi Hautakangas*[a], *Aki Havulinna*[a,b], *Tuomo Kiiskinen*[a], *Ida Surakka*[a,b], *Matti Pirinen*[a,c,e], *Hannele Laivuori*[a], *Elisabeth Widén*[a], *Samuli Ripatti*[a,c,d]

[a]FIMM – Institute for Molecular Medicine Finland, Helsinki, [b]THL – National institute for Health and Welfare, Helsinki, [c]Department of Public Health, University of Helsinki, Helsinki, Finland; [d]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; [e]Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Heritability analyses have yielded important insights into complex disease architecture, and electronic health registry (EHR) data represent a novel resource for studying heritability of many traits and diseases that are not typically accessible. Genome-wide profiles together with disease event histories created from routinely collected healthcare data has the potential for studies of genetic effects along the life histories including sequences of diseases and comorbidities. Here we report liability-scale SNP heritability estimates (h^2_g) of various International Disease Codes (ICD) using 30,000 Finns collected in population-based health surveys with genome-wide variation data linked to a nationwide hospital discharge registry and death registry data. We estimated h^2_g by first applying BOLT-REML directly to observed case-control status obtaining raw observed-scale h^2_{g-cc} estimates and then converting h^2_{g-cc} to liability-scale h^2_g using linear transformation and supporting multi-component modeling to partition SNP-heritability. We illustrate the potential of these data by presenting heritability estimates for the main ICD10 classes and a selection of more detailed events using data harmonized over the ICD8, 9 and 10 classifications. The main class heritability point estimates varied between 8% (CI: 3–13; certain infectious and parasitic diseases) and 29% (CI: 22–36; endocrine, nutritional and metabolic diseases). While illustrating the potential of the approach, we also discuss some of the limitations of the approach.

Author Index

Human Heredity

Numbers refer to page numbers

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

Hum Hered 2016;81:211–238
DOI: 10.1159/000469638