Quantify Genomic Heritability Through a Prediction Measure

Arthur Frouin, Edith Le Flocch, Christophe Ambroise

Human Heredity

Hum Hered 2017–18;83:1–29
DOI: 10.1159/000488519

Published online: April 18, 2018

46th European Mathematical Genetics Meeting (EMGM) 2018

Cagliari, Italy, April 18–20, 2018

Abstracts

Guest Editors
Justo Lorenzo Bermejo, Heidelberg
Marcella Devoto, Philadelphia, PA/Rome
Christine Fischer, Heidelberg

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

Human Heredity

Principal Components Analysis in the Phenotypic Selection of Inbred Lines

David Almorza[1], *María Victoria Kandús*[2], *Juan Carlos Salerno*[2–3]

[1]University of Cádiz, Cádiz, Spain; [2]INTA – IGEAF – CICVYA, Argentina; [3]USAL, Argentina
Correspondence to: david.almorza@uca.es

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, and represents a continuation of the one published in Almorza, Kandus and Salerno (2016).

An PCA (principal component analysis) was carried out to evaluate the performance components of different inbred maize lines, determining the following variables: Ear Length (LE); Number of rows of grains per spike (NH); Grain depth (TG); Percent of cob (% Marlo); Weight of 100 seeds (P100K) and Grain yield (KGHA).

The Infostat software was used to carried out the PCA and the biplot graphics, determining the correlation among the variables studied (coefficients and probability); selfvalues and selfvectors and the correlation between the PCs and the original variables.

In turn, the cofenetics correlation was calculated, which is a measure of the quality of the reduction achieved with the data. PC1 and PC2 explained approximately 80% of the variability of the data. A positive and significant correlation was found between grain yield (KGHA) and the following variables: Grain depth (TG), Weight of 100 grains (P100K) and Ear length (LE). The variables: grain yield (KGHA), weight of 100 seeds (P100K), grain depth (TG) and ear length (LE) were associated with PC1 positively, while NH and MARLO% were positively associated with PC2.

Reference

Almorza D, Kandus MV, Salerno JC: Implication of Multivariate Analysis to Correlate Variables in the Selection of Specific Characters of Interest. Human Heredity 2016;81(4)232–233.

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

Mila D. Anasanti, *Marika Kaakinen, Inga Prokopenko*

Section of Genomics and Common Disease, Department of Medicine, Imperial College London, London, UK
Correspondence to: m.anasanti15@imperial.ac.uk

Multi-phenotype genome-wide association studies (MP-GWAS) play an important role in improving the power for locus discovery. However, joint analysis of multiple phenotypes increases the proportion of missingness, leading the standardly applied complete case (CC) analysis to become inefficient. We investigated the properties of single imputation (SI), multiple imputation (MI), left-censored method (QRILC), k-Nearest Neighbour (k-NN), and Random Forest (RF) within the MP-GWAS framework, and compared them with the CC analysis using simulation studies.

We simulated genetic data for 5,000/50,000/500,000 individuals using Hapgen2, and highly (r = 0.64) and moderately correlated (r = 0.33) phenotypes (3/30/120) for these individuals in the statistical package R. We randomly chose common (minor allele frequency (MAF) >5%), low-frequency (1%< MAF <5%) and rare (MAF <1%) variants to be associated with the simulated phenotypes. We considered 1/10/20/50% missingness under the three mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

The resulting betas and standard errors (SEs) from the MP-GWAS after applying the selected methods were compared to the true values from full data analysis as well as to those from the CC analysis. These analyses showed that MI/QRILC/KNN/RF perform the best, followed by SI, even under the scenario of MNAR, although SI/MI assume at least MAR. For the CC analysis, the performance worsened when the number of phenotypes was increased, while the other approaches were not influenced by the number of phenotypes nor differences in phenotype correlations or MAF.

In conclusion, we recommend MI/QRILC/KNN/RF imputation approaches over the commonly applied CC analysis, especially QRILC/RF under MNAR.

KARGER

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

A Bayesian Joint Fine Mapping Approach That Shares Information Between Related Autoimmune Diseases Increases Accuracy and Identifies Novel Associations

Jennifer L. Asimit[1], *Mary D. Fortune*[1,2], *Daniel B. Rainbow*[3], *Linda S. Wicker*[3], *Chris Wallace*[1,2]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK; [2]Department of Medicine, University of Cambridge, Cambridge, UK; [3]JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, University of Oxford, Oxford, UK
Correspondence to: jennifer.asimit@mrc-bsu.cam.ac.uk

Hundreds of genetic variants have been identified as associated with a spectrum of diseases, but the fine-mapping of causal variants has been complicated by extended linkage disequilibrium (LD) and finite sample sizes. We propose to leverage information between diseases through joint analysis of data from related diseases in a novel Bayesian multinomial stochastic search framework, where prior model probabilities are formulated to favour combinations of models with a degree of sharing of causal variants between diseases. We use simulations and real data examples to illustrate the improved accuracy in comparison to a marginal analysis of each disease. That is, in simulations of two diseases that each have two causal variants, of which one is shared, we find that marginal disease analyses may fail to identify both causal variants for each disease. However, our multinomial framework tends to detect shared variants that are missed by marginal analyses. We jointly fine-map association signals for six diseases and of particular interest is *IL2RA*, which is known to be associated with several autoimmune diseases, including multiple sclerosis (MS), type 1 diabetes (T1D), autoimmune thyroid disease (AITD) and coeliac disease. Our proposed approach is computationally efficient and adds only five minutes overhead to the fine mapping of individual diseases.

Exact Model Comparisons When Determining Parent-of-Origin Effects

Stefan Böhringer[1], *Dietmar Lohmann*[2]

[1]Department of Biomedical Data Sciences, Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands; [2]Institut für Humangenetik, Universitätsklinikum Essen, Essen, Germany
Correspondence to: s.boehringer@lumc.nl

The plausibility framework (Martin 2016, JASA) is a generalization of Fisher's exact test to a wide class of parametric models and allows for goodness-of-fit testing for which test sizes are guaranteed for finite sample size. It is more difficult in this framework to compare different models for better fit. We propose to extend the plausibility framework by re-weighing the probability mass of observations. Weights are determined by a test statistic such as a likelihood-ratio to compare two models when plausibility is evaluated under the null-model. We show that the exact guarantees of the plausibility framework can be maintained under re-weighting. We also evaluate a parametric bootstrap scheme under which models can be compared.

We illustrate our methods with a Retinoblastoma (RB) data set. In RB, different mutations in the RB1 gene lead to different severeness of RB which is quantified by the number of affected eyes in each individual. Additionally, parent-of-origin effects play a role. The data set is modeled using a binomial, ascertained likelihood and is analyzed in the plausibility framework. We show, that mutation type can explain some part of the variability of RB. We conduct simulations to verify the exactness of the procedure and to evaluate power under several scenarios.

We conclude, that the plausibility framework can be used for goodness-of-fit testing and model comparisons as an alternative to bootstrapping or permutations when small sample sizes are involved.

Reference

Martin R: Plausibility Functions and Exact Frequentist Inference. J Am Stat Assoc 2015;110:512, 1552–1561.

Pleiotropic Effects of Genetic Variants on Gallstone Disease and Gallbladder Cancer in Europeans and Chileans

Carol Barahona Ponce, *Felix Boekstegers*, *Justo Lorenzo Bermejo*

Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany
Correspondence to: barahona@imbi.uni-heidelberg.de

Gallstones are considered the main risk factor for gallbladder cancer (GBC) development. The prevalence of gallstones has a strong ethnic component – more than 40% of Chilean Native Americans are affected by gallstones, compared to 15–20% of Europeans. In a meta-analysis of genome-wide association studies, Joshi et al. identified six genetic variants associated to gallstone disease.

In order to examine the possible pleiotropic, combined effect of identified variants in both gallstone disease and GBC risk, we examined genotype data from Chileans (295 GBC cases and 2267 controls) and Europeans (100 cases and 134 controls). We constructed weighted genetic risk scores for gallstone disease based on the reported log odds ratios for the six variants and individual genotypes. The association between the genetic risk of gallstone disease and GBC risk was assessed by logistic regression with the individual risk score as independent variable adjusting for potential confounders (e.g. age and gender).

We found an association between the genetic risk of gallstone disease and GBC risk in both Chileans (1.20 per standard deviation in the genetic risk score, 95% CI 1.05–1.36) and Europeans (OR = 1.82, 95% CI 1.29–2.55). Two variants in the *ABCG8* gene showed a strong GBC association in Europeans: rs11887534 (OR = 2.52, 95% CI 1.17–5.43) and rs4245791 (OR = 1.81, 95% CI 1.12–2.93). Among the six variants associated with gallstone disease, only rs11887534 was associated with GBC risk in Chileans (OR = 1.38, 95% 1.04–1.83). We are currently working on the validation of GBC findings from an Indian association study in Chileans and

Europeans, and on Mendelian randomization analyses to evaluate the causal relationship between gallstones and GBC risk. Updated results will be presented at the meeting.

Optimal Selection of Genetic Variants for Adjustment of Population Stratification in European Association Studies

Regina Brinster, Dominique Scherer, Justo Lorenzo Bermejo

Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany
Correspondence to: brinster@imbi.uni-heidelberg.de

Population stratification due to differential environmental exposures combined to differential genetic composition of the study subpopulations may lead to spurious findings in genetic association studies. A well-established method to correct association results for stratification relies on principal component analysis (PCA) of genome-wide genotype data. The estimated principal genetic components are subsequently included as covariates in regression models that investigate the genotype-phenotype relationship. Alternative correction approaches rely on the use of genetic variants which show large differences in allele frequency among subpopulations, so-called ancestry informative markers (AIMs), as covariates. In contrast to the commonly applied PCA correction, usually a small number of AIMs is needed for stratification adjustment.

We compared several approaches for AIM identification and subsequent stratification adjustment including the method proposed by Bauchet et al. [1]. We simulated case-control studies using the European subset of the Population Reference Sample (POPRES) [2] as reference, and evaluated the correction performance of the different approaches by means of the type I error rate and the genomic inflation factor. Investigated parameters included the case-control distribution in each subpopulation, the number of study populations and their genetic heterogeneity.

Preliminary results show that AIMs identified based on PCA have the potential to correct for population stratification. Details on the applied methodology and results for the most promising examined approaches will be presented during the meeting.

References

1 Bauchet M, et al: Measuring European Population Stratification with Microarray Genotype Data. Am J Hum Genet 2008;80(5):948–956.

2 Nelson MR, et al: The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. Am J Hum Genet 2008;83(3):347–358.

Ancient DNA Study Reveals HLA Susceptibility Locus for Leprosy in Medieval Europeas

Ben Krause-Kyora[1,2], Amke Caliebe[3], Andre Franke[1], Jesper L. Boldsen[4], Tobias L. Lenz[5], Michael Nothnagel[6], Almut Nebel[1]

[1]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany; [2]Max Planck Institute for the Science of Human History, Jena, Germany; [3]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [4]Unit of Anthropology (ADBOU), University of Southern Denmark, Odense, Denmark; [5]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Jena, Germany; [6]Cologne Center for Genomics (CCG), University of Cologne, Cologne, Germany
Correspondence to: b.krause-kyora@ikmb.uni-kiel.de

Leprosy, a chronic infectious disease caused by *Mycobacterium leprae* (*M. leprae*), was very common in Europe till the 16th century. Here, we perform an ancient DNA study on medieval skeletons from Denmark that show lesions specific for lepromatous leprosy (LL). First, we test the remains for *M. leprae* DNA to confirm the infection status of the individuals and to assess the bacterial diversity. Second, we evaluate whether the HLA allele DRB1*15:01, a strong LL susceptibility factor in modern populations, also predisposed medieval Europeans to the disease. The comparison of genotype data from 69 *M. leprae* DNA-positive LL cases with those from contemporary and medieval controls reveals a statistically significant association in both instances. In addition, we observe that DRB1*15:01 always occurs with DQB1*06:02 on a haplotype that is a strong risk factor for inflammatory diseases today.

Cell-Specific Somatic Mutation Detection from Single-Cell RNA-Sequencing

Nghia Vu[1], Stefano Calza[1,2], Yudi Pawitan[1]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden; [2]Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
Correspondence to: stefano.calza@unibs.it

Motivation: The single-cell RNA- and DNA-sequencing are generally different techniques, but both have been applied widely. To study genomic mutations of a cell, single-cell DNA-sequencing is the method of choice. However, the task is still challenging due to the limited input materials, amplification biases, technical noises. Generally, single-cell RNA-sequencing with a higher input amount has better data quality. There exists various methods for detecting mutation from bulk-cell sequencing. However these methods are usually not directly applied for single-cell RNA-sequencing due to the data noises and artefacts.

Results: We develop SCmut, a novel and robust statistical method to detect cell-specific mutations from single-cell RNA-sequencing. In particular, SCmut extracts single nucleotide variants of cells from the RNA-sequencing data and statistically detects mutations of cells based on the variant alternative fraction and the coverage using the two-dimensional local false discovery rate method. We apply SCmut to single-cell RNA-sequencing datasets of primary tumors and lymph nodes of two breast cancer patients and two data-sets (in different conditions) from the breast cancer cell line MDA-MB-231. The results show that traditional mutation detection methods for bulk-cell sequencing data do not work well for the single-cell data, where they introduce a lot of false positives. For breast cancer datasets which have a high level of cell-to-cell heterogeneity, SCmut identifies significant cell-specific mutations mostly from tumor cells which are well separated from non-tumor cells. For breast cancer cell line datasets which have highly homogeneous cell populations, the detected cell-specific mutations are highly consistent in the two independent data-sets.

Prediction of Treatment Response from Genome-Wide SNP Data in Rheumatoid Arthritis Patients

Svetlana Cherlin[1], *Heather J. Cordell*[1], *MATURA Consortium*[2]

[1]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK; [2]Maximising Therapeutic Utility in Rheumatoid Arthritis
Correspondence to: heather.cordell@newcastle.ac.uk

Although a number of treatments are available for Rheumatoid Arthritis (RA), each of them shows a significant non-response rate in patients. Therefore, predicting a priori the likelihood of treatment response would be of great benefit. Here we conducted a comparison of a variety of statistical methods for predicting the change in C-reactive protein score (CRP), in 28 swollen joint count score (SJC28) and in erythrocyte sedimentation rate (ESR) between baseline and 3 or 6 months using genome-wide SNP data from RA patients available from the MAximising Therapeutic Utility in Rheumatoid Arthritis (MATURA) consortium. Two different treatments (tumor necrosis factor α inhibitors and methotrexate) and nine different methods (lasso, ridge, elastic net, random forests, support vector regression, sparse partial least squares, genome-wide complex trait analysis, Bayesian sparse linear mixed model, and neural network) were evaluated. We used 10-fold cross validation to assess predictive performance, with nested 10-fold cross validation used to tune the model parameters when required. Overall, we found that SNPs add very little prediction information to that obtained using clinical information only, such as baseline trait value. This observation can be explained by the lack of strong genetic effects and the relatively small sample sizes available. However, methods that assume a complex underlying genetic architecture of the trait were able to extract some information about prediction, in comparison to methods that assume a simplified genetic architecture. Additionally, methods that are consistent with the genetic architecture of the trait were able to achieve better prediction than methods that are not.

An Improved Bioinformatics Tool for Rare Disease Variant Prioritization: The Exomiser 9.0.1 in Clinical Practice

Valentina Cipriani[1–4], *Nikolas Pontikos*[2,3], *Gavin Arno*[2,3], *Andrew R. Webster*[2,3], *Anthony T. Moore*[2,3,5], *Keren J. Carss*[6], *Lucy F. Raymond*[6], *Daniel Danis*[7], *Peter N. Robinson*[7], *Julius O.B. Jacobsen*[1], *Damian Smedley*[1]

[1]William Harvey Research Institute, Queen Mary University of London, UK; [2]UCL Institute of Ophthalmology, University College London, UK; [3]Moorfields Eye Hospital NHS Foundation Trust, UK; [4]UCL Genetics Institute, University College London, UK; [5]Ophthalmology Department, UCSF School of Medicine, CA, USA; [6]NIHR Bioresource – Rare Disease study, UK; [7]Jackson Laboratory for Genomic Medicine, CT, USA
Correspondence to: v.cipriani@qmul.ac.uk

The Exomiser was launched by Robinson et al. in 2014 [1] as a freely available Java program to annotate, filter and prioritise variants from next-generation sequencing (NGS) projects for disease gene discovery or differential diagnostics of Mendelian disease. The software requires i) a variant call format (VCF) file with the called variants of a rare disease patient and ii) a set of patient's phenotypes encoded using the Human Phenotype Ontology (HPO) [2]. A range of user-defined variant filtering criteria can be applied based on JANNOVAR functional annotation [3], frequency and expected inheritance pattern. The filtered variants are then prioritised according to their rarity and algorithm-predicted pathogenicity, combined with the phenotypic similarity between the patient's HPO terms and those used to annotate genes in known human diseases, mouse/zebrafish model organisms and/or relevant protein-protein interactions. The Exomiser was able to prioritise causative variants as top candidates in 97% of simulated whole-exomes [4].

Here, we present the latest Exomiser version 9.0.1 with a number of new features, including the use of additional publicly available datasets (e.g., gnomAD), the choice of Ensembl, UCSC or RefSeq gene transcript definitions, and the analysis of non-coding [5] and mitochondrial variants. We assessed the software performance using a set of 132 whole-exomes from patients with a range of different rare retinal diseases and a confirmed molecular diagnosis. The causative variant(s) were ranked first/up to 5th/up to 10th in 74%/87%/89% of the patients's exomes. The Exomiser is an effective variant prioritization tool for the integrative analysis of Mendelian HPO-encoded clinical diagnoses and NGS data.

References

1 Robinson PN, Kohler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, et al: Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 2014;24:340–348.

2 Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, et al: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 2014;42:D966–D974.

3 Jager M, Wang K, Bauer S, Smedley D, Krawitz P, Robinson PN: Jannovar: a java library for exome annotation. Hum Mutat 2014;35:548–555.

4 Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, Flynn ED, Girdea M, Godfrey R, Golas G, et al: Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet Med 2016;18:608–617.

5 Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, Jager M, Hochheiser H, Washington NL, McMurry JA, et al: A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet 2016;99:595–606.

Genome-Wide Data Manipulation, Association Analysis and Heritability Estimates in R with Gaston 1.5

Claire Dandine-Roulland[1], *Hervé Perdry*[2]

[1]Centre National de Recherche en Génomique Humaine, Institut François Jacob, CEA, Université Paris-Saclay, Evry, France; [2]CESP, Inserm, Univ. Paris-Sud, Université Paris-Saclay, Villejuif, France
Correspondence to: dandine@cng.fr

The R package Gaston, available on the CRAN website, allows to manipulate large SNP matrices in an efficient manner, through well-documented functions. Most functions are implemented in C++ for better performances, and many are multi-threaded. It can perform all classical Quality Control steps, including Principal Component Analysis.

Gaston can read genotype data in the PLINK format (bed/bim/fam) and VCF files. Genotypes are stored compactly (2 bits per genotypes). It computes efficiently Linkage Disequilibrium matrices, Genomic Relationship Matrices (GRMs) and Dominance Matrices (allowing to compute broad heritability estimates, cf abstract by *Herzig et al).*

An implementation of the AIREML algorithm for dense variance matrices is included, which allows to fit Linear Mixed Models, and thus to obtain genomic heritability estimates. Genome-Wide Association testing can be done under a variety of models, in particular using the Linear Mixed Model to account for population stratification.

The versatility of the R environment and the good performances of Gaston make it a viable alternative to classical standalone genetic softwares. Companion packages allowing Gene x Environment association testing, rare variants analysis, or the manipulation of so-called "dosage" genotypes (as produced by imputation softwares) are currently being developed.

An Efficient Test for Gene-Environment Interaction in Generalized Linear Models with Family Data

Mariza de Andrade[1], *Mauricio A.M. Lopera*[2], *Brandon J. Coombes*[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; [2]School of Statistics, National University of Colombia, Medellin, Antioquia, Colombia
Correspondence to: mandrade@mayo.edu

Gene-environment (GE) interaction has important implications in the etiology of complex diseases that are caused by a combination of genetic factors and environment variables. Several authors have developed GE analysis in the context of independent subjects or longitudinal data using a gene-set.

In this presentation, we propose a method to analyze GE interaction for discrete and continuous phenotypes in family studies by incorporating the relatedness among the relatives for each family into a generalized linear mixed model (GLMM) and by using a gene-based variance components test.

In addition, we deal with collinearity problems arising from linkage disequilibrium among single nucleotide polymorphisms (SNPs) by considering their coefficients as random effects under the null estimation. We show that the best linear unbiased predictor (BLUP) of such random effects in the GLMM is equivalent to the ridge regression estimator. This equivalence provides a simple method to estimate the ridge penalty parameters in comparison to other computationally-demanding estimation approaches based on cross-validation schemes. We evaluated the proposed test using simulation studies and applied it to real data from the Baependi Heart Study consisting of 76 families. Using our approach, we identified an interaction between BMI and the Peroxisome Proliferator Activated Receptor Gamma (PPARG) gene associated with diabetes.

A Comparison of Univariate and Multivariate GWAS Methods for Analysis of Multiple Dichotomous Phenotypes

Yasmmyn D. Salinas[1], *Andrew T. DeWan*[1], *Zuoheng Wang*[2]

[1]Department of Chronic Disease Epidemiology, Yale School of Public Health, Yale University, 60 College St, New Haven, Connecticut, United States of America; [2]Department of Biostatistics, Yale School of Public Health, Yale University, 60 College St, New Haven, Connecticut, United States of America
correspondence to: andrew.dewan@yale.edu

Analysis of multiple phenotypes in genome-wide association studies (GWASs) has the potential to enhance statistical power and allows for exploration of pleiotropy. Multi-trait analyses can be conducted using both univariate and multivariate methods. To select an analytic approach, it is important to understand the performance of available methods. However, comparative evaluations of multi-trait methods have primarily focused on the analysis of quantitative traits. Therefore, this study aimed to evaluate the performance of multivariate GWAS methods for analysis of dichotomous (case/control) phenotypes using simulated data. We focused on three methods implemented through R statistical packages—MultiPhen, generalized estimating equations (GEEs), and generalized linear mixed models (GLMMs)—and also compared them to the standard univariate GWAS. We simulated data (N = 20,000) for one bi-allelic SNP and two case/control phenotypes assuming a classical liability threshold model, and varied the number of traits associated with the SNP, degree of association, trait-specific prevalences, and cross-phenotype correlation. We generated 10,000 replicates and evaluated power using a genome-wide significance level of $5x10^{-8}$. Our results show that, in the absence of pleiotropy, multivariate methods outperform the univariate when there are strong, positive cross-phenotype correlations, but that, in the presence of pleiotropy, the univariate approach tends to outperform multivariate methods when the cross-phenotype correlation is positive. GEEs outperformed MultiPhen and GLMMs across most scenarios. This suggests that, to maximize GWAS discovery, the use of univariate and multivariate (GEE-based) approaches in parallel can be recommended. This study provides researchers with empirical guidelines for the application of these methods to real data.

Multi-Phenotype Epigenome-Wide Association Analysis of Fasting Glucose and Insulin in 1,105 Finnish Individuals

Harmen Draisma[1], *Matthias Wielscher*[2], *Saqib Hassan*[1], *Zhanna Balkhiyarova*[1], *Marjo-Riitta Jarvelin*[2,4], *Marika Kaakinen*[1,3], *Inga Prokopenko*[1]

[1]Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London, UK; [2]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK; [3]Centre for Pharmacology and Therapeutics/Division of Experimental Medicine, Imperial College London, London, UK; [4]Center for Life-Course Health Research and Northern Finland Cohort Center, Biocenter Oulu, University of Oulu, Oulu, Finland
Correspondence to: h.draisma@imperial.ac.uk

Background: Multi-phenotype genome-wide association studies (MP-GWAS) of correlated traits have greater power to detect genotype–phenotype associations than single-trait GWAS. In our previously-developed MP-GWAS method, implemented in the SCOPA software, single-nucleotide polymorphism (SNP) genotype dosage is 'reversely' regressed on a linear combination of phenotypes. Considering the epidemiologic correlations between fasting plasma glucose (FG) and fasting insulin (FI) levels, we aimed to use a SCOPA adaptation for multi-phenotype epigenome-wide association analysis (MP-EWAS) for these two traits in non-diabetic individuals.

Methods: We developed the "methylSCOPA" software, extending the SCOPA approach to MP-EWAS using DNA hyper/hypo-methylation data, and applied it to FG, FI and Illumina Infinium HumanMethylation450K BeadChip array data from the Northern Finland Birth Cohorts (NFBC) 1966/1986. We quality-controlled the data, regressed out the effects of measured (potential) confounders, and normalized the methylation signal intensity and FI data. The MP-EWAS included data for 670/435 individuals from NFBC1966 and NFBC1986, respectively. We meta-analyzed the cohort-specific MP-EWAS results, mapped genomic locations to CGCh37/hg19, and adopted $p < 1\times10^{-7}$ to denote epigenome-wide significance.

Results: The strongest association in MP-EWAS meta-analysis (433,848 methylation probes) at cg05063096 (chr3:143,689,810) within *C3orf58* ($p = 2.3\times10^{-7}$) was driven by its effect on FI ($\beta = -5.1\times10^{-3}$, $SE = 1.1\times10^{-3}$, $p = 2.4\times10^{-6}$) while demonstrating only marginal impact on FG ($\beta = 1.4\times10^{-3}$, $SE = 8\times10^{-4}$, $p = 0.08$). We detected a nominal association within the FG well-established *ABCG1* locus at cg06500161 (chr21:43,656,587; $p = 0.03$). This variant was associated with FG ($\beta = -3.2\times10^{-3}$, $SE = 1.7\times10^{-3}$, $p = 0.06$) but not with FI.

Conclusion: We implemented MP-EWAS in methylSCOPA, and demonstrated its enhanced power over single-trait EWAS for correlated phenotypes in large-scale data.

Integration of Polygenic Risk Estimates into Personalized Risk Prediction Algorithms – Experience from the Estonian Biobank

Krista Fischer[1], *Kristi Läll*[1,2], *Tõnu Esko*[1]

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia
Correspondence to: Krista.Fischer@ut.ee

We will discuss the process of development and validation of algorithms for personalized prediction of complex disease risk. One aims to capture the genetic component of the risk in a Genetic (polygenic) Risk Score (GRS). Usually the GRS is defined as a linear combination of effect allele counts of several or even several thousands of Single Nucleotide Polymorphisms (SNPs), whereas the SNPs and their corresponding weights are based on results of a large-scale meta-analysis of Genome-Wide Association Study (GWAS).

In this talk we demonstrate how the effects of phenotypic risk factors and GRS are combined to an overall risk score and illustrate how the absolute risks can be calculated. We will discuss the prediction of the risk of three common complex diseases: Type 2 Diabetes, Coronary Artery Disease and Breast Cancer. There are several statistical challenges that are related to specific features of the population-based biobank data, such as left-truncation for some outcomes, mix of retrospective and prospective data for some others, etc. We will discuss how such features can be accommodated in the analysis process.

Next we will discuss the issues of absolute risk prediction and possible ways to accommodate competing risks in this process. Finally, we demonstrate how this methodology is implemented in the process of giving feedback to the biobank cohort participants.

Quantify Genomic Heritability Through a Prediction Measure

Arthur Frouin[1], *Edith Le Floch*[1], *Christophe Ambroise*[2,3]

[1]Institut de Biologie François Jacob, CEA, Evry France; [2]UMR 8071 LaMME, UEVE CNRS ENSIIE USC INRA, France; [3]UMR MIA-Paris, AgroParisTech INRA Universite Paris-Saclay, Jouy-en-Josas, France
Correspondence to: frouin@cng.fr

Many associations have been identified by Genome-wide association studies (GWAS) between common genetic variants and complex phenotypes. However, these variants only explain a limited amount of the genetic variability of these phenotypes. Several hypotheses have been suggested to explain this missing heritability: the influence of rare variants, gene-gene and gene-environment interactions or the additivity of many weak effects of common variants not detected in GWAS. The literature is teeming with methodology based on the latter hypothesis.

Genomic additive heritability is currently defined via the mixed model as the part of phenotypic variance explained by genetic markers. Weak genetic effects are modeled as distributed according to the same Gaussian law. In this context of high-dimensional statistics, the estimators of heritability may show great variability at fixed sample size.

This work proposes an alternative measure of heritability based on the ability to predict a quantitative phenotype. The proposed approach is illustrated with ridge regression on simulations. The use of ridge regression makes it possible to exhibit a link between our predictive approach and the usual approach based on variance ratio. Generalized cross-validation is used for an effective and easy estimation of the penalization parameter of ridge regression.

Using an existing link between the penalization parameter of ridge regression and the estimation of variance components in the random-effect model, a method for estimating the standard heritability defined as a variance ratio is also proposed using ridge regression. Simulations have shown an equivalent performance between these two approaches.

Improving Gene Expression Prediction Accuracy in Transcriptome-Wide Association Studies

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

[1]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK; [2]Department of Biostatistics, University of Liverpool, Liverpool, UK
Correspondence to: j.j.fryett@newcastle.ac.uk

In transcriptome-wide association studies (TWAS), gene expression values are predicted from genotype data, then tested for association with a phenotype. The power of this approach to detect associations relies on the accuracy of the gene expression prediction. Here we compare the prediction accuracy of six different statistical models – LASSO, Ridge regression, Elastic net, Best Linear Unbiased Predictor (BLUP), a Bayesian Sparse Linear Mixed Model (BSLMM) and Random Forests – using data from the Geuvadis Project. Using cross-validation, we investigate how accurately these models predict gene expression, and the impact that choice of statistical model has on detection of expression-trait associations. We also examine prediction accuracy at different sample sizes, and when the population used to train the model is different from the population for whom expression is being predicted. We find that expression cannot be accurately predicted for most genes with any model. Overall, sparse statistical models tend to predict expression better than polygenic models, with the BSLMM showing the best average prediction accuracy. When applied to type 1 diabetes data from the Wellcome Trust Case Control Consortium, all statistical models find similar expression-trait associations. We observe that all statistical models perform less well when the model training set size is reduced, and when predicting expression into a population different from the population used to train the model. We conclude that using sparse statistical models and increasing the reference panel size will lead to better prediction of gene expression, and may lead to detection of more associations in TWAS.

Association Mapping of Multivariate Phenotypes in the Presence of Missing Data

Saurabh Ghosh, *Kalins Banerjee*

Human Genetics Unit, Indian Statistical Institute, Kolkata, India
Correspondence to: saughosh@gmail.com

Clinical end-point traits are often characterized by quantitative and/or qualitative precursors and it has been argued that it may be statistically a more powerful strategy to analyze a multivariate phenotype comprising these precursor traits to decipher the genetic architecture of the underlying complex end-point trait. We (Majumdar et al., 2015) recently developed a Binomial Regression framework that models the conditional distribution of the allelic count at a SNP given a vector of phenotypes. The model does not require a priori assumptions on the probability distributions of the phenotypes. Moreover, it provides the flexibility of incorporating both quantitative and qualitative phenotypes simultaneously. However, it may often arise in practice that data may not be available on all phenotypes for a particular individual. In this study, we explore methodologies to estimate missing phenotypes conditioned on the available ones and carry out the Binomial Regression based test for association on the "complete" data. We partition the vector of phenotypes into three subsets: continuous, count and categorical phenotypes. For each missing continuous phenotype, the trait value is estimated using a conditional normal model. For each missing count phenotype, the trait value is estimated using a conditional Poisson model. For each missing categorical phenotype, the risk of the phenotype status is estimated using a conditional logistic model. We carry out simulations under a wide spectrum of multivariate phenotype models and assess the effect of the proposed imputation strategy on the power of the association test vis-a-vis the ideal situation with no missing data as well as analyses based only on individuals with complete data. We illustrate an application of our method using data on Coronary Artery Disease.

Shared Genetic Ancestry of Scotland and Ireland Reveals Fine Scale Structure

Edmund Gilbert[1], *James F. Wilson*[2,3], *Gianpiero L. Cavalleri*[1,4]

[1]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin, Ireland; [2]MRC Human Genetics Unit, University of Edinburgh, Edinburgh, United Kingdom; [3]Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, Scotland; [4]FutureNeuro Research Centre, Royal College of Surgeons in Ireland, Dublin, Ireland
Correspondence to: gcavalleri@rcsi.ie

Scotland and Ireland are separated in places by less than 20 kilometres of sea. They share the Gaelic language and similar frequencies of particular alleles and phenotypes, hinting at shared ancestry. The population structure within England and Ireland have previously been described. However, the extent of structure within the majority of Scotland, its surrounding islands, and their links to Ireland are currently unknown. We present an analysis of the British Isles and Ireland using a combined and comprehensive sample (n = 2,556) of all major regions within – expanding coverage in mainland Scotland (n = 567), the Hebrides (n = 57), the Isle of Man (n = 40), Orkney (n = 111) and Shetland (n = 172). By analysing individuals with extended ancestry from specific regions, we demonstrate extensive structure in all regions of the British Isles and Ireland, as well as some of the finest scale structure observed worldwide within Orkney. We; illustrate common genetic ancestries between Ireland and Mainland Scotland, confirm the strongest differentiation of Orkney and Shetland from other populations, show the major differentiation in Mainland Scotland is between the south-west and the north-east, and reveal the distinctiveness of the Hebrides and the Isle of Man. We also show different Norwegian ancestries across Scottish islands, and different ancient ancestries across Britain and Ireland. Our work represents the first comprehensive description of genetic structure in the British Isles and Ireland greatly expands the knowledge of genetic stratification within the north of the British Isles, informing on the study of rare genetic variants and genetic trait associations in these populations.

Classification of CAD Status Using Machine Learning Approaches

Damian Gola[1,2], *Till Andlauer*[3], *Nazanin Mirza-Schreiber*[3], *Lingyao Zeng*[4], *Graciele Delgado*[5], *Marcus Kleber*[5], *Ingrid Gergei*[5], *Winfried März*[5], *Stavroula Kanoni*[6], *Panos Deloukas*[6–8], *Heribert Schunkert*[4], *Nilesh Samani*[9], *Jeanette Erdmann*[2,10,11], *Bertram Müller-Myhsok*[3], *Inke R. König*[1,2,13]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]German Centre for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Lübeck, Germany; [3]Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany; [4]Deutsches Herzzentrum München, Technische Universität München, München, Germany; [5]Medical Clinic V, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany; [6]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom; [7]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [8]Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah, Saudi Arabia; [9]Deparment of Cardiovascular Sciences University of Leicester and NIHR Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, United Kingdom; [10]Institute for Cardiogenetics, Universität zu Lübeck, Lübeck, Germany; [11]University Heart Center Lübeck, Lübeck, Germany; [12]DZHK, partner site Munich Heart Alliance, Munich, Germany; [13]Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL)
Correspondence to: inke.koenig@imbs.uni-luebeck.de

Coronary artery disease (CAD) is the leading global cause of mortality and has substantial heritability with a polygenic architecture. Recent approaches of risk estimation and risk prediction have not utilized the complete genetic information available, but rather focused on genetic loci found to be associated to CAD. We benchmarked simple genomic risk scores (RS), logistic (penalized) regression, naïve Bayes (NB), random forests (RF), support vector machines (SVM) and gradient boosting (GB) on a dataset of 7736 CAD cases and 6774 controls from Germany to identify the algorithms for most accurate classification of CAD patients. After training, the final models were validated on an internal dataset of 527 CAD cases and 473 controls and two external datasets from UK (1874 CAD cases, 1874 controls) and Germany, France and UK (365 CAD cases, 401 controls).

We found RS using 50 633 genetic markers to be the most suitable algorithm for CAD classification, yielding an area under the receiver operating curve (AUC) of 0.9131 in the benchmark and a cross-validated AUC of 0.9106 in training. NB and SVM performed better than RF and GB, with AUC of about 0.81 and 0.76, respectively. These fairly good performances were confirmed in the internal validation dataset. However, on the external datasets all classification models had dramatically reduced AUC of 0.5–0.6. We conclude that using all available genetic information can boost classification performance, although external validation of prediction models is crucial to assess their usability in populations different from those used to build the models.

Allele Counts: A Good Alternative for Testing Genetic Association in Next Generation Sequence Data

Rosa González Silos, *Justo Lorenzo Bermejo*

Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany
Correspondence to: gonzalez@imbi.uni-heidelberg.de

Called genotypes are typically used to investigate the relationship between a phenotype of interest and a particular genetic variant. Genotypes are called using probabilistic methods that rely on quality scores and allele counts computed after base-calling and alignment. Here we investigate an alternative approach that takes into account the uncertainly of called genotypes. We directly use allele counts to test genetic association.

We took advantage of next generation sequence (NGS) data on chromosome 20 from 1417 HapMap individuals to simulate association studies. At each genetic position, phenotypes were simulated independently of, and depending on individual genotypes (*phenotype~N(#alternative alleles,* $\sigma = 6.5$*)* subsequently median-dichotomized). We grouped NGS-data into 18 categories according to the individual genotype, the reference allele and the alternative allele, and simulated genotype quality scores and allele counts relying on NGS data from eight participants in the Personal Genome Project. First, quality scores were randomly drawn according to observed frequencies for each genotype class. Then, allele counts were simulated according to a bivariate normal distribution which reflected observed counts. Finally, genotypes were called with the Haplotype Caller.

Different regression models were used to explore the relationships "called genotype-phenotype" and "allele counts-phenotype", and the type-I-error rate and statistical power were evaluated. For illustration, the attained statistical power was 0.69 using called genotypes and 0.75 using allele counts, and the two methods revealed no type-I-error rate inflation. Details on simulations complemented with the analysis of a real dataset, applied methodology and results will be presented at the meeting.

Genetic Causal Pathways: Mendelian Randomization, Fine-Mapping and Colocalization

Qian Liu[1,2], *Hui Guo*[1]

[1]Centre for Biostatistics, School of Health Sciences, The University of Manchester, Manchester, UK; [2]School of Mathematics and Statistics, Xidian University, Xi'an, P.R. China
Correspondence to: hui.guo@manchester.ac.uk

A number of single nucleotide polymorphisms (SNPs) are found to be associated with certain clinical outcomes and exposures. To date, several statistical approaches have been developed with the aim to understand genetic causal pathways, which are of clinical importance.

Existing methods, such as Mendelian randomization, fine-mapping and colocalization, make the use of summary statistics from SNP-exposure and SNP-outcome association studies. Mendelian randomization is designed for estimating causal effect of the exposure on the outcome, using exposure associated SNPs as instruments. In recent applications, Fine-mapping and colocalization are tailored to quantify the likelihood of SNPs causal to both the exposure and the outcome in genetic regions. Fine-mapping assumes at least one SNP is causal in each region while colocalization assumes there is at most one common causal SNP. Despite the different underlying assumptions between fine-mapping and colocalization, we see similarities of the two approaches.

Mendelian randomization alone is insufficient to identify causal pathways starting from the SNP level. Recent research has proposed a method by combining Mendelian randomization with fine-mapping. We aim to compare the results from Mendelian randomization and fine-mapping (MR_FM) with the former coupled with colocalization (MR_COLOC). In particular, we will make the use of publicly available summary data of BMI and diabetes from large studies, applying both MR_FM and MR_COLOC, to investigate potential causal pathways from genotypes to diabetes through BMI, which will help further explore biological mechanisms of diabetes.

Effect of Reference Panel in ID Score Regression Analysis of Lipid Traits in the Finnish Population

Heidi Hautakangas[1], *Aki S. Havulinna*[1,2], *Veikko Salomaa*[2], *Samuli Ripatti*[1,3], *Matti Pirinen*[1,3,4]

[1]Institute for Molecular Medicine Finland, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland; [2]National Institute of Health and Welfare, Helsinki, Finland; [3]Department of Public Health, University of Helsinki, Helsinki, Finland; [4]Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
Correspondence to: heidi.hautakangas@helsinki.fi

LD Score Regression (LDSC) uses summary statistics from a genome-wide association study (GWAS) and LD information from an external reference panel to estimate heritability. In particular, LDSC does not require access to the original individual-level data. However, mismatch between LD estimates and summary statistics is known to cause problems in some applications such as fine-mapping. Here we evaluate whether similar problems occur with LDSC.

We estimate heritability of four blood serum lipid levels; high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides (TG) and total cholesterol (TC) in the National FINRISK Study. As LD reference panels, we use three different subsets of 1000 Genomes Project Phase 3 (1KG) data (Finns n = 99, non-Finnish Europeans n = 405 and all Europeans n = 504) and the LD information from a subset of n = 10,659 individuals from the underlying GWAS data. Using the LD from the GWAS data, the heritability estimates were 0.08 (s.e. 0.03) for HDL, 0.21 (0.04) for LDL, 0.22 (0.04) for TC and 0.18 (0.04) for TG.

In general, variation in the heritability estimates between different LD reference panels was small in all four lipid traits and estimates of external LD reference panels were consistent with the estimates obtained from the GWAS data. The LD information from the GWAS data produced the smallest standard errors.

What Insights Can Be Gained Through Comparisons of Broad-Sense Heritability Estimates in Isolated and Outbred Populations?

Anthony F. Herzig[1,2], *Teresa Nutile*[3], *Daniela Ruggiero*[3], *Marina Ciullo*[3,4], *Hervé Perdry*[5], *Anne-Louise Leutenegger*[1,2]

[1]Université Paris-Diderot, Sorbonne Paris Cité, U946, F-75010 Paris, France; [2]Inserm, U946, Genetic variation and Human diseases, F-75010 Paris, France; [3]Institute of Genetics and Biophysics A. Buzzati-Traverso – CNR, Naples, Italy; [4]IRCCS Neuromed, Pozzilli, Isernia, Italy; [5]Université Paris-Saclay and CESP, Villejuif, France
Correspondence to: anthony.herzig@inserm.fr

Inconsistencies have been observed between published estimations of non-additive heritability for many phenotypes between studies of isolated populations and of unrelated individuals. This prompted us to investigate the methods used for estimating broad-sense heritability in order to discern whether such inconsistencies could be indicative of particular trait ætiologies, result from specific population characteristics, or stem from non-equivalence between interpretations of heritability in differing study settings. To achieve this, we analyse simulated data mimicking either an isolate or a large sample of unrelated individuals, as well as data from the isolates of Cilento, with a range of cardiovascular related traits.

We focus on linear mixed modelling techniques for trait variance decomposition. We discuss the use of identity by descent (IBD) based relatedness matrices or genetic correlation matrices (GRMs) and show the strengths of GRMs against even explicit knowledge of IBD-sharing in an isolate. In the case of trait architectures that involve many low-frequency causal variants, we demonstrate that analysing population isolates avoids the downward bias of heritability estimates that is known to occur in studies of outbred populations.

Our analyses of the Cilento isolates produced significant estimates for non-additive components for traits such as body-mass index and low-density lipoprotein level. These analyses, in conjunction with our simulation study, have aided us firstly to clarify the disparities in published heritability results from differing populations and estimation methods and secondly to suggest how such arrays of results should be interpreted. Furthermore, we advocate that for many traits, non-additive genetic models should be revisited.

Imputation of Missing Data for Bayesian Network Analyses of Complex Biological Data

Richard Howey, *Heather Cordell*

Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK
Correspondence to: richard.howey@ncl.ac.uk

Bayesian networks have been proposed as a way to identify possible causal relationships between measured variables based on their conditional dependencies and independencies, particularly in complex scenarios with many variables. When there is missing data, the standard approach is to remove individuals with missing data before performing Bayesian network analyses. This is undesirable when there are many individuals with missing data, perhaps with only one variable missing. Thus, imputation of the missing data is a natural choice. We present a new imputation approach designed to increase the power to detect causal relationships whilst accounting for model uncertainty. This method uses a version of nearest neighbour imputation, whereby missing data from one individual is replaced with data from another individual, their nearest neighbour. An important feature of this approach is that it can be used with both discrete and continuous data. For each individual with missing data, a single bootstrap iteration of the complete data is used to estimate a preliminary Bayesian network. Subsets of variables that have close connections to the missing variables are then chosen to find the nearest neighbour. We show that use of our imputation method increases the power to detect the correct model in simulated data by as much as over 50%. Such increases may be possible in real data when most individuals have missing data due to cost or practical reasons. Thus, the use of our imputation method has great potential to boost the power of Bayesian network analyses to identify possible causal relationships.

A Semi-Supervised Approach for Predicting Cell Type/Tissue Specific Functional Consequences of Non-Coding Variation Using Massively Parallel Reporter Assays

Zihuai He[1], *Linxi Liu*[2], *Kai Wang*[3], *Iuliana Ionita-Laza*[1]

[1]Department of Biostatistics, Columbia University, New York, NY, USA; [2]Department of Statistics, Columbia University, New York, NY, USA; [3]Department of Biomedical Informatics, Columbia University, New York, NY, USA
Correspondence to: ii2135@cumc.columbia.edu

Predicting the functional consequences of genetic variants is a challenging problem, especially for variants in non-coding regions. Projects such as ENCODE and Roadmap Epigenomics make available various epigenetic features, including histone modifications and chromatin accessibility, genome-wide in over a hundred different tissues and cell types. In addition, recent developments in high-throughput assays to assess the functional impact of variants in regulatory regions (e.g. massively parallel reporter assays, CRISPR/Cas9-mediated in situ saturating mutagenesis) can lead to the generation of high quality data on the functional effects of selected variants. We propose here a semi-supervised approach, GenoNet, to jointly utilize experimentally confirmed regulatory variants (labeled variants), millions of unlabeled variants genome-wide, and more than a thousand cell type/tissue specific epigenetic annotations to predict functional consequences of non-coding genetic variants. Through the application to several experimental datasets, including massively parallel reporter assay validated variants, and sets of eQTLs and dsQTLs, we demonstrate that the proposed method significantly improves prediction accuracy compared to existing functional prediction methods, both at the organism and tissue/cell type level. We further show that eQTLs and especially dsQTLs in specific tissues tend to be significantly enriched among variants with high GenoNet scores, and how the

GenoNet scores can help in the discovery of disease associated genes through an integrative analysis of lipid phenotypes using a Metabochip dataset on 12,281 individuals.

FastNGSadmix: Admixture Proportions and Principal Component Analysis of a Single Low-Depth Sequencing Sample

E. Jørsboe[1], K. Hanghøj[2,3], A. Albrechtsen[1]

[1]The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen N, Denmark; [2]Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen K, Denmark; [3]Université de Toulouse, University Paul Sabatier (UPS), Laboratoire AMIS, CNRS UMR 5288, Toulouse, France

We present fastNGSadmix, a method for fast and easy estimation of admixture proportions and principal component analysis (PCA), for a single low-depth next generation sequencing (NGS) sample, using a panel of reference populations, with population specific allele frequencies.

We show that fastNGSadmix has increased accuracy compared to established methods for estimating admixture using reference panels, such as iAdmix and ADMIXTURE. fastNGSadmix corrects for the bias, introduced by having a limited size reference panel, which is a substantial problem with other methods. fastNGSadmix works for samples with very low sequencing depth, we show through down-sampling that fastNGSadmix works for samples with depth of less than 0.005X. This is because the method uses genotype likelihoods, thereby incorporating the uncertainty of the genotypes in the model.

fastNGSadmix works by maximising the likelihood of the sequencing data given the admixture proportions and the population specific allele frequencies, using the EM algorithm. A bootstrapping approach has been implemented for the admixture estimation, meaning the uncertainty on the admixture estimates can easily be obtained.

We use the estimated admixture proportions to perform PCA incorporating both population structure and genotype uncertainty. Existing PCA methods based on NGS data do not model population structure. This method models population structure via the estimated admixture proportions from fastNGSadmix.

This method has been applied to ancient DNA samples. Ancient DNA is usually characterised by low quality data, why it is crucial to take genotype uncertainty into account. In addition, the samples are modelled independently, which allows for analysing related samples.

The Genomic Basis of Human Lifespan

Paul R.H.J. Timmers[1], Ninon Mounier[2,3], Kristi Läll[4,5], Krista Fischer[4], Zheng Ning[6], Xiao Feng[7], Andrew Bretherick[8], David W. Clark[1], Xia Shen[1,6], Tõnu Esko[4], Zoltán Kutalik[2,3], James F .Wilson[1,8], Peter K. Joshi[1,2]*

[1]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom; [2]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland; [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4]Estonian Genome Center, University of Tartu, Tartu, Estonia; [5]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [7]State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-sen University, Guangzhou, China; [8]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom

Correspondence to: peter.joshi@ed.ac.uk

Motivation: Whilst of great interest to us all, investigation into the genomic basis of longevity has been hampered by limited sample sizes. As a result, until very recently, only 4 genome-wide significant loci had been discovered and replicated, limiting the inferences that be made about its genetic basis1.

Results: Here using an independent replication cohort, we examine 20 published2 but unreplicated genome-wide significant loci for longevity, validating associations at or near CDKN2B-AS1, ATXN2/BRAP, FURIN/FES, FOXO3A, 5q33.3/EBF1, ZW10, PSORS1C3, 13q21.31, and provide evidence against previous findings near CLU, CHRNA4, PROX2, and d3-GHR. In a GWAS using all data combined, totalling over 1m lifespans, we next find 15 further loci at genome-wide significance. Of the life lengthening loci significant in our analyses, we find many protect against cardiovascular, metabolic and neurological diseases (CGND), but not cancer. Using our GWAS, we then create polygenic scores for survival in independent sub-cohorts, and are able to partition populations, using DNA alone, into deciles of expectation of life, with a difference in excess of five years from top to bottom decile.

Conclusion: It seems that natural selection has been more effective in purging common variants affecting lifespan through cancer, but in our modern obesogenic and long lived environment has yet to as effectively purge variants affecting CGND. Materially accurate predictions of lifespan can now be made from DNA.

DOI: 10.1159/000488519

Genome-Wide Association Study of 1,124 Protein Levels in Pulmonary Arterial Hypertension Patients Identifies a Novel *trans*-pQTL at *ELK2AP* for Death Receptor 3

Marika Kaakinen[1,2], *Christopher J. Rhodes*[1], *NIHR BioResource for Rare Diseases*, *Inga Prokopenko*[2], *Martin Wilkins*[1]

[1]Pharmacology and Therapeutics, Imperial College London, United London, Kingdom; [2]Genomics of Common Disease, Imperial College London, London, United Kingdom
Correspondence to: m.kaakinen@imperial.ac.uk

Recent genome-wide association studies (GWAS) on blood plasma proteome in healthy individuals have identified hundreds of protein quantitative trait loci, pQTLs, both *in cis* and *in trans*. Dissecting the mechanisms of protein level variability in unhealthy individuals may help to reveal novel disease-specific pathways. Pulmonary arterial hypertension, PAH, is a rare disease leading to premature death. We conducted a GWAS using whole-genome sequencing data and 1,124 blood plasma protein levels measured with the SOMAscan platform in 128 British (discovery) and 79 French (replication) PAH patients. Each protein level was natural logarithm transformed, adjusted for age, sex and four principal components, and the resulting residuals were further inverse-normal transformed to assure normality. We discovered 21 cis and 6 trans-pQTLs at genome-wide significance corrected for multiple testing ($P < 4.45\times10^{-11}$) that replicated with nominal significance and directional consistency. These contain four novel *trans*-pQTLs at/near *ELK2AP* in the immunoglobulin heavy locus (IGH) for death receptor 3 (DR3); *MIR4435–1* for Complement component 1 subcomponent r (C1r); *RETNLB* for Hepatocyte growth factor activator (HGFA); and *TMEM215* for Properdin. Our study shows that the combination of intermediate phenotypes with high coverage genomic data in a patient cohort of a relatively small sample size is already well-powered to detect dozens of signals at genome-wide significance and to provide novel biological clues into the rare disease of PAH. Future work will need to address the correlation between the proteins to avoid unnecessary penalisation from multiple testing and to further boost the power for association detection.

Population-Specific Imputation Reference Panel as a Tool for GWAS Analysis

Mart Kals[1,2], *Priit Palta*[1,3], *Reedik Mägi*[1]

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [3]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
Correspondence to: mart.kals@ut.ee

Over last decade genome-wide association studies (GWAS) have been most frequent tool to identify genes involved in common diseases. Input of GWA studies requires high-quality genetic imputation and imputation accuracy depends on underlying imputation reference panel (IRP). In the current study, we compared GWAS results of common diseases of two IRPs: 1) a population-specific IRP (N = 2,279 Estonians and N = 1,856 Finns) based on high-coverage (30x) whole genome sequences and 2) publically available 1000G IRP (N = 2,504) containing diverse set of populations.

GWA study individuals (N = 36,716, unrelated) originate from the Estonian Biobank of the Estonian Genome Center, University of Tartu (EGCUT). EGCUT is a population-based biobank, containing almost 52,000 samples of the adult Estonian population. Population-specific IRP contained 38 M variants (SNVs and short indels) and 1000G IRP 84 M variants. 2/3 of the samples were genotyped using Illumina Global Screening Array, and the rest 1/3 using HumanOmniExpress, HumanCoreExome or CNV370 microarrays. IRPs and microarray were phased using EAGLE2 and imputed with BEAGLE. For association analysis, we used Firth test, adjusted for age, sex, microarrays and PC1-10. Finally, we applied FINEMAP for fine-mapping in genomic regions.

After filtering remained 10.9 M high-quality imputed variants for EGCUT IRP and 8.7 M for 1000G IRP. Results indicated that we obtained slightly more genome-wide significant (P-value $<5e^{-8}$) and fine-mapped hits using population-specific IRP compared to 1000G IRP. EGCUT IRP is more powerful for low-frequency (MAF <0.5%) variants and for variants which are more frequent in local population compared to European average.

Assessing Novel Anthropometric Indices and their Predictive Efficacy Over Metabolic Health and Disease

Katherine A. Kentistou[1,2], *Peter K. Joshi*[1], *James F. Wilson*[1,3]

[1]Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, EH8 9AG, UK; [2]Centre for Cardiovascular Sciences, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, EH16 4TJ, Scotland; [3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
Correspondence to: k.kentistou@ed.ac.uk

Cardiometabolic disease (CMD) remains a major public health concern and is exacerbated by the modern diets and lifestyles. Obesity, insulin resistance and hypertension are key components of CMD. Fat distribution *per se* influences the risk for CMD. Namely, visceral adiposity, i.e. adipose tissue that surrounds the abdominal organs, is associated with increased CMD risk, whereas lower-body adiposity has been linked to a reduced risk [1]. Thus, recently effort has gone into developing novel indices that combine several anthropometric measurements in order to create more accurate reflections of visceral adiposity [2–4] and thus CMD risk.

However, it is often unclear whether such indices remain equally informative for individuals across all age-ranges, sexes and socioeconomic backgrounds and whether they can enhance the predictive powers of simple anthropometric measures.

This work uses phenotypic and genetic data from 500,000 individuals to assess the predictive efficacy such indices over several aspects of metabolic health or disease. First, we explore the genetic architecture of the association signals that arise for each index, in terms of strength with known obesity and adiposity associations. Differential signals are then examined in terms of enrichment for specific biological pathways. Finally, each index is correlated with measurements indicative of metabolic health to assess their success in doing so, and also whether they are superior to more conventional measurements, such as BMI.

Our findings show how composite anthropometric indices can classify cases of CMD and whether novel insights can be made from their utilisation in genetic and epidemiological studies.

References

1 Karpe F, Pinnick KE: Biology of upper-body and lower-body adipose tissue [mdash] link to whole-body phenotypes. Nature reviews Endocrinology 2015;11(2):90–100.

2 Krakauer NY, Krakauer JC: A new body shape index predicts mortality hazard independently of body mass index. PloS one 2012;7(7):e39504.

3 Thomas DM, et al: Relationships between body roundness with body fat and visceral adipose tissue emerging from a new geometrical model. Obesity 2013;21(11):2264–2271.

4 Bergman RN, et al: A better index of body adiposity. Obesity 2011;19(5): 1083–1089.

Evaluating Robustness and Geographic Differences in Polygenic Risk in Finland

Sini Kerminen[1], Jukka Koskela[1], Aki S. Havulinna[1,2], Ida Surakka[1,3], Alicia R. Martin[4–6], Aarno Palotie[1,4,5,7,8], Markus Perola[1,2], Veikko Salomaa[2], Mark J. Daly[1,4–6], Samuli Ripatti[1,9], Matti Pirinen[1,9,10]

[1]Institute for Molecular Medicine Finland, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland; [2]National Institute of Health and Welfare, Helsinki, Helsinki, Finland; [3]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA; [4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA; [5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, USA; [6]Program in Medical and Population Genetics, Broad Institute, Cambridge, USA; [7]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, USA; [8]Department of Neurology, Massachusetts General Hospital, Boston, USA; [9]Department of Public Health, University of Helsinki, Helsinki, Finland; [10]Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Correspondence to: sini.kerminen@helsinki.fi

The main population structure in Finland appears between eastern and western subpopulations. Our goal is to determine how much of the geographic phenotypic differences this genetic structure explains. As a model trait, we consider adult height which shows a difference of 1.6 cm between the subpopulations and for which multiple large GWASs are available to generate polygenic risk scores (PRS). We use 2,376 geographically well-defined samples from the National FINRISK study to evaluate these PRSs in Finland.

First, we generated a PRS using summary statistics from the GIANT consortium (GWAS n > 250,00). This score explained 14% of the variance of height and showed a suspiciously high, 3.5 cm, predicted difference between the subpopulations. Second, we built PRS using the UK Biobank summary statistics (GWAS n > 400,000) that explained 27% of the height variance in Finland but, surprisingly, showed no statistically significant differences between the subpopulations. Third, we built a PRS from the FINRISK data where all our 2,376 target individuals were excluded. Despite of the small GWAS n = 25,000, this FINRISK-score was able to explain 15% of the variance and predicted 1.4 cm height difference between the subpopulations.

We discuss the predictive power and possible biases of PRSs as a function of similarities in population genetics, sample overlap and sample sizes between the GWAS summary statistics and the target sample. We also relate these results to the observed PRS patterns of several complex diseases in Finland. A thorough understanding of these topics is crucial before taking PRSs to clinical use.

A Joint Analysis of Adiposity Genetics Unravels Subtypes with Different Metabolic Implications

Thomas W. Winkler[1], Felix Günther[1,2], Simon Höllerer[1], Martina Zimmermann[1], Ruth J.F. Loos[3–5], Zoltán Kutalik[6,7,], Iris M. Heid[1,*]*

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany; [3]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, USA; [4]The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, USA; [5]The Mindich Child health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, USA; [6]Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland; [7]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

*Equal contribution

Correspondence to: zoltan.kutalik@unil.ch

The genetics of related phenotypes is often tackled by analyzing adjusted model traits, which may be subject to statistical biases. Here, we developed a joint analysis and applied it to adiposity traits. Analyzing GIANT consortium data (up to 322,154 subjects), we identified 159 single nucleotide polymorphisms (SNPs) robustly ($P < 5x10^{-8}$) associated with at least one classical measures of obesity (body-mass-index (BMI), waist-to-hip-ratio (WHR), WHR adjusted for BMI). These variants were then classified into four groups according to their effect direction for BMI and WHR: increasing both BMI and WHR (via increased waist) [BMI+WHR+]; having opposite effect on BMI and WHR (via hip-increase) [BMI+WHR–]; being associated only with BMI (via proportional waist- and hip-increase) [BMI-only]; WHR-only (via waist-increase and hip-decrease). Deeper examination revealed that these classes represent meaningful genetic subgroups of obesity. First, BMI-only and BMI+WHR– classes impact mostly subcutaneous fat, while WHR-only SNPs change visceral/adipose fat ratio. Class-specific Mendelian randomization provided evidence that 1 SD (~4.5 kg/m^2) of genetically increased BMI driven by waist-gain (via the 82 BMI+WHR+ variants) confers 2.5 higher odds for type 2 diabetes (T2D) ($P_{MR} = 5x10^{-29}$). However, when the same BMI-increase is coupled with hip-gain (via the 24 BMI+WHR– variants) it reduces the T2D odds by >80% ($P_{MR} = 7x10^{-12}$). Finally, tissue enrichment analyses (DEPICT, FUMA) implicate the involvement of the gastrointestinal tract and sigmoid colon for the WHR-only subclass. Our results demonstrate that a joint view of related phenotypes genetics can unravel obesity sub-types with variable fat depots and vastly different metabolic health (lipids, coronary artery disease, T2D) consequences.

Comparison of Genetic Risk Scores in the Estonian Biobank Cohort

Kristi Läll[1,2], *Reedik Mägi*[1], *Marili Palover*[1], *Krista Fischer*[1]

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia
Correspondence to: kristi.lall@ut.ee

So far, published genetic risk scores (GRSs) for breast cancer have been based on only a limited number of SNPs, as the summary statistics of large meta-analysis have not been publicly available. Last year, this was rectified by two large-scale studies, giving an opportunity to systematically develop and test polygenic versions of GRSs.

Estonian Biobank cohort consists of more than 32000 women with both genotype and phenotype data available. A case-control sample including 317 prevalent BC cases was formed to develop different versions of GRSs based on both UK biobank GWAS and OncoArray Consortium results. Using logistic regression models, optimal GRSs from both studies (denoted as GRS_{UK} and GRS_{ONCO}) were chosen. Two already published GRSs with 70 GRS_{70} and 77 SNPs GRS_{75} were also calculated. The predictive ability of all four scores and combination of the scores were investigated using incident breast cancer data (n = 308).

Combining several GRSs together (including almost 1000 SNPs) into one score (GRS_{meta2}) resulted the strongest predictor. Women in the top 5% GRS_{meta2} category have four times higher hazard (HR = 4.2, 95% CI 2.84–6.2) of developing breast cancer than women in the lower half of the GRS. Furthermore, women on the top 5% GRS_{meta2} category have almost 3 times higher hazard (HR = 2.73, 95% CI 1.92–3.9) of developing breast cancer compared to the rest of the population.

We have showed that including hundreds of individually insignificant SNPs improves the predictive ability of GRS for breast cancer, enabling the better targeting of individuals for prevention.

A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late Onset Alzheimer's Disease Whole Genome Sequence

Linhai Zhao[1], *Di Zhang*[1], *Carl A. Broadbent*[1], *Gao T. Wang*[2], *Alison M. Goate*[3], *Richard Mayeux*[4], *Suzanne M. Leal*[1]

[1]Center for Statistical Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA; [3]Department of Neuroscience and Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA; [4]Department of Neurology, Taub Institute on Alzheimer's Disease and the Aging Brain, and Gertrude H. Sergievsky Center, Columbia University, New York, NY 10027, USA
Correspondence to: sleal@bcm.edu

Nonparametric-linkage (NPL) methods were applied to common variants to analyze complex familial traits. Although linkage could be detected, gene identification was usually not successful due to large mapped regions, e.g. >50 Mb. Motivated by rare-variant (RV) aggregate methods used to analyze complex trait sequence data, we developed the RV-NPL for analysis of pedigrees. The RV-NPL, collapses variants within a region e.g., gene and tests for sharing of rare-haplotypes. The RV-NPL has some definite advantages compared to population and pedigree RV association methods: 1.) increased power to detect causal-variants with familial aggregation due to larger effect sizes; 2.) only necessary to analyze affected individuals which increases power, since unaffected individuals can be susceptibility variant carriers; 3.) can be applied to noncoding regions; and 4.) robust to population substructure, locus and allelic heterogeneity, and inclusion of nonpathogenic variants. To evaluate the RV-NPL method, we simulated exome data using non-Finnish European minor allele frequencies (MAFs) from ExAC. Power was estimated by the proportion of tested genes that were significant ($\alpha = 2.5 \times 10^{-6}$). For all tested scenarios, the RV-NPL was more powerful than the NPL, e.g., for 100 extended-pedigrees power is 86% (RV-NPL) and 74% (NPL) and for 2,000 affected-sib-pairs power is 87% (RV-NPL) and 65% (NPL) when variants with MAF ≤0.01 with odds ratio = 5 were analyzed. Results from analyzing Alzheimer's disease pedigrees and extensive simulation studies demonstrate the power of RV-NPL and its ability to detect linkage to individual genes, making it an ideal method to elucidate the genetic etiology of complex familial diseases.

An Elston-Stewart Algorithm for Computing Exact Derivatives of the Likelihood in Pedigrees

Alexandra Lefebvre, *Gregory Nuel*

Stochastics and Biology Group, Probability and Statistics Lab (LPSM, CNRS 8001), Sorbonne Université, Paris, France
Correspondence to: alexandra.lefebvre@sorbonne-universite.fr

Estimating parameters in genetic diseases requires efficient algorithms to compute likelihood of genetic models in pedigrees. The Elston-Stewart algorithm allows to compute pedigree likelihoods with a complexity $O(n \times g^{tw})$ where n is the number of individuals, g is the number of genotypes, tw is the tree-width of the pedigree (tw = 3 to 5 for standard families). Computing first and second derivatives of the likelihood function is of great interest both to maximise the likelihood more efficiently, and to obtain confidence intervals on parameters. These derivatives can be computed numerically but this approach is slow and might lead to unstable computations.

In this work, we present an extension of the Elston-Stewart algorithm combining Mendelian laws and polynomial arithmetic in order to obtain exact derivatives of the likelihood function. For a univariate model (one parameter to estimate) our algorithm computes derivatives up to the d^{th} order with a multiplicative complexity factor of $(d+1)(d+2)/2=O(d^2)$ (3 for $d = 1$, 6 for $d = 2$). For a multivariate model with p parameters, we obtain the likelihood, the gradient, and the Hessian with a complexity factor of $O(p^2)$.

We illustrate the interest of our algorithm with two classical models in genetic epidemiology: 1) in segregation analysis, order 2 multivariate likelihood derivatives allows to compute confidence interval jointly for penetrance parameters and disease allele frequencies; 2) for two-point linkage we establish the distribution of recombination rate estimates and derive pointwise confidence intervals for LOD scores.

New Quality Measure for CNV: A Multi-Omics Approach

Maarja Lepamets[1,2], *Kaido Lepik*[3], *Mart Kals*[1,4], *Cristian Carmeli*[5], *Annique Claringbould*[6], *Murielle Bochud*[5], *Silvia Stringhini*[5], *Cisca Wijmenga*[6], *Lude Franke*[6], *Reedik Mägi*[1,7], *Zoltán Kutalik*[5,7]

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; [3]Institute of Computer Science, University of Tartu, Tartu, Estonia; [4]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [5]Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, Switzerland; [6]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, Netherlands;
[7]These authors supervised this project equally
Correspondence to: maarja.lepamets@ut.ee

Copy number variants (CNV) are associated with several human traits, but their detection remains challenging. Analysing signal intensities from genotyping arrays with software such as PennCNV is a widely used approach for CNV detection. To minimize false positive calls from PennCNV, various metrics built on CNV and sample parameters have been proposed.

True CNV can affect the expression levels of genes (GE) and the methylation intensity (MET) in regions nearby. Thus, we scored CNV by measuring the fluctuation of GE/MET in CNV carriers compared to non-carriers in LLDeep (N = 1,383) and EGCUT (N = 979), and by the fraction of calls validated by whole-genome sequencing in EGCUT. The scores were concordant: MET-WGS correlation was r = 0.66/0.75 (deletions/duplications) and for MET-GE r = 0.45/0.55. MET-scores provided the best generalization across different cohorts.

Next, PennCNV parameters were used to predict MET-scores in the combined LLDeep+EGCUT dataset. We used CNV of related individuals from UK Biobank (UKB; N = 50,733) and SKIPOGH (N = 816) for validation, observing significant differences between median scores of familial and non-familial CNV in both datasets ($P_{UKB} < 1x10^{-149}$ and $P_{SKIPOGH} < 1x10^{-36}$). Finally, we tested the association of published trait-associated CNV using the UK Biobank (N = 398,623) and EGCUT (N = 30,917). We improved several association signals by an order of magnitude (e.g. $P_{UKB} = 6x10^{-10}$ to $P_{UKB} = 5x10^{-11}$ and $P_{EGCUT} = 0.04$ to $P_{EGCUT} = 0.002$ for the 16p11.2-BMI association) when using the MET-score compared to the best published CNV score. Therefore, using multiomics data for identifying the quality of CNV increases statistical power and enables us to discover new associations.

Genome-Wide Association Analysis of 225 Metabolites in Isolated Populations

Erin Macdonald-Dunlop[1], *Peter K. Joshi*[1], *James F. Wilson*[1,2]

[1]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom; [2]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom
Correspondence to: e.macdonald.dunlop@ed.ac.uk

Circulating serum metabolites' involvement in countless biological pathways makes their abundance a risk factor for numerous diseases, particularly those that are a burden to the health service in the UK such as cardiovascular and metabolic disease (1, 2). The emergence of high-throughput technologies such as nuclear magnetic resonance (NMR) spectroscopy for the profiling and quantification of small molecules has facilitated the identification of disease biomarkers (3–5). Here we present results from GWAS on 225 metabolites on individuals from two isolated population cohorts, ORCADES and Korcula. The use of isolated populations increases the number of rare variants available for association testing and are not likely to be present in other, more heterogeneous populations. ORCADES is a family-based cohort consisting of individuals from the Orkney isles who have at least two grandparents also from Orkney. Korcula is a Croatian cohort based on individuals from the Dalmatian island of the same name. We performed GWAS and meta-analysis of these two cohorts on 225 metabolites in a total of 2,981 individuals (ORCADES = 1,937, Korcula = 1,044) at 11,802,652 genotyped and imputed variants. We report 41 loci that associate with at least one metabolite that surpass Bonferroni corrected genome-wide significance ($p < 2.222222e^{-10}$). All significant loci that are resistant to multiple testing correction have been previously reported to be associated to metabolite levels or metabolic syndrome, validating our results.

References

1 Schlessinger BS, Wilson FH, Milch LJ: Serum Parameters as Discriminators Between Normal and Coronary Groups. Circulation 1959;19:265–268.

2 Stumvoll M, Goldstein BJ, van Haeften TW: Type 2 diabetes: principles of pathogenesis and therapy. Lancet 2005;365:1333–1346.

3 Roberts LD, Gerszten RE: Toward New Biomarkers of Cardiometabolic Diseases. Cell Metabolism 2013;18:43–50.
4 Shah SH, Kraus WE, Newgard CB: Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases: form and function. Circulation 2012;126:1110–1120.
5 Nicholson JK, et al: Metabolic phenotyping in clinical and surgical environments. Nature 2012;491:384–392.

How Well Does Parental Genetic Risk Predict Early Menopause Genetic Risk in Offspring?

Triin Laisk-Podar[1–3], *Reedik Mägi*[1]

[1]Estonian Genome Center, Institute of Genomic Medicine, University of Tartu, Tartu, Estonia; [2]Department of Obstetrics and Gynecology, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia; [3]Competence Centre on Health Technologies, Tartu, Estonia
Correspondence to: reedik.magi@ut.ee

Reproductive aging is a heritable trait. Early menopause (before the age of 45) affects 5% of women, putting them at risk of earlier reproductive senescence and necessitating the search for markers usable for screening. Currently, maternal age at menopause is considered one of the best markers for assessing the respective risk in offspring. The aim of this study was to evaluate if high genetic risk for early menopause in one parent (mother) is an adequate predictor of daughter's genetic risk.

Data for 51,886 samples from the Estonian Biobank was used to calculate early menopause genetic risk scores based on 745 variants associated with reproductive aging. High genetic risk for early menopause was defined as the top 20% of risk scores, whereas medium risk encompassed the 20–80% deciles. Genetic risk scores for 10,000 offspring were simulated using combinations of parental genetic risk (high, medium, random).

If one of the parents has a high genetic risk for early menopause, the offspring has a 3.9%, 52.4% and 43.7% chance of having a low, medium or high genetic risk for early menopause, respectively.

If one of the parents (mother) has a high genetic risk for early menopause, the offspring still has a high likelihood of having medium or even low genetic, therefore parental genetic risk for early menopause is not an accurate predictor in offspring and individual testing is recommended.

A Cluster-Randomized Trial on Personalized Feedback on Genetic Risks: Effects on Treatment Compliance of Patients with Hypertension

Merli Mändul[1,2], *Krista Fischer*[1], *Kristi Läll*[1,2], *Liis Leitsalu*[1], *Helene Alavere*[1], *Marika Tammaru*[3], *Andres Metspalu*[1]

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [3]East-Tallinn Central Hospital, Tallinn, Estonia
Correspondence to: merli.mandul@gmail.com

There is controversial evidence on the potential effects of personalized risk prediction in general practice. To study various aspects of applicability of the genetic risk prediction and feedback to patients and get initial estimates of its short- and long-term effects, a small-scale cluster-randomized trial has been conducted in Estonia. The trial enrolled 238 male participants of 42 family doctors. The participants (age 18–65) were diagnosed with primary hypertension and received their first prescription of anti-hypertensive medication during the first trial visit. The family doctors were randomized into intervention and control arms. The participants of intervention arm doctors received information on their genetic risk for 4 diseases (type 2 diabetes, coronary artery disease, stroke, atrial fibrillation) during their second clinic visit (6 weeks after the initial visit) in addition to a standardized assessment of their non-genetic risk for those diseases. The control group participants received only the non-genetic risk assessment at the second visit, the genetic risk estimates were provided 6 months after the baseline visit. Altogether, the patients were scheduled 5 clinic visits during the one-year trial period.

Using generalized linear mixed models that accommodate random cluster effects and repeated measurements, we analysed the effects of genetic risk prediction on compliance to the anti-hypertensive treatment as well as systolic and diastolic blood pressure. The results indicate that genetic risk prediction is likely to have a short-term effect on lowering blood pressure, but the long-term effects remain unclear, pointing to the need for larger studies in this field.

Inferring Population Structure and Admixture Proportions in Low Depth Next-Generation Sequencing Data

Jonas Meisner, *Anders Albrechtsen*

The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
Correspondence to: jonas.meisner@bio.ku.dk

Population genetic studies usually consist of individuals of diverse ancestries, and inference of population structure therefore plays an important role in population genetics and association studies. structure importance. Here we present PCAngsd, a framework for analyzing low depth next-generation sequencing (NGS) data in heterogeneous populations using principal component analysis (PCA). NGS methods provide large amounts of genetic data but are associated with statistical uncertainty for low depth sequencing data which is used in large-scale population studies due to cost limitations. Probabilistic methods have therefore been developed to take this uncertainty into account when estimating population genetic parameters by using genotype likelihoods and external information. We have developed two new methods for inferring population structure. The first method is using an Empirical Bayes method to estimate individual allele frequencies based on genotype dosages in an iterative approach of inferring population structure. The estimated individual allele frequencies are then used as prior information to estimate a covariance matrix and perform PCA. The second method uses the estimated individual allele frequencies of the first method to estimate admixture proportions based on a fast non-negative matrix factorization (NMF) algorithm. The method for performing PCA outperforms existing methods in both simulated and real low depth NGS datasets, while the method for estimating admixture proportions produces comparable results to other methods with shorter run-times.

A Statistical Method for Alignment-Free Analysis of Sequencing Reads with Applications in Copy Number Determination and Plasmid Integration Detection

Märt Möls

Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia
Correspondence to: martm@ut.ee

The analyzes of data from second generation sequencing experiments can produce heavy computational loads – if the traditional workflows are used. To speed up the computations often fast k-mer based alignment-free methods are used. However, the k-mer counts can behave badly in statistical analyses because the k-mer counts from nearby regions are statistically dependent. Some applications of k-mers also depend on the availability of reference genome – which sometimes might not be complete nor error free. A statistical method to analyze the dependent k-mer counts will be introduced and some possibilities to make the analyses more robust to the possible errors in reference genome will be discussed. The proposed methodology will be illustrated with two examples – an analyze to determine the copy number of a genomic region will be presented and a statistical test for detecting plasmid integration into bacterial genome will be described.

Bayesian Genome-Wide Association Study to Discover Novel Lifespan-Associated Loci

Ninon Mounier[1,2], Paul R.H.J. Timmers[3], Kristi Läll[4,5], Krista Fischer[4], Zheng Ning[6], Xiao Feng[7], Andrew Bretherick[8], David W. Clark[3], eQTLGen Consortium, Xia Shen[1,6], Tõnu Esko[4], James F. Wilson[3,8], Peter K. Joshi[3], Zoltán Kutalik[1,2]

[1]Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne 1010, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; [3]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom; [4]Estonian Genome Center, University of Tartu, Tartu, Estonia; [5]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [7]State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-sen University, Guangzhou, China; [8]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom
Correspondence to: mounier.ninon@gmail.com

Genome-Wide Association Studies (GWASs) are often underpowered for traits with weak genetic associations and for which large datasets are particularly difficult to gather, such as human lifespan. An alternative way to improve GWAS power is to leverage independent sources of information and include them as priors in a Bayesian analysis.

We developed a framework for informed GWAS (and implemented it in the R package bGWAS) that accounts for prior information by comparing the observed Z-scores from a conventional GWAS to prior effects. One way to derive prior effects is to combine summary statistics of GWASs for related traits with their causal effect on the trait of interest using Mendelian Randomization. We applied this method to improve the power of a lifespan GWAS based on >1 million parental lives. Our approach identified 10 additional genome-wide significant variants, notably in the long-suspected *ABO* gene (beta = 2.6 months/allele, $P = 5.5x10^{-10}$) and in *LPL* (beta = 2 months/allele, $P = 9.7x10^{-9}$). Most of the loci identified showed pleiotropic effects, such as a variant near *POM121C* (beta = 2 months/allele, $P = 2.2x10^{-9}$) which has not been significantly associated with any of the risk factors previously but might be affecting lifespan through moderate effects on insulin, body mass index, smoking and coronary artery diseases.

Interestingly, we observed that variants regulating genes whose expression had been shown to vary with age are 3.5 times more likely to be associated with lifespan (at $P < 5x10^{-8}$). This suggests that genes whose expression level is age-associated are not only a biomarker of aging, but can also causally influence lifespan.

A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits

Zheng Ning[1], *Youngjo Lee*[2], *Peter K. Joshi*[3], *James F. Wilson*[3,4], *Yudi Pawitan*[1], *Xia Shen*[1,3]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Department of Statistics, Seoul National University, Seoul, South Korea; [3]Center for Population Health Sciences, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom; [4]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom
Correspondence to: xia.shen@ed.ac.uk

In recent years, as a secondary analysis in genome-wide association studies (GWAS), conditional and joint multi-variant analysis (GCTA-COJO) has been successful in discovering additional association signals within detected loci. This suggests that many loci mapped in GWAS harbor more than a single causal variant. We develop a penalized selection operator (SOJO) within each mapped locus, based on LASSO regression derived from summary association statistics. We show that SOJO achieves better variable selection accuracy and prediction performance. Our empirical results indicate that human height is not only a highly polygenic trait, but also has high allelic heterogeneity within its established hundreds of loci.

Impact of Pathway Structures on Allelic Spectra of Diseases

George Kanoungi, *Michael Nothnagel*

Cologne Center for Genomics, University of Cologne, Cologne, Germany
Correspondence to: michael.nothnagel@uni-koeln.de

Complex diseases are frequently modeled as following an additive model. However, metabolic and signaling pathways are a widespread phenomenon in disease etiology. We explored the impact of three basic pathway motifs on the relationship between epidemiological parameters that characterize a disease, including prevalence, relative and sibling recurrence risk, causal variant number and allele frequency, by use of forward population simulations. We found that some but not all pathway motifs can shift the relationships between these parameters in comparison to the classical additive liability threshold model. The strongest deviations were observed with linear motifs that form an integral part of many reported pathways. As a real-world example, we modeled maturity-onset diabetes of the young (MODY) as a combination of different basic pathway motifs and observed a good concordance in epidemiological parameter values between our simulated data under this model and those reported in the literature. Our results support the notion that non-additive interaction modeling of genetic variants should become an additional standard approach in analyzing human genetic data.

The eQTLs Catalog and LinDA Browser: A Platform for Determining the Effects on Transcription of GWAS Variants

Mauro Pala[1], *Stefano Onano*[1,2], *Francesco Cucca*[1,2]

[1]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche Monserrato, Cagliari, Italy; [2]Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, Sassari, Italy
Correspondence to: mauro.pala@.irgb.cnr.it

The expression Quantitative Traits Loci (eQTLs) are genetic polymorphisms associated with changes in gene expression levels. They have been successfully used to prioritize the target genes of the variants associated with complex traits and diseases (GWAS variants). Up to date a few eQTLs databases exist and they collect only a small portion of the available datasets.

We thus planned to build the largest publically available catalog of eQTLs, coupled with a browser, to optimize and simplify their interrogation.

We collected and manually curated 50 eQTL public studies ranging from 2007 to date, corresponding to more than 100 sample types and 25 human populations for a total of 259,176 cis-eQTLs and 32,929 genes with at least one cis-eQTL (cis-eGenes). Most of the eQTLs studies were conducted in blood samples from healthy individuals of European ancestry. We found that for 93% of the known protein-coding genes were eGenes, 20% of them intersecting ($r^2 \geq 0.8$) with the NHGRI-EBI GWAS Catalog and 26% of whom considered as druggable. Furthermore, for those GWAS variants for which an eGene was known, we found that the NHGRI-EBI GWAS Catalog proposed the same gene as candidate target only for the 60% of the times.

Our eQTL-Catalog can be used as a reference to measure the degree of novelty for future eQTLs studies; it is provided within a platform with a web interface (LinDA) that we plan to implement with other types of quantitative traits (i.e. epigenetic, proteomic, metabolomics and microbiota) to better dissect the pleiotropy of the GWAS variants.

Comparison of the Performance of Polygene Scores and Artificial Neural Networks in the Classification of Disease Status

Carlos Pinto, *Michael Gill*, *Elizabeth Heron*

Neuropsychiatric Genetics Group, Department of Psychiatry, University of Dublin, Dublin, Ireland
Correspondence to: capinto@tcd.ie

Genome wide association studies (GWAS) can potentially be used to identify affected individuals on the basis of their genetic profile alone, even in the absence of detailed knowledge of the disease mechanism.

The current method of choice for classification is the polygene score, which measures the cumulative effect of SNPs, each SNP be-

DOI: 10.1159/000488519

ing weighted by the strength of the association of that SNP with the disease.

Genetic classification can also be viewed as a pattern recognition problem; a class of problems for which artificial neural networks (ANNs) are particularly well suited.

ANNs present challenges, however, both in terms of the tuning of the parameters of the network and in the selection of the correct architecture.

Two conditions are required for the effective use of ANNs in genetic classification; the availability of datasets with a large number of learning instances and significant computational resources. Both these requirements have been met in recent years.

We have previously compared the performance of an ANN with the polygene score on a small subset of a large schizophrenia dataset generated by the Psychiatric Genomic Consortium (PGC).

Here we present the results of a scaled up version of that study containing all the data available to us. In addition to the classification accuracy, we also consider the variation of performance with sample size and number of SNPS, the effects of optimisation of the ANN and the computational resources required.

The Exposural Landscape of Coronary Artery Disease Gives New Insight in Its Aetiology and Missing Heritability

Nicola Pirastu[1], *Peter K. Joshi*[1], *James F. Wilson*[1,2]

[1]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, Scotland; [2]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland

Correspondence to: nicola.pirastu@ed.ac.uk

Coronary Artery disease (CAD) is one of the leading causes of death world-wide. Given its importance understanding its aetiological causes is extremely important. Recently Mendellian Randomization studies have become really popular and have helped uncovering some of the factors which cause CAD. In this study, we propose to use a 3-step approach to uncovering casual relationships between CAD and its causal exposures. We first show that several common traits and disease are causally related to CAD mostly linked to blood lipid levels, including both HDL cholesterol which exhibit protective effects. Furthermore, we show that estimated effect sizes can be used to combine the exposure's polygenic risk score with that estimated using SNPs arising from the CAD GWAS. The combined score can explain 1.5% more variance of CAD than that estimated solely from the CAD GWAS SNPs, corresponding to 5.7% of explained heritability. Our results show that studying multiple potential causal exposure to CAD increases the power and accuracy of the causal inferences. Furthermore, we show that heritability is partly due to the effect of heritable causal exposures on CAD suggesting a new neglected source of missing heritability.

Mendelian Randomization Combining GWAS and eQTL Data Reveals New Loci, Extensive Pleiotropy and Genetic Determinants of Complex and Clinical Traits

Eleonora Porcu[1,2] , *Sina Rueger*[2,3], *Eqtlgen Consortium, Federico A. Santoni*[4], *Alexandre Reymond*[1], *Zoltán Kutalik*[2,3]

[1]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3]Institute of Social and Preventive Medicine, CHUV and University of Lausanne, Lausanne, Switzerland; [4]Endocrine, Diabetes, and Metabolism Service, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne 1011, Switzerland

correspondence to: Eleonora.Porcu@unil.ch

Interpretation of GWAS results is challenging, as most of the associated variants fall into regulatory regions and overlap with expression-QTLs (eQTLs), indicating their potential involvement in gene expression regulation.

To address this challenge, we propose a summary statistics-based Mendelian Randomization (MR) approach that uses multiple SNPs jointly as instruments and multiple gene expression traits as simultaneous exposures. Such an approach should be more robust to violations of MR assumptions than state-of-the-art tools (GSMR, TWAS). When applied to 43 human phenotypes it uncovered 3,233 putative genes causally associated with at least one phenotype resulting in 8,388 gene-trait associations; of note 5,982 of these loci were missed by GWAS. For example, expression of *CRIPT,* previously associated with a Mendelian syndrome with short stature (OMIM:615789), is causally associated with height in the general population. Similarly, expression of the RIDDLE syndrome-associated *RNF168* (OMIM:611943) correlates with educational attainment. Our analysis found 58% (1,866/3,233) of genes having pleiotropic causal effect, impacting up to 20 traits. Notably, *ANKRD55* showed directionally consistent causal effect on rheumatoid arthritis, Crohn's disease and inflammatory bowel disease.

Using eQTLs from multiple tissues (GTEx) revealed numerous tissue-specific causal effects, often driven by differences in gene-expression. In line with previous findings, LDL level is driven specifically by *SORT1* liver expression.

Our method identifies loci missed by conventional GWAS and pinpoints likely functionally relevant disease genes in known regions. Our findings unravel tissue-specific, causal gene-expression networks shared among a range of phenotypes. They shed light onto key biological mechanisms underlying complex clinically important traits.

Investigating Pleiotropic Architecture of Plasma Proteins Using Multivariate Methods

Linda Repetto[1], Zheng Ning[2], Andrew Bretherick[3], James F. Wilson[1,3], Xia Shen[1,2]

[1]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom; [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom
Correspondence to: linda.repetto@ed.ac.uk or xia.shen@ed.ac.uk

Motivation: Multivariate genome-wide association analyses have been proven to provide higher power than standard univariate analyses in terms of discovering novel pleiotropic loci for a range of complex traits, e.g. omics phenotypes.

Results: We measured a panel of 92 cardiovascular disease-related plasma proteins in an isolated Scottish population (ORCADES) and conducted a 92-trait combined multivariate genome scan, using summary association statistics from 92 single-trait scans. We discovered 8 novel pleiotropic pQTLs for multiple proteins which were not detectable using standard univariate analysis, nor using multivariate test directly on individual-level data. We replicate the multivariate signals and examine their pleiotropic genetic effects using an independent cohort from Croatia (VIS).

Conclusion: Our study provides new insights into shared genetic architecture across a set of human plasma proteins. The analysis pipeline emphasizes the importance of secondary multivariate analysis of summary association statistics, as one can gain substantial statistical power that is difficult to achieve even with multivariate analysis on individual-level data directly.

Genetic Modifiers of Radon Induced Lung Cancer Risk – A Genome-Wide Interaction Study in Former Uranium Miners

Albert Rosenberger[1], Heike Bickeböller[1], ILCCO/TRICL-Consortium, Maria Gomolka[2]

[1]Department of Genetic Epidemiology, University Medical Center, Georg August University Göttingen, Göttingen, Germany; [2]Unit Biological Radiation Effects, Biological Dosimetry, Department of Radiation Protection and Health, Federal Office for Radiation Protection, BfS, Neuherberg, Germany
Correspondence to: arosenb@gwdg.de

Purpose: Radon exposure is a risk factor for lung cancer, strongest in underground uranium miners. A genome-wide interaction analysis was carried out to identify genomic loci, genes or gene sets that modify the susceptibility to lung cancer given occupational exposure to the radioactive gas radon.

Methods: Samples from 28 studies provided by the International Lung Cancer Consortium were pooled with samples of former uranium miners collected by the German Federal Office of Radiation Protection. The total sample consists of 15,077 Caucasian cases and 13,522 controls. Single-marker and multi-marker models were fitted. An explorative gene-set analysis was performed in addition. The methodological challenge consisted of harmonizing inconsistent phenotype data, integrating non-representative samples, and the multiple test problem in the face of a relatively small informative number of radon-exposed lung cancer cases (n = 49 of 15,077 cases).

Results: We discovered a genome-wide significant interaction of the marker rs12440014 within the gene *CHRNB4* (OR = 0.26 95% CI: 0.11–0.60, p = 0.0386 corrected for multiple testing). Suggestive significant interaction were observed at the chromosomal regions 18q21.23 ($p = 1.2x10^{-6}$), 5q23.2 ($p = 2.5x10^{-6}$), 1q21.3 ($p = 3.2x10^{-6}$), 10p13 ($p = 1.3x10^{-5}$) and 12p12.1 ($p = 7.1x10^{-5}$). Genes belonging to term "*DNA dealkylation involved in DNA repair*" (GO:0006307; p = 0.0139) or the gene-family HGNC:476 "*microRNAs*" (p = 0.0159) were enriched with LD-block-wise significance.

Conclusion: The well-established association of the genomic region 15q25 to lung cancer might be influenced by exposure to radon among uranium miners. Further, lung cancer susceptibility is related to the functional capability of DNA damage signaling via ubiquitination processes and repair of radiation-induced double-strand breaks by the single-strand annealing mechanism.

Drugs for Specific Diseases Tend to Target Genes Whose Expression is Causally Linked to Those Diseases

Sina Rüeger[1,2], Eleonora Porcu[2,3], Zoltán Kutalik[1,2]

[1]Institute of Social and Preventive Medicine, University Hospital (CHUV), Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
Correspondence to: sina.rueger@unil.ch

While GWASs alone are limited to SNP-trait associations, in combination with gene expression QTLs via Mendelian randomization (MR) studies they have the potential to pinpoint key causal genes for diseases.

Our MR analysis provided 603'404 causal effect estimates (of which 8'388 are significantly non-zero) between whole blood expression of 15'985 genes and 43 traits/diseases), which were combined with known drug target genes using the Drugbank.ca database. For this, we developed a score to quantify for each drug-trait pair how well the drug targets correspond to the set of genes causally implicated for the given trait.

For 15 diseases (among the 43 traits) we compiled a curated list of known treatments, 309 *disease-specific drugs* in total. First, for a given disease, we compared the score of drugs known to treat that disease to that of all other drugs, and found that disease-specific drugs score significantly ($P < 0.05$) higher for 6 out of 15 diseases, with depressive symptoms and schizophrenia showing the clearest separation ($P = \mathit{8x10^{-5}}$ and $P = 6x10^{-4}$, respectively). Second, for each drug we calculated the rank of its score with the disease for which it is clinically used relative to the scores with the other 42 traits. We found that for 54 out of 309 drugs the score for the relevant disease ranked in the top five (out of 43). Notably, certain drugs for fasting insulin, HDL, LDL (Mifepristone, Saxagliptin, Niacin, Ezetimibe) ranked first.

Given these promising findings, our drug-disease score could provide guidance for repositioning existing drugs to off-target diseases.

Phenome Wide Scan of ANGPTL4 E40K Mutation Reveals New Insights for Future Drug Development

Sanni Ruotsalainen[1], *Ida Surakka*[1,4], *Veikko Salomaa*[3], *Samuli Ripatti*[1,2]

[1]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; [2]Department of Public Health, University of Helsinki, Helsinki, Finland; [3]National institute for Health and Welfare, Helsinki, Finland; [4]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA
Correspondence to: sanni.ruotsalainen@helsinki.fi

Recent publications have reported mutations that protect against cardiovascular disease which makes them interesting targets for future drug testing. In this study, we have further characterized the well-known E40K missense mutation in ANGPTL4 gene that has previously been reported to have protective effect on cardiovascular diseases, mainly by altering blood lipid levels.

We tested the association for E40K mutation with over 600 disease endpoints, 60 biochemical and physiological measurements and drug consumption in over 27,000 Finns from population-based FINRISK Study with genome-wide genotype data linked to nationwide health registries. We used logistic regression for prevalent endpoints, cox proportional hazard model for incident endpoints and linear regression for quantitative measurements. For drug use we used both Firth regression and Fisher's exact test.

In the phenome-wide scan, we were able to validate the previously reported associations with blood lipid levels (beta for TG = –0.1995, p = 2.91e^{-12}, beta for HDL cholesterol = 0.1846, p = 3.172e^{-11}), and coronary artery disease risk (OR = 0.792, p = 0.043). Surprisingly, this variant is associated with lower risk of "Any cancer" (OR = 0.729, p = 0.00035), and higher risk for incident mental and behavioral disorders due to alcohol and substance use (HR = 1.5, p = 0.01) and carriers of this variant used more drugs used in addictive disorders (OR = 1.61, p = 0.0098). Our results show how phenome-wide registry data can be used to test drug safety prior to medical testing and give us hint of possible side-effects of Angptl4 inhibitors.

Comparison and Assessment of Family- and Population-Based Genotype Imputation Methods in Large Pedigrees

Ehsan Ullah[1], *Raghvendra Mall*[1], *Mostafa M. Abbas*[1], *Khalid Kunji*[1], *Alejandro Q. Nato Jr*[2], *Halima Bensmail*[1], *Ellen M. Wijsman*[2,3], *Mohamad Saad*[1]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, USA; [3]Department of Biostatistics, University of Washington, Seattle, USA
Correspondence to msaad@hbku.edu.qa

Genotype imputation is widely used in genome-wide association studies to boost variant density, allowing increased power in association testing. Many studies currently include pedigree data due to increasing interest in rare variants coupled with the availability of appropriate analysis tools. The performance of population-based (subjects are unrelated) imputation methods is well established. However, the performance of both family-based and population-based imputation methods on family data has been subject to much less scrutiny. Here, we extensively compare several family-based and population-based imputation methods on family data of large pedigrees with both European and African ancestry. Our comparison includes many widely used population and family-based tools and another method, Ped_Pop, which combines both family- and population-based imputation results. We also compare four subject selection strategies for full sequencing to serve as the reference panel for imputation: GIGI-Pick, ExomePicks, PRIMUS, and Random selection. Moreover, we compare two imputation accuracy metrics: the Imputation Quality Score and Pearson correlation R^2 for predicting association analysis power using the imputation results. Our results show that: GIGI outperforms MERLIN, family-based imputation outperforms population-based imputation for rare variants but not for common ones, using both pedigree- and population-based imputation outperforms all imputation approaches for all minor allele frequencies, GIGI-Pick gives the best selection strategy, and R^2 provides a more useful measure of imputation accuracy for use in downstream association analysis. Our study is the first to extensively evaluate the imputation performance of many available family- and population-based tools on the same family data, and provides guidelines for future studies.

Genetic Imbalance Affects Functional Outcome After Ischemic Stroke

Kristina Schlicht[1], *Dorothea Pfeiffer*[2], *Caspar Grond-Ginsbach*[2], *Michael Krawczak*[1], *Sandra Freitag-Wolf*[1], *Didier Leys*[3], *Stephanie Debette*[4], *Alessandro Pezzini*[5], *Stefan Engelter*[6,10], *Turgut Tatlisumak*[7,11], *Steven Kittner*[8], *John Cole*[8], *Arne Lindgren*[9]

[1]Institute of Medical Statistics and Informatics, University of Kiel, Kiel, Germany; [2]Department of Neurology, University of Heidelberg, Heidelberg, Germany; [3]Department of Neurology, University of Lille, Lille, France; [4]Department of Public Health, Bordeaux University Hospital, Bordeaux, France; [5]Department of Clinical and Experimental Sciences, Neurology Clinic, University of Brescia, Brescia, Italy; [6]Department of Neurology and Stroke Center, University Hospital Basel, Basel, Switzerland; [7]Department of Neurology, Sahlgrenska University Hospital, Gothenburg, Sweden; [8]Department of Neurology, Veterans Affairs Medical Center and Department of Neurology, University of Maryland School of Medicine, Baltimore, USA; [9]Department of Clinical Sciences, Lund, Neurology, Lund University, Lund, Sweden; [10]Neurorehabilitation Unit, University Center for Medicine of Aging and Rehabilitation Basel, Felix Platter Hospital, University of Basel, Basel, Switzerland; [11]Department of Neurology, Helsinki Unviersity Central Hospital, Helsinki, Finland
Correspondence to: schlicht@medinfo.uni-kiel.de

Genetic imbalance occurs when a protein-coding gene has more or fewer copies than the two copies of a normal diploid genome. Such imbalance has been associated with various disease phenotypes. Ohnologs are genes with pronounced dose-sensitivity. We explored the impact of genetic imbalance with and without ohnologs on outcome after ischemic stroke (IS) in two independent study samples using logistic regressions. Copy number variation (CNV) was detected by PennCNV analysis of GWAS-microarrays. CNV Findings were individually validated after noise reduction [1]. Genetic imbalance was studied in IS patients with favorable (mRS 0–2) and unfavorable (mRS 3–6) outcome 3 months after stroke from the CADISP[2] study (n = 816; age = 44 ± 10 years). To validate the findings, similar analyses were performed in IS patients from seven SiGN/GISCOME[2] centers (n = 2498; age = 68 ± 14 years). Genetic imbalance was analysed as a continuous variable by counting the number of imbalanced protein-coding genes per patient. Models were adjusted for age, sex, stroke severity (NIHSS), stroke subtype (TOAST) and ancestry-informative principal components.

The number of imbalanced genes was negatively associated with favorable outcome in CADISP (p = 0.001; OR = 0.89; 95% CI = 0.82–0.95) as well as in SiGN/GISCOME (p = 0.003; OR = 0.94; 95% CI = 0.91–0.98). Upon subgroup analysis, Imbalance affecting ohnologs was negatively associated with good outcome in the SiGN/GISCOME sample (p = 0.002; OR = 0.93; 95% CI = 0.89–0.98) as well as in the CADISP sample (p = 0.002; OR = 0.88; 95% CI = 0.80–0.95). In contrast, such association was lacking for imbalances without ohnologs in both studies (SiGN/Giscome p = 0.9, CADISP p = 0.4).

References

1 http://noise-free-cnv.sourceforge.net/.

[2] Results are presented on behalf of the CADISP (Cervical Artery Dissections and Ischemic Stroke Patients), SiGN and GISCOME (Genetics of Ischaemic Stroke Functional Outcome) studies as well as the International Stroke Genetics Consortium ISGC.

Combining Surrogate Variables and Minimal Depth Variable Importance

Stephan Seifert, *Sven Gundlach*, *Silke Szymczak*

Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany
Correspondence to: seifert@medinfo.uni-kiel.de

Is has been shown that random forests can be successfully applied to omics data, such as gene expression data for classification or regression, and to select variables that are important for prediction. Currently applied variable selection techniques, however, evaluate the different variables individually and do not take into account that the complex assocations with the outcome are strongly defined by pathways and networks and, hence, the relationship between the different variables. Here we show that surrogate variables, that were introduced to compensate for missing values in the data [1], can be analyzed as proxies for variable relationships. Furthermore, we introduce a new strategy to determine variable importance incorporating surrogate variables into the concept of minimal depth [2]: Surrogate minimal depth (SMD). Applying SMD to simulated data we show that simulated correlation patterns can be recreated and that the increased consideration of variable relationships improves variable selection. Compared with existing state-of-the-art methods SMD has higher empirical power to identify causal variables while the resulting variable lists are equally stable. In conclusion, SMD is a promising approach to get more insight into the complex interplay of predictor variables and outcome in a high dimensional data setting.

References

1 Breiman L: Random Forests. Mach Learn 2001;45:5–32.
2 Ishwaran H, et al: High-Dimensional Variable Selection for Survival Data. J Am Stat Assoc 2010;105:205–217.

High-Definition Likelihood Inference for Heritability and Genetic Correlation Using GWAS Summary Statistics

Zheng Ning[1], *Yudi Pawitan*[1], *Xia Shen*[1,2]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom
Correspondence to: xia.shen@ed.ac.uk

Motivation: In the past three years, LD Score Regression has been an influential technique, allowing estimation of heritability and genetic correlation using only GWAS summary statistics. LD

Score Regression was also reported to be able to distinguish polygenicity from population structure in a GWAS study. However, the definition of LD score only uses part of the linkage disequilibrium information, thus the current LD Score Regression method has limited efficiency of estimation.

Results: In order to improve LD Score Regression as a method based on a "low-definition" likelihood (LDL) function, we develop a full high-definition likelihood (HDL) model that accounts for substantially more information in the LD structure. Simulation studies show that HDL has about 7 times efficiency compared to LD Score Regression, e.g. for genetic correlation estimation. With this, HDL is able to report accurate heritability and genetic correlation estimates with much smaller standard errors, identifying more genetically correlated complex traits. With the full likelihood method, bootstrap implementation is avoided so that computational efficiency is also improved.

Conclusion: The HDL method, as a full likelihood model of summary association statistics and LD structure, substantially improves the accuracy of heritability and genetic correlation estimation. Some other remaining issues of LD Score Regression can also be solved directly when the full likelihood is considered.

Region-Based Association Analysis of Summary Statistics Using Principal Components and Functional Linear Regression Models

Gulnara Svishcheva[1,2], *Nadezhda Belonogova*[1], *Tatiana Axenovich*[1,3]

[1]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; [2]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia; [3]Novosibirsk State University, Novosibirsk, Russia
Correspondence to: gulsvi@mail.ru

In recent years, a large number of SNP-level summary statistics (effect sizes and corresponding p-values) obtained by GWAS and meta-analyses became available in databases. It has been shown that this data can be used for gene-based association analysis. Several popular methods have been adapted to analyze the summary statistics: the score-based tests (Burden, SKAT, and SKAT-O), and the multiple linear regression test (MLR). However, methods constructed on truncated MLR models, namely the principal components regression model (PC) and functional linear regression model (FLM), have not been adapted to the association analysis using summary statistics. We solved this problem. We analytically obtained the formulas for the region-based test statistics and the parameters of their distributions. We numerically demonstrated that the results of analysis performed by our method are identical to the results obtained on individual genotype and phenotype data. Together with the summary statistics, the PC- and FLM-based methods use SNP-by-SNP correlations. Usually the researchers working with summary statistics do not have individual genotype data to calculate these correlations. In this case, reference samples of individual genotypes, for example, 1000 G, can be used for the estimation of SNP-by-SNP correlations. Such technique is widely used in all known methods of gene-based association analysis using summary statistics. We implemented the new methods in the sumFREGAT R package (https://cran.r-project.org/web/packages/sumFREGAT). The work was supported by RFBR (18-04-00076 and 16-04-00360) and FASO (0324-2018-0017).

Empirically Derived Dietary Patterns – Characteristics, Phenotypic Background, and Associations with NMR Metabolites and Health Outcomes in the Estonian Biobank Cohort

Nele Taba[1,2], *Tõnu Esko*[1], *Andres Metspalu*[1], *Krista Fischer*[1]

[1] Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Riia 23b, 51010, Estonia; [2] PhD student
Correspondence to: nele.taba@ut.ee

Dietary patterns are potentially affecting several health-related outcomes, such as Coronary Artery Disease (CAD), Type II Diabetes (T2D), metabolic profile and mortality. Better understanding of the association structures that involve dietary factors may lead to reduction in the incidence of chronic diseases and premature mortality via targeted nutrition information and educational interventions. Furthermore, detecting novel links between dietary factors and metabolic profiles can broaden the knowledge on ways to obtain and maintain a healthy metabolic profile.

The aim of this study is to cluster 49 276 adults from the population-based Estonian Biobank cohort based on their dietary pattern, and to investigate the associations of metabolic profile (using ANOVA) and health-related outcomes (using Cox proportional hazards modeling) with the dietary-clusters. We analyze the data of 17 items from the Food-Frequency Questionnaire using k-means clustering algorithm. All these items are measured on a 4-point scale indicating consumption frequency per week, or coded relevantly. The 155 NMR biomarkers were log-transformed and scaled after multiple imputation of the missing values. The NMR data was analyzed for 9248 individuals.

We show that the dietary clusters clearly differ from each other by their metabolic profile, and by the long-term health outcomes. The differences between the clusters are very evident in the citrate levels, concentration of large HDL particles, docosahexaeonic acid levels, and omega-3 fatty acids levels in general. After adjusting for gender, smoking status, and education level (using age as timescale) clusters clearly differ in their risk for T2D, CAD, CAD mortality and overall mortality.

External Evaluation of Population Pharmacokinetic for Vancomycin in Neonates with DosOpt

Tõnis Tasa[*1,2], *Riste Kalamees*[3], *Jaak Vilo*[1] , *Irja Lutsar*[2], *Tuuli Metsvaht*[3]

[1]Institute of Computer Science, University of Tartu, J. Liivi 2, 50409, Tartu, Estonia; [2]Department of Microbiology, University of Tartu, Ravila 19, 50411, Tartu, Estonia; [3]Clinic of Anaesthesiology and Intensive Care, Tartu University Hospital, L. Puusepa 8, 51014, Tartu, Estonia
Correspondence to: ttasa@ut.ee

Numerous vancomycin population pharmacokinetic (PK) models for neonates have been published. We aimed to evaluate externally these models and assess inter- and intra-model differences with variable number of individual concentrations guiding the Bayesian PK estimation.

We implemented and evaluated 12 PK models from literature and used their estimates as priors in Bayesian dose optimisation tool DosOpt. Goodness-of-fit was assessed with posterior predictive checks. Model generalisation was assessed with error metrics on predictions of patient concentrations in a retrospective dataset. We created PK model rankings based on linear mixed models on probabilities of target attainment (PTA) of trough concentration ranges of 10–15 mg/L and 10–20 mg/L.

The majority of tested models showed predictive bias on forecasts made with population model based PK. Pharmacokinetics estimation with a single input concentration improved both precision and accuracy. The best performing models attained predicted concentrations within 20% and 30% of the observed values in around 40% and 60% of cases respectively. For most models mean absolute percentage error values remained between 23%–30% and mean absolute error values between 2–2.5 mg. The best performing model was Zhao et al. (2013) On average it resulted in PTA of 40.4% (CV 0.5%) in 10–15 mg/L and 62.9% (CV 0.4%) in 10–20 mg/L target range. Second input concentration did not additionally improve PTA.

Bayesian forecasting with informative concentration inputs improved PK model based error estimates. Based on our retrospective evaluation, there were significant differences between the models. The model by Zhao et al. (2013) performed the best in PTA simulations.

Ascertainment Corrections for Secondary Phenotypes Analysis in Family Studies

Renaud Tissier[1], *Jeanine Houwing-Duistermaat*[2]

[1]Educational and Developmental Psychology, Faculty of Social Sciences, Leiden University, Leiden, The Netherlands; [2]School of Mathematics, Leeds University, Leeds, United Kingdom
Correspondence to: r.l.m.tissier@fsw.leidenuniv.nl

In numerous studies, in addition to the primary phenotype a number of additional traits, known as secondary phenotypes, are routinely recorded and modelled. Analyzing secondary phenotypes lead to biased effect estimates, especially when the covariate is also associated with the primary phenotype and when the primary and secondary phenotypes are correlated. Therefore, ascertainment corrections are needed. In family studies, we can recognize various types of ascertainment. The most common ones are the proband design, i.e. selection of families via a specific member(s), and the multiple cases design, i.e. families are recruited for having at least a certain number of cases. We previously proposed an approach for multiple cases family studies based on the retrospective likelihood and joint modelling of the primary and secondary phenotypes. Here, we will compare via mathematical formula and simulations the performance of our novel approach on family studies with the proband design. We consider the current state of the art method of conditioning the likelihood on proband values using the SOLAR-eclipse software. In addition, we will apply our methodology to family studies on Social Anxiety Disorder (SAD) and on Cardiovascular disease (Proband Design).

Our conclusion is that for secondary phenotype the conditional likelihood given the proband does not perform well, while the naive approach omitting the probands does perform well but leads to power loss. Our approach performed well in all settings.

Searching for Genetic Variants Matching a Given Multivariate Target Profile

Hande Topa[1], *Aki S. Havulinna*[1,2], *Veikko Salomaa*[2], *Mika Kähönen*[3,4], *Terho Lehtimäki*[5], *Olli T. Raitakari*[6], *Marjo-Riitta Järvelin*[7,8], *Mika Ala-Korpela*[9,10], *Peter Würtz*[11], *Matti Pirinen*[1,12]

[1]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; [2]National Institute for Health and Welfare, Helsinki, Finland; [3]The Department of Clinical Physiology, Tampere University Hospital, Tampere, Finland; [4]Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland; [5]Department of Clinical Chemistry, Fimlab Laboratories, Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland; [6]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland; [7]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK; [8]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland; [9]Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia; [10]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK; [11]Nightingale Health Ltd., Helsinki, Finland; [12]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
Correspondence to: hande.topa@helsinki.fi

Modern high-throughput technologies can provide highly informative multivariate profiles of an exposure or an intervention. An example of such a “target profile” is the change in 53 circulating metabolites associated with initiating statin therapy, as measured by a nuclear magnetic resonance (NMR) platform. We define a multivariate “candidate profile” for each genetic variant by its effect on the same 53 traits in a genome-wide association study

(GWAS). Our task is to rank the candidates (genetic variants) by their similarities to the target profile (statin effect) and that way reveal target genes of statin therapy using existing GWAS data and a statin profile.

The conventional way is to rank the variants by their correlation to the target profile by considering each trait as an independent observation. However, this approach does not account for uncertainty of the estimated effect sizes or the correlations among traits. Here, we develop a Bayesian approach which incorporates these additional sources of information when comparing two profiles. In a simulation study, we show that the Bayesian approach outperforms the correlation-based ranking especially when the number of traits is small.

We use GWAS data from Finnish cohorts (n = 10,753) at 156 lipid-associated SNPs to search for genetic variants which have similar metabolic profile to starting statin therapy. Top five variants matching with the statin profile are in/near genes *LDLR*, *APOC1*, *HMGCR*, *SORT1* and *APOB*, which is consistent with known function of statins. We conclude that our method seems a promising approach to reveal genetics related to interventions or exposures.

Coverage of the OncoArray

Viola Tozzi, *Heike Bickeböller*

Institute of Genetic Epidemiology, University Medical Center, Georg-August University of Göttingen, Göttingen, Germany
Correspondence to: viola.tozzi@med.uni-goettingen.de

The OncoArray by Illumina includes approximately 570K SNPs with a genome wide backbone of 275K tag SNPs. Additional SNPs on the chip associated with five cancers (breast, ovarian, prostate, colon-rectal and lung cancer) were identified by genome wide association studies, fine-mapping of known susceptibility regions, sequencing studies and other approaches. The aim of this study was to investigate global and local coverage of the OncoArray taking as example some regions in the genome concerning inflammatory and immunological pathways, with the related genes. The global and local coverage rates are the proportion of SNPs in the genome, or in a specific region, detected by the chip, as defined by Barret & Cardon (2006) and Li et al. (2008). We used chip data (only health individuals) from the International Lung Cancer Consortium (ILCCO) and the European reference population from the 1000 Genome Project (Version 3 April 2012, NCBI Build 37). All SNPs under analysis have MAF ≥1% and the threshold of linkage disequilibrium is $r^2 \geq 0.8$. The global coverage rate estimated for the OncoArray is 12% for all SNPs included and 10% for the backbone SNPs. Local coverage was estimated for 15 genes of interest and the range of coverage estimates varies from 11% to 82%. Regarding the local coverage for the eight inflammatory and immunological pathways taken into account, we obtained estimates between 8% and 11%. Hence the subset of genome wide backbone SNPs covers the largest part of all the coverage rates estimated for the total SNPs in the OncoArray.

Combined CNV and SNP Ancestry Analysis

Steffen Uebe, *Mandy Krumbiegel*, *Arif B. Ekici*, *André Reis*

Institute of Human Genetics, University of Erlangen-Nuremberg, Erlangen, Germany
Correspondence to: steffen.uebe@uk-erlangen.de

Principal Component Analysis of SNP data has been a mainstay of ancestry analysis in human population genetics. In this study, we combine both copy number variation (CNV) and SNP data of 2872 individuals of mixed European and American provenance, gathered using the Affymetrix CytoScan HD array. CNV genotypes were incorporated as additional information into the principal component analysis. Furthermore, the correlation of the copy number alleles with both provenance and likely ancestry (as determined from SNP genotypes) was evaluated. Finally, the performance of our ancestry model with SNP data alone and with combined SNP and CNV data was compared. While copy number variations alone are no substitute for SNP data in determining ancestry, their incorporation into a comprehensive ancestry analysis has the potential to increase the precision of the model by adding components of more recent genetic variation events than those examined by looking at SNP data alone.

Is the Healthy Migrant Effect Heritable? Population Structure and Demographic History of Resettlers and the Autochthone German Population

Maren Vens, *Heiko Becher*

Universitätsklinikum Hamburg-Eppendorf, Insitut für Medizinische Biometrie und Epidemiologie, Martinistraße 52, 20246 Hamburg, Germany
Correspondence to: m.vens@uke.de

Resettlers are ancestors of German emigrants, who settled in Eastern Europe before the 20th century. Between 1950 and 2009, more than two million resettlers migrated from the former Soviet Union (FSU) to Germany. Since the mortality rate due to cardio- and cerebrovascular diseases (CVD) is about four times higher in Russia compared to Germany, mortality rates in resettlers were expected to be higher compared to the German population.

Surprisingly, in retrospective cohort studies CVD mortality was lower among resettlers compared to the German population. This fact could neither be explained by commonly known life style risk factors nor the healthy migrant effect since almost all ethnic Germans in Russia resettled to Germany rather than just a selection of relatively healthy migrants.

We assume a genetic background contributing to a lower CVD mortality in resettlers and a healthy migrant effect for the ancestors of the resettlers migrating to the FSU. In addition, the group of ethnic Germans in the FSU can be considered as an isolate population. Thus, we hypothesize that the healthy migrant effect was inherited to the today's resettlers and that the genetic background differs to the autochthone Germans.

Studies have shown extensive migration and gene flow among European populations. We will focus on elucidation of ancient and recent population differentiation between resettlers and the autochthone German population. Therefore, we will use PCA for the analysis of population structure and analyse patterns of identity-by-descent patterns.

Beyond Burden Testing – Association Analysis of Ultra-Rare Variants

Adam Waring[1], *Martin Farrall*[1,2]

[1]Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [2]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom
Correspondence to: adam.waring@msdtc.ox.ac.uk

In the context of human genetic variation, the definition of the word rare is continually evolving. As both sequencing power and sample size increase so does the range of variation we can investigate. However, current methods for the association analysis of rare exonic variants are suboptimal for variants with MAF <0.01%. Furthermore, standard gene-based analyses consider only allele counts, either by testing an aggregated burden or the variance between sites but ignore the linear position of variants within a protein. We hypothesis that causal variants will cluster in specific contiguous regions of the gene, that correspond to specific functional domains. Here we utilise both burden and positional information in a rapid gene-based association method aimed at ultra-rare variants.

Using Monte-Carlo simulations to explore various parameter sets, we show that our test is conservative, yet has more power than six previously published methods. We then applied our method to a cohort of hypertrophic cardiomyopathy (HCM) patients sequenced at a panel of candidate genes in an association analysis using population controls from the exome aggregation database GNOMAD. In several genes where a burden signal was evident, the signal was enhanced by the simultaneous consideration of amino-acid position. In one gene without a strong burden signal (PRKAG2) we identify a significant association using the combined test. We followed up the results by exploring a non-linear modelling framework to map disease risk as a function of amino-acid position that is able to simultaneously identify protective and deleterious variant clusters.

Sex-Specific Inbreeding Depression in Humans

David W. Clark[1], *Peter K. Joshi*[1], *Tõnu Esko*[2–4], *James F. Wilson*[1,5], *on behalf of the ROHgen Consortium*

[1]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland; [2]Estonian Genome Centre, University of Tartu, Tartu, Estonia; [3]Broad Institute, Cambridge, Massachusetts, USA; [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA; [5]MRC Human Genetics Unit, University of Edinburgh, Edinburgh, Scotland
Correspondence to: jim.wilson@ed.ac.uk

In many plant and animal species the offspring of related parents suffer reduced reproductive success – a phenomenon known as inbreeding depression. In humans, the importance of this effect remains unclear, partly because reproduction between close relatives is both rare, and frequently associated with confounding cultural factors. To address these difficulties, we have performed a large-scale meta-analysis of more than one million individuals, from over 100 culturally diverse populations. In each individual, dense genotype data was used to identify autozygous genomic segments caused by both recent, and more distant, parental relatedness. We find that increased autozygosity is associated with apparently deleterious changes in 14 of 44 complex traits analysed, including components of reproductive success (age at first sex, age at first birth, number of opposite-sex partners, parity) and medically important traits associated with survival (birth weight, total cholesterol, lymphocyte percentage, haemoglobin concentration). As well as providing insight into the genetic architecture of these traits, our results have direct relevance to highly autozygous individuals. For example, we find that the offspring of first cousins are 1.6 [95% CI 1.4–1.9] times more likely to be childless than their outbred peers, apparently due to reduced fertility. Intriguingly, we also find that, for many traits, the effect of autozygosity is significantly greater in men than in women, suggesting a contribution of sexual selection to human evolution. We present evidence that the observed effects are caused by rare genetic variants and not by unknown environmental confounders. In particular, the effects of autozygosity are consistent across diverse demographic origins, and linear relationships are observed between autozygosity and trait means.

Human
Heredity

Author Index

Numbers refer to page numbers

KARGER

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