STATISTICAL METHODS FOR GENETICS & GENOMICS
- RESEARCH SEMINAR AND JOURNAL CLUB
2017-2018

TIME and PLACE:

Fall term 10am – 11am Friday (Health Sciences Room 614)
Winter term 10am – 11am Friday (Health Sciences Room TBA)
Small Group Discussion: 11am – 12 noon.

Co-ORGANIZERS:

Shelley Bull
Professor, DLSPH
Senior Scientist,
Lunenfeld-Tanenbaum Research Institute
60 Murray Street, Box 18
Room 5-226
Email: bull@lunenfeld.ca
Phone: 416-586-8245

Andrew Paterson
Professor, DLSPH
Senior Scientist,
Hospital for Sick Children Research
PGCRL 686 Bay Street,
Room 12.9710
Email: andrew.paterson@utoronto.ca
Phone: 416-813-6994

September 22 10 am – IGES Reports / Organization for Journal Club

September 29 10 am – Statistical & Computational Methods
Topic: Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome
Abstract: https://goo.gl/jzZ4h5
Presenter: Mehran Karimzadeh, Michael Hoffman, MedBiophysics & PMH

October 6 12 noon – STAGE International Speaker Seminar
Speaker: Cristen Willer, University of Michigan
http://www.stage.utoronto.ca/home/isss

Topic: Insights into Human Health & Cardiometabolic Disease Using Large-Scale Genomics

October 13 10 am – Journal Club
Presenter: Andrew Paterson, SickKids
Reading: An Expanded View of Complex Traits: From Polygenic to Omnigenic, *Boyle, Li & Pritchard (2017)*
http://www.sciencedirect.com.myaccess.library.utoronto.ca/science/article/pii/S0092867417306293

October 20 *No Seminar* * ASHG October 17-21, Orlando *

October 27 10 am – Reports from ASHG Meeting
(Z Baskurt, L Briollais, N Panjwani, D Roshandel, B Xiao)

November 3 12 noon – STAGE International Speaker Seminar
Speaker: Matthew Stephens, University of Chicago
http://www.stage.utoronto.ca/home/isss
Title: "Come join the multiple testing party!"

Two key references for the talk:
https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxw041
https://www.biorxiv.org/content/early/2017/05/09/096552

November 17 10 am – Statistical & Computational Methods

Topic: Statistical Analysis of Imaging Genetics Data

Presenter: Dehan Kong, Statistical Sciences

Biosketch: Dr. Kong received his PhD degree in statistics from North Carolina State University in 2013. He did his postdoc in biostatistics at the University of North Carolina at Chapel Hill from 2013 to 2016. His research focuses on object oriented data analysis and big data analysis. In particular, his research areas include neuroimaging data analysis, imaging genetics, functional data analysis, high dimensional data analysis and statistical machine learning. The statistical methods in these areas have a wide range of applications such as neuroimaging, genetic and genomic studies. Dr. Kong is interested in building a mutually beneficial interdisciplinary research program with those interested collaborators.

Abstract: In this talk, I will first briefly introduce my research on object oriented data analysis with application to neuroimaging. I will then talk about a detailed project on imaging genetics. In this project, we develop a high-dimensional matrix linear regression model to correlate 2D imaging responses with high-dimensional genetic covariates. We propose a fast and efficient screening procedure based on the spectral norm to deal with the case that the dimension of scalar covariates is much larger than the sample size. We develop an efficient estimation procedure based on the nuclear norm regularization, which explicitly borrows the matrix structure of coefficient matrices. We examine the finite-sample performance of our methods using simulations and large-scale imaging genetic datasets collected by the Alzheimer's Disease Neuroimaging Initiative study and the Philadelphia Neurodevelopmental Cohort.

November 24 10 am – Seminar/Journal Club

Topic: Post-selection inference following aggregate tests

Presenter: Shelley Bull, Lunenfeld-Tanenbaum

Reading: Heller, Meir, Chatterjee (2017) https://arxiv.org/pdf/1711.00497.pdf

Background Reading: Taylor & Tibshirani (2015) Statistical learning & selective inference, PNAS, 112(25), 7629-7634, 23 June 2015

http://www.pnas.org.myaccess.library.utoronto.ca/content/112/25/7629.full

December 1 12 noon – STAGE International Speaker Seminar

Speaker: Joseph H Lee, Epidemiology, Columbia University

http://www.stage.utoronto.ca/home/isss

December 8 10 am – Seminar/Journal Club

Topic: Application of Gtex

Presenter: Sara Good, SickKids

Primary Reading:

A Powerful Framework for Integrating eQTL & GWAS Summary Data, *Xu et al (2017)*

https://www.ncbi.nlm.nih.gov/pubmed/28893853

http://www.genetics.org.myaccess.library.utoronto.ca/content/207/3/893

Secondary Reading: Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues, *Wheeler et al (2016)*

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006423

December 15 10 am – Statistical & Computational Methods

Topic: X-chromosome Association on Microbiome Data

Presenters: Osvaldo Espin-Garcia, Wei Xu, Biostatistics & PMH

Background Readings:

An EM algorithm for regression analysis with incomplete covariate information, *Zhang & Rockette (2007)*

http://resolver.scholarsportal.info.myaccess.library.utoronto.ca/resolve/00949655/v77i0002/163_aeafrawici.xml

X-Chromosome Genetic Association Test Accounting for X-Inactivation, Skewed X-Inactivation, and Escape from X-Inactivation, *Wang, Yu, Shete (2014)*

http://resolver.scholarsportal.info.myaccess.library.utoronto.ca/resolve/07410395/v38i0006/483_xgatafsxaefx.xml

********** 2018 **********

* first week of classes Jan 8-12

January 12 10 am – Educational Session

Presenter: Lei Sun, Statistical Sciences

Topic: A Tutorial on Multiple Hypothesis Testing beyond the Usual Suspects

Abstract:
The multiple hypothesis testing issue is inherent in whole-genome scans. Facing a million or more p-values, we are becoming quite familiar with the concepts of controlling family-wide error rate (FWER) or the false discovery rate (FDR), as well as providing the corresponding QQ-plot. In this tutorial, I invite you to have a critical look at the your p-values. Through numerical illustration, I attempt to address the following questions:

- Can we intuitively explain that the distribution of p-values under the null is Unif (0,1)? And what is the effect of LD?
- What is the distribution of the p-values under the alternative? We know it would depend on power, but given high power, would we only see very small p-values? And conversely, given low or moderate power, do we expect to see only moderately small p-values?
- Why should a QQ-plot always be accompanied by a histogram?
- Suppose there are two sets of p-values, and set 1 contains 8% truly associated SNPs while set 2 contains only 3% signals. Can the QQ-plot for set 1 look more "flat" than the QQ-plot for set 2?
- If there were 5% truly associated SNPs in your whole-genome scan, would the "expected" genomic control lambda value be bigger than one? In other words, would the 5% signals "contaminate" the median that we use to calculate the GC lambda?

Finally, the p-value itself has some inherent pitfalls

January 19 10 am – Statistical & Computational Methods

Presenters: Laurent Briollais & Amelia Xu, Biostatistics & LTRI

Topic: A Novel Bayesian Approach for Region-Based Analysis of Next Generation Sequencing Data
Background reading: Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol. 2009 Jan;33(1):79-86.
http://onlinelibrary.wiley.com.myaccess.library.utoronto.ca/doi/10.1002/gepi.20359/full

Abstract: The discovery of rare genetic variants through Next Generation Sequencing (NGS) is becoming a very challenging issue in the field of human genetics. We propose a novel region-based statistical test based on a Bayes Factor (BF) approach to assess evidence of association between a set of rare variants located on the same genomic region and a disease outcome. Marginal likelihoods are computed under the null and alternative hypotheses assuming a binomial distribution for the rare variants count in the region and a Beta or mixture of Dirac and Beta prior distribution for the probability of rare variant in the region. The hyper-parameters are estimated empirically from the data. We derive the theoretical null distribution of the BF under our prior setting and study its statistical properties in the context of genome-wide inference. We show that a Bayesian control of the False Discovery Rate (FDR), using the BF as test statistic, can be used for genome-wide inference. We develop a simulation program, sim1000G, to simulate rare variants data similar to the 1,000 genomes sequencing data. Our simulations studies showed that the new BF statistic outperforms standard methods (SKAT, Burden test) under most situations considered. Our real data application to a lung cancer case-control study found enrichment for rare variants in novel genes. In conclusion, the use of our BF approach along with a Bayesian control of FDR offers a comprehensive framework for region-based analysis of NGS data.

January 26 10 am – Journal Club

Presenter: Shelley Bull, LTRI

Topic: GWAS & Polygenic Epidemiology

Readings: Dudbridge (2016) Polygenic epidemiology. Genet Epidemiol. 2016 May;40(4):268-72.
http://onlinelibrary.wiley.com.myaccess.library.utoronto.ca/doi/10.1002/gepi.21966/full
Visscher et al (2017) 10 years of GWAS discovery: Biology, function, & translation. AJHG 2017July;101:5-12

http://www.sciencedirect.com/science/article/pii/S0002929717302409?via%3Dihub

February 2 12 noon – STAGE International Speaker Seminar

Speaker: Julia Bailey, Epidemiology, UCLA

http://www.stage.utoronto.ca/home/isss

February 9 10 am – Journal Club

Presenter: Sangook Kim, Biostatistics & SickKids

Topic: Methods for Polygenic Scores from Summary Statistics

Reading: Mak et al (2017) Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 41:469-480.
http://onlinelibrary.wiley.com.myaccess.library.utoronto.ca/doi/10.1002/gepi.22050/full

March 2 12 noon – STAGE International Speaker Seminar

Speaker: Daniel Schaid, Biomedical Statistics and Informatics, Mayo Clinic

http://www.stage.utoronto.ca/home/isss

Title:
Statistical Fine-mapping: Methods to Move Beyond Genome-Wide Association Studies
http://www.stage.utoronto.ca/wp-content/uploads/2018/02/Daniel-Schaid-seminar.pdf

March 9 10 am – Research Seminar

Presenter: Wei Deng, Statistical Sciences

Reading: Paré, Mao, Deng (2017) A machine-learning heuristic to improve gene score prediction of polygenic traits.
Scientific Reports. 2017 Oct 4;7(1):12665. doi: 10.1038/s41598-017-13056-1. PMID: 28979001
https://www.nature.com/articles/s41598-017-13056-1

March 23 10 am – Journal Club

Presenter: Shelley Bull, Lunenfeld-Tanenbaum Research Institute

Reading: Aschard et al (2017) Covariate selection for association screening in multiphenotype genetic studies, Nature Genetics ;49:1789–1795. doi:10.1038/ng.3975
http://www.nature.com.myaccess.library.utoronto.ca/articles/ng.3975
Suppl:
https://media-nature-com.myaccess.library.utoronto.ca/original/nature-assets/ng/journal/v49/n12/extref/ng.3975-S1.pdf

April 6 12 noon – STAGE International Speaker Seminar

Speaker: Nilanjan Chatterjee, Johns Hopkins University

http://www.stage.utoronto.ca/home/isss

Title: Complex Model Building for Precision Medicine using Summary-Level Information from Big Data Sources

April 13 10 am – Journal Club

Presenter: Andrew Paterson, SickKids Research Institute

Readings:
Wijmenga & Zhernakova (2018) The importance of cohort studies in the post-GWAS era. Perspective in *Nature Genetics* 50:322–328. doi:10.1038/s41588-018-0066-3
http://www.nature.com.myaccess.library.utoronto.ca/articles/s41588-018-0066-3
Bycroft et al (2017) Genome-wide genetic data on ~500,000 UK Biobank participants
bioRxiv: **doi:** https://doi.org/10.1101/166298

https://www.biorxiv.org/content/early/2017/07/20/166298

April 20 10 am – Statistical & Computational Methods

Presenter: Emery Goossens, Purdue University

Title: Data integration via BOSS: Best Ordered-Subset Selection analysis for testing a global null hypothesis

Abstract:
Pathway and PheWAS, among other multi-dimensional genetic association studies, are forms of data integration. When only summary statistics are available, a common approach is set-based analysis which involves combining evidence across ALL analytical units (e.g. SNPs in pathway and phenotypes in PheWAS) to test the global null hypothesis of no association. When only a low percentage of variants or phenotypes are expected to be associated, however, combining evidence across subsets is a natural choice. A number of authors have investigated this approach, including Dudbridge and Koeleman (2003) on Rank Truncated Product (RTP), Chen et al (2013) on adaptive RTP, and the recent Bhattacharjee et al. (2012) on a subset-based approach. But, as reviewed in Su et al. (2016), several important analytical issues remain. In this work, we derive the theoretical adjustment factor to account for the inherent selection or 'data-dredging' bias, when all tests are independent of each other under the null. In the presence of correlation, we develop efficient Monte Carlo sampling methods. Through analytical and simulation studies, we also demonstrate that while BOSS and other adaptive data-driven approaches can increase power as compared to the set-based analysis in some settings, they are NOT guaranteed to have the highest or optimal power against all alternatives.

Reading:

Su, Y.C., Gauderman, W.J., Berhane, K., and Lewinger, J.P. (2016).
Adaptive Set-Based Methods for Association Testing.
Genet Epidemiol 40, 113-122.
https://doi-org.myaccess.library.utoronto.ca/10.1002/gepi.21950

April 27 10 am – Journal Club

Presenter: Myriam Brossard, Lunenfeld-Tanenbaum

Reading:

Yang et al (2017) A scalable Bayesian method for integrating functional information in GWAS.
American Journal of Human Genetics 101(3):404–416.
https://doi.org/10.1016/j.ajhg.2017.08.002
https://www.sciencedirect.com/science/article/pii/S0002929717303245

May 4 12 noon – STAGE International Speaker Seminar

Speaker: Andrey Rzhetsky Medicine & Human Genetics, University of Chicago

http://www.stage.utoronto.ca/home/isss

Title: Adventures in the land of complex disease etiology with large clinical datasets

May 25 10 am - Statistical & Computational Methods

Presenter: Qiang Sun, Statistical Sciences, Toronto

Title: Manifold learning for discovering change points patterns in dynamic functional brain connectivities

Abstract: In neuroscience, functional connectivity describes the connectivity between brain regions that share functional properties. It is often characterized by a time series of covariance matrices between functional measurements of distributed neuron areas. An effective statistical model for functional connectivity and its changes over time is critical for better understanding neurological diseases. To this end, we propose a log-mean model with an additive heterogeneous noise for modeling random symmetric positive definite matrices that lie in a Riemannian manifold. We introduce heterogeneous error terms to capture the non-Euclidean structure of the manifold. A scan statistic is then developed for the purpose of multiple change point detection. Theoretical results are provided. Simulation studies and an application to the Human Connectome Project lend further support to the proposed methodology. We will also explore other alternatives.

Suggestions for Background Reading:

The Statistical Analysis of fMRI Data *Lindquist. Statistical Science*, 23, Number 4 (2008), 439-464
https://projecteuclid.org/euclid.ss/1242049389
Non-euclidean Statistics for Covariance Matrices *Dryden et al. The Annals of Applied Statistics*
Vol. 3, No. 3 (Sep., 2009), pp. 1102-1123 http://www.jstor.org/stable/30242879
Meta-analysis of functional neuroimaging data: current & future directions
Wager et al. Social Cognitive and Affective Neuroscience, 2(2), 2007,150–158,
https://academic.oup.com/scan/article/2/2/150/2736775
Wikipedia page on fMRI and DFC
https://en.wikipedia.org/wiki/Dynamic_functional_connectivity

June 1 12 noon – STAGE International Speaker Seminar
Speaker: Carl Langefeld, Wake Forest University
http://www.stage.utoronto.ca/home/isss

June 8 10 am – Research Seminar
Presenter: Apostolos Dimitromanolakis, Statistical Sciences
Topic: "Computational methods for uncovering distant relatives in the 1000 Genomes project data"

* Statistical Society of Canada, June 3 - 6, Montreal *
* Canadian Human & Statistical Genetics Meeting, June 10 - 13, Harrison Hot Springs, BC *

Journal Club Suggestions
20 September 2017

Concepts, estimation and interpretation of SNP-based heritability.
Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM.
Nat Genet. 2017 Aug 30;49(9):1304-1310. doi: 10.1038/ng.3941.
PMID: 28854176

Abstract

Narrow-sense heritability (h2) is an important genetic parameter that quantifies the proportion of phenotypic variance in a trait attributable to the additive genetic variation generated by all causal variants. Estimation of h2 previously relied on closely related individuals, but recent developments allow estimation of the variance explained by all SNPs used in a genome-wide association study (GWAS) in conventionally unrelated individuals, that is, the SNP-based heritability (). In this Perspective, we discuss recently developed methods to estimate for a complex trait (and genetic correlation between traits) using individual-level or summary GWAS data. We discuss issues that could influence the accuracy of , definitions, assumptions and interpretations of the models, and pitfalls of misusing the methods and misinterpreting the models and results.

An Expanded View of Complex Traits: From Polygenic to Omnigenic
Evan A. Boyle,1,* Yang I. Li,1,* and Jonathan K. Pritchard1,2,3,*
http://dx.doi.org/10.1016/j.cell.2017.05.038

A central goal of genetics is to understand the links between genetic variation and disease. Intuitively,
one might expect disease-causing variants to cluster into key pathways that drive disease etiology. But for complex traits, association signals tend to be spread across most of the genome—including near many genes without an obvious connection to disease. We propose that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways. We refer to this hypothesis as an ‘‘omnigenic’’ model.

10 Years of GWAS Discovery: Biology, Function, and Translation.
Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, **Yang J**.
Am J Hum Genet. 2017 Jul 6;101(1):5-22. doi: 10.1016/j.ajhg.2017.06.005. Review.
PMID: 28686856

Abstract

Application of the experimental design of genome-wide association studies (GWASs) is now 10 years old (young), and here we review the remarkable range of discoveries it has facilitated in population and complex-trait genetics, the biology of diseases, and translation toward new therapeutics. We predict the likely discoveries in the next 10 years, when GWASs will be based on millions of samples with array data imputed to a large fully sequenced reference panel and on hundreds of thousands of samples with whole-genome sequencing data.

False discovery rates: a new deal
Matthew Stephens
Biostatistics, Volume 18, Issue 2, 1 April 2017, Pages 275–294,

https://doi.org/10.1093/biostatistics/kxw041
Published: 17 October 2016
Summary
We introduce a new Empirical Bayes approach for large-scale hypothesis testing, including estimating false discovery rates (FDRs), and effect sizes. This approach has two key differences from existing approaches to FDR analysis. First, it assumes that the distribution of the actual (unobserved) effects is unimodal, with a mode at 0. This "unimodal assumption" (UA), although natural in many contexts, is not usually incorporated into standard FDR analysis, and we demonstrate how incorporating it brings many benefits. Specifically, the UA facilitates efficient and robust computation—estimating the unimodal distribution involves solving a simple convex optimization problem—and enables more accurate inferences provided that it holds. Second, the method takes as its input two numbers for each test (an effect size estimate and corresponding standard error), rather than the one number usually used (*p* value or *z* score). When available, using two numbers instead of one helps account for variation in measurement precision across tests. It also facilitates estimation of effects, and unlike standard FDR methods, our approach provides interval estimates (credible regions) for each effect in addition to measures of significance. To provide a bridge between interval estimates and significance measures, we introduce the term "local false sign rate" to refer to the probability of getting the sign of an effect wrong and argue that it is a superior measure of significance than the local FDR because it is both more generally applicable and can be more robustly estimated. Our methods are implemented in an R package ashr available from http://github.com/stephens999/ashr.

Genet Epidemiol. 2017 May;41(4):320-331. doi: 10.1002/gepi.22038. Epub 2017 Apr 10.
Inclusion of biological knowledge in a Bayesian shrinkage model for joint estimation of SNP effects. Pereira M1, Thompson JR2, Weichenberger CX3, Thomas DC4, Minelli C1.
https://www.ncbi.nlm.nih.gov/pubmed/28393391
Abstract
With the aim of improving detection of novel single-nucleotide polymorphisms (SNPs) in genetic association studies, we propose a method of including prior biological information in a Bayesian shrinkage model that jointly estimates SNP effects. We assume that the SNP effects follow a normal distribution centered at zero with variance controlled by a shrinkage hyperparameter. We use biological information to define the amount of shrinkage applied on the SNP effects distribution, so that the effects of SNPs with more biological support are less shrunk toward zero, thus being more likely detected. The performance of the method was tested in a simulation study (1,000 datasets, 500 subjects with ~200 SNPs in 10 linkage disequilibrium (LD) blocks) using a continuous and a binary outcome. It was further tested in an empirical example on body mass index (continuous) and overweight (binary) in a dataset of 1,829 subjects and 2,614 SNPs from 30 blocks. Biological knowledge was retrieved using the bioinformatics tool Dintor, which queried various databases. The joint Bayesian model with inclusion of prior information outperformed the standard analysis: in the simulation study, the mean ranking of the true LD block was 2.8 for the Bayesian model versus 3.6 for the standard analysis of individual SNPs; in the empirical example, the mean ranking of the six true blocks was 8.5 versus 9.3 in the standard analysis. These results suggest that our method is more powerful than the standard analysis. We expect its performance to improve further as more biological information about SNPs becomes available.

iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis.

Huang YT, Liang L, Moffatt MF, Cookson WO, Lin X.

Genet Epidemiol. **2015** Jul;39(5):347-56. doi: 10.1002/gepi.21905. Epub **2015** May 22.

PMID: 25997986

https://www.ncbi.nlm.nih.gov/pubmed/25997986

Abstract

Genome-wide association studies (GWAS) have been a standard practice in identifying single nucleotide polymorphisms (SNPs) for disease susceptibility. We propose a new approach, termed integrative GWAS (iGWAS) that exploits the information of gene expressions to investigate the mechanisms of the association of SNPs with a disease phenotype, and to incorporate the family-based design for genetic association studies. Specifically, the relations among SNPs, gene expression, and disease are modeled within the mediation analysis framework, which allows us to disentangle the genetic effect on a disease phenotype into two parts: an effect mediated through a gene expression (mediation effect, ME) and an effect through other biological mechanisms or environment-mediated mechanisms (alternative effect, AE). We develop omnibus tests for the ME and AE that are robust to underlying true disease models. Numerical studies show that the iGWAS approach is able to facilitate discovering genetic association mechanisms, and outperforms the SNP-only method for testing genetic associations. We conduct a family-based iGWAS of childhood asthma that integrates genetic and genomic data. The iGWAS approach identifies six novel susceptibility genes (MANEA, MRPL53, LYCAT, ST8SIA4, NDFIP1, and PTCH1) using the omnibus test with false discovery rate less than 1%, whereas no gene using SNP-only analyses survives with the same cut-off. The iGWAS analyses further characterize that genetic effects of these genes are mostly mediated through their gene expressions. In summary, the iGWAS approach provides a new analytic framework to investigate the mechanism of genetic etiology, and identifies novel susceptibility genes of childhood asthma that were biologically meaningful.

Ainsworth HF, Shin S-Y, Cordell HJ. **A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements**. Genet Epidemiol. 2017;00:1–10. https://doi-org.myaccess.library.utoronto.ca/10.1002/gepi.22061

http://onlinelibrary.wiley.com.myaccess.library.utoronto.ca/doi/10.1002/gepi.22061/full

ABSTRACT

Genome wide association studies (GWAS) have been very successful over the last decade at identifying genetic variants associated with disease phenotypes. However, interpretation of the results obtained can be challenging. Incorporation of further relevant biological measurements (e.g. ‘omics’ data) measured in the same individuals for whom we have genotype and phenotype data may help us to learn more about the mechanism and pathways through which causal genetic variants affect disease. We review various methods for causal inference that can be used for assessing the relationships between genetic variables, other biological measures, and phenotypic outcome, and present a simulation study assessing the performance of the methods under different conditions. In general, the methods we considered did well at inferring the causal structure for data simulated under simple scenarios. However, the presence of an unknown and unmeasured common environmental effect could lead to spurious inferences, with the methods we considered displaying varying degrees of robustness to this confounder. The use of causal inference techniques to integrate omics and GWAS data has the potential to improve biological

understanding of the pathways leading to disease. Our study demonstrates the suitability of various methods for performing causal inference under several biologically plausible scenarios.

Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data

Eu-ahsunthornwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2, Jeronimo SMB, Blackwell JM, Cordell HJ (2014) PLOS Genetics 10(7): e1004445. https://doi.org/10.1371/journal.pgen.1004445

Abstract

Approaches based on linear mixed models (LMMs) have recently gained popularity for modelling population substructure and relatedness in genome-wide association studies. In the last few years, a bewildering variety of different LMM methods/software packages have been developed, but it is not always clear how (or indeed whether) any newly-proposed method differs from previously-proposed implementations. Here we compare the performance of several LMM approaches (and software implementations, including EMMAX, GenABEL, FaST-LMM, Mendel, GEMMA and MMM) via their application to a genome-wide association study of visceral leishmaniasis in 348 Brazilian families comprising 3626 individuals (1972 genotyped). The implementations differ in precise details of methodology implemented and through various user-chosen options such as the method and number of SNPs used to estimate the kinship (relatedness) matrix. We investigate sensitivity to these choices and the success (or otherwise) of the approaches in controlling the overall genome-wide error-rate for both real and simulated phenotypes. We compare the LMM results to those obtained using traditional family-based association tests (based on transmission of alleles within pedigrees) and to alternative approaches implemented in the software packages MQLS, ROADTRIPS and MASTOR. We find strong concordance between the results from different LMM approaches, and all are successful in controlling the genome-wide error rate (except for some approaches when applied naively to longitudinal data with many repeated measures). We also find high correlation between LMMs and alternative approaches (apart from transmission-based approaches when applied to SNPs with small or non-existent effects). We conclude that LMM approaches perform well in comparison to competing approaches. Given their strong concordance, in most applications, the choice of precise LMM implementation cannot be based on power/type I error considerations but must instead be based on considerations such as speed and ease-of-use.

Quantifying the extent to which index event biases influence large genetic association studies

Hanieh Yaghootkar Michael P. Bancks Sam E. Jones Aaron McDaid Robin Beaumont Louise Donnelly Andrew R. Wood Archie Campbell Jessica Tyrrell Lynne J. Hocking ... Show more

Hum Mol Genet (2017) 26 (5): 1018-1030. **DOI:** https://doi.org/10.1093/hmg/ddw433

Published: 30 December 2016

https://academic.oup.com/hmg/article-abstract/26/5/1018/2749608/Quantifying-the-extent-to-which-index-event-biases

Abstract

As genetic association studies increase in size to 100 000s of individuals, subtle biases may influence conclusions. One possible bias is ‘index event bias’ (IEB) that appears due to the stratification by, or enrichment for, disease status when testing associations between

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