STATISTICAL METHODS IN GENETIC/GENOMIC STUDIES

03 Jan 2022–14 Jan 2022, Institute for Mathematical Sciences, National University of Singapore

OVERVIEW

In this two-week program, we will focus on the advances in statistical genetics/genomics from both the applied and theoretical perspectives. The specific goal of the proposed workshop will be to equip researchers in genetic-related fields with the cutting-edge methods to better conduct genetic studies as biomedical research including genetic/genomic studies is one of the primary research areas supported by Singapore.

Organizers:

Jialiang Li, Jin Liu, Indranil Mukhopadhyay

Book of Abstracts:

Table of Contents

Xihong Lin, Harvard University, USA

January 3, 8 am – 9 am
Title: Integrative Analysis of Large-Scale Biobanks and Whole Genome Sequencing Studies

Abstract: Big data from genome, exposome, and phenome are becoming available at a rapidly increasing rate with no apparent end in sight. Examples include Whole Genome Sequencing data, digital data, and Electronic Health Records (EHRs). A rapidly increasing number of large scale national and institutional biobanks have emerged worldwide. Biobanks integrate genotype, electronic health records, and epidemiological and biomarker data, and is the trend of health science research. In this talk, I will discuss opportunities, analytic tools and resources, and challenges presented by analyzing large scale biobanks and population-based Whole Genome Sequencing (WGS) studies of common and rare genetic variants and EHRs by integrating WGS data with functional multi-omic data. The discussions are illustrated using ongoing large scale whole genome sequencing studies of the Genome Sequencing Program of the National Human Genome Research Institute and the Trans-Omics Precision Medicine Program from the National Heart, Lung and Blood Institute, and the UK Biobank and FinnGen, as well as rare-variant meta- analysis. I will also discuss integrative analysis of different types of data using whole genome causal mediation analysis.

Nancy Cox, Vanderbilt University Medical Center, USA

January 3, 9 am – 10 am
Title: Methods for Integrating Phenome and Genome Across Electronic Health Records

Abstract: As genome data becomes a part of medicine, every medical center becomes a biobank allowing investigation of genome to phenome relationships at unprecedented scale. We describe here recent extensions for the use of phenome risk scores (PheRS) (Bastarache et al., 2018 Science) for systematizing phenome information to improve our ability to discover genome to phenome relationships in a variety of contexts, from N=1 investigations, to large-scale investigations of both common and rare phenotypes. We utilize the BioVU, the biobank at Vanderbilt, with DNA samples on > 285,000 subjects, and electronic health records (EHR) data going back an average of 10-15 years on these subjects as well as another 1.8 million patients receiving healthcare at our medical center,

to provide examples of how extensions to the PheRS can be used to enhance discovery of genome to phenome relationships.

Peter X.K. Song, University of Michigan, USA

January 3, 10 am – 11 am
Title: Analyzing high-dimensional mediators by mixed integer optimization

Abstract: I will introduce an extension of the best-subset regularization to perform a high-dimensional mediation analysis in the framework of directed acyclic graphs (DAGs). This new methodology allows a simultaneous operation of parameter clustering and estimation in structural equation models to search causal mediation pathways. The double regularization on homogeneity fusion and sparsity is formulated as a mixed integer optimization (MIO) problem, in the hope to minimize estimation bias and give rise to an appealing setting for post-variable selection inference. We develop a fast and reliable algorithm, Alternating Penalization Operator for L-zero Loss Optimization (APOLLO), to implement the MIO problem for numerical solutions, which is shown to be superior to existing commercial integer programming solver Gurobi. APOLLO algorithm begins with an upper bound search for warm start, followed by a lower bound search via cutting planes. The proposed MIO estimator is rigorously investigated for its key theoretical guarantees. Numerical examples are used to illustrate the performance of the proposed MIO solver in simulation experiments and in motivating scientific studies.

Lin Chen, University of Chicago, USA

January 3, 11 am – 12 am
Title: Integrating multi-tissue multi-omics QTL with GWAS summary statistics for elucidation of the dynamic molecular mechanisms underlying disease genetics

Abstract: In the post-GWAS era, tens of thousands of unique associations between single nucleotide polymorphisms (SNPs) and complex diseases/traits were discovered. Many of them may affect complex diseases and traits through their effects on expression levels and/or other molecular traits (i.e., omics traits). Extensive evaluations of genetic effects on omics traits have revealed an abundance of quantitative trait loci (QTLs), and disease-associated QTLs often have a dynamic regulatory pattern with effects depending on tissue/cell types and contexts. In order to further understand the biological mechanisms underlying trait-associated SNPs, many efforts have been made to integrate GWAS summary statistics with expression-QTL (eQTL) and QTL statistics for other omics data types. In this talk, motivated by the dynamic tissue-specific effects and regulatory patterns of disease-associated QTLs for multiple omics data types, we will discuss several recent works in integrating GWAS with multi-tissue multi-omics QTL statistics. We proposed integrative association methods for characterizing the effects and regulatory patterns of multi-omics QTLs in different tissue types/conditions, and two-sample Mendelian Randomization methods for mapping putative risk factors for complex diseases. We analyzed GWASs with multi-tissue eQTL and methylation QTL (meQTL) statistics from the Genotype-tissue Expression (GTEx) project.

Jeanine Houwing-Duistermaat, University of Leeds, UK

January 3, 4 pm – 5 pm
Title: Probabilistic partial least squares methods for data integration

Abstract: Many studies collect multiple omics datasets to gather novel insights about different stages of biological processes and to associate omic features with outcome variables. For joint modelling of omic datasets, several data integration methods have been developed. These methods address high dimensionality, within and across datasets correlation, and the presence of heterogeneity among datasets due to representing different biological processes and using different measurement technologies. Most methods, neither provide statistical evidence for a relationship between the datasets nor identify relevant variables that contribute to this relationship. We propose a probabilistic latent variable modelling framework for inferring the relationship between two omics datasets. These methods reduce dimensionality and capture correlations by forming components that are linear combinations of the variables. The correlation structure is modelled by joint and data specific components. We also propose a model for the relationship between these joint components and an outcome variable. Model parameters are estimated using maximum likelihood. Test statistics are proposed for the null hypothesis of no relationship. We evaluate our methods via simulations. Under the null hypothesis, the test statistics appear to approximately follow the normal distribution. Our method appears to outperform existing methods for small and heterogeneous datasets in terms of selecting relevant variables and prediction accuracy. We illustrate the methods by analysing omics datasets from a population cohort and a case control study.

Weiwei Zhai, Chinese Academy of Sciences, China

January 3, 5 pm – 6 pm
Title: Harness tumor heterogeneity and evolution for understanding ethnic differences and patient prognosis in cancer

Abstract: Intra-tumor heterogeneity (ITH) lies at the center of tumor evolution and treatment response. In this talk, I will use lung and liver cancers as two examples to illustrate how we can use computational and evolutionary principles to dissect the genomic landscape and pinpoint ethnic differences in (lung) cancer. In addition, we made several interesting explorations harnessing multi-omic tumor heterogeneity for patient prognosis and treatment in cancer.

Hongyu Zhao, Yale University, USA

January 4, 8 am – 9 am
Title: Genetic Correlations across Traits and Populations

Abstract: Genome-wide association study (GWAS) has achieved remarkable success and has identified numerous single-nucleotide polymorphisms (SNPs) associated with complex human traits and diseases. Multi-trait and multi-population modeling has undergone rapid developments, leading to the emergence of numerous methods that study the shared genetic basis across multiple phenotypes and populations. Among these methods, genetic correlation analysis is a statistically

powerful and biologically interpretable approach to quantifying the genetic similarity of two traits, both across the genome and in local chromosomal regions. It has gained popularity in the field, provided new insights into the shared genetics of many phenotypes and different populations, and has a variety of downstream applications, including association analyses, genetic risk prediction, and causal inference and mediation analysis. In this presentation, we discuss various methods that have been developed for characterizing genetic correlations across populations and phenotypes, using both individual data and summary statistics.

Yun Li, The University of North Carolina at Chapel Hill, USA

January 4, 9 am – 10 am
Title: MAST-Decon: Smooth Cell-type Deconvolution Method for Spatial Transcriptomics data

Abstract: Spatial transcriptomics (ST) technologies have gained increasing popularity due to their ability to provide positional context of gene expressions in a tissue. One major limitation of current commercially available ST methods such as the 10X Genomics Visium platform is that they cannot yet reach single cell resolution. The number of cells within a spatial spot may range from 1 to 200 depending on the biological tissue and ST platform. Therefore, any downstream analysis such as spatially variable gene detection could be confounded by differential cell type compositions across spots. Cell type deconvolution for ST data is highly needed in order to fully reveal underlying biological mechanisms. Existing ST data deconvolution methods share two common limitations: first, few of them utilize spatial neighborhood information. Existing methods such as RCTD and SPOTlight intrinsically treat each spatial spot as independent of neighboring spots, although we anticipate nearby spots to share similar cell type compositions based on clinical evidence of tissue structures. Such limitation could be amplified when sequencing depths at single spots are relatively low so that borrowing information from neighboring spots is necessary in order to obtain reliable deconvolution results. Second, although Visium data provide us with a histological image which could add additional information regarding spot heterogeneity, most existing methods do not utilize this H&E image. To solve these two limitations, we developed Multiscale Adaptive ST Deconvolution (MAST-Decon), a smooth deconvolution method for ST data. MAST-Decon uses a weighted likelihood approach and incorporates both gene expression data, spatial neighborhood information and H&E image features by constructing different kernel functions to obtain a smooth deconvolution result. We showcased the strength of MAST-Decon through simulations based on real data, including sub-cellular resolution seqFISH+ data. By introducing spatial smoothness, we were able to correct several spots erroneously inferred by RCTD when the average UMI count per spot is at ~2.5k level, commonly observed in current ST data. Overall, we were able to improve the Spearman correlation between deconvolution results and ground truth cell-type proportions from 0.757 (RCTD) to 0.823 (MAST-Decon), and reduce RMSE between deconvolution results and ground truths from 0.0491 (RCTD) to 0.0334 (MAST-Decon).

Jian Huang, University of Iowa, USA

January 4, 10 am – 11 am
Title: A Deep Generative Approach to Conditional Sampling

Abstract: Conditional distribution is a fundamental quantity in statistics and machine learning that provides a full description of the relationship between a response and a predictor. There is a vast literature on conditional density estimation. A common feature of the existing methods is that they seek to estimate the functional form of the conditional density. We propose a deep generative approach to learning a conditional distribution by estimating a conditional generator, so that a random sample from the target conditional distribution can be obtained by transforming a sample from a reference distribution. The conditional generator is estimated nonparametrically with neural networks by matching appropriate joint distributions using a discrepancy measure. There are several advantages of the proposed generative approach over the classical methods for conditional density estimation, including: (a) there is no restriction on the dimensionality of the response or predictor, (b) it can handle both continuous and discrete type predictors and responses, and (c) it is easy to obtain estimates of the summary measures of the underlying conditional distribution by Monte Carlo. We conduct numerical experiments to validate the proposed method and using several benchmark datasets, including the California housing, the MNIST, and the CelebA datasets, to illustrate its applications in conditional sample generation, uncertainty assessment of prediction, visualization of multivariate data, image generation and image reconstruction.

Mengjie Chen, University of Chicago, USA

January 4, 11 am – 12 am
Title: Demystifying the drop-outs in single cell RNA-seq data

Abstract: Droplet-based single-cell RNA-sequencing (scRNA-seq) methods have changed the landscape of genomics research in complex biological systems by producing single cell resolution data at affordable costs. In the state-of-the-arts protocols, a step called barcoding unique molecular identifiers (UMI) has been introduced to remove amplification bias and further improve data quality. Recent literature suggests that barcoding has led to a different error structure in the count data with much less technical noise. Regardless, many tools do not acknowledge the differences between the read count data and UMI count data, still assuming that both suffer from excessive technical noise. In this presentation, I will make a brief overview of scRNA-seq data analysis pipelines and then present extensive analyses of publicly available UMI data sets that challenge the assumptions of most existing pre-processing tools. Our results suggest that resolving cell-type heterogeneity should be the foremost step of the scRNA-seq analysis pipeline. Normalizing or imputing the data set before resolving the heterogeneity can lead to adversary consequences in downstream analysis. As a result, we provide a new perspective on scRNA-seq data analysis by fully integrating pre-processing and clustering, which was classified as part of the downstream analysis. The proposed procedures have been implemented in software, HIPPO. If time permits, I will also talk about other single cell analysis tools developed in my group, VIPER, an imputation method for SMART-sea data, and dmatch, an alignment tool for multiple scRNA-seq samples batch correction.

Can Yang, The Hong Kong University of Science and Technology, China

January 4, 3 pm – 4 pm

Title: A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits

Abstract: The development of polygenic risk scores (PRS) has proved useful to stratify the general population into different risk groups for the European population. However, PRS is less accurate in non-European populations due to genetic differences across different populations. To improve the prediction accuracy in non-European populations, we propose a cross-population analysis framework for PRS construction with both individual-level (XPA) and summary-level (XPASS) GWAS data. By leveraging trans-ancestry genetic correlation, our methods can borrow information from the Biobank-scale European population data to improve risk prediction in the non-European populations. Our framework can also incorporate population-specific effects to further improve the construction of PRS. With innovations in data structure and algorithm design, our methods provide a substantial saving in computational time and memory usage. Through comprehensive simulation studies, we show that our framework provides accurate, efficient, and robust PRS construction across a range of genetic architectures. In a Chinese cohort, our methods achieved 7.3%-198.0% accuracy gain for height and 19.5%-313.3% accuracy gain for body mass index (BMI) in terms of predictive R-squared compared to existing PRS approaches, respectively. We also show that XPA and XPASS can achieve substantial improvement for the construction of height PRS in the African population, suggesting the generality of our framework across global populations.

Lin Hou, Tsinghua University, China

January 4, 4 pm – 5 pm
Title: Detecting Local Genetic Correlations with Scan Statistics

Abstract: Genetic correlation analysis has quickly gained popularity in the past few years and provided insights into the genetic etiology of numerous complex diseases. However, existing approaches oversimplify the shared genetic architecture between different phenotypes and cannot effectively identify precise genetic regions contributing to the genetic correlation. In this work, we introduce LOGODetect, a powerful and efficient statistical method to identify small genome segments harboring local genetic correlation signals. LOGODetect automatically identifies genetic regions showing consistent associations with multiple phenotypes through a scan statistic approach. It uses summary association statistics from genome-wide association studies (GWAS) as input and is robust to sample overlap between studies. Applied to seven phenotypically distinct but genetically correlated neuropsychiatric traits, we identify 227 non-overlapping genome regions associated with multiple traits, including multiple hub regions showing concordant effects on five or more traits. Our method addresses critical limitations in existing analytic strategies and may have wide applications in post-GWAS analysis.

Zhixiang Lin, The Chinese University of Hong Kong, China

January 4, 5 pm – 6 pm
Title: Statistical method for integrative analysis in single-cell genomics

Abstract: The recent advancements in single-cell technologies, including single-cell chromatin accessibility sequencing (scCAS), have enabled profiling the epigenetic landscapes for thousands of individual cells. However, the characteristics of scCAS data, including high dimensionality, high degree of sparsity and high technical variation, make the computational analysis challenging. Reference-guided approaches, which utilize the information in existing datasets, may facilitate the analysis of scCAS data. Here, we present RA3 (Reference-guided Approach for the Analysis of single-cell chromatin Accessibility data), which utilizes the information in massive existing bulk chromatin accessibility and annotated scCAS data. RA3 simultaneously models (1) the shared biological variation among scCAS data and the reference data, and (2) the unique biological variation in scCAS data that identifies distinct subpopulations. We show that RA3 achieves superior performance when used on several scCAS datasets, and on references constructed using various approaches. Altogether, these analyses demonstrate the wide applicability of RA3 in analyzing scCAS data.

Hongzhe Lee, University of Pennsylvania, USA

January 5, 8 am – 9 am
Title: A Unified Approach to Robust Inference for Genetic Covariance

Abstract: Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits. Many complex traits are found to have shared genetic etiology. Genetic covariance is defined as the underlying covariance of genetic values and can be used to measure the shared genetic architecture. The data of two outcomes may be collected from the same group or different groups of individuals and the outcomes can be of different types or collected based on different study designs. This paper proposes a unified approach to robust estimation and inference for genetic covariance of general outcomes that may be associated with genetic variants nonlinearly. We provide the asymptotic properties of the proposed estimator and show that our proposal is robust under certain model mis-specification. Our method under linear working models provides robust inference for the narrow-sense genetic covariance, even when both linear models are mis-specified. Various numerical experiments are performed to support the theoretical results. Our method is applied to an outbred mice GWAS data set to study the overlapping genetic effects between the behavioral and physiological phenotypes. The real data results demonstrate the robustness of the proposed method and reveal interesting genetic covariance among different mice developmental traits.

Pei Wang, Icahn Medical School at Mount Sinai, USA

January 5, 9 am – 10 am
Title: DAGBagM: Learning directed acyclic graphs of mixed variables with an application to identify protein biomarkers for treatment response in ovarian cancer

Abstract: Gene/protein regulatory networks inferred by applying directed acyclic graph (DAG) models to proteogenomic data has been shown effective for detecting causal biomarkers of clinical outcomes. However, there remain unsolved challenges in DAG learning to jointly model clinical outcome variables, which often take binary values, and biomarker measurements, which usually are continuous variables. Therefore, in this paper, we propose a new tool, DAGBagM, to learn DAGs with

both continuous and binary nodes. By using appropriate models for continuous and binary variables, DAGBagM allows for either type of nodes to be parents or children nodes in the learned graph. DAGBagM also employs a bootstrap aggregating strategy to reduce false positives in edge inference. Moreover, the aggregation procedure provides a flexible framework to robustly incorporate prior information on edges. As shown by simulation experiments, DAGBagM performs better in identifying edges between continuous and binary nodes, as compared to commonly used strategies of either treating binary variables as continuous or discretizing continuous variables, as well as a recent method for learning DAGs with mixed types of nodes. Moreover, DAGBagM outperforms several popular DAG structure learning algorithms including the score-based hill climbing (HC) algorithm, constraint-based PC-algorithm (PC-alg), and the hybrid method max-min hill climbing (MMHC) when learning DAGs with only continuous nodes. The HC algorithm implementation in the R package DAGBagM is much faster than that in a widely used DAG learning R package bnlearn as well as mDAG. When applying DAGBagM to proteomics datasets from ovarian cancer studies, we identify potential protein biomarkers for platinum refractory/resistant response in ovarian cancer.

Judy Hua Zhong, New York University, USA

January 5, 10 am – 11 am
Title: Fair Generalized Linear Models

Abstract: Despite recent advances in algorithmic fairness, methodologies for achieving fairness with generalized linear models (GLMs) have yet to be explored, even though GLMs are very widely used in practice. In this paper we introduce two fairness criteria based on equalizing expected outcomes or log-likelihoods for GLMs. We prove that in the case of GLMs, both criteria can be achieved via a convex penalty term based solely on the linear components of the GLM, thus permitting efficient optimization. We also show the theoretical properties of the resulting fair GLM estimator. To empirically demonstrate the efficacy of the proposed fair GLM, we compare it with other well-known fair prediction methods on an extensive set of benchmark datasets for binary classification and regression. In addition, we demonstrate that the fair GLM can generate fair predictions for a wide range of response variables, other than binary and continuous outcomes.

Xiang Zhou, University of Michigan, USA

January 5, 11 am – 12 am
Title: Statistical analysis of spatial expression pattern for spatially resolved transcriptomic studies

Abstract: Identifying genes that display spatial expression patterns in spatially resolved transcriptomic studies is an important first step towards characterizing the spatial transcriptomic landscape of complex tissues. Here, we developed a statistical method, SPARK, for identifying such spatially expressed genes in data generated from various spatially resolved transcriptomic techniques. SPARK directly models spatial count data through the generalized linear spatial models. It relies on newly developed statistical formulas for hypothesis testing, providing effective type I error control and yielding high statistical power. With a computationally efficient algorithm based on penalized quasi-likelihood, SPARK is also scalable to data sets with tens of thousands of genes measured on tens of thousands of samples. In four published spatially resolved transcriptomic data

sets, we show that SPARK can be up to ten times more powerful than existing methods, revealing new biology in the data that otherwise cannot be revealed by existing approaches. SPARK is recently extended towards non-parametric modeling via a new method SPARK-X for rapid and effective detection of spatially expressed genes in large spatial transcriptomic studies. SPARK-X not only produces effective type I error control and high power but also brings orders of magnitude computational savings compared to SPARK, thus representing the only current spatial expression analysis method for large-scale spatial transcriptomics studies.

Zhonghua Liu, The University of Hong Kong, China

January 5, 3 pm – 4 pm
Title: On Mendelian Randomisation Mixed-Scale Treatment Effect Robust Identification (MR MiSTERI) and Estimation for Causal Inference

Abstract: Standard Mendelian randomization analysis can produce biased results if the genetic variant defining instrumental variable (IV) is confounded and/or has a horizontal pleiotropic effect on the outcome of interest not mediated by the treatment. We provide novel identification conditions for the causal effect of a treatment in the presence of unmeasured confounding by leveraging an invalid IV for which both the IV independence and exclusion restriction assumptions may be violated. The proposed Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification (MR MiSTERI) approach relies on (i) an assumption that the treatment effect does not vary with the invalid IV on the additive scale; and (ii) that the selection bias due to confounding does not vary with the invalid IV on the odds ratio scale; and (iii) that the residual variance for the outcome is heteroscedastic and thus varies with the invalid IV. Although assumptions (i) and (ii) have, respectively appeared in the IV literature, assumption (iii) has not; we formally establish that their conjunction can identify a causal effect even with an invalid IV subject to pleiotropy. MR MiSTERI is shown to be particularly advantageous in the presence of pervasive heterogeneity of pleiotropic effects on the additive scale. For estimation, we propose a simple and consistent three-stage estimator that can be used as preliminary estimator to a carefully constructed one-step-update estimator, which is guaranteed to be more efficient under the assumed model. In order to incorporate multiple, possibly correlated and weak IVs, a common challenge in MR studies, we develop a MAny Weak Invalid Instruments (MR MaWII MiSTERI) approach for strengthened identification and improved estimation accuracy. Both simulation studies and UK Biobank data analysis results demonstrate the robustness of the proposed MR MiSTERI method.

Yingying Wei, The Chinese University of Hong Kong, China

January 5, 4 pm – 5 pm
Title: Meta-clustering of Genomic Data

Abstract: Like traditional meta-analysis that pools effect sizes across studies to improve statistical power, it is of increasing interest to conduct clustering jointly across datasets to identify disease subtypes for bulk genomic data and discover cell types for single-cell RNA-sequencing (scRNA-seq) data. Unfortunately, due to the prevalence of technical batch effects among high-throughput

experiments, directly clustering samples from multiple datasets can lead to wrong results. The recent emerging meta-clustering approaches require all datasets to contain all subtypes, which is not feasible for many experimental designs. In this talk, I will present our Batch-effects-correction-with-Unknown-Subtypes (BUS) framework. BUS is capable of correcting batch effects explicitly, grouping samples that share similar characteristics into subtypes, identifying features that distinguish subtypes, and enjoying a linear-order computational complexity. We prove the identifiability of BUS for not only bulk data but also scRNA-seq data whose dropout events suffer from missing not at random. We mathematically show that under two very flexible and realistic experimental designs—the "reference panel" and the "chain-type" designs—true biological variability can also be separated from batch effects. Moreover, despite the active research on analysis methods for scRNA-seq data, rigorous statistical methods to estimate treatment effects for scRNA-seq data—how an intervention or exposure alters the cellular composition and gene expression levels—are still lacking. Building upon our BUS framework, we further develop statistical methods to quantify treatment effects for scRNA-seq data.

Xingjie Shi, East China Normal University, China

January 5, 5 pm – 6 pm
Title: Spatial clustering for identifying tissue regions defined by spatial transcriptomics

Abstract: Spatial transcriptomics has been emerging as a powerful technique for resolving gene expression profiles while retaining tissue spatial information. These spatially resolved transcriptomics make it feasible to examine the complex multicellular systems of different microenvironments. To answer important scientific questions with spatial transcriptomics, the first challenge is to identify cell clusters by integrating the available spatial information. In this talk, I will present and discuss two methods to handle this challenge. The first method is SC-MEB, an empirical Bayes approach for spatial clustering analysis using a hidden Markov random field. In contrast to BayesSpace, a recently developed method, SC-MEB is not only computationally efficient and scalable to large sample sizes but is also capable of choosing the smoothness parameter and the number of clusters. In SC-ME and many other existing methods, dimension reduction and (spatial) clustering are sequentially conducted as two consecutive stages. Low-dimensional embeddings estimated in the dimension reduction stage may not necessarily be relevant to class labels. This may endanger the performance of the follow-up clustering and downstream analysis. Our second method, DR-SC, is a unified and principled method that performs dimensional reduction and spatial clustering simultaneously. DR-SC can improve both the spatial and non-spatial clustering performance, resolve a low-dimensional representation with improved visualization, and facilitate the downstream analysis, such trajectory inference. Both SC-MEB and DR-SC provide valuable computational tools for investigating the structural organizations of tissues from spatial transcriptomic data.

Shuangge Ma, Yale University, USA

January 6, 8 am – 9 am
Title: Gaussian graphical model-based heterogeneity analysis via penalized fusion

Abstract: Heterogeneity is a hallmark of many complex diseases. This study has been motivated by the unsupervised heterogeneity analysis for complex diseases based on molecular and imaging data, for which, network-based analysis can be more informative than that limited to mean, variance, and other simple distributional properties. In the literature, there has been limited research on network-based heterogeneity analysis, and a common limitation shared by the existing techniques is that the number of subgroups needs to be specified a priori or in an ad hoc manner. We develop a novel approach for heterogeneity analysis based on the Gaussian graphical model. It applies penalization to the mean and precision matrix parameters to generate regularized and interpretable estimates. A fusion penalty is imposed to “automatedly” determine the number of subgroups. The heterogeneity analysis of non-small-cell lung cancer based on single-cell gene expression data of the Wnt pathway and that of lung adenocarcinoma based on histopathological imaging data not only demonstrate the practical applicability of the proposed approach but also lead to interesting new findings.

Li Hsu, Fred Hutchinson Cancer Research Center, USA

January 6, 9 am – 10 am
Title: A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomic Data

Abstract: Genome-wide association studies (GWAS) have successfully identified thousands of genetic variants for many complex diseases; however, these variants explain only a small fraction of the heritability. Recent developments in transcriptome-wide association studies have shown promises for discovering novel loci by leveraging genetically regulated molecular phenotypes (e.g., gene expression, methylation, proteomics) derived from external reference genotype- omics studies. However, there is a limitation in the existing approaches. Some variants can individually influence disease risk through alternative functional mechanisms. Existing approaches of testing only the association of imputed molecular phenotypes will potentially lose power. To tackle these challenges, we consider a unified mixed effects model that formulates the association of intermediate phenotypes such as imputed gene expression through fixed effects, while allowing residual effects of individual variants to be random. We consider a set-based score testing framework, MiST (Mixed effects Score Test), and propose data-driven combination approaches to jointly test for the fixed and random effects. We also provide p-values for fixed and random effects separately to enhance interpretability over GWAS. Recently, we extend the MiST to depend on only GWAS summary statistics instead of individual level data, allowing for a broad application of MiST to GWAS data. Extensive simulations demonstrate that MiST is more powerful than existing approaches and summary statistics-based MiST (sMiST) agrees well those obtained from individual level data with substantively improved computational speed. We apply sMiST to a large-scale GWAS of colorectal cancer using summary statistics from >120, 000 study participants and gene expression data from the Genotype-Tissue Expression (GTEx) project. We identify several novel and secondary independent genetic loci.

Hongkai Ji, Johns Hopkins University, USA

January 6, 10 am – 11 am

Title: A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples

Abstract: Pseudotime analysis with single-cell RNA-sequencing (scRNA-seq) data has been widely used to study dynamic gene regulatory programs along continuous biological processes. While many computational methods have been developed to infer the pseudo-temporal trajectories of cells within a biological sample, methods that compare pseudo-temporal patterns with multiple samples (or replicates) across different experimental conditions are lacking. Lamian is a comprehensive and statistically-rigorous computational framework for differential multi-sample pseudotime analysis. It can be used to identify changes in a biological process associated with sample covariates, such as different biological conditions, and also to detect changes in gene expression, cell density, and topology of a pseudo temporal trajectory. Unlike existing methods that ignore sample variability, Lamian draws statistical inference after accounting for cross-sample variability and hence substantially reduces sample-specific false discoveries that are not generalizable to new samples. Using both simulations and real scRNA-seq data, including an analysis of differential immune response programs between COVID-19 patients with different disease severity levels, we demonstrate the advantages of Lamian in decoding cellular gene expression programs in continuous biological processes.

Wei Pan, University of Minnesota, USA

January 6, 11 am – 12 am
Title: Robust Mendelian Randomization via Constrained Maximum Likelihood

Abstract: With the increasing availability of large-scale GWAS summary data on various complex traits and diseases, there have been tremendous interests in applications of Mendelian randomization (MR) to investigate causal relationships between pairs of traits using SNPs as instrumental variables (IVs) based on observational data. In spite of the potential significance of such applications, the validity of their causal conclusions critically depend on some strong modeling assumptions required by MR, which may be violated due to the widespread (horizontal) pleiotropy. Although many MR methods have been proposed recently to relax the assumptions by mainly dealing with uncorrelated pleiotropy, only few can handle correlated pleiotropy, in which some SNPs/IVs may be associated with hidden confounders, such as some heritable factors shared by both traits. Here we propose a simple and effective approach based on constrained maximum likelihood applicable to GWAS summary data. To deal with more challenging situations with many invalid IVs with only weak pleiotropic effects, we modify and improve it with data perturbation. Extensive simulations demonstrated that the proposed methods could control the type I error rate better while achieving higher power than other competitors. Applications to 48 risk factor- disease pairs based on large-scale GWAS summary data of three cardio-metabolic diseases (coronary artery disease, stroke and type 2 diabetes), asthma and 12 risk factors confirmed its superior performance. The major part of this talk will be based on the joint work with Haoran Xue and Xiaotong Shen as published in AJHG (2021, 108(7): 1251-1269; https://pubmed.ncbi.nlm.nih.gov/34214446/), and we will outline some of our related and on-going work.

Andrew Xiaohua Zhou, Peking University, China

January 6, 3 pm – 4 pm
Title: Causal Inference with Truncation by Death in Observational Study

Abstract: In this talk, we introduce a new method to deal with unmeasured confounding when the outcome is truncated by death in an observational clinical study. We first propose a new method to identify the heterogeneous conditional survivor average causal effect based on a substitutional variable under monotonicity. Then, we show the proposed method has both asymptotic and good finite-sample properties. Finally we illustrate the application of the proposed method in a real-world example.

Heather J. Cordell, Newcastle University, UK

January 6, 4 pm – 5 pm
Title: A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships

Abstract: 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. Here we propose two improvements to an existing method for Bayesian network analysis, designed to increase the power to detect potential causal relationships between variables (including potentially a mixture of both discrete and continuous variables). Our first improvement relates to the treatment of missing data. 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 that 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, and we illustrate the approach through application to data from a recent study investigating the causal relationship between methylation and gene expression in early inflammatory arthritis patients. We also describe a second improvement in the form of a pseudo-Bayesian approach for upweighting certain network edges, which can be useful when there is prior evidence concerning their directions.

Yuling Jiao, Wuhan University, China

January 6, 5 pm – 6 pm
Title: Deep Nonparametric Estimation

Abstract: In this talk, I will present some theoretical results in deep nonparametric estimation including regression, classification and GANs.

Pei-Fen Kuan, Stony Brook University, USA

January 7, 8 am – 9 am
Title: Deciphering the Transcriptome of the World Trade Center Disaster-Related PTSD via RNA-Seq

Abstract: Post-traumatic stress disorder (PTSD) is a debilitating psychiatric condition that is very common among responders to the 9/11 World Trade Center (WTC) terrorist attack: incidence since the disaster is 10-20%. PTSD in WTC responders is often chronic, persisting two decades later and can lead to cognitive, social, and occupational impairment. As such, it is of high importance to understand the etiological factors that maintain PTSD. Gene expression has emerged as a promising biomarker for PTSD. However, current PTSD gene expression studies have not fully explored the transcriptomic landscape of PTSD. In this talk, we will focus on the RNA-Seq transcriptomic studies conducted at the Stony Brook WTC Health Program. We will describe our proposed statistical models to enhance our understanding of the genes and premature aging associated with PTSD.

Fei Zou, University of North Carolina, USA

January 7, 9 am – 10 am
Title: Joint Gene Network Construction by Single-Cell RNA Sequencing Data

Abstract: In contrast to differential gene expression analysis at the single-gene level, gene regulatory network(GRN) analysis depicts complex transcriptomic interactions among genes for better understanding of underlying genetic architectures of human diseases and traits. Recent advances in single-cell RNA sequencing (scRNA-seq) allow constructing GRNs at a much finer resolution than bulk RNA-seq and microarray data. However, scRNA-seq data are inherently sparse, which hinders direct application of the popular Gaussian graphical models (GGMs).Furthermore, most existing approaches for constructing GRNs with scRNA-seq data only consider gene networks under one condition. To better understand GRNs under different but related conditions with single-cell resolution, we propose to construct Joint Gene Networks with scRNA-seq data using the GGMs framework. In this talk, we will first introduce a hybrid imputation procedure to efficiently impute zero-inflated counts resulted from technical artifacts. We then discuss how to transform the imputed data via a nonpara-normal transformation, based on which we perform joint GGM constructions. We demonstrate our approach and assess its performance using synthetic data and two cancer clinical studies on medulloblastoma and glioblastoma.

Hui Zou, School of Statistics, University of Minnesota, USA

January 7, 10 am – 11 am
Title: Sparse Convoluted Rank Regression in High Dimensions

Abstract: Wang et al. (2020, JASA) studied the high-dimensional sparse penalized rank regression and established its nice theoretical properties. Compared with the least squares, rank regression can

have a substantial gain in estimation efficiency while maintaining a minimal relative efficiency of 86.4%. However, the computation of penalized rank regression can be very challenging for high-dimensional data, due to the highly nonsmooth rank regression loss. In this work we view the rank regression loss as a non-smooth empirical counterpart of a population level quantity, and a smooth empirical counterpart is derived by substituting a kernel density estimator for the true distribution in the expectation calculation. This view leads to the convoluted rank regression loss and consequently the sparse penalized convoluted rank regression (CRR) for high-dimensional data. Under the same key assumptions for sparse rank regression, we establish the rate of convergence of the l1-penalized CRR for a tuning free penalization parameter and prove the strong oracle property of the folded concave penalized CRR. We further propose a high-dimensional Bayesian information criterion for selecting the penalization parameter in folded concave penalized CRR and prove its selection consistency. We derive an efficient algorithm for solving sparse convoluted rank regression that scales well with high dimensions. Numerical examples demonstrate the promising performance of the sparse convoluted rank regression over the sparse rank regression. Our theoretical and numerical results suggest that sparse convoluted rank regression enjoys the best of both sparse least squares regression and sparse rank regression.

Jingyi Jessica Li, University of California at Los Angeles, USA

January 7, 11 am – 12 am
Title: Clipper: a general statistical framework for p-value-free FDR control in large-scale feature screening

Abstract: Large-scale feature screening is ubiquitous in high-throughput biological data analysis: identifying the features (e.g., genes, mRNA transcripts, and proteins) that differ between conditions from numerous features measured simultaneously. The false discovery rate (FDR) is the most widely-used criterion to ensure the reliability of screened features. The most famous Benjamini-Hochberg procedure for FDR control requires valid, high-resolution p-values, which are, however, often hardly achievable because of the reliance on reasonable distributional assumptions or large sample sizes. Motivated by the Barber-Candes procedure, Clipper is a general statistical framework for large-scale feature screening with theoretical FDR control and without p-value requirement. Extensive numerical studies have verified that Clipper is a versatile and effective tool for correcting the FDR inflation crisis in multiple bioinformatics applications, including peak calling from ChIP-seq data and differentially expressed gene identification from bulk and single-cell RNA-seq data.

Mark van der Laan, University of California, Berkeley, USA

January 10, 8 am – 9 am

Title: Targeted Learning with Applications to Genomic Studies

Abstract: We consider data sets in which one observes a large number of covariates, including genomic measurements, possibly a treatment of interest, and a subsequent outcome. We present the roadmap of targeted learning involving specifying the statistical model, and a collection of

statistical queries about the probability distribution of the data to formally define the statistical estimation problem. We suggest both causal effect queries as well as variable importance queries that can handle both discrete variables as well as continuous variables. We present the general targeted maximum likelihood estimation methodology, incorporating super-learning and highly adaptive lasso as a particularly powerful machine learning algorithm, and corresponding statistical inference. The statistical inference and multiple testing adjustment can be based on a multivariate normal limit distribution, or a nonparametric bootstrap method, thereby utilizing dependence among test statistics. We also present a cross-validated TMLE for data adaptively determined target estimands/statistical queries. Methods are demonstrated with simulations and data sets.

Michael Newton, University of Wisconsin-Madison, USA

January 10, 9 am – 10 am

Title: Empirical Bayes and the false discovery rate, revisited

Abstract: Large-scale hypothesis testing and prioritization problems occur in many contemporary statistical applications. Well developed and widely studied methods and computational tools are available to report the most interesting testing units subject to control on the rate of false discoveries. Motivated by examples from single-cell RNA-sequencing, antibody profiling, and brain imaging I will review recent work on empirical Bayesian approaches to this problem, noting in many cases that standard tools control false discovery rates but at the expense of deficiencies in other operating characteristics.

George C. Tseng, University of Pittsburgh, USA

January 10, 10 am – 11 am

Title: Recent advances in p-value combination methods with emphasis on omics applications

Abstract: The issue of combining p-values is a long-standing question in statistics and has many scientific applications. In this talk, I will give a brief review of methods for combining *K* p-values from a classical setting of fixed *K* to a modern sparse and weak signal setting when *K* goes to infinity. I will then present some recent development of methods for combining p-values with heterogeneous signals or combining dependent p-values with emphasis on omics applications. Specifically, I will discuss several versions of adaptively weighted Fisher's method and demonstrate their asymptotic optimality properties, such as Little and Folks' asymptotically Bahadur optimality (ABO) and Donoho and Jin's detection boundary. Simulations and applications help show usefulness of these adaptive weighting methods in terms of statistical power and biological interpretations. For combining dependent p-values, we will present recently developed Cauchy and harmonic mean methods, their effectiveness in SNP association applications and some theoretical foundations of why they work.

Śaunak Sen, The University of Tennessee Health Science Center, USA

January 10, 11 am – 12 pm

Title: Sparse bilinear models for structured high-throughput data

Abstract: Rapid generation of high-throughput biological data have transformed biomedical research; they can be used to address novel scientific questions in broad research areas. Examples include genomewide transcriptomics, metabolomics, and microbiome profiling. These data can be thought of as a large matrix with covariates annotating both rows and columns of this matrix. Bilinear models provide a convenient framework for modeling such data taking into account the correlations and known relationships. In many situations, sparse estimation of these models is desired. We present fast methods for sparse estimation using elastic net regularization and consider the case when the response matrix and the covariate matrices are large. Due to data size, standard methods for estimation of these penalized regression models fail if the problem is converted to the corresponding univariate regression problem. By leveraging matrix properties in the structure of our model, we have developed several fast estimation algorithms. We have evaluated their performance in simulated data, and applications to high-throughput chemical genetic screens, eQTL (expression quantitative loci) detection, and metabolomic profiling. Our algorithms have been implemented in three packages in the Julia programming language available at https://senresearch.github.io.

Samsiddhi Bhattacharjee, National Institute of Biomedical Genomics, India

January 10, 3 pm – 4 pm

Title: Statistical Approaches to Incorporating Prior Knowledge for Multiple-Testing and Variable-selection in Genomic Studies

Abstract: The 'curse of dimensionality' impacts genomics studies in two ways. Firstly, it reduces the power to make discoveries due to considerable multiple-testing burden in genome-wide searches. Secondly, it makes it difficult to distinguish and select the most-important variables among many correlated variables, particularly with limited sample sizes. There is vast amount of prior knowledge (e.g., from public databases or other omics studies) that can be utilized to potentially boost power in such situations. However, the methods to incorporate such priors in a scalable and flexible manner are currently lacking. Bayesian methods are more successful in this regard. However, computational efficiency can often be a concern particularly for high-dimensional priors. We discuss hierarchical mixture models with empirical-bayes estimation for multiple-testing and variable selection problems. Our scalable 'prediction approach' fits the prior model efficiently using standard high-dimensional penalized regression techniques. We conducted re-weighted GWAS analysis of Posirasis, SLE, Type-2 Diabetes and Coronary Artery Disease, with various kinds of gene-level prior knowledge from pathway databases, GTEx, RegulomeDB etc. For the regression-likelihood in high dimensions, we use a novel MCMC-based strategy to allow sequential variable selection by

efficiently computing arbitrary posterior probabilities of variable subsets. We compare the variable-selection performance of our methods to some existing alternatives using simulations.

Anbupalam Thalamuthu, Centre for Healthy Brain Ageing, University of New South Wales, Australia

January 10, 4 pm – 5 pm

Title: Phenotypic and genetic profiling of exceptionally long-lived individuals

Abstract: Life expectancy has increased steadily in the last century due to improvements in lifestyle factors such as advancements in medical care. Consequently, there is a significantly higher proportion of older individuals in developed countries. However, aging is associated with functional decline, disease and comorbidity which in turn places a greater social, medical and economic burden on the community. Thus, better understanding the risk as well as uncovering protective factors for healthy ageing is vital. Exceptionally long-lived individuals can be thought of as model organisms to study successful healthy ageing. Here we describe cognitive and metabolic health profiles of participants in the Sydney Centenarian Study (SCS, n = 395, > 95 years). Also, we have examined the polygenic risk profiles for exceptional longevity, cardiovascular health and metabolic syndrome of the SCS cohort in relation to younger controls and compared and contrasted these results to other cohorts.

Partha P. Majumder, National Institute of Biomedical Genomics, India

January 10, 5 pm – 6 pm

Title: Joint analysis of multi-omics data: some applications

Abstract: In our quest to identify the major genomic forces that drive cancer of the oral cavity – a predominant form of cancer in India, we have generated data on genome-wide genomic and epigenomic alterations and also expression dysregulation of genes. We have carried out analyses of these individual data sets and have also performed a joint analysis of these data. I shall present the results of our analyses and also emphasize that the methods that we have used and are currently being use elsewhere are ad hoc from a statistical point of view. Better statistical modeling and methodology to carry out joint analysis of these data are required.

Yang Ni, Texas A&M University, USA

January 11, 8 am – 9 am

Title: Ordinal Causal Discovery for Reverse-Engineering Gene Regulatory Networks

Abstract: Categorical data frequently arise in multi-omics such as single-nucleotide polymorphisms or discretized gene/protein expression. Causal discovery for purely observational, categorical data is a long-standing challenging problem. The vast majority of existing methods focus on inferring the Markov equivalence class, which leaves the direction of some causal relationships undetermined. This paper proposes an identifiable ordinal causal discovery method that exploits the ordinal information contained in many real-world applications including genomic studies to uniquely identify causal structure. Score-and-search algorithms are developed for structure learning. The proposed method is applicable beyond ordinal data via data discretization. Through real-world genomic applications and synthetic experiments, we demonstrate that the proposed ordinal causal discovery method has favorable and robust performance compared to state-of-the-art alternative methods in both ordinal categorical and non-categorical data.

Matthew Stephens, University of Chicago, USA

January 11, 9 am – 10 am

Title: A simple new approach to variable selection in regression, with application to genetic fine-mapping

Abstract: We introduce a simple new approach to variable selection in linear regression, and to quantifying uncertainty in selected variables. The approach is based on a new model -- the ``Sum of Single Effects'' (SuSiE) model -- which comes from writing the sparse vector of regression coefficients as a sum of ``single-effect'' vectors, each with one non-zero element. We also introduce a corresponding new fitting procedure -- Iterative Bayesian Stepwise Selection (IBSS) -- which is a Bayesian analogue of stepwise selection methods. IBSS shares the computational simplicity and speed of traditional stepwise methods, but instead of selecting a single variable at each step, IBSS computes a *distribution* on variables that captures uncertainty in which variable to select. The method leads to a convenient, novel, way to summarize uncertainty in variable selection, and provides a Credible Set for each selected variable. Our methods are particularly well suited to settings where variables are highly correlated and true effects are sparse, both of which are characteristics of genetic fine-mapping applications. We demonstrate through numerical experiments that our methods outperform existing methods for this task.

Suprateek Kundu, The University of Texas MD Anderson Cancer Center, USA

January 11, 10 am – 11 am

Title: Scalable Bayesian Variable Selection for Structured High-Dimensional Data

Abstract: Variable selection for structured covariates lying on an underlying known graph is a problem motivated by practical applications, and has been a topic of increasing interest. However, most of the existing methods may not be scalable to high dimensional settings involving tens of thousands of variables lying on known pathways such as the case in genomics studies. We propose an adaptive Bayesian shrinkage approach which incorporates prior network information by smoothing the shrinkage parameters for connected variables in the graph, so that the corresponding

coefficients have a similar degree of shrinkage. We fit our model via a computationally efficient expectation maximization algorithm which is scalable to high dimensional settings (p~100,000). Theoretical properties for fixed as well as increasing dimensions are established, even when the number of variables increases faster than the sample size. We demonstrate the advantages of our approach in terms of variable selection, prediction, and computational scalability via a simulation study, and apply the method to a cancer genomics study.

Guolian Kang, St. Jude Children's Research Hospital, USA

January 11, 11 am – 12 pm

Title: Fancy and powerful study design is cost beneficial only coupled with valid or versatile statistical approaches

Abstract: Genome-wide association studies (GWAS) or next generation sequencing (NGS) studies have been successful in the last decades to identify genetic variants associated with common or rare diseases. One of the powerful study designs commonly used is extreme phenotype sequencing/genotyping design (EPS) for studying an ordinal or continuous phenotype as the primary outcomes of interest, such as the well-known National Heart, Lung, and Blood Institute Exome Sequencing Project. Besides the primary outcome, extensive data on vital clinical, treatment, and environmental factors etc are readily available. Secondary data analyses provide a mechanism for researchers to address high impact questions that would otherwise be prohibitively expensive and time-consuming to study.

Defining a clear and clinically relevant research question or topic might be easy based on the existing data but the statistical analysis approaches needed to answer the questions are not straightforward for EPS design. The naïve methods lead to biased estimates for secondary data analysis if the GWAS/NGS samples are not a random representative sample for the outcome of interest based on the research question or topic. Therefore, the critical question is how to conduct secondary data analysis in post-GWAS/NGS era on the data collected under EPS design?

Here, I will discuss three main scientific questions: 1) how to conduct whole-genome secondary genetic association analysis under EPS (STEPS); 2) how to conduct genome-wide mediation analysis under EPS (GMEPS); 3) how to conduct mendelian randomization analysis under EPS (MREPS). Three novel valid and versatile statistical approaches were proposed to tackle the EPS design issue for these three different statistical analyses. Extensive simulations and real data analysis showed the striking superiority of the proposed approaches over their alternatives under EPS and demonstrated compatible capabilities under the general random sampling framework. All these proposed approaches could also be readily applied to tackle relevant questions in any data collected under extreme-value sampling design in any modern epidemiology or clinical studies.

Terry Speed, Walter and Eliza Hall Institute for Medical Research, Australia

January 11, 3 pm – 4 pm

Title: RUV-III: Removing Unwanted Variation in III steps

Abstract: The method we call RUV-III was introduced several years ago in a soon to be published monograph written jointly with Johann Gagnon-Bartsch and Laurent Jacob. There we demonstrated its effectiveness with microarray gene expression data. The method makes essential use of technical replicates and negative controls, both of which yield information about unwanted variation. It was first published in 2019 jointly with Ramyar Molania and others in a paper about normalizing gene expression data with from the Nanostring platform. In that paper the use of pseudo-replicates was shown to be an effective supplement to, even replacement of technical replicates. More recently Molania and others showed the value of the method for normalizing bulk RNA-seq gene expression data, where pseudo-samples were also used. At the same time Salim and others obtained a version of RUV-III suitable for single cell RNA-seq data called RUV-III-NB. There pseudo-replicates are a necessity, and there also is a reasonable notion of pseudo-cell generalizing bulk pseudo-samples. In this talk I will review the theory of RUV-III and discuss the key choices that must be made to use it effectively.

Agus Salim, Melbourne School of Population and Global Health, University of Melbourne, Australia

January 11, 4 pm – 5 pm

Title: RUV-III-NB: Normalization of single-cell RNA-seq Data

Abstract: Despite numerous methodological advances, the normalization of single cell RNA-seq (scRNA-seq) data remains a challenging task. The performance of different methods can vary greatly across datasets. Part of the reason for this is the different kinds of unwanted variation, including library size, batch and cell cycle effects, and the association of these with the biology embodied in the cells. A normalization method that does not explicitly account for cell biology risks removing some of the signal of interest. Here we propose RUV-III-NB, a statistical method that can be used to adjust counts for library size and batch effects. The method uses the concept of pseudo-replicates to ensure that relevant features of the unwanted variation are only inferred from cells with the same biology.

Results: Using five publicly avail-able datasets that encompass different technological platforms, kinds of biology and levels of association between biology and unwanted variation, we show that RUV-III-NB manages to remove library size and batch effects, strengthen biological signals, improve differential expression analyses, and lead to results exhibiting greater concordance with independent datasets of the same kind. The performance of RUV-III-NB is consistent across the five datasets and is not sensitive to the number of factors assumed to contribute to the unwanted variation. It also shows promise for removing other kinds of unwanted variation such as platform effects.

Availability: The method is implemented as a publicly available R package available from https://github.com/limfuxing/ruvIIInb.

Contact: salim.a@unimelb.edu.au, terry@wehi.edu.au

Ramyar Molania, Walter and Eliza Hall Institute for Medical Research, Australia

January 11, 5 pm – 6 pm

Title: Removing unwanted variation from large-scale cancer RNA-sequencing data

Abstract: The accurate identification and effective removal of unwanted variation are essential to derive meaningful biological results from RNA-seq data, especially when the data come from large and complex studies. We have used The Cancer Genome Atlas (TCGA) RNA-seq data to show that library size, batch effects, and tumor purity are major sources of unwanted variation across all TCGA RNA-seq datasets and that existing gold standard approaches to normalizations fail to remove this unwanted variation. Additionally, we illustrate how different sources of unwanted variation can compromise downstream analyses, including gene co-expression, association between gene expression and survival outcomes, and cancer subtype identifications. Here, we propose the use of a novel strategy, pseudo-replicates of pseudo-samples (PRPS), to deploy the Removing Unwanted Variation III (RUV-III) method to remove different sources of unwanted variation from large and complex gene expression studies. Our approach requires at least one roughly known biologically homogenous subclass of samples shared across sources of unwanted variation. To create PRPS, we first need to identify the sources of unwanted variation, which we will call batches in the data. Then the gene expression measurements of biologically homogeneous sets of samples are averaged within batches, and the results called pseudo-samples. Pseudo-samples with the same biology and different batches are then defined to be pseudo-replicates and used in RUV-III as replicates. The variation between pseudo-samples of a set pseudo-replicates is mainly unwanted variation. We illustrate the value of our approach by comparing it to the TCGA normalizations on several TCGA RNA-seq datasets. RUV-III with PRPS can be used for any large genomics project involving multiple labs, technicians, or platforms.

Eleanor Feingold, University of Pittsburgh, USA

January 12, 8 am – 9 am

Title: Title: Are there still open statistical questions in modern genomics?

Abstract: Genomic studies are increasingly driven by bioinformatic data. This is an incredibly valuable advance in the field - a terrific change from the days when purely statistical approaches left scientists with enormous linkage or association regions and no way to narrow them down. But what is the place of statistics in this new genomic era? I will discuss several important practical problems in contemporary genomic studies that require statistical thought and statistical approaches. These

include approaches to combining samples and datasets, and to understanding the implications of performing millions of tests across the genome.

Veera Baladandayuthapani, University of Michigan, USA

January 12, 9 am – 10 am

Title: Personalized Integrated Network Estimation

Abstract: Personalized (patient-specific) approaches have recently emerged with a precision medicine paradigm that acknowledges the fact that molecular pathway structures and activity might be considerably different within and across patient populations. In the context of cancer, the functional cancer genome and proteome provide rich sources of information to identify patient-specific variations in signaling pathways and activities within and across tumors; however, current analytic methods lack the ability to exploit the diverse and multi-layered architecture of these complex biological networks. We consider the problem of modeling conditional independence structures in heterogenous data using Bayesian graphical regression techniques that allows patient-specific network estimation and inferences. We propose a novel specification of a conditional (in)dependence function of patient-specific covariates—which allows the structure of a graph to vary flexibly with the covariates; imposes sparsity in both edge and covariate selection; produces both subject-specific and predictive graphs; and is computationally tractable.

Cheng Cheng, St. Jude Children's Research Hospital, USA

January 12, 10 am – 11 am

Title: Genomic Determination Index

Abstract: In this presentation we introduce and discuss a new concept called Genomic Determination Index (GeDI), to address the questions of how much variability in a phenotype can be determined by large sets of diverse omics factors totaling possibly up to a million, and which specific omics factors are largely responsible for the explained phenotypic variation. In a way similar to heritability, GeDI is a measurement of the proportion of the phenotype variance attributable to the variations in a set of omics factors under an assumed population model. No existing large-scale sparse regression method or mixed effect models can effectively and fully address this problem. A method to estimate GeDI will be presented. This method consists of three steps: initial variable screening, regression modelling with forward variable selection drive by increments in GeDI, and a permutation analysis to correct selection bias. The entire development will be illustrated and evaluated by a diverse dataset from a study of ex vivo sensitivity of acute lymphoblastic leukemia cells to glucocorticoid treatment. If time permits, connection with the recent work on variable importance (Williamson et al. (2021) Biometrics, 77:9-22) will also be discussed.

Saumyadipta Pyne, University of California Santa Barbara, USA

January 12, 11 am – 12 pm

Title: Systematic spatial modeling of the heterogeneity in tumor signaling landscapes

Abstract: Intratumor heterogeneity is a complex phenomenon that could enable tumor cells to develop new therapy-resistant phenotypes and lead to poor outcomes for patients with cancer. Tumor stroma, for instance, comprises, in addition to the extracellular matrix, various cell types including cancer-associated fibroblasts (CAF), which represent a highly heterogeneous cell population due to their diverse cell types of origin. Dense solid tumors often have high CAF content, which can stimulate tumor cell proliferation by providing various growth factors, hormones and cytokines in a dynamic and context-dependent manner. Previously, we have shown how the spatial proximity to CAFs impacts the local molecular features and therapeutic sensitivity of breast cancer cells, thus influencing clinical outcomes. Recently, we have developed a computational framework for multivariate spatial mixture modeling of different CAF signaling landscapes in a given tumor. It allows us to test for local enrichment of key oncogenic pathways induced by CAF phenotypes that are specific to microenvironments. Further, the fitted spatial distributions provide a precise and parametric understanding of intratumor heterogeneity at single cell level.

Arnab Kumar Maity, Pfizer Inc, USA

January 12, 3 pm – 4 pm

Title: Bayesian structural equation modeling in multiple omics data integration with application to circadian genes

Abstract: It is well known that the integration among different data-sources is reliable because of its potential of unveiling new functionalities of the genomic expressions, which might be dormant in a single-source analysis. Moreover, different studies have justified the more powerful analyses of multi-platform data. Toward this, in this study, we consider the circadian genes' omics profile, such as copy number changes and RNA-sequence data along with their survival response. We develop a Bayesian structural equation modeling coupled with linear regressions and log normal accelerated failure-time regression to integrate the information between these two platforms to predict the survival of the subjects. We place conjugate priors on the regression parameters and derive the Gibbs sampler using the conditional distributions of them. Our extensive simulation study shows that the integrative model provides a better fit to the data than its closest competitor. The analyses of glioblastoma cancer data and the breast cancer data from TCGA, the largest genomics and transcriptomics database, support our findings.

Mayetri Gupta, University of Glasgow, UK

January 12, 4 pm – 5 pm

Title: Bayesian hierarchical mixture-based clustering for non-normal, noisy genomic datasets

Abstract: Clustering to find subgroups with common features is often a necessary first step in the statistical modelling and analysis of large and complex genomic datasets. Although follow-up analyses often make use of complex statistical models that are appropriate for the specific application, most popular clustering approaches are either nonparametric, or based on Gaussian mixture models and their variants, often for reasons of computational efficiency. Certain characteristics in the data, such as the presence of outliers, or non-ellipsoidal cluster shapes, that are common in datasets from genomics, often lead these methods to fail to detect the cluster components accurately. In this talk, we present three new efficient and robust Bayesian approaches to cluster datasets overcoming these limitations– (i) a model-based "tight" clustering approach to cluster points in the presence of outliers, (ii) a robust hierarchical scale-mixture-based approach for severely non-normal components, and (iii) a clustering approach based on a recently introduced family of geometric skew normal distributions, that can handle skewness as well as multimodality in clusters. The performance of these methods is illustrated through simulation studies and genomics applications, including the genotyping of single nucleotide polymorphisms (SNPs) in genome-wide association studies.

Sounak Chakraborty, University of Missouri, USA

January 12, 5 pm – 6 pm

Title: Title: Bayesian Nonlinear EM Based Approach for Analysis of Multi-Platform Genomics Data

Abstract: More recently different platforms have been brought together on the same patient set. Each output from different platforms could provide a different and complementary view of the whole genome. Thus, borrowing information across platforms has become more crucial. Therefore, in a given multi-platform genomic data, identifying important genes that have significant association with the clinical outcome through integration of mRNA expression level and other output of different types of platforms is of the greatest interest. In this paper we focus on building a nonlinear model that can simultaneously incorporate the multi-platform information and identify significant linear and non-linear gene effects associated with the clinical outcome. For gene-selection from a large set we consider the EMVS methodology and Bayesian LASSO with modified NEG (Normal-Exponential-Gamma) prior, both of these method provide solution for large p scenarios. In addition to that each of these methods solve different issues as well, for the EMVS, it reduces the computation time drastically, and for the Bayesian LASSO with modified NEG prior, it is adapted to incorporate genetic grouping information.

Jeffrey Morris, University of Pennsylvania, USA

January 13, 8 am – 9 am

Title: Top-Down Integrative Genomics for Colon Cancer Precision Therapeutics

Abstract: In recent decades, the world of cancer research has become completely transformed by the development of hypersensitive technologies taking detailed biological measurements of quantities previously unmeasurable, including multi-platform genomics data containing complementary molecular information at the DNA, RNA, protein, and epigenetic levels.

These data contain biological insights into molecular-based diseases like cancer and the extraction of this knowledge from the data can lead to novel precision therapeutic strategies with the potential to change clinical practice. Successful knowledge extraction from these data depends on integrative methods that can combine information across these complementary modalities to paint a more complete picture of the underlying biology. I will discuss several integrative genomic methods in the context of colorectal cancer. Our research group has used a top-down integrative learning approach to discover, validate, and translate new precision therapeutic concepts for colorectal cancer.
These efforts heavily depend on our ability to use integrative genomics methods to deeply characterize the molecular characteristics of four recently discovered and validated subtypes of colorectal cancer (Guinney, et al. 2015 Nature Medicine, >2000 citations already) that are reshaping how the biomedical community defines and studies colorectal cancer. I will describe the big picture schema underlying our integrative learning approach that can serve as a template for other settings and give examples of some innovative statistical methods we have developed, are developing, and will develop and use in this process.

Sanjay Shete, MD Anderson Cancer Center, USA

January 13, 9 am – 10 am

Title: Quantifying the reversible relationship between obesity and diabetes using bidirectional mediation models and Approaches for bi-directional Mendelian Randomization

Abstract: Obesity and diabetes are both major public health issues and are risk factors for each other. Recent genome wide studies have identified several genetic variants associated with either obesity, diabetes, or both. Because of the known interdependence between obesity and diabetes we hypothesize that some of these SNPs that are associated with both obesity and diabetes may be mediated through obesity to affect diabetes or through diabetes to affect obesity. Identifying these mediated relationships would further enhance our knowledge about diabetes and obesity. We propose a framework for performing bidirectional mediation analyses. Through simulations, we showed the statistical properties of our methods. In many scenarios, the residuals of mediator and outcome are correlated because of unmeasured predictors not included in the mediation model. We showed that the proposed model gives unbiased estimates in this scenario too. We also propose two novel Mendelian randomization methods for bidirectional model: BiRatio and BiLIML extended from Ratio and limited information maximum likelihood (LIML) methods, respectively. Our simulations show that BiRatio and BiLIML methods provide accurate estimation of causal effects with unidirectional or bidirectional underlying causal relationships when strong IVs are used for estimations. When weak IVs are used, the BiLIML method provides the least biased estimation of causal effects. We applied the proposed model to investigate the bidirectional relationship between the diabetes and obesity using the genome wide association study data from the MESA cohort. We identified 6 SNPs that were associated with both diabetes and obesity. Two SNPs (rs3752355 and rs6087982) had indirect effects on obesity mediated through diabetes (0.28; 95% CI [0.01, 0.67] and

0.36; 95% CI [0.08, 0.85], respectively). The remaining four SNPs (rs7969190, rs4869710, rs10201400 and rs12421620) directly affect diabetes and obesity without any mediation effects.

Junmin Peng, St. Jude Children's Research Hospital, USA

January 13, 10 am – 11 am

Title: High Throughput Proteomics to Basic and Clinic Research

Abstract: Our mission is to develop novel mass spectrometry-based technologies, including proteomic/metabolomic profiling, protein modification analysis, structural proteomics and bioinformatics software. We also actively apply these technologies in the fields of Alzheimer's disease, cancer and ubiquitin biology. We seek to obtain the full spectra of temporal, spatial and structural omics data from cellular/animal/clinical samples. Integration of such multi-omics data offer a systems or holistic view for unbiased identification of central disease networks, functional modules and master regulators. To validate the derived hypotheses from the big data analysis, we develop mouse models for perturbation with genetic methods for mechanistic and functional studies. These studies provide novel insights into the pathogenesis for therapeutic intervention and may discover disease biomarkers for precision medicine.

Derek Gordon, Rutgers University, Piscataway, USA

January 13, 11 am – 12 pm

Title: Impact of Heterogeneity on Genetic Association Sample size

Abstract:

Background: It is well-known that heterogeneity, or mixtures, can cause power loss, or equivalently, increase in minimal sample size, to detect genetic association for commonly used statistical methods. The purpose of our research is to quantify the power loss and/or sample size in the presence of genetic heterogeneity. We include parameters from genetic model-based and genetic model-free formulas in addition to the heterogeneity parameter.

Results: Heterogeneity dominates the power loss/minimal sample size. In fact, in a regression analysis, the heterogeneity parameter is essentially the only parameter needed to estimate minimal sample size in the presence of heterogeneity.

Conclusions: It is vitally important to consider heterogeneity parameters when designing studies of genetic association.

Qi Yan, Columbia University Irving Medical Center, USA

January 14, 8 am – 9 am

Title: Multi-omics data analysis in complex human diseases

Abstract: Genome-wide association study (GWAS) has been widely used to identify common single nucleotide polymorphisms (SNPs) associated with complex human diseases. However, this single-marker association test is not powerful to detect rare variants. To increase power, gene-based tests have been developed. In addition to SNPs, other omics data, such as gene expression and DNA methylation can be obtained from the same individuals, which offers researchers to study multi-omics data in an integrative manner. While large consortia (e.g., conducting GWAS or eQTL) usually do not share individual level data, it is relatively easy to access summary level data. More recently, researchers have developed methods to utilize these summarized genetic associations for different purposes, such as (1) polygenic risk score (PRS); (2) transcriptome-wide association study (TWAS); and (3) Mendelian randomization (MR) based causal inference.

Susmita Datta, University of Florida, USA

January 14, 9 am – 10 am

Title: Statistical Analysis of single cell RNA sequencing (ScRNA-seq) data

Abstract: Transcriptomic studies such as in bulk RNA-sequencing, one can examine transcript abundance measurements averaged over bulk populations of thousands (or even millions) of cells. While these measurements have been valuable in countless studies, they often conceal cell-specific heterogeneity in expression signals that may be paramount to new biological findings. Fortunately, with single cell RNA-sequencing (scRNA-Seq), transcriptome data from individual cells are now accessible, providing opportunities to investigate functional states of cells, identify rare cell populations and uncover diverse gene expression patterns in cell populations that seemed homogeneous. Most importantly it provides an unprecedented resolution to the characterization of cellular clinical isolates. However, there are challenges analyzing such scRNA-Seq data. Amongst many challenges the most significant are the bimodal or multimodal distribution, sparsity and tremendous heterogeneity in the data. Consequently, we will describe potential ways of statistical modeling of such data, finding differentially expressed genes and methods for constructing gene-gene interaction network using this data.

Saonli Basu, University of Minnesota, USA

January 14, 10 am – 11 am

Title: Estimating SNP heritability in presence of population substructure in biobank-scale datasets

Abstract: SNP heritability of a trait is measured as the proportion of total variance explained by the additive effects of genome-wide single nucleotide polymorphisms (SNPs). Linear mixed models are routinely used to estimate SNP heritability for many complex traits. This approach requires estimation of 'relatedness' among individuals in the sample, which is usually captured by estimating a genetic relationship matrix (GRM). Heritability is estimated by the restricted maximum likelihood (REML) or method of moments (MOM) approaches, and this estimation relies heavily on the GRM computed from the genetic data on individuals. The presence of population substructure in the data could significantly impact the GRM estimation and may introduce bias in heritability estimation. The common practice of accounting for such population substructure is to adjust for the top few principal components of the GRM as covariates in the linear mixed model. Here we propose an alternative way of estimating heritability in multi-ethnic studies. Our proposed approach is a MOM estimator derived from the Haseman-Elston regression and gives an asymptotically unbiased estimate of heritability in presence of population stratification. It introduces adjustments for population stratification in a second orgorder estimating equation and allows for the total phenotypic variance to vary by ethnicity. We study the performance of different MOM and REML approaches in presence of population stratification through extensive simulation studies. We estimate the heritability of height, weight, and other anthropometric traits in the UK Biobank cohort to investigate the impact of subtle population substructure on SNP heritability estimation.

This is joint work with Zhaotong Lin and Souvik Seal.

Qian Li, St. Jude Children's Research Hospital, USA

January 14, 11 am – 12 pm

Title: A joint nested random effects model for metagenomic trajectory analysis with disease outcome in matched sets

Abstract: Many pediatric studies had found operational taxonomy units (OTUs) in children's gut microbiota predictive of hosts growth and disease status. Most metagenomic trajectory analyses aimed to compare OTUs between exposure groups instead of disease outcomes and did not use matched sets. We proposed a joint mixed effect model (JointMatch) to detect OTUs longitudinally associated with disease outcome in matched participants. The disease status was predicted by logistic regression with two random effects nested in matched sets and subjects. The longitudinal relative abundance per OTU was fitted by a zero-inflated generalized mixed effect model with the above random effects. We used a matched-set-specific marginal likelihood and a Wald statistic to test the OTU-disease association. We simulated disease outcome and temporal high-dimensional metagenomic counts in matched sets to compare JointMatch with LMM and ZIBR, showing JointMatch yielding higher detection power and lower type I error rate. An application to the longitudinal gut microbiota of TEDDY participants showed that JointMatch identified more taxa in infant-age stool samples signaling the hosts' autoimmune status.