STATISTICAL METHODS FOR GENETICS & GENOMICS
- RESEARCH SEMINAR AND JOURNAL CLUB
2019-2020

TIME and PLACE:

Fall term 10am – 12noon Friday (Health Sciences Room 614)
Winter term 10am - 12noon Friday (Health Sciences Room 614)
Seminar: 1 hour; Small Group Discussion: 1 hour.
http://www.dlsph.utoronto.ca/students/current-students/timetables/

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

PRELIMINARY SEMINAR SCHEDULE (subject to revision)

September 13 10 am – Organizational Meeting

September 20 10 am – Seminar/Journal Club – Nanwei Wang, LTRI
Topic: Linear mixed models with application in GWAS

Abstract: Due to the sample relatedness population stratification in Genetic Association Studies, simple tests will end up with high false discovery rates. Nowadays, linear mixed models are widely used to detect the significant SNPs which are correlated to some phenotype. However, researchers have found that the marginal effects of significant SNPs can't explain the total variability of phenotype. Interactions or even more complex genetic structures may contain useful information to help us understand how the SNPs affect phenotype. In this presentation, I will talk about how to test the interaction effects between SNPs and how to build the genetic network to discover more complex genetic structures.

Reading:

Lippert, Christoph, et al. "FaST linear mixed models for genome-wide association studies." Nature methods 8.10 (2011): 833. https://www-nature-com.myaccess.library.utoronto.ca/articles/nmeth.1681

Ganjgahi, Habib, et al. "Fast and powerful genome-wide association of dense genetic data with high dimensional imaging phenotypes." Nature Communications 9.1 (2018): 3254. https://www-nature-com.myaccess.library.utoronto.ca/articles/s41467-018-05444-6

Fang, Gang, et al. "Discovering genetic interactions bridging pathways in genome-wide association studies." bioRxiv (2017): 182741. https://www.biorxiv.org/content/10.1101/182741v1

October 4 10 am – Journal Club - PRS Methods– Andrew Paterson, SickKids

Reading: Janssens (2019) Human Molecular Genetics
Validity of polygenic risk scores: are we measuring what we think we are?
https://doi.org/10.1093/hmg/ddz205
https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddz205/5555564?redirectedFrom=fulltext

Abstract: Polygenic risk scores (PRS) have become the standard for quantifying genetic liability in the prediction of disease risks. PRSs are generally constructed as weighted sum scores of risk alleles using effect sizes from genome-wide association studies as their weights. The construction of PRSs is being improved with more appropriate selection of independent single nucleotide polymorphisms (SNPs) and optimized estimation of their weights, but is rarely reflected upon from a theoretical perspective, focusing on the validity of the risk score. Borrowing from psychometrics, this paper discusses the validity of PRSs and introduces the three main types of validity that are considered in the evaluation of tests and measurements: construct, content, and criterion validity. This introduction is followed by a discussion of three topics that challenge the validity of PRS, namely their claimed independence of clinical risk factors, the consequences of relaxing SNP inclusion thresholds, and the selection of SNP weights. This discussion of the validity of PRS reminds us that we need to keep questioning if weighted sums of risk alleles are measuring what we think they are in the various scenarios in which PRSs are used and that we need to keep exploring alternative modeling strategies that might better reflect the underlying biological pathways.

Background Reading: Khera et al (2018)
Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, *Nature Genetics* 50:1219-1224
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41588-018-0183-z

October 18 *No Seminar* * IGES/ASHG October 13-18, Houston *

October 25 12 noon – CANSSI Ontario STAGE International Speaker Seminar

Speaker: Mait Metspalu, Director of Institute of Genomics, University of Tartu
Title: Digging genomes for propelling medicine and understanding our past
Abstract

November 1 12 noon – CANSSI Ontario STAGE International Speaker Seminar

Speaker: Stephen Chanock, US National Cancer Institute
Title: Complexity of cancer susceptibility
https://stage.utoronto.ca/home/isss

November 8 10 am – Research Seminar/Journal Club
Topic: IGES/ASHG Highlights
Multi–speakers: Myriam Brossard, Osvaldo Espin-Garcia, Jerry Lin, Andrew Paterson

November 22 10 am – Journal Club - PRS Methods – Shelley Bull, LTRI

Reading: Mostafavi et al (2019) biorxiv paper
Variable prediction accuracy of polygenic scores within an ancestry group
https://www.biorxiv.org/content/10.1101/629949v1

Abstract: Fields as diverse as human genetics and sociology are increasingly using polygenic scores based on genome-wide association studies (GWAS) for phenotypic prediction. However, recent work has shown that polygenic scores have limited portability across groups of different genetic ancestries, restricting the contexts in which they can be used reliably and potentially creating serious inequities in future clinical applications. Using the UK Biobank data, we demonstrate that even within a single ancestry group, the prediction accuracy of polygenic scores Depends on characteristics such as the age or sex composition of the individuals in which the GWAS and the prediction were conducted, and on the GWAS Study design. Our findings highlight both the complexities of interpreting polygenic scores and underappreciated obstacles to their broad use.

Background Reading: Torkamani et al (2018)
The personal and clinical utility of polygenic risk scores.
Nature Reviews Genetics, 19(9):581.
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41576-018-0018-x

November 29 10 am – Research Seminar – Hamed Heydari, Molecular Genetics

Topic: Discovery of genetic interactions between functional modules from genotype data

Abstract: Genome-Wide Association Studies (GWAS) have been increasingly successful at identifying Single Nucleotide Polymorphisms (SNP) and gene sets associated to diseases. However, these studies have not been able to fill in the gap between the disease risks explained by the discovered loci and the estimated total heritable disease risk based on familial aggregation, a problem often referred to as "missing heritability". This problem, which is defined as the inability to explain heritability of genetic diseases and phenotypes based on mutations in a single gene, can be partially explained by interactions between genes in a functional network. Here I present a method, called BridGE, for detecting interactions between/within functional modules from the genotype data. We have applied this method to various cohorts and also data from model organism and were able to discover and replicate the results in independent cohorts. I will also talk about ongoing efforts towards functional validation of the discoveries in model organism.

Background Reading:
A unified framework for variance component estimation with summary statistics in genome-wide association studies
https://www.ncbi.nlm.nih.gov/pubmed/29515717
http://xzlab.org/papers/2017_Zhou_AOAS.pdf
Genotypic Context and Epistasis in Individuals and Populations
https://www.sciencedirect.com/science/article/pii/S0092867416308558#
Discovering genetic interactions bridging pathways in genome-wide association studies
https://www.nature.com/articles/s41467-019-12131-7

December 6 12 noon – CANSSI Ontario STAGE International Speaker Seminar

Speaker: Joseph Petrosino, Baylor College of Medicine

Title: The impact of early microbiome exposures on early life development and disease

Poster and Abstract https://stage.utoronto.ca/home/isss

********** 2020 **********

January 10 10 am – Research Seminar

Guest Speaker: Quan Long, University of Calgary

Topic: Inferring haplotypes from mixtures by integrating genomics and genetics

Abstract: Deconvoluting haplotypes, or more generally genetic subtypes, from mixtures is a frequently revisited problem in many fields. Examples include phasing parental haplotypes from genotyping data of diploid individuals, reconstructing quasispecies in viral sequencing data, as well as figuring out strain-level polymorphisms in metagenomics data. The algorithms in different fields utilize different aspects of information. Phasing algorithms rely on population genetic models to explicitly model coalescence (using MCMC) or recombination (using HMM). Tools reconstructing quasispecies utilize genomic information from sequencing reads. By in-depth analyzing the nature of the problem, we have figured out a general mathematical model that flexibly integrates evidences of both genomic sequencing reads and genetic sharing, leading to a novel tool, PoolHapX, that outperforms state-of-the-art tools in multiple fields including viruses, bacteria, humans and metagenomics. For the first time, PoolHapX enables researchers to leverage bulk-sequencing data to characterize individual haplotype-level dynamics within ecosystems, opening a new avenue to analyze within-host evolution.

Moreover, PoolHapX is enacting the promise of single-cell DNA sequencing. Although the current 10X linked reads can isolate single-cell DNA into a single drop, it suffers from the tiny amount of DNA materials therefore can only generate sparse coverage (~10Kb per drop) even if one has sufficient funds to sequence to saturation. The statistical framework of PoolHapX seamlessly compensates for this limitation of sparseness, making the envisioned future of real long-range single-molecule into reality.

As an ongoing work, we are using deep neural networks to automatically train the parameters of PoolHapX to adopt it to more fields, such as bisulfite sequencing data (for DNA methylation), Hi-C data (for 3-D genomics), etc. This automation will allow quickly adaptation of the tool to other fields without expertise into the complicated PoolHapX algorithm itself.

January 24 10 am – Journal Club - PRS Methods
Presenter: Shelley Bull, LTRI

Journal Article: Meisner et al (2019)
Case-Only Analysis of Gene-Environment Interactions Using Polygenic Risk Scores,
American Journal of Epidemiology, 188(11), 2013–2020.
https://doi-org.myaccess.library.utoronto.ca/10.1093/aje/kwz175

Abstract: Investigations of gene (*G*)-environment (*E*) interactions have led to limited findings to date, possibly due to weak effects of individual genetic variants. Polygenic risk scores (PRS), which capture the genetic susceptibility associated with a set of variants, can be a powerful tool for detecting global patterns of interaction. Motivated by the case-only method for evaluating interactions with a single variant, we propose a case-only method for the analysis of interactions with a PRS in case-control studies. Assuming the PRS and *E* are independent, we show how a linear regression of the PRS on *E* in a sample of cases can be used to efficiently estimate the interaction parameter. Furthermore, if an estimate of the mean of the PRS in the underlying population is available, the proposed method can estimate the PRS main effect. Extensions allow for PRS-*E* dependence due to associations between variants in the PRS and *E*. Simulation studies indicate the proposed method offers appreciable gains in efficiency over logistic regression and can recover much of the efficiency of a cohort study. We applied the proposed method to investigate interactions between a PRS and epidemiologic factors on breast cancer risk in the UK Biobank (United Kingdom, recruited 2006–2010).

Background Reading: Gauderman et al (2017)
Update on the state of science for analytical methods for gene-environment interactions.
American Journal of Epidemiology, 186(7), 762–770.
https://academic-oup-com.myaccess.library.utoronto.ca/aje/issue/186/7

January 31 10 am – Journal Club
Presenter: Lidija Latifovic, Epidemiology

Journal Article: Yoon et al (2018)

Efficient pathways enrichment & network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Res.; 46(10):e60
https://www.ncbi.nlm.nih.gov/pubmed/29562348
https://academic.oup.com/nar/article/46/10/e60/4942469

Abstract: Pathway-based analysis in genome-wide association study (GWAS) is being widely used to uncover novel multi-genic functional associations. Many of these pathway-based methods have been used to test the enrichment of the associated genes in the pathways, but exhibited low powers and were highly affected by free parameters. We present the novel method and software GSA-SNP2 for pathway enrichment analysis of GWAS *P*-value data. GSA-SNP2 provides high power, decent type I error control and fast computation by incorporating the random set model and SNP-count adjusted gene score. In a comparative study using simulated and real GWAS data, GSA-SNP2 exhibited high power and best prioritized gold standard positive pathways compared with six existing enrichment-based methods and two self-contained methods (alternative pathway analysis approach). Based on these results, the difference between pathway analysis approaches was investigated and the effects of the gene correlation structures on the pathway enrichment analysis were also discussed. In addition, GSA-SNP2 is able to visualize protein interaction networks within and across the significant pathways so that the user can prioritize the core subnetworks for further studies. GSA-SNP2 is freely available at https://sourceforge.net/projects/gsasnp2.

Background Reading: White et al (2019)
Strategies for pathway analysis using GWAS and WGS data.
Current Protocols in Human Genetics, 100, e79. doi: 10.1002/cphg.79
https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cphg.79

February 28 10 am – Research Seminar
Speaker: Amelia Jingxiong Xu, LTRI/SickKids

Topic: Bayes Factor Approaches for Region-Based Analysis of Rare Variants from Next Generation Sequencing Studies

Abstract: The emergence of new high-throughput genotyping technologies, such as Next Generation Sequencing (NGS), allows the study of the human genome at an unprecedented depth and scale. The discovery of germline rare variants (RVs) through NGS is a very challenging issue in the field of human genetics. Since RVs have extremely low frequencies, traditional strategies that analyze one variant at a time are underpowered for detecting associations with RVs. Gene-level or region-based statistics can provide a first step in the analysis of RVs that can lead to further experimental validation. Bayesian analysis is not well developed for RV analysis. Our goal is to develop such approaches and show their interest for germline RV analyses in the context of case-control studies. We propose a novel region-based statistical approach based on a Bayes Factor (BF) to assess evidence of association between a set of rare variants (RVs) 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 RV count in the region and a beta or mixture of Dirac and beta prior distribution for the probability of RV. We derive the theoretical null distribution of the BF under our prior setting and show that a Bayesian control of the False Discovery Rate (BFDR) can be obtained for genome wide inference. Informative priors are introduced using prior evidence of association from a Kolmogorov-Smirnov test statistic. Our simulation studies show that the new BF statistic outperforms standard methods (SKAT, SKAT-O, Burden test) in case-control studies with moderate sample sizes and is equivalent to them under large sample size scenarios. Finally, we further extend the BF approach to integrate individual-level and variant-level covariates by using a Bayesian regression approach and inference based on the Integrated Nested Laplace Approximation (INLA). Furthermore, our methodological developments are illustrated by data applications to a lung cancer case-control study seeking RV association with known and novel cancer genes.
Link to related paper: http://arxiv.org/abs/2002.08505

Background Reading:
Bayes Factors Kass and Raftery (1995)
JASA 90; June 1995, 773-795. DOI: 10.2307/2291091
https://www-jstor-org.myaccess.library.utoronto.ca/stable/2291091
Wen (2016) Robust Bayesian FDR Control using Bayes Factors, with Applications to Multi-tissue eQTL Discovery https://arxiv.org/abs/1311.3981v2

March 13 No Seminar

March 20 10 am – Research Seminar – On-line
Speaker: Livia Loureiro, SickKids

Topic: Genome-wide scans and studies of genomic evolution in autism

Abstract: Autism is a human condition that encompasses characteristics that help to define our species, namely, our ability to socially interact and demonstrate empathy, communicate non-verbally and with language, and to exhibit higher-level cognitive functioning. When an individual demonstrates impairment in two or more domains of function within these categories, and comes to medical attention, they can receive a diagnosis of autism (now formally called Autism Spectrum Disorder, or ASD). ASD is very genetic heterogenous. Already, ~100 genes and recurrent copy number variants (CNVs) have been discovered that when altered by mutation become susceptibility factors for ASD development. However, much is still to be investigated. The genomewide scans for natural selection (GWSS), in which

anomalous patterns of genetic diversity are linked to selective events, have produced a number of important results. The survey of GWSS is meant to identify loci positively selected and generate a list of a priori candidate genes based on potential selective pressures, which are used to filter the list of significant hits a posteriori. The main advantages to this type of approach is that it links the otherwise anonymous list of putative selective targets with ecological and biological information and may apply directly to the discovery of disease genes. Therefore, my working hypothesis is that 'studying genetic features in genes that demonstrate some evolutionary difference between individuals, or species, that delineate autism etiology, will lead to new understanding of this complex behavioral condition'. Using known ASD genes and new ones coming from the GWSS I am exploring how heritable genomic regions related to typically harmful mental illness may have survived the fitness-maximizing process of evolution by natural selection, and what kind of selective pressures are acting upon these variants to maintain them in a population.

Background Reading:
Oleksyk et al (2010). Phil Trans R Soc B, 365, 185-205.
Genome-wide scans for footprint of natural selection, Review.
doi: 10.1098/rstb.2009.0219
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842710/

March 27 10 am – Research Seminar – On-line
Speaker: Andrew Paterson, Sick Kids
Topic: Powerful approaches to identify likely causal rare variants for human disease & traits

Abstract: The field of human genetics has made amazing strides in the last decades. Starting from the mapping of variations to particular chromosomes, to being able to sequence the whole genomes of individuals at reasonable cost, such advances have resulted in the identification of variations that influence the risk of many rare, Mendelian diseases. In addition, genome-wide association studies (GWAS) have identified 10,000s of loci associated with the risk of common diseases and/or quantitative traits. However, challenges still remain: Many rare Mendelian diseases remain still to be solved; evidence that specific rare variants are likely causal is a major challenge, both in research and clinical settings; much of the variation in common disease risk is as yet unidentified. I will focus on study design, specifically taking advantage of family structure and multiple individuals within a family, as well as cryptic relatedness to improve the evidence that specific variants are likely causal for a disease
** last week of classes: March 30 – April 3 **

April 3 10 am – Research Seminar – On-line

Association of DNA Methylation Age with Diabetic Complications & Their Risk Factors in Type 1 Diabetes
Delnaz Roshandel, MD PhD, Bioinformatician, SickKids

Genetic Risk Association of Cardiovascular Disease in People with Type 1 Diabetes
Sareh Keshavarzi, PhD, Biostatistician, UHN

Background Reading:
Lachin JM, White NH, Hainsworth DP, Sun W, Cleary PA, Nathan DM. Effect of intensive diabetes therapy on the progression of diabetic retinopathy in patients with type 1 diabetes: 18 years of follow-up in the DCCT/EDIC. *Diabetes*. 2015;64(2):631–42.

Chen Z, Miao F, Paterson AD, Lachin JM, Zhang L, Schones DE, et al. Epigenomic profiling reveals an association between persistence of DNA methylation and metabolic memory in the DCCT/EDIC type 1 diabetes cohort. *Proc Natl Acad Sci U S A*. 2016;113(21):E3002–11.

April 10 *No Seminar* – Good Friday Holiday

April 17 10 am – Journal Club - PRS Methods – On-line
Presenter: Laurent Briollais, LTRI

Topic: Bayesian PRS methods – Ge et al (2019), Nature Communications 10:1776
Polygenic Prediction via Bayesian regression and continuous shrinkage priors
https://doi.org/10.1038/s41467-019-09718-5

Abstract: Polygenic risk scores (PRS) have shown promise in predicting human complex traits and diseases. Here, we present PRS-CS, a polygenic prediction method that infers posterior effect sizes of single nucleotide polymorphisms (SNPs) using genome-wide association summary statistics and an external linkage disequilibrium (LD) reference panel. PRS-CS utilizes a high-dimensional Bayesian regression framework, and is distinct from previous work by placing a continuous shrinkage (CS) prior on SNP effect sizes, which is robust to varying genetic architectures, provides substantial computational advantages, and enables multivariate modeling of local LD patterns. Simulation studies using data from the UK Biobank show that PRS-CS outperforms existing methods across a wide range of genetic architectures, especially when the training sample size is large. We apply PRS-CS to predict six common complex diseases and six quantitative traits in the Partners HealthCare Biobank, and further demonstrate the improvement of PRS-CS in prediction accuracy over alternative methods.

Background Reading: Vilhja´lmsson et al (2015), AJHG 97, 576-592
Modeling linkage disequilibrium increases accuracy of polygenic risk scores
https://www.cell.com/ajhg/fulltext/S0002-9297(15)00365-1

April 24 10 am – Research Seminar – On-line
Presenter: Lei Sun, Statistical Sciences & Biostatistics

Topic: INDIRECT GxG and GxE interaction analyses via tests for variance heterogeneity: a tutorial and some recent research

Abstract and Suggested Readings:
Direct GxG or GxE interaction analyses may not be desirable because of substantially increased multiple hypothesis testing burden in a genome-wide GxG association scan, or feasible because of missing data on the interacting environmental factor.

What is the idea of variance testing for indirect interaction analyses?
https://www.ncbi.nlm.nih.gov/pubmed/20585554
Pare et al. (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects. *PLoS Genetics* 6:e1000981.

How do you implement this variance test for a X-chromosomal variant? Well, we need to discuss a method that was developed for something else.
https://www.ncbi.nlm.nih.gov/pubmed/28099998
Soave and Sun (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics* 73(3):960-971.

Now let us talk about the X-chromosome
https://www.ncbi.nlm.nih.gov/pubmed/31332826
Deng et al (2019). Analytical strategies to include the X-chromosome in variance heterogeneity analyses: Evidence for trait-specific polygenic variance structure. *Genetic Epidemiology* 43(7):815-830

How about joint analyses of both G main and (indirect) GxE interaction effects?
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4572492/
Soave et al. (2015). A joint location-scale test improves power to detect associated SNPs, gene sets, and pathways. *The American Journal of Human Genetics*. 97(1): 125–138

Any pitfalls? Many: method implementation, result interpretation and assumptions, e.g.
https://www.ncbi.nlm.nih.gov/pubmed/25152454

Dudbridge and Fletcher (2014). Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics Am J Hum Genet. 95*(3):301-307.

To end on a more positive note, this analytical strategy is methodologically interesting and practically useful in identifying promising trait-associated candidates that might have been missed by the standard GWAS approach.

May 1 10 am – Journal Club – On-line
Presenter: Andrew Paterson, SickKids

Topic: NHLBI TOPMed program: 97,256 deeply sequenced genomes and imputation server

Suggested Readings:
Taliun et al., (March 06, 2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
Biorxiv, doi:10.1101/563866.
https://www.biorxiv.org/content/10.1101/563866v1
and:
https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/about

Summary: The Trans-Omics for Precision Medicine (TOPMed) program seeks to elucidate the genetic architecture and disease biology of heart, lung, blood, and sleep disorders, with the ultimate goal of improving diagnosis, treatment, and prevention. The initial phases of the program focus on whole genome sequencing of individuals with rich phenotypic data and diverse backgrounds. Here, we describe TOPMed goals and design as well as resources and early insights from the sequence data. The resources include a variant browser, a genotype imputation panel, and sharing of genomic and phenotypic data via dbGaP. In 53,581 TOPMed samples, >400 million single-nucleotide and insertion/deletion variants were detected by alignment with the reference genome. Additional novel variants are detectable through assembly of unmapped reads and customized analysis in highly variable loci. Among the >400 million variants detected, 97% have frequency <1% and 46% are singletons. These rare variants provide insights into mutational processes and recent human evolutionary history. The nearly complete catalog of genetic variation in TOPMed studies provides unique opportunities for exploring the contributions of rare and non-coding sequence variants to phenotypic variation. Furthermore, combining TOPMed haplotypes with modern imputation methods improves the power and extends the reach of nearly all genome-wide association studies to include variants down to ~0.01% in frequency.

May 8 10 am – Research Seminar – On-line
Speaker: Hamed Heydari, Molecular Genetics

Topic: Efficient multi-kernel mixed models on millions of samples

Abstract: Linear Mixed Model (LMM) has become an important method in statistical genomics toolbox for analysis of various types of genomic data. In genetics, LMM plays a major role in GWAS analysis (SNP/Gene/Gene-set), Sequence Kernel Association Test (SKAT), heritability estimation, GxG and GxE interaction tests, and genetic prediction (polygenic risk scores). Despite their importance and popularity, LMM cannot be applied to large scale genetic data due to it computational (cubic) and memory (quadratic) complexity, in sample size. While development of scalable LMM has been an active area of research over the past few year, proposed approaches sacrifice certain properties of the method (ie only single kernel, low rank approximation, bias, etc). In this presentation, I will talk about our novel approach for scaling LMM and reducing its complexity (to near linear in sample size) without losing statistical efficiency of the method. Our approach is based on Minimum Norm Quadratic Estimation (MINQE) proposed by Rao in a series of papers in 1970s. Specifically we are proposing a scalable iterative method of moments for estimating variance components. In order to avoid costly computation (inversion or a decomposition such as cholesky/eigen - cubic complexity in sample size) we utilize Conjugate Gradient along with Stochastic Trace Estimation (a Monte-Carlo approach for unbiased estimation of trace) that results in linear complexity without loss of efficiency. Moreover, our approach does not

require full kernels (no GRM) and would directly utilize the genotype data in binary format (such as plink file) ie avoiding costly computation of the Genetic Relationship Matrix.

May 15 10 am – Research Seminar – On-line

Speaker: Myriam Brossard, Lunenfeld-Tanenbaum Research Institute

Topic: Joint modelling of multiple time-to-event traits and multiple longitudinal risk factors in genetic association studies

Abstract: With the increasing availability of large population biobanks with genetic data and many longitudinal, environmental risk factors and time-to-event traits collected on each individual, there is a need for methodological development for joint modelling of these traits: to increase efficiency of SNP effect estimates, to improve power for SNP discovery and to characterize the direct &/or indirect SNP associations on these related traits. Over the last two decades, joint models of longitudinal and time-to-event outcomes have emerged as a powerful approach accounting for measurement errors & informative dropouts in longitudinal traits while flexibly modeling their latent association structures, leading to more efficient parameters estimates than conventional analysis methods (separate analysis of each trait or Cox PH model adjusted on longitudinal risk factors as time-varying covariates). Despite their increasing popularity in clinical trials and observational studies, they have received limited attention in genetic association studies.

In this talk, I will present a joint model of multiple time-to-event traits and multiple longitudinal risk factors we formulated for genetic association studies. This model consists of: (i) a linear mixed model for the multiple longitudinal traits that describes the trajectory of each trait as a function of time, SNP effects and includes subject random effects accounting for dependences within/between traits; (ii) a frailty Cox PH model for the time-to-event outcomes that depends on SNP & trajectory effects from (i) and includes a subject frailty term to account for dependence between time-to-event traits. We used two-stage inference where (i) & (ii) are fitted sequentially with a bootstrap estimate of the covariance matrix. We developed hypothesis testing methods to assess global SNP association with all (or a subset) of the traits for SNP discovery; and to characterize direct/indirect SNP association with each time-to-event trait. I will present the performance of our approach in a realistic simulation study designed to mimic the genetic architecture of Type 1 diabetes complications and illustrate results obtained from an application to the Diabetes Control and Complications Trial (DCCT) study.

May 22 10 am – Research Seminar - PRS Methods – On-line

Speaker: Matt Warkentin, Epidemiology and LTRI

Topic: Integrating Functional Annotations for Polygenic Risk in Lung Cancer: Preliminary Results

Summary: High-risk stratification based on polygenic risk and other important risk factors has shown promise for many diseases. Prioritizing high-risk individuals for lung cancer screening can improve early cancer detection and screening program efficiency. The objective of this project is to develop a lung cancer risk prediction model based on established risk factors and polygenic risk. To improve polygenic risk scoring, we aim to integrate genome-wide functional annotations using a machine learning approach. Preliminary results will be presented.

May 29 10 am – Journal Club – On-line

Presenter: Kieran Campbell, LTRI & U of Toronto

Reading: Zhu, Lei, Devlin, Roeder (2017) A unified statistical framework for single cell and bulk RNA sequencing data. https://arxiv.org/pdf/1609.08028.pdf

Abstract: Recent advances in technology have enabled the measurement of RNA levels for individual cells. Compared to traditional tissue-level bulk RNA-seq data, single cell sequencing yields valuable insights about gene expression profiles for different cell types, which is potentially critical for understanding many complex human diseases. However, developing quantitative tools for such data remains challenging because of high levels of technical noise, especially the “dropout” events. A “dropout” happens when the RNA for a gene fails to be amplified prior to sequencing, producing a “false” zero in the observed data. In this paper, we propose a

Unified RNA-Sequencing Model (URSM) for both single cell and bulk RNA-seq data, formulated as a hierarchical model. URSM borrows the strength from both data sources and carefully models the dropouts in single cell data, leading to a more accurate estimation of cell type specific gene expression profile. In addition, URSM naturally provides inference on the dropout entries in single cell data that need to be imputed for down-stream analyses, as well as the mixing proportions of different cell types in bulk samples. We adopt an empirical Bayes approach, where parameters are estimated using the EM algorithm and approximate inference is obtained by Gibbs sampling. Simulation results illustrate that URSM outperforms existing approaches both in correcting for dropouts in single cell data, as well as in deconvolving bulk samples. We also demonstrate an application to gene expression data on fetal brains, where our model successfully imputes the dropout genes and reveals cell type specific expression patterns.

June 5 12 noon – CANSSI Ontario STAGE International Speaker Seminar

Speaker: Nancy Saccone, Washington University School of Medicine, Genetics

https://stage.utoronto.ca/home/isss

Title: SNPs, Statistics, and Society: Lessons from Addiction Genetics Research

June 12 10 am – Research Seminar: **Statistical Methods - Two Presenters**

Presenter 1**: Divya Sharma, Post-doctoral Fellow, UHN Research, Toronto**

Title: Combining human and artificial intelligence: Ensemble of convolutional neural networks for disease prediction from microbiome data

Abstract: Research supports the potential use of microbiome as a predictor of some diseases. Motivated by the findings that microbiome data is complex in nature and there is an inherent correlation due to hierarchical taxonomy of microbial Operational Taxonomic Units (OTUs), we propose a novel machine learning method incorporating a stratified approach to group OTUs into phylum clusters. Convolutional Neural Networks (CNNs) were used to train within each of the clusters individually. Further, through an ensemble learning approach, features obtained from each cluster were then concatenated to improve prediction accuracy. Our two-step approach comprising of stratification prior to combining multiple CNNs, aided in capturing the relationships between OTUs sharing a phylum efficiently, as compared to using a single CNN ignoring OTU correlations. We used simulated datasets containing 168 OTUs in 200 cases and 200 controls for model testing. Thirty-two OTUs, potentially associated with risk of disease, were randomly selected and interactions between three OTUs were used to introduce non-linearity. We also implemented this novel method in two human microbiome studies: (i) cirrhosis with 118 cases, 114 controls; (ii) type 2 diabetes with 170 cases, 174 controls; to demonstrate the model's effectiveness. Extensive experimentation and comparison against conventional machine learning techniques yielded encouraging results. We obtained mean AUC values of 0.88, 0.92, 0.75, showing a consistent increment (5%, 3%, 7%) in simulations, cirrhosis and type 2 diabetes data respectively, against the next best performing method, Random Forest.

Presenter 2: **Razvan Romanescu, U of Manitoba (formerly Post-doctoral Fellow, LTRI)**

Title: Gene-based and pathway-based testing for rare-variant association in affected sib pairs

Abstract: Next generation sequencing technologies have made it possible to investigate the role of rare variants (RVs) in disease etiology. Because RVs associated with disease susceptibility tend to be enriched in families with affected individuals, study designs based on affected sib pairs (ASP) can be more powerful than case-control studies. We construct tests of RV-set association in ASPs for single genomic regions as well as for multiple regions. Single-region tests can efficiently detect a gene region harboring susceptibility variants, while multiple-region extensions are meant to capture signals dispersed across a biological pathway, potentially as a result of locus heterogeneity. Within ascertained ASPs, the test statistics contrast the frequencies of duplicate rare alleles (usually appearing on a shared haplotype) against frequencies of a single rare allele copy (appearing on a non-shared haplotype); we call these allelic parity tests. Incorporation of minor allele frequency estimates from reference populations can markedly improve test efficiency. Under various genetic penetrance models,

application of the tests in simulated ASP datasets demonstrates good type I error properties as well as power gains over approaches that regress ASP rare allele counts on sharing state, especially in small samples. We discuss robustness of the allelic parity methods to the presence of genetic linkage, misspecification of reference population allele frequencies, sequencing error and de novo mutations, and population stratification. As proof of principle, we apply single- and multiple-region tests in a motivating study dataset consisting of whole exome sequencing of sisters ascertained with early onset breast cancer.
https://onlinelibrary.wiley.com/doi/10.1002/gepi.22291

June 19 10 am – Research Seminar – On-line

Presenter: Lin Zhang, PhD candidate, Statistical Sciences

Title: Jointly testing association and Hardy-Weinberg equilibrium in case-control studies

Abstract: In a case-control association study, deviation from Hardy-Weinberg equilibrium (HWE) or Hardy-Weinberg dis-equilibrium (HWD) in the control group is usually considered as evidence for potential genotyping error, and the corresponding SNP is then removed from the study. On the other hand, assuming HWE holds in the study population, a truly associated SNP is expected to be out of HWE in the case group. Efforts have been made in combining association tests with tests of HWE in the cases to increase the power of detecting disease susceptibility loci (Song and Elston (2006), Wang and Shete (2010)). However, these existing methods are ad-hoc and sensitive to model assumptions. Utilizing the recent robust allele-based (RA) regression model for conducting allelic association tests (Zhang and Sun (2020)), here we propose a joint RA test that naturally integrates association evidence from the traditional association test and a test that evaluates the difference in HWD between the case and control groups. The proposed test is robust to genotyping error, as well as to potential HWD in the population attributed to factors that are unrelated to phenotype-genotype association. We provide the asymptotic distribution of the proposed test statistic so that it is easy to implement, and we demonstrate the accuracy and efficiency of the test through extensive simulation studies and an application.

Reading: Zhang and Sun (2020). A generalized robust allele-based genetic association test.
https://www.biorxiv.org/content/10.1101/2020.03.12.989004v1

June 26 10 am – Research Seminar - PRS Methods – On-line

Presenter: Yanyan Zhao, Post-doctoral Fellow, Statistical Sciences

Title: A stable and adaptive polygenic signal detection method based on repeated sample splitting

Abstract: Using polygenic risk scores to perform trait association analysis and disease prediction is paramount for genetic studies of complex traits. To achieve valid inference, polygenic analyses utilize sample splitting, or more recently external data, to obtain the set of genetic variants to be evaluated along with their weights. The use of external data has been popular, but recent work increasingly calls its use into question due to adverse effects of (subtle) data heterogeneity. Thus, our study here adheres to the original sampling-splitting principle but does so repeatedly to increase stability of the inference. To accommodate different polygenic structures, we develop an adaptive test under generalized linear models for testing high dimensional data. We show the asymptotic null distributions of the proposed test for both fixed and diverging number of variants. We also show the asymptotic properties of the proposed test under local alternatives, providing insights on why efficiency loss due to sample reduction associated with sample splitting can be compensated by power gain, attributed to variable selection and weighting. We support our analytical findings through extensive simulation studies and an application.