STATISTICAL METHODS FOR GENETICS & GENOMICS
- RESEARCH SEMINAR AND JOURNAL CLUB 2021-2022

TIME and PLACE:
10am – 12noon Friday (On-line Zoom – details distributed by weekly email)
Seminar: 1 hour starting at 10:00am
Small Group Discussion: up to 1 hour starting at 11:00am.

Co-organizers:
Shelley Bull, bull@lunenfeld.ca Lunenfeld-Tanenbaum Research Institute & Dalla Lana School of Public Health
Andrew Paterson, andrew.paterson@sickkids.ca SickKids Research Institute & Dalla Lana School of Public Health

To be added to the e-distribution list:
please email Teresa MacKinnon <mackinnon@lunenfeld.ca>

SEMINAR SCHEDULE

September 17 10 am – Organizational Meeting re topics & themes for the Seminar/Journal Club this academic year

September 24 10 am – Research Seminar – Delnaz Roshandel, SickKids

Topic: A cystic fibrosis lung disease modifier locus harbors tandem repeats associated with gene expression

Introduction: The largest genome-wide association study (GWAS) of cystic fibrosis (CF) lung disease identified two independent signals on chromosome 5 (Chr5:403,462-686,129) in 5' and 3' of a previously reported CF modifier gene, *SLC9A3*. The locus displays a high density of CpG islands and variable number of tandem repeats (VNTRs). The exact boundaries of these VNTRs have not been well defined as they often expand hundreds of bps difficult to capture by short-read sequencing.
Methods: We used long-read PacBio phased data and multiple alignment to identify the boundaries of common (> 2%) VNTRs in the region (N = 22). Association between the identified VNTRs and gene expression in the region (*AHRR*, *EXOC3*, *SLC9A3*, *CEP72* & *TPPP*) was then investigated using RNA-seq of CF nasal epithelia (NE; N = 46). Subsequently, the lengths of VNTRs were estimated in 10X Genomics (10XG) linked-read technology by dividing the number of reads aligned to the location of each VNTR by the average sequencing depth: we confirmed a high correlation between estimated lengths from short-read (10XG) sequencing and lengths from long-read (PacBio) sequencing in 53 subjects with both 10XG and PacBio. Therefore, the same strategy was used to estimate the VNTR lengths using short-read sequencing from the Genotype-Tissue Expression (GTEx) to investigate VNTR associations with gene expression in all 49 GTEx tissues.
Results: At the locus, 54 VNTRs were identified. A VNTR in the last intron of *SLC9A3* overlapping a CpG island was associated with *SLC9A3* expression in NE (p = 5E-4) and forty GTEx tissues including lung (p = 8E-15). This VNTR was partially tagged by rs72711364, the top associated GWAS SNP 3' of *SLC9A3* (Spearman correlation coefficient = 0.28). Its repeat sequence was ≈100bp including 7 CpGs. Subjects had 2-10 copies of the repeat (mean = 5) per haplotype. Another VNTR in the 3' UTR of *TPPP* also overlapping a CpG island was associated with both *TPPP* (p = 9E-5) and *SLC9A3* (p = 2E-3) expression in NE, and multiple GTEx tissues. This VNTR had a 31 bp repeat sequence including 1-4 CpGs and multiple SNPs some changing CpG count. Subjects had 3-21 copies of the repeat per haplotype (mean = 9). The long/short version of this VNTR was perfectly tagged by C/T alleles of rs72703083, a GWAS SNP 5' of *SLC9A3*. These two VNTRs together explained 22% and 9% of variation in *SLC9A3* expression in CF NE and GTEx lung, respectively.
Conclusion: We used long-read sequencing to identify the precise boundaries of common VNTRs at the *SLC9A3* locus. Two of these VNTRs, tagged by genome-wide associated CF lung disease SNPs from two independent GWAS peaks in the region, account for almost a quarter of *SLC9A3* expression in the NE model of CF airway.

Background Reading: Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis, *Nat Commun*. 2015 Sep 29;6:8382. https://pubmed.ncbi.nlm.nih.gov/26417704/

October 1 12 noon – CANSSI STAGE International Speaker Seminar

Speaker: Benjamin Neale
Director, Genomics of Public Health Initiative, Analytic and Translational Genetics Unit,
Massachusetts General Hospital; Associate Professor, Harvard Medical School
Institute Member and Director, Genetics, Stanley Center for Psychiatric Disease, Broad Institute
https://canssiontario.utoronto.ca/event/stage_isss_ben_neale/

Talk Title: Genetic methods for biology and epidemiology

Abstract: Dr. Neale will provide an overview of LD score and the various methods and extensions to estimate heritability and extract insights from genetic datasets, including estimation of genetic correlation, boosting prediction performance using correlated traits and partitioning heritability to develop biological hypotheses.

October 8, 15, 22 *No Seminars* ** IGES October 11-15 ** ASHG October 17-22 **

October 29 10 am – Research Seminar/Journal Club – Jianhui Gao, Biostatistics

Topic: Current methods integrating variant functional annotation scores have limited capacity to improve the power of genome-wide association studies

Abstract: Functional annotations have the potential to increase the power of genome-wide association studies (GWAS) by prioritizing variants according to their biological function. Focusing on variant-specific annotation meta-scores including CADD (Kircher et al., 2014) and Eigen (Ionita-laza et al., 2016), we broadly examined GWAS summary statistics of 1,132 traits from the UK Biobank (Sudlow et al., 2015) using the weighted p-value approach (Genovese et al., 2006) and stratified false discovery control (sFDR) method (Sun et al., 2006). These 1,132 traits were rated by Benjamin Neale's lab, Broad Institute as having medium to high confidence for their heritability estimates.

Averaged across the 1,132 UK Biobank traits, sFDR was more robust to uninformative meta-scores, but the weighted p-value method identified more variants using CADD or Eigen, based on performance measures that included type I error control, recall, precision, and relative efficiency. Our application results were consistent with those from an extensive simulation study using three different designs, including leveraging the real genetic data combined with simulated genomic data and vice versa.

We also considered the recent FINDOR method (Kichaev et al., 2019), which leverages a set of individual 75 functional annotations into GWAS. An earlier application of FINDOR to 27 traits selected from the z7 category (SNP-heritability p-value < 1.27×10^{-12} by Nealelab) detected 13%-20% additional genome-wide significant loci as compared to the standard annotation-free GWAS, which we confirmed. Moreover, across all 438 traits in the z7 category, 46,631 out 59,764 (80%) significant loci discovered are common across the three data-integration methods.

However, across all the 1,132 UK Biobank traits examined, the median [Q1,Q3] of the total numbers of new, genome-wide significant independent loci were 0 [0, 3] by FINDOR, 0 [0, 2] by weighted p-value, and 0 [0, 0] by sFDR. Notably, 162 traits (89%) in the nonsig trait category (SNP-heritability p-value > 0.05, "likely reflecting limited statistical power rather than a true lack of heritability" by Nealelab) had no new discoveries after data-integration by any of the three methods. Our findings suggest that more informative scores or new data integration methods are warranted to further improve the power of GWAS by leveraging the variant functional annotations.

Background Reading: Leveraging Polygenic Functional Enrichment to Improve GWAS Power, *AJHG*, 3 January 2019, 104(1):65-75. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323418/
`` https://www.biorxiv.org/content/10.1101/2020.11.25.396721v2

November 5 12 noon – CANSSI STAGE International Speaker Seminar

Speaker: Jordana Bell, King's College London
Senior Lecturer, Head of Epigenomics Research Group;
Department of Twin Research & Genetic Epidemiology; Faculty of Life Sciences & Medicine
https://canssiontario.utoronto.ca/event/stage_isss_jordana_bell/

Talk Title: Genetic impacts on human methylome variation

Abstract: Studies of the human methylome have identified several drivers of DNA methylation variability. Over the past decade multiple efforts have found that a proportion of the human methylome is under genetic control. I will present applications of twin studies and population cohort-based study designs to identify genetic impacts on the human blood methylome. An overview of the extent of human methylome variation that is estimated to be under genetic control will be presented, based on recent large-scale datasets profiled on the Illumina Infinium HumanMethylation450 and MethylationEPIC BeadChips. Further analyses exploring the relationship between genetic influences on DNA methylation and human phenotypes has identified a series of co-localisation events, giving some insights into how these signals relate to human health. Identifying genetic impacts on DNA methylation can improve our understanding of pathways that underlie gene regulation and disease risk.

November 19 10 am – Journal Club – Andrew Paterson, Sickkids

Reading: Mapping the human genetic architecture of COVID-19, *Nature* (July 2021).
https://www.nature.com/articles/s41586-021-03767-x_reference.pdf
News & Views: https://www.nature.com/articles/d41586-021-01773-7

Abstract: The genetic makeup of an individual contributes to susceptibility and response to viral infection. While environmental, clinical and social factors play a role in exposure 1,2to SARS-CoV-2 and COVID-19 disease severity , host genetics may also be important. Identifying host-specific genetic factors may reveal biological mechanisms of therapeutic relevance and clarify causal relationships of modifiable environmental risk factors for SARS-CoV-2 infection and outcomes. We formed a global network of researchers to investigate the role of human genetics in SARS-CoV-2 infection and COVID-19 severity. We describe the results of three genome-wide association meta-analyses comprised of up to 49,562 COVID-19 patients from 46 studies across 19 countries. We reported 13 genome-wide significant loci that are associated with SARS-CoV-2 infection or severe manifestations of COVID-19. Several of these loci correspond to previously documented associations to lung or autoimmune and 3–7 inflammatory diseases . They also represent potentially actionable mechanisms in response to infection. Mendelian Randomization analyses support a causal role for smoking and body mass index for severe COVID-19 although not for type II diabetes. The identification of novel host genetic factors associated with COVID-19, with unprecedented speed, was made possible by the community of human genetic researchers coming together to prioritize sharing of data, results, resources and analytical frameworks. This working model of international collaboration underscores what is possible for future genetic discoveries in emerging pandemics, or indeed for any complex human disease.

November 26 10 am – Journal Club – Osvaldo Espin-Garcia, UHN

Reading: A Penalized Regression Framework for Building Polygenic Risk Models Based on Summary Statistics from Genome-Wide Association Studies and Incorporating External Information, *JASA* (March 2021)
https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/01621459/v116i0533/133_aprffbasaiei.xml
DOI: 10.1080/01621459.2020.1764849

Abstract: Large-scale genome-wide association studies (GWAS) provide opportunities for developing genetic risk prediction models that have the potential to improve disease prevention, intervention or treatment. The key step is to develop polygenic risk score (PRS) models with high predictive performance for a given disease, which typically requires a large training dataset for selecting truly associated single nucleotide polymorphisms (SNPs) and estimating effect sizes accurately. Here, we develop a comprehensive penalized regression for fitting l 1 regularized regression models to GWAS summary statistics. We propose incorporating pleiotropy and annotation information into PRS (PANPRS) development through suitable formulation of penalty functions and associated tuning parameters. Extensive simulations show that PANPRS performs equally well or better than existing PRS methods when no functional annotation or pleiotropy is incorporated. When functional annotation data and pleiotropy are informative, PANPRS substantially outperforms existing PRS methods in simulations. Finally, we applied our methods to build PRS for type 2 diabetes and melanoma and found that incorporating relevant functional annotations and GWAS of genetically related traits improved prediction of these two complex diseases. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

December 3 12 noon – CANSSI STAGE International Speaker Seminar

Speaker: Antonis Antoniou, University of Cambridge
Professor of Cancer Risk Prediction; Academic Course Director MPhil in Epidemiology;
Department of Public Health and Primary Care
https://canssiontario.utoronto.ca/event/stage_isss_antonis_antoniou/

Talk Title: CanRisk: Personalising cancer risk prediction for prevention and early detection

Abstract: Much more reliable and powerful risk prediction for breast and ovarian cancer can be achieved by combining data on all genetic, lifestyle and hormonal risk factors for the diseases. We have recently enabled multifactorial breast and ovarian cancer risk-assessment through the CanRisk tool (www.canrisk.org) which allows healthcare professionals to obtain personalised cancer risks easily. The presentation will review the CanRisk development process, the challenges in combining the effects of rare pathogenic variants in known susceptibility genes, polygenic risk scores, questionnaire-based risk factors, mammographic density and family history into multifactorial cancer risk prediction algorithms; and will review the efforts to assess the clinical validity of the predicted risks in large independent studies. The presentation will finally discuss ongoing and planned efforts for the implementation of multifactorial cancer risk assessment in routine clinical practice for enabling cancer risk stratification and the better targeting of early detection and prevention approaches to those most likely to benefit.

********** **2022** **********

January 21 10 am – Research Seminar – Changchang Xu, Biostatistics

Topic: Mixture Cure Modelling in Molecular Genetic Prognosis

Background Reading:
Yilmaz *et al*. Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. J Clin Oncol. 2013 Jun 1;31(16):2047-54. https://pubmed.ncbi.nlm.nih.gov/23630217/
https://ovidsp-dc2-ovid-com.myaccess.library.utoronto.ca/ovid-a/ovidweb.cgi?&S=ALFCFPOJDBEBJCDJJPOJKFHGMKFLAA00&Link+Set=S.sh.22.23.27.31%7c23%7csl_10&Counter5=TOC_article%7c00005083-201331160-00023%7covft%7covftdb%7covfto

Heinze & Schemper. A solution to the problem of monotone likelihood in Cox regression. Biometrics. 2001 Mar;57(1):114-9 https://pubmed.ncbi.nlm.nih.gov/11252585/
https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/0006341x/v57i0001/114_asttpomlicr.xml
Kosmidis. Bias in parametric estimation: reduction and useful side-effects. WIREs Comput Stat 2014, 6:185–196. https://doi.org/10.1002/wics.1296

January 28 10 am – Journal Club – Ziang Zhang, Statistical Sciences

Topic: Analysis of genetic dominance in the UK Biobank

Reading: Palmer et al (2021), bioRxiv https://www.biorxiv.org/content/10.1101/2021.08.15.456387v2

Abstract: Classical statistical genetic theory defines dominance as a deviation from a purely additive effect. Dominance is well documented in model organisms and plant/animal breeding; outside of rare monogenic traits, however, evidence in humans is limited. We evaluated dominance effects in >1,000 phenotypes in the UK Biobank through GWAS, identifying 175 genome-wide significant loci ($P < 4.7 \times 10^{-11}$). Power to detect non-additive loci is low: we estimate a 20-30 fold increase in sample size is required to detect dominance loci to significance levels observed at additive loci. By deriving a new dominance form of LD-score regression, we found no evidence of a dominance contribution to phenotypic variance tagged by common variation genome-wide (median fraction 5.73×10^{-4}). We introduce dominance fine-mapping to explore whether the more rapid decay of dominance linkage disequilibrium can be leveraged to find causal variants. These results provide the most comprehensive assessment of dominance trait variation in humans to date.

February 4 3:30 pm – CANSSI STAGE International Speaker Seminar

Speaker: David Balding, University of Melbourne
Honorary Professor of Statistical Genetics; Director: Melbourne Integrative Genomics;
https://canssiontario.utoronto.ca/event/stage_isss_david_balding/

Talk Title: How are the causes of complex disease distributed in the human genome?

Abstract: The advent of very large, richly-phenotyped and high-quality human genomics datasets, together with the development of models that allow joint analyses of all GWAS test statistics, have led to big advances in understanding the genomic architecture of complex traits. However, models for the analysis of genome-wide SNPs, particularly for analyses based on association test statistics rather than individual genotype data, often rest on simplistic assumptions about the distribution of causal variation across the genome. Different approaches have led to discordant results about the genomic architecture of complex traits. I will review recent progress in genome-wide models of the heritability of complex human traits. In particular we look at the relationship between heritability and a range of genome annotation features, as well as linkage disequilibrium and minor allele fraction (MAF). The relationship between MAF and heritability is informative about the effects of negative or purifying selection, for different traits and in different genome regions. I will also discuss how the heritability models that arise from our work can be used to improve genomic prediction. This work leads to improved insights into the genomic architecture of complex traits.

February 11 10 am – Journal Club – Xiaoyu Men, CAMH

Topic: GRAF-pop: A Fast Distance-Based Method to Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis

Reading: Jin, Schaffer, Feolo, Holmes, Kattman (2019), G3 (Bethesda) Aug 8;9(8): 2447-61
https://pubmed.ncbi.nlm.nih.gov/31151998/
https://academic.oup.com/g3journal/article/9/8/2447/6026822
Supplemental material available at FigShare: https://doi.org/10.25387/g3.8061485.

Abstract: Inferring subject ancestry using genetic data is an important step in genetic association studies, required for dealing with population stratification. It has become more challenging to infer subject ancestry quickly and accurately since large amounts of genotype data, collected from millions of subjects by thousands of studies using different methods, are accessible to researchers from repositories such as the database of Genotypes and Phenotypes (dbGaP) at the National Center for Biotechnology Information (NCBI). Study-reported populations submitted to dbGaP are often not harmonized across studies or may be missing. Widely-used methods for ancestry prediction assume that most markers are genotyped in all subjects, but this assumption is unrealistic if one wants to combine studies that used different genotyping platforms. To provide ancestry inference and visualization across studies, we developed a new method, GRAF-pop, of ancestry prediction that is robust to missing genotypes and allows researchers to visualize predicted population structure in color and in three dimensions. When genotypes are dense, GRAF-pop is comparable in quality and running time to existing ancestry inference methods EIGENSTRAT, FastPCA, and FlashPCA2, all of which rely on principal components analysis (PCA). When genotypes are not dense, GRAF-pop gives much better ancestry predictions than the PCA-based methods. GRAF-pop employs basic geometric and probabilistic methods; the visualized ancestry predictions have a natural geometric interpretation, which is lacking in PCA-based methods. Since February 2018, GRAF-pop has been successfully incorporated into the dbGaP quality control process to identify inconsistencies between study-reported and computationally predicted populations and to provide harmonized population values in all new dbGaP submissions amenable to population prediction, based on marker genotypes. Plots, produced by GRAF-pop, of summary population predictions are available on dbGaP study pages, and the software, is available at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/Software.cgi.

March 4 12 noon –CANSSI STAGE International Speaker Seminar

Speaker: André Uitterlinden, Erasmus Medical Centre, Rotterdam, The Netherlands
Professor of Complex Genetics; Head, Genetic Laboratory and Human Genomics Facility
https://canssiontario.utoronto.ca/event/stage_isss_andre_uitterlinden/

Talk Title: Towards implementing (complex) genetics in health care settings

Abstract: Most if not all human diseases and risk factors have a genetic component, implying that variance among individuals in susceptibility, treatment response and/or progression, is determined in part by genetic variation. Human genome sequencing has uncovered hundreds of millions of genetic variants, while DNA analysis technology has progressed to allow sequencing a human genome in <24hours, and to analyze millions of SNPs in millions of DNA samples using arrays. Together with data from large longitudinal cohort studies and biobanks, in particular the array/chip technology has identified tens of thousands of genetic factors for common disease by Genome Wide Association Studies (GWAS). GWAS has led to global collaborative consortia, leading to a new scientific research culture producing robust results. World-wide several large-scale sequencing projects are now ongoing, such as the European 1 million Genomes (1MG) Project, including several nation-wide genome programs mostly based on array technology. Together, this has led to genetic information now entering the hospital clinic in a broad sense, whereby –in theory- all patients can be assessed by cheap array technology (at <$30/sample) for mutations and polygenic risk scores, next to pharmacogenetics and blood group/HLA typing for example, to help clinicians in decision making for diagnosis and treatment, and to provide self-empowerment for patients for prevention. Such a program, called GOALL (Genotyping On ALL patients) is currently established at Erasmus MC, The Netherlands. However, also outside of the (academic) hospital setting applications of using genetic information are explored, such as in population screening programs, e.g. for breast cancer. I will describe aspects of these developments, highlight examples, and provide an outlook to the future.

March 11 10 am – Journal Club – Henry Lu, Biostatistics

Topic: Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease

Reading: Xu et al (2022). *Cell Genomics* 2(1), 100086.
https://pubmed.ncbi.nlm.nih.gov/35072137/
https://reader.elsevier.com/reader/sd/pii/S2666979X21001075?token=2A98BB2F13AF0BE7104E3887A05046328DB2C3DC605A33B2402C87AE1BEF501BAF8D9B33ACF130DBE7014962C0C1942F&originRegion=us-east-1&originCreation=20220227214244

Abstract: Genetic association studies for blood cell traits, which are key indicators of health and immune function, have identified several hundred associations and defined a complex polygenic architecture. Polygenic scores (PGSs) for blood cell traits have potential clinical utility in disease risk prediction and prevention, but designing PGS remains challenging and the optimal methods are unclear. To address this, we evaluated the relative performance of 6 methods to develop PGS for 26 blood cell traits, including a standard method of pruning and thresholding (P + T) and 5 learning methods: LDpred2, elastic net (EN), Bayesian ridge (BR), multilayer perceptron (MLP) and convolutional neural network (CNN). We evaluated these optimized PGSs on blood cell trait data from UK Biobank and INTERVAL. We find that PGSs designed using common machine learning methods EN and BR show improved prediction of blood cell traits and consistently outperform other methods. Our analyses suggest EN/BR as the top choices for PGS construction, showing improved performance for 25 blood cell traits in the external validation, with correlations with the directly measured traits increasing by 10%–23%. Ten PGSs showed significant statistical interaction with sex, and sex-specific PGS stratification showed that all of them had substantial variation in the trajectories of blood cell traits with age. Genetic correlations between the PGSs for blood cell traits and common human diseases identified well-known as well as new associations. We develop machine learning-optimized PGS for blood cell traits, demonstrate their relationships with sex, age, and disease, and make these publicly available as a resource.

March 18 10 am – Seminar/Journal Club – Daniel Felsky, CAMH

Topic: Considering population stratification and admixture in modern GWAS; concepts and tools

.

Background Reading:
Atkinson *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power, *Nat Genet* **53,** 195–204 (2021).
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7867648/
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41588-020-00766-y
https://doi-org.myaccess.library.utoronto.ca/10.1038/s41588-020-00766-y

Abstract: Admixed populations are routinely excluded from genomic studies due to concerns over population structure. Here, we present a statistical framework and software package, Tractor, to facilitate the inclusion of admixed individuals in association studies by leveraging local ancestry. We test Tractor with simulated and empirical two-way admixed African–European cohorts. Tractor generates accurate ancestry-specific effect-size estimates and *P* values, can boost genome-wide association study (GWAS) power and improves the resolution of association signals. Using a local ancestry-aware regression model, we replicate known hits for blood lipids, discover novel hits missed by standard GWAS and localize signals closer to putative causal variants.

Review: Korunes KL, Goldberg A (2021) Human genetic admixture. *PLoS Genet* 17(3): e1009374.

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009374

Abstract: Throughout human history, large-scale migrations have facilitated the formation of populations with ancestry from multiple previously separated populations. This process leads to subsequent shuffling of genetic ancestry through recombination, producing variation in ancestry between populations, among individuals in a population, and along the genome within an individual. Recent methodological and empirical developments have elucidated the genomic signatures of this admixture process, bringing previously understudied admixed populations to the forefront of population and medical genetics. Under this theme, we present a collection of recent PLOS Genetics publications that exemplify recent progress in human genetic admixture studies, and we discuss potential areas for future work.

March 25 10 am – Seminar/Journal Club – Kieran Campbell, LTRI

Topic: Methods for cell type assignment across single-cell technologies

Reading: Geuenich, Hou, Lee, Ayub, Jackson, Campbell. Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data, *Cell Systems* (2021),12(12):1173-1186.e5,
https://pubmed.ncbi.nlm.nih.gov/34536381/
https://www-sciencedirect-com.myaccess.library.utoronto.ca/science/article/pii/S2405471221003355

Abstract: A major challenge in the analysis of highly multiplexed imaging data is the assignment of cells to *a priori* known cell types. Existing approaches typically solve this by clustering cells followed by manual annotation. However, these often require several subjective choices and cannot explicitly assign cells to an uncharacterized type. To help address these issues we present Astir, a probabilistic model to assign cells to cell types by integrating prior knowledge of marker proteins. Astir uses deep recognition neural networks for fast inference, allowing for annotations at the million-cell scale in the absence of a previously annotated reference. We apply Astir to over 2.4 million cells from suspension and imaging datasets and demonstrate its scalability, robustness to sample composition, and interpretable uncertainty estimates. We envision deployment of Astir either for a first broad cell type assignment or to accurately annotate cells that may serve as biomarkers in multiple disease contexts. A record of this paper's transparent peer review process is included in the supplemental information.

Background Reading re single-cell technologies:
Jackson *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578,** 615–620 (2020).
https://pubmed.ncbi.nlm.nih.gov/31959985/
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41586-019-1876-x
Zhang *et al*. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods* 2019 Oct;16(10):1007-1015
https://pubmed.ncbi.nlm.nih.gov/31501550/
https://www-nature-com.myaccess.library.utoronto.ca/articles/s41592-019-0529-1

April 8 10 am – Research Seminar – Celia Greenwood, Lady Davis Research Institute and McGill University

Topic: A Bayesian hierarchical model for improving measurement of 5-methylcytosine and 5-hydroxymethylcytosine levels: Towards revealing associations between phenotypes and methylation states

Abstract: 5-hydroxymethylcytosine (5hmC) is a methylation state linked with gene regulation, commonly found in cells of the central nervous system. 5hmC is associated with demethylation of cytosines from 5-methylcytosine (5mC) to the unmethylated state. The presence of 5hmC can be inferred by a paired experiment involving bisulfite and oxidative-bisulfite treatments on the same sample, followed by a methylation assay using a platform such as the Illumina Infinium MethylationEPIC BeadChip (EPIC). Existing methods for analysis of the resulting EPIC data are not ideal. Most approaches ignore the correlation between the two experiments and any imprecision associated with DNA damage from the additional treatment.

Here I will describe a hierarchical Bayesian method called Constrained HYdroxy Methylation Estimation (CHYME) to simultaneously estimate 5mC/5hmC signals as well as any associations between these signals and covariates or phenotypes, while accounting for the potential impact of DNA damage and dependencies induced by the experimental design.

Background Reading: Rakyan *et al*. Epigenome-wide association studies for common human diseases, *Nature Reviews Genetics* 12, p 529-541 (2011)
https://www-nature-com.myaccess.library.utoronto.ca/articles/nrg3000.pdf
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3508712/

April 22 10 am – Research Seminar – Jinko Graham, Simon Fraser University

Topic: Relatedness, gene genealogies and DNA sharing

Background Reading:
Karunarathna, Graham: *perfectphyloR*: An R package for reconstructing perfect phylogenies. *BMC Bioinformatics* **20,** 729 (2019). https://doi.org/10.1186/s12859-019-3313-4
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3313-4
Karunarathna, Graham: Using gene genealogies to localize rare variants associated with complex traits in diploid populations. *Hum Hered* 2018;83:30-39. doi: 10.1159/000486854
https://doi-org.myaccess.library.utoronto.ca/10.1159/000486854

April 29 10 am – Seminar/Journal Club – Michael Hoffman, UHN

Topic: Reproducibility standards for machine learning in the life sciences

Abstract: To make machine-learning analyses in the life sciences more computationally reproducible, we propose standards based on data, model and code publication, programming best practices and workflow automation. By meeting these standards, the community of researchers applying machine-learning methods in the life sciences can ensure that their analyses are worthy of trust.

Reading: Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. “Reproducibility standards for machine learning in the life sciences.” *Nat Methods* 2021Oct; 18(10):1132-35
https://doi.org/10.1038/s41592-021-01256-7

May 6 12 noon –CANSSI STAGE International Speaker Seminar

Speaker: David Conti, University of Southern California
Division of Biostatistics, Department of Population and Public Health Sciences
https://canssiontario.utoronto.ca/event/stage_isss_david_conti/

Title: Development of a Trans-Ancestry Genetic Risk Score for Prostate Cancer

Abstract: Prostate cancer is a highly heritable disease with large disparities in incidence rates across ancestry populations. I will present a recent multi-ancestry meta-analysis of prostate cancer genome-wide association studies and the methodological issues surrounding the construction of a genetic risk score (GRS) that is effective in multiple ancestry groups. The top GRS decile is associated with a 4-fold increase in risk for men of European, African, and Asian ancestry and for Latino men. These risks, combined with population-specific incidence rates lead to a 26-38% lifetime risk of prostate cancer across populations. These findings support the role of germline variation contributing to population differences in prostate cancer risk, with the GRS offering a tool for risk stratification.

May 13 10 am – Seminar/Journal Club
Speakers: Andrew Paterson, SickKids & Lei Sun, Statistical Sciences

Topic: The X chromosome: Sex differences in allele frequency; Hardy-Weinberg equilibrium.

Abstract:

Andrew will provide an overview of the X chromosome (Xchr), including the recent observations that a proportion of Xchr SNPs show sex differences in allele frequency (sdMAF). The sdMAF significance was set, conservatively, at the genome-wide level of 5e-8, the sdMAF analysis was performed, extensively, using the high coverage whole genome sequence data of the 1000 Genomes Project and the gnomAD V 3.1.2, and the observations were made, consistently, between the datasets and populations.

Lei will discuss the sdMAF test and related Hardy-Weinberg disequilibrium (HWD) test for a Xchr variant. To facilitate the discussion, let pf and pm be the female-MAF and male-MAF sample estimates. The numerator of an appropriate sdMAF test statistic, intuitively, should be (pf-pm). The denominator, intuitively, should contain MAF*(1-MAF), but several questions arise. 1) Do we use sex-pooled or sex-stratified MAF estimate? 2) Should we include a HWD factor and how? 3) How to deal with samples from multiple populations? And finally, 4) How to perform HWD-based quality control of the Xchr, and in the context of sdMAF?

Recommended Reading: Wang et al (2022). https://www.biorxiv.org/content/10.1101/2021.10.27.466015v1 (a revised version to appear in *PLoS Genetics*)

May 27 10 am – Research Seminar – Alexandre Bureau, Universite Laval

Title: Multivariate extension of penalized regression on summary statistics to construct polygenic risk scores for correlated traits

Authors: Meriem Bahda, Jasmin Ricard, Simon Girard, Michel Maziade, Maripier Isabelle, Alexandre Bureau

Abstract: Genetic correlations between human traits such as schizophrenia (SZ) and bipolar disorder (BD) diagnoses are well established. Improved prediction of individual traits has been obtained by combining predictors of multiple correlated traits derived from summary statistics produced by genome-wide association studies, compared to single trait predictors. We extend this idea to penalized regression on summary statistics in Multivariate Lassosum, where regression coefficients for the multiple traits on single nucleotide polymorphisms (SNPs) are treated as correlated random effects, as in multi-trait summary statistic best linear unbiased predictors (BLUP). We conducted simulations with two dichotomous traits having polygenic architecture similar to SZ and BD, using genotypes from 12 000 subjects from the CARTaGENE cohort. Multivariate Lassosum produced polygenic risk scores (PRS) more strongly correlated to the true genetic risk predictor than univariate sparse PRS (Lassosum, sparce LDpred2 and the standard clumping-thresholding). Application of Multivariate Lassosum to predict SZ, BD and related psychiatric traits in the Eastern Quebec SZ and BD kindred study revealed stronger associations with every trait than those obtained with univariate sparse PRS.

Background Reading:

Maier, R.M., Zhu, Z., Lee, S.H. *et al.* Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun* **9,** 989 (2018)
https://doi.org/10.1038/s41467-017-02769-6