STATISTICAL METHODS FOR GENETICS & GENOMICS
- RESEARCH SEMINAR AND JOURNAL CLUB
2020-2021

TIME and PLACE:

Fall term 10am – 12noon Friday (On-line Synchronous)
Winter term 10am - 12noon Friday
Seminar: 1 hour; Small Group Discussion: 1 hour.
http://www.dlsph.utoronto.ca/students/current-students/timetables/
https://www.dlsph.utoronto.ca/course/statistical-methods-for-genetics-genomics-research-seminar-and-journal-club/

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@sickkids.ca
Phone: 416-813-6994

SEMINAR SCHEDULE

September 25 10 am – Organizational Meeting re topics & themes for the Seminar/Journal Club this academic year

October 2 12 noon – CANSSI STAGE International Speaker Seminar
Speaker: Joan E. Bailey-Wilson, NIH/NHGRI

Title: Detecting germline risk variants for complex diseases with illustrations for rare variants

Abstract: Historically, human geneticists have used both family-based and population-based study designs to detect germline genetic variants that increase risk for diseases and traits. These study design approaches have different characteristics and differ in their power to detect risk variants of different types. Population-based association studies are most powerful to detect common risk variants which tend to have small effects on risk of most diseases. Family-based linkage studies are powerful to detect regions of the genome harboring variants with moderate to large effects on risk of disease, which tend to be rare in the population. However, in the past it could be very difficult to identify the causal gene/variant within these large linked regions. I will briefly compare various study designs and discuss their utility for today's extremely dense genotype, whole exome sequencing and whole genome sequencing data. I will then present illustrations of family-based approaches for detecting rare, highly penetrant risk variants in several of my own studies.

October 16 10 am – Journal Club – Andrew Paterson, SickKids
Topic: A brief history of human disease genetics

Reading: Claussnitzer et al. Review Nature 2020 Jan;577(7789):179-189.

https://pubmed.ncbi.nlm.nih.gov/31915397/
https://www.nature.com/articles/s41586-019-1879-7

Abstract. A primary goal of human genetics is to identify DNA sequence variants that influence biomedical traits, particularly those related to the onset and progression of human disease. Over the past 25 years, progress in realizing this objective has been transformed by advances in technology, foundational genomic resources and analytical tools, and by access to vast amounts of genotype and phenotype data. Genetic discoveries have substantially improved our understanding of the mechanisms responsible for many rare and common diseases and driven development of novel preventative and therapeutic strategies. Medical innovation will increasingly focus on delivering care tailored to individual patterns of genetic predisposition.

October 23 10 am – Seminar/Journal Club - PRS Methods

Topic: **Statistical Issues in Developing and Evaluating Polygenic Risk Models**

Source: **IBC 2020 Online Learning Series** Session IS.10

Organizer & Chair: **Celia Greenwood**, **Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University**; *Session Discussant:* **Peter Kraft, Harvard T.H. Chan School of Public Health**

Speakers: **Frank Dudbridge, University of Leicester** *Evaluating risk prediction of multiple outcomes;* **Sohee Park, Yonei University** *Recent advances in individualized cancer risk prediction models in Korea*; **Alicia Martin, Massachusetts General Hospital** *Polygenic risk scores for the world: current applications, limitations, and promise*

Reading: Chatterjee N, Shi J and Garcia-Closas M.
Developing and evaluating polygenic risk prediction models for stratified disease prevention.
Nat Rev Genet 2016; 17: 392–406.
https://www-nature-com.myaccess.library.utoronto.ca/articles/nrg.2016.27
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6021129/

October 30 *No Seminar* * ASHG October 27-31 *

November 6 12 noon – CANSSI STAGE International Speaker Seminar

Speaker: Geneva I. Allen, Rice University

Title: Data Integration: Data-Driven Discovery from Diverse Data Sources

Abstract: Data integration, or the strategic analysis of multiple sources of data simultaneously, can often lead to discoveries that may be hidden in individual analyses of a single data source. In this talk, we present several new techniques for data integration of mixed, multi-view data where multiple sets of features, possibly each of a different domain, are measured for the same set of samples. This type of data is common in heathcare, biomedicine, national security, multi-senor recordings, multi-modal imaging, and online advertising, among others. In this talk, we specifically highlight how mixed graphical models and new feature selection techniques for mixed, mutli-view data allow us to explore relationships amongst features from different domains. Next, we present new frameworks for integrated principal components analysis and integrated generalized convex clustering that leverage diverse data sources to discover joint patterns amongst the samples. We apply these techniques to integrative genomic studies in cancer and neurodegenerative diseases to make scientific discoveries that would not be possible from analysis of a single data set.

November 20 10 am – Seminar/Journal Club - **Daniel Felsky, CAMH**

Title: "Linking central and peripheral inflammation in Alzheimer's disease using genetics and transcriptomics"

Abstract: Neurodegeneration due to Alzheimer's disease is the most common cause of major cognitive impairment in late life and is characterized by the presence of amyloid plaques, neurofibrillary tangles, and chronic inflammation in the brain. The brain's resident immune cells, microglia, mediate the destructive series of events leading to neuron loss. However, these cells are inaccessible in living humans. Peripheral blood monocytes, on the other hand, bear morphological and functional similarity to microglia and are directly testable. A series of published and unpublished results seeking to clarify the roles of microglia and monocytes in Alzheimer's disease will be presented, showcasing approaches to studying the integrative multi-'omic, ante- and post-mortem datasets collected on participants from the Rush University Religious Orders Study and Memory and Aging Project.

Readings:
1. Felsky et al (2019) https://www.nature.com/articles/s41467-018-08279-3
2. Bennett et al (2018) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6380522/
https://content-iospress-com.myaccess.library.utoronto.ca/articles/journal-of-alzheimers-disease/jad179939?resultNumber=0&totalResults=448&start=0&q=Bennett&resultsPageSize=10&rows=10

November 27 10 am – Seminar/Journal Club - **Angelo Canty, McMaster**

Topic: Robust Mendelian randomization

Readings:
1. Slob & Burgess (2020) A comparison of robust Mendelian Randomization methods using summary data, *Genetic Epidemiology* June 2020, 313-329 https://onlinelibrary.wiley.com/doi/10.1002/gepi.22295
2. Qi & Chatterjee (2020) A Comprehensive Evaluation of Methods for Mendelian Randomization Using Realistic Simulations and an Analysis of 38 Biomarkers for Risk of Type-2 Diabetes https://www.biorxiv.org/content/10.1101/702787v2

December 4 12 noon – CANSSI STAGE International Speaker Seminar
Speaker: Kathryn Roeder, Carnegie Mellon University

Title: Statistics and Genetics Offer a Window into Autism

Abstract: Recently the largest exome sequencing study to date of autism spectrum disorder (ASD) implicated 102 genes in risk. An innovative statistical approach was required to obtain this breakthrough for a disorder that has been unusually challenging to unravel. This risk gene set serves as a springboard for additional explorations into the etiological pathways of ASD, which can guide in the hunt for therapeutics. Quantification of gene expression, both single cell and bulk RNA-sequencing of brain tissues, can be a critical step in such investigations. We describe our statistical approaches to understand how cells develop in the brain, identifying both when and where these risk genes are primarily active. Together, our methods and results can broaden our understanding of the neurobiology of ASD.

December 11 10 am – Seminar/Journal Club – **Laurent Briollais, LTRI**

Title: Application of the Polygenic Risk Score to Longitudinal Studies: Examples based on Growth Modeling in Child Cohorts

Readings:
1) Wu et al. Exclusive breastfeeding can attenuate body-mass-index increase among geneticallysusceptible children: A longitudinal study from the ALSPAC cohort. https://pubmed.ncbi.nlm.nih.gov/32525877/.
2) Craig et al. Polygenic risk score based on weight gain trajectories is predictive of childhood obesity. https://www.biorxiv.org/content/10.1101/606277v2.full.

********** **2021** **********

January 15 10 am – Seminar/Journal Club - **Osvaldo Espin-Garcia, UHN**

Topic: Incorporating functional annotations in variant set whole-genome sequencing and polygenic risk score analyses

Abstract:
Biological function information contained across annotation databases has leveraged genomic association analyses. These annotations provide a series of variant-specific qualitative and/or quantitative functional features such as epigenetic function, evolutionary conservation, protein function and local nucleotide diversity, among others. With the increasing availability of biobank-scale datasets, computationally efficient algorithms that incorporate functional annotations are warranted. In this talk, I will review two available methods: STAAR and LDpred-funct, which are designed to integrate such biological information in variant set association studies and polygenic risk construction, respectively. On the one hand, STAAR calculates a set of multiple candidate test statistics using different annotation weights under a particular testing approach (burden, SKAT, ACAT-V) to later combine the resulting p-values using the aggregated Cauchy association test method. On the other hand, LDpred-funct builds on previous work on genome-wide polygenic risk scores and stratified LD score regression to compute functionally informed scores. This integration is achieved by modifying the prior specification of the original LDpred formulation by incorporating variant-specific heritability estimates

Background readings:
Li, X., Li, Z., Zhou, H. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat Genet 52, 969–983 (2020)
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41588-020-0676-4
Carla Márquez-Luna et al. Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv 375337
https://www.biorxiv.org/content/10.1101/375337v1
Finucane, H., Bulik-Sullivan, B., Gusev, A. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47, 1228–1235 (2015)
https://www-nature-com.myaccess.library.utoronto.ca/articles/ng.3404

January 22 10 am – Seminar/Journal Club - **Dongyang Yang , Biostats**

Topic: "Compositional Mediation Analysis for Microbiome Studies" by Sohn and Li (Annals of Applied Statistics 2019)
https://www.biorxiv.org/content/10.1101/149419v3

Background Reading:
Imai et al (Statistical Science 2010) "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects"
https://arxiv.org/abs/1011.1079;
https://projecteuclid.org/download/pdfview_1/euclid.ss/1280841733
Cho and Blaser (Nature Reviews Genetics 2012) "The human microbiome: at the interface of health and disease"
https://pubmed-ncbi-nlm-nih-gov.myaccess.library.utoronto.ca/22411464/.

January 29 No Seminar/Journal Club

February 5 10 am – Research Seminar – **Andrew Paterson, SickKids**

Title: Using GWAS arrays to do a GWAS of a molecular phenotype: Nuclear Genome-wide Associations with Mitochondrial Heteroplasmy

Abstract:
The role of the nuclear genome in maintaining the stability of the mitochondrial genome (mtDNA) is incompletely known. mtDNA sequence variants can exist in a state of heteroplasmy, which denotes the coexistence of organellar genomes with different sequences. Heteroplasmic variants that impair mitochondrial capacity cause disease and the state of heteroplasmy itself is deleterious. However mitochondrial heteroplasmy may provide an intermediate state in the emergence of novel mitochondrial haplogroups. We utilized genome-wide genotyping data from 982,072 European ancestry individuals to evaluate variation in mitochondrial heteroplasmy and to identify the regions of the nuclear genome that affect it. Age, sex and mitochondrial haplogroup were associated with the extent of heteroplasmy. GWAS identified 20 loci for heteroplasmy that exceeded genome-wide significance. This included a region overlapping mitochondrial transcription factor A (TFAM), which has multiple roles in mtDNA packaging, replication and transcription. These results show that mitochondrial heteroplasmy has a heritable nuclear component.

Reading: Main Paper and Supplementary Materials available here: https://sickkidsca-my.sharepoint.com/:f:/g/personal/andrew_paterson_sickkids_ca/EkaYJq3NZOxMrEuHBSj8D80BkjC-B3xJmxmi2WZiN6lfUQ?e=OJSqmA

February 12 10 am – Seminar/Journal Club - **Shelley Bull, LTRI**
Topic: Polygenic Risk Scores and Family-based Studies
Reading:
1) Fahed, A.C., Wang, M., Homburger, J.R. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun* **11,** 3635 (2020).
https://doi-org.myaccess.library.utoronto.ca/10.1038/s41467-020-17374-3

2) Lello, L., Raben, T.G. & Hsu, S.D.H. Sibling validation of polygenic risk scores and complex trait prediction. *Sci Rep* **10,** 13190 (2020). https://doi-org.myaccess.library.utoronto.ca/10.1038/s41598-020-69927-7

February 19 **No Seminar (Reading Week)**

February 26 10 am – Seminar/Journal Club - **Jennifer Brooks, DLSPH**
Topic: PRS for risk stratified breast cancer screening: PERSPECTIVE I&I

Reading:
Mavadatt et al (2019) Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics*, 104(1), 21-34. https://escholarship.org/uc/item/18w5121s
Lee et al (2019) BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors, *Genetics in Medicine* 21, 1708-1718.
https://www.nature.com/articles/s41436-018-0406-9
Background Reading:
Pashayan et al (2020). Personalized early detection and prevention of breast cancer: ENVISION consensus statement.*Nature Reviews Clinical Oncology* **17**, 687–705.
https://www.nature.com/articles/s41571-020-0388-9

March 5 **12 noon – CANSSI/STAGE International Speaker Seminar Series**
Speaker: Olufunmilayo I. Olopade, MD, FACP

Dr. Walter L. Palmer Distinguished Service Professor of Medicine and Human Genetics
Director, Center for Clinical Cancer Genetics & Global Health
The University of Chicago Medicine
https://canssiontario.utoronto.ca/event/canssi_ontario_stage_isss_olopade_olufunmilayo/

Title: What African Genomes Tell Us About the Origins of Breast Cancer

Abstract:
Analysis of cancer genomes has provided fundamental insights into the process of malignant transformation, and cancer genomes have rapidly become an integral part of the practice of clinical oncology, with implications for diagnosis, prognosis, treatment and prevention. Inherited and sporadic cancers often share common mutational events. Work from our group and others have defined the genomic landscape of common cancers such as breast, colon and prostate cancers. Using high throughput whole genome strategies, including genome-wide association studies, whole exome sequencing, and whole genome sequencing, we are deeply exploring the most foundational instigators of the most aggressive forms of breast cancer across the African Diaspora. To accelerate progress and promote health equity, we have embarked on innovative interventions that couple genomic analysis for risk prediction with innovative interventions to reduce the high mortality from aggressive young onset cancers in low resource settings in the US and Nigeria. I will present our recent findings and future directions for genetic epidemiology research in underserved and understudied African ancestry populations.

March 12 10 am – Journal Club – **Alexandra Bushby, PhD Student, DLSPH**
Topic: Polygenic transmission disequilibrium

Reading: Weiner et al (2017) Nature Genetics 49(7), 978-985
https://pubmed.ncbi.nlm.nih.gov/28504703/
https://www-nature-com.myaccess.library.utoronto.ca/articles/ng.3863

March 19 10 am – Journal Club – **Abhinav Thakral , MSc Student, DLSPH**
Topic: Horizontal pleiotropy and Mendelian randomization

Reading: Verbanck et al (2018) *Nature Genetics* 50(5), 693-698
Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases
https://pubmed.ncbi.nlm.nih.gov/29686387/
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41588-018-0099-7

Background Reading: Davies et al (2018). *BMJ* 2018;362:k601
Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians
https://doi.org/10.1136/bmj.k601
https://www.bmj.com/content/362/bmj.k601

March 26 No Seminar/Journal Club

April 2 No Seminar/Journal Club - Good Friday Holiday

April 9 10 am – Journal Club – **Shelley Bull, LTRI**
Topic: Exome Sequencing in UK Biobank

Readings:
Jurgens et al (2020) *Rare genetic variation underlying human diseases and traits: Results from 200,000 individuals in the UK Biobank*
https://www.biorxiv.org/content/10.1101/2020.11.29.402495v1.full
Wang et al (2020) *Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants*
https://www.biorxiv.org/content/10.1101/2020.12.13.422582v1.full

Background Reading: Van Hout et al (2018). *Nature* 2020;586:749-756
Exome sequencing and characterization of 49,960 individuals in the UK Biobank
https://pubmed.ncbi.nlm.nih.gov/33087929/
https://www.nature.com/articles/s41586-020-2853-0

April 16 10 am – PRS Research Seminar – **Linbo Wang, Statistical Sciences**

Topic: Ultrahigh Dimensional Learning of Polygenic Risk Scores for Mendelian Randomization Studies

Abstract: Mendelian randomization is a method by which genetic variants are leveraged as instrumental variables to investigate causal relationships between modifiable exposure or risk factor and a clinically relevant outcome from observational data. One key step is to identify valid instrumental variables among all genetic variants. Current methods work well when the number of variants is of moderate size. However, for the identification of valid IVs from ultrahigh dimensional genetic variants, empirical evidence implies that existing procedures may miss many or even all the valid instrumental variables, due to the inclusion of irrelevant variables which have nonignorable sample correlation with the exposure. To overcome this challenge, we propose a novel approach to remove irrelevant variables from candidate instruments and apply existing work to the remaining set for causal effect estimation. Extensive simulation studies demonstrate that the proposed procedure outperforms existing methods under ultrahigh-dimensional setting. The advantages of the new approach are also illustrated through analysis of the causal effect of Tau protein on Alzheimer's disease.

Background Reading: Guo, Z., Kang, H., Cai, T. T., & Small, D. S. (2018). Confidence Interval for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4), 793-815. https://rss-onlinelibrary-wiley-com.myaccess.library.utoronto.ca/doi/10.1111/rssb.12275

April 23 No Seminar

April 30 10 am – Research Seminar – **Gengming He, Biostatistics/Sick Kids**

Topic: Comparing gene expression across paired human airway models for cystic fibrosis precision medicine

Abstract: Cultured human bronchial epithelia (HBE) are the gold standard model for assessing the efficacy of small molecule therapies in cystic fibrosis (CF). However, these cells are difficult to access, especially in young children. Human nasal epithelia (HNE) are used as a surrogate model although it is unknown whether HNE recapitulate the gene expression properties of HBE. To investigate the similarities and differences in the transcriptome between HBE and HNE with a focus on modifier genes of CF lung disease identified by genome-wide association studies, RNA-sequencing was conducted on paired cultured and fresh HNE and HBE (n=71 samples) collected from 21 patients with CF who underwent lung transplantation. Gene expression was first compared across cultured and fresh samples to assess the impact of the culturing process, then compared between cultured HNE and HBE. Co-expression relationships of CF modifier genes were compared between cultured HNE and HBE. The culturing process had little impact on the expression level of CF lung disease modifier genes. These genes also showed significant equivalent expression between cultured HNE and HBE. Their co-expression relationships in cultured HNE and HBE overlapped significantly, suggesting the corresponding biological processes are consistent across the two tissues. CF lung disease modifier genes have similar expression profiles across cultured HNE and HBE. This supports the use of HNE as a surrogate airway model to investigate CF lung disease, modifier genes and enable investigation of CF precision medicine.

May 7 **12 noon – CANSSI Ontario STAGE International Speaker Seminar Series**

Clare Turnbull, Professor of Translational Cancer Genetics,
Institute of Cancer Research, UK

Breast Cancer Susceptibility Genetics: expansion of testing and emerging challenges

As optimisim has waned that 'precision-oncology' will be panacea for decreasing the bulk cancer-related mortality associated with most common solid tumors, attention has re-focused on cancer prevention and

early detection. Any such strategy is likely a priori to have more impact if targeted to those at the highest risk of developing cancer, namely the genetically predisposed. With recent technology advances facilitating cheap high throughput sequencing, and massive political imperative to leverage 'the power of genomics' to rescue us from the rocketing cost of healthcare, pressure for delivery of results in cancer susceptibility genetics is unprecedented. I shall discuss leverageable opportunities for expansion of polygenic and monogenic testing for breast cancer genetic susceptibility along with the challenges of complex genomic architecture, risk and variant pathogenicity.

May 14 *No Seminar*

May 21 10 am – Research Seminar – **Sarah Gagliano, Universite de Montreal**
Topic: Leveraging dense genotype imputation for disease-associated rare variant discovery

Abstract: The TOPMed genotype imputation reference panel consists of 97K multi-ethnic deep (average 38X) whole-genome sequences. The panel contains 308 million genetic variants (SNPs and short insertions and deletions) of which 94% are very rare (alternate allele frequency <0.5%). To date, it is the largest publicly available deep whole-genome sequencing panel for genotype imputation. To illustrate the potential for rare variant discovery, we imputed genotypes in the UK Biobank using the TOPMed panel. Specifically, (1) we assessed the concordance between imputed genotype calls to calls from whole-exome sequencing, and (2) we conducted association analyses of putative loss-of-function variants in the TOPMed-imputed UK Biobank using single-variant and gene burden tests.
Background Reading: Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program**,** *Nature* **590**, 290–299 (2021)
https://www.nature.com/articles/s41586-021-03205-y

May 28 10 am – Research Seminar – **Jun Young Park, Statistical Sciences**
Topic: Bidimensional Linked Matrix Decomposition for Pan-Omics Pan-Cancer Analysis

Abstract: Several modern applications require the integration of multiple large data matrices that have shared rows and/or columns. For example, cancer studies that integrate multiple omics platforms across multiple types of cancer, pan-omics pan-cancer analysis have extended our knowledge of molecular heterogenity beyond what was observed in single tumor and single platform studies. However, these studies have been limited by available statistical methodology. We propose a flexible approach to the simultaneous factorization and decomposition of variation across such bidimensionally linked matrices, BIDIFAC+. This decomposes variation into a series of low-rank components that may be shared across any number of row sets (e.g., omics platforms) or column sets (e.g., cancer types). This builds on a growing literature for the factorization and decomposition of linked matrices, which has primarily focused on multiple matrices that are linked in one dimension (rows or columns) only. Our objective function extends nuclear norm penalization, is motivated by random matrix theory, gives an identifiable decomposition under relatively mild conditions, and can be shown to give the mode of a Bayesian posterior distribution. We apply BIDIFAC+ to pan-omics pan-cancer data from TCGA, identifying shared and specific modes of variability across 4 different omics platforms and 29 different cancer types. https://arxiv.org/abs/2002.02601

Background Reading: Park & Lock (2019) https://onlinelibrary-wiley-com.myaccess.library.utoronto.ca/doi/10.1111/biom.13141

June 4 **12 noon – CANSSI Ontario STAGE International Speaker Seminar Series**

Heather J. Cordell, Professor of Statistical Genetics, Newcastle University, UK

A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships

Bayesian networks can be used to identify possible causal relationships between variables based on their conditional dependencies and independencies, which can be particularly useful in complex biological scenarios with many measured variables. When there is missing data, the standard approach is to remove every individual with any missing data before performing analysis. This can be wasteful and undesirable when there are many individuals with missing data, perhaps with only one or a few variables missing. This motivates the use of imputation. We present a new imputation method designed to increase the power to detect causal relationships, where the data may include a mixture of both discrete and continuous variables. Our method uses a version of nearest neighbour imputation, whereby missing data from one individual is replaced with data from another individual, their nearest neighbour. For each individual with missing data, the subsets of variables to be used to select the nearest neighbour are chosen by sampling without replacement the complete data and estimating a best fit Bayesian network. We show that this approach leads to marked improvements in the recall and precision of directed edges in the final network identified. We illustrate the relationship between methylation and gene expression in early inflammatory arthritis patients.

June 11 10 am – Research Seminar – **Divya Sharma, UHN**

Topic: Machine learning methodologies in microbiome based disease prediction

Abstract: Microbiome inherently is dynamic in nature, attributing to the presence of interactions among microbes, microbes and the host, and with the environment. Researchers have shown that the microbiome can be altered over time, either transiently or long term, by infections or medical interventions such as antibiotics. In this presentation, I will be discussing disease prediction using longitudinal microbiome data and will shed light on how advanced neural networks such as Convolutional Neural Networks (CNNs) and Long Short Term Memory networks (LSTMs) can be used for feature extraction and temporal dependency analysis in longitudinal microbiome data. I will further discuss how stratification of microbiome data based on taxonomic information, followed by application of CNNs, aids in capturing the relationships between OTUs efficiently, and how the LSTMs help in mitigating challenges associated with longitudinal data such as, missing time point information and variable sequence lengths. Finally, I will demonstrate the proposed model's effectiveness on two real longitudinal human microbiome studies and discuss future scope of the analysis.

Background reading:

1. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction, https://www.frontiersin.org/articles/10.3389/fgene.2019.00579/full
2. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977
3. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction, https://academic.oup.com/bioinformatics/article/36/17/4544/5843784
4. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2833-2

June 18 10 am – Journal Club – **Jerry Lin, Biostatistics**

Topic: SNPnet ("*A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank*")

Reading: Qian, Tanigawa, Du, Aguirre, Chang Tibshirani, et al. (2020) PLoS Genet 16(10): e1009141. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009141

June 25 10 am – Research Seminar – **Sangook Kim, Biostatistics**
Topic: A robust genome-wide association test leveraging latent genetic interactions: Application to cystic fibrosis lung disease

Abstract: For complex traits such as lung disease in cystic fibrosis (CF), Gene x Gene or Gene x Environment interactions can impact disease severity, but these remain largely unknown and/or unmeasured. Unaccounted-for genetic interactions introduce heterogeneity in the variance of the quantitative trait across the genotypic groups; thus, a GWAS using variance testing can identify variants putatively involved in genetic interactions. Joint tests of both mean and variance (or full distributional differences) across genotype groups can account for unknown genetic interactions and increase power for gene identification. Parametric joint location and scale tests that assume a Gaussian trait are accessible for GWAS, but under departures from normality these can suffer from type I error inflation or be inefficient after data transformation. Aschard et al. (2013) proposed a variant of the Kolmogorov-Smirnov test but its computational requirements limit implementation genome-wide. Here, we overcome previous limitations by developing a quantile regression-based JLS method (qJLS) without any distributional assumptions; it is robust to outliers, and computationally efficient for GWAS. A simulation study compares the statistical power of qJLS to other methods in the literature, investigating the effect of varying interaction magnitudes and phenotypic distributions including skewed and heavy-tailed data. We apply the qJLS test in a GWAS of CF lung disease in the Canadian CF Gene Modifier consortium.

Background Reading:
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000981
https://www.cell.com/ajhg/fulltext/S0002-9297(15)00201-3

July 2 No Seminar

July 9 10 am – Journal Club – **Boxi Lin, Biostatistics**
Topic: *Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies*

Citation: McCaw ZR, Lane JM, Saxena R, Redline S, Lin X. Biometrics. 2020 Dec;76(4):1262-1272. doi: 10.1111/biom.13214. Epub 2020 Jan 13. PMID: 31883270.
https://onlinelibrary-wiley-com.myaccess.library.utoronto.ca/doi/10.1111/biom.13214
https://pubmed.ncbi.nlm.nih.gov/31883270/

Abstract: Quantitative traits analyzed in Genome-Wide Association Studies (GWAS) are often nonnormally distributed. For such traits, association tests based on standard linear regression are subject to reduced power and inflated type I error in finite samples. Applying the rank-based inverse normal transformation (INT) to nonnormally distributed traits has become common practice in GWAS. However, the different variations on INT-based association testing have not been formally defined, and guidance is lacking on when to use which approach. In this paper, we formally define and systematically compare the direct (D-INT) and indirect (I-INT) INT-based association tests. We discuss their assumptions, underlying generative models, and connections. We demonstrate that the relative powers of D-INT and I-INT depend on the underlying data generating process. Since neither approach is uniformly most powerful, we combine them into an adaptive omnibus test (O-INT). O-INT is robust to model misspecification, protects the type I

error, and is well powered against a wide range of nonnormally distributed traits. Extensive simulations were conducted to examine the finite sample operating characteristics of these tests. Our results demonstrate that, for nonnormally distributed traits, INT-based tests outperform the standard untransformed association test, both in terms of power and type I error rate control. We apply the proposed methods to GWAS of spirometry traits in the UK Biobank. O-INT has been implemented in the R package RNOmni, which is available on CRAN.