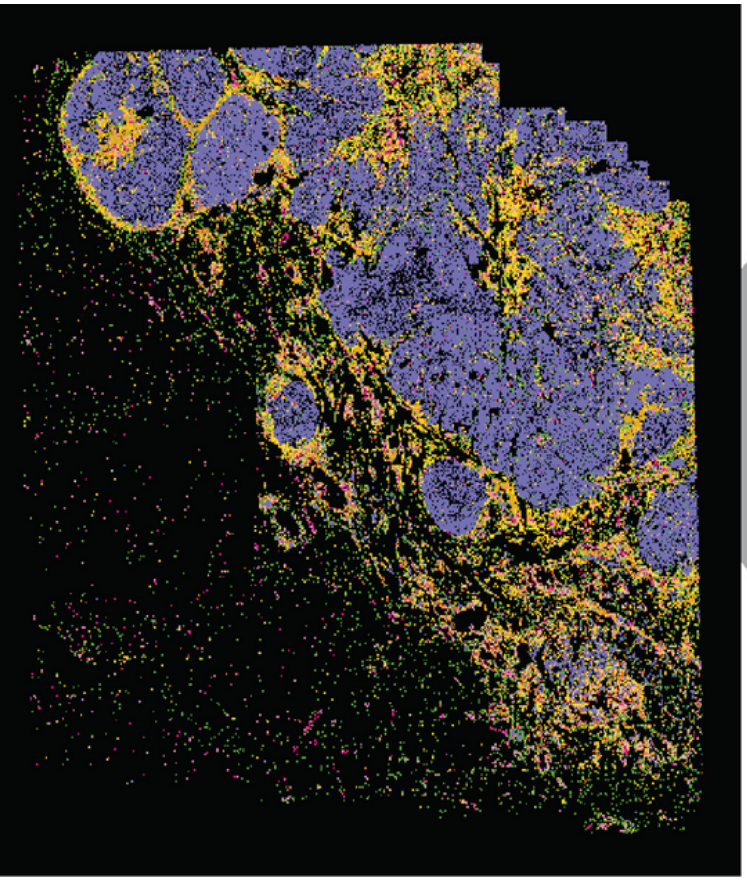
Single-Cell Plus: Data Science Challenges in Single-Cell Research

July 2-7, 2023

Organised by Jean Yang, Hongyu Zhao, Sunduz Keles, Sara Mostafavi, and Joshua Ho



Banff International Research Station
for Mathematical Innovation and Discovery

About the workshop

Cells are the fundamental building blocks of life. Recent advancement in biotechnology has allowed us to peek inside this every cell for better understanding of biology and human disease. Single-cell technology also generates big and complex data and brings about new data science challenges for computational and life scientists. The objective of this workshop is to bring together international leaders in diverse disciplines including mathematical, statistical, computational, biological and medical in a collaborative atmosphere to develop the collaborative capacity that will tackle various underlying data science challenges in single-cell research.

About the Banff International Research Station

The Banff International Research Station for Mathematical Innovation and Discovery (BIRS) is a collaborative Canada-US-Mexico venture that provides an environment for creative interaction as well as the exchange of ideas, knowledge, and methods within the Mathematical Sciences, with related disciplines and with industry. The research station is located at The Banff Centre in Alberta and is supported by Canada's Natural Science and Engineering Research Council (NSERC), the U.S. National Science Foundation (NSF), Alberta's Advanced Education and Technology, and Mexico's Consejo Nacional de Ciencia y Tecnología (CONACYT).

Communication platform

We will be using slack as a communication platform for this week. Please join us on https://join.slack.com/t/singlecellplus/shared_invite/zt-1yozj53ps-XnyEG~bTSQnBiUK_5AAGNg

Monday 3 July

System biology, gene regulation and epigenomics
Sunduz Keles

Greater than the sum of the parts: Learning relationships between histone modifications in single cells

Jake Yeung
Institute of Science and Technology Austria (ISTA), Austria

Abstract: Detecting histone modifications in single cells by sequencing is still in its infancy but has the potential to unlock the full spectrum of different chromatin states in the genome of individual cells. More established single-cell sequencing technologies, such as mRNA-seq and ATAC-seq, interrogate only a tiny fraction of the genome. Progress therefore hinges on both new measurement methods that can map multiple epigenetic marks in single cells as well as new analysis methods that connect different modalities together. I will first present sortChIC, a method to map histone modifications in single cells from rare cell populations. During hematopoiesis, we find that repressive chromatin dynamics are qualitatively different from active ones: active chromatin states mainly distinguish mature blood cell types, while changes in repressive states mainly distinguish HSCs and mature cell types. These findings suggest that hematopoiesis requires overcoming heterochromatin barriers in a cell fate-independent manner. Next I will present scChIX-seq, a framework to generate multimodal histone modification data in single cells. We develop multimodal analysis methods that reveal genome regulation that would otherwise be missed from unimodal data. Overall, these integrated experimental and computational methods reveal dynamic relationships between chromatin states, and how those relationships change during differentiation.

The broad question of Jake Yeung's research is asking how different cells regulate their genomes to enable specializations into distinct cell types while maintaining robust gene regulatory dynamics that are preserved across nearly all cells in the body. To tackle this problem, he and his team combine experimental and computational techniques to measure, infer, and model how different histone marks modify the chromatin and regulate the genome.

Probabilistic modelling of single-cell methylation sequencing data reveals regions that are informative of cell type and cell state

Keegan Korthauer
University of British Columbia, Canada

Regions that exhibit heterogeneity in DNA methylation (DNAm) across cells may play a role in processes such as gene regulation, disease susceptibility and environmental influences. They may also act as predictive signatures of cell type or cell state. Single-cell bisulfite sequencing (scBS-seq) provides measurements of DNAm in individual cells, but the data are extremely sparse, typically with greater than 80% missing rate. We propose a novel computational tool for detection of

variably methylated regions (VMRs) in scBS-seq data. Our approach uses a probabilistic model to (1) leverage the correlation structure of nearby DNAm sites, and (2) pool information across cells to overcome the challenges of sparsity. Compared to VMRs detected by previous methods, our approach demonstrates increased clustering accuracy in simulations and a case study of mouse neuronal cells.

Keegan Korthauer is an assistant professor in the Department of Statistics at the University of British Columbia and an investigator at the British Columbia Children's Hospital Research Institute. Previously, she was a Postdoctoral Research Fellow at Dana-Farber Cancer Institute and Harvard T. H. Chan School of Public Health. She earned her PhD in Statistics from the University of Wisconsin Madison. Her research lies at the intersection of statistics and biology, and her group focuses on developing novel frameworks and rigorous inferential procedures that exploit the increased scope and scale of high-throughput sequencing data, with the ultimate goal of uncovering new molecular signals in cancer, child health, and development.

Modelling gene regulation via integrative analysis of single cell multi-omics data

Zhana Duren
Clemson University, USA

The accurate inference of context-specific Gene Regulatory Networks (GRNs) from genomics data is a crucial task in computational biology. However, current methods have limitations, including relying solely on gene expression data, lower resolution from bulk data, and limited data availability for certain cellular systems. To address these challenges, we developed a new method based on a lifelong neural network to infer high accuracy GRNs (LINGER) from single cell gene expression and chromatin accessibility data by leveraging atlas-scale external data. The LINGER model proposes a metric called the "pioneer index" to quantify the ability of transcription factors (TFs) to initiate chromatin remodeling, improving the accuracy and interpretability of the GRN. The LINGER method achieved 3 times higher accuracy compared to currently available methods and provided insights into the interpretation of disease-associated variants and genes, offering a comprehensive tool for inferring gene regulation from genomics data.

Zhana Duren is an Assistant Professor in the Department of Genetics and Biochemistry at Clemson University and has been affiliated with the Center for Human Genetics since 2020. He received his BS in Applied Mathematics from Beihang University in China in 2012, followed by a Ph.D. in Operational Research from the Applied Mathematics Institute at the Chinese Academy of Sciences in 2017, under the supervision of Dr. Yong Wang. From 2015 to 2020, he worked as a visiting Ph.D. student (2015-2017) and postdoctoral research fellow (2017-2020) in the lab of Professor Wing Hung Wong at Stanford University. Dr. Duren's research interests include computational regulatory genomics and statistical machine learning. In addition to his research, Dr. Duren is committed to mentoring and training the next generation of scientists and regularly participates in outreach activities to promote science education and engagement.

Scalable test of statistical significance for protein-DNA binding changes with insertion and deletion of bases in the genome

Sunyoung Shin
Pohang University of Science and Technology (POSTECH), South Korea

Mutations in the noncoding DNA, which represents approximately 99% of the human genome, have been crucial to understanding disease mechanisms through dysregulation of disease-associated genes. One key element in gene regulation that noncoding mutations mediate is the binding of proteins to DNA sequences. Insertion and deletion of bases (InDels) are the second most common type of mutations, following single nucleotide polymorphisms, that may impact protein-DNA binding. However, no existing methods can estimate and test the effects of InDels on the process of protein-DNA binding. We develop a novel test of statistical significance, namely the binding change test (BC test), using a Markov model to evaluate the impact and identify InDels altering protein-DNA binding. The test predicts binding changer InDels of regulatory significance with an efficient importance sampling algorithm generating background sequences in favor of large binding affinity changes. Simulation studies demonstrate its excellent performance. The application to human leukemia data uncovers candidate pathological InDels on modulating MYC binding in leukemic patients. We develop an R package atIndel, which is available on GitHub.

Sunyoung Shin is Associate Professor in the Department of Mathematics at Pohang University of Science and Technology (POSTECH), South Korea. Prior to that, she worked in the Department of Mathematical Sciences at University of Texas at Dallas as an Assistant Professor for five years. Her research interests focus on developing and analyzing statistical models and methods that can be used for big data problems. She also develops computational and statistical tools for large-scale genomic data analysis and integration to uncover biological mechanisms of diseases and traits.

scTIE: a unified framework for data integration and inference of gene regulation using single-cell temporal multimodal data

Rachel Wang
University of Sydney, Australia

Single-cell technologies offer unprecedented opportunities to dissect gene regulatory mechanisms in context-specific ways. Although an increasing number of computational methods have been developed for inferring gene regulatory relationships from scRNA-seq and scATAC-seq data, the data integration problem, essential for accurate cell type identification, has been mostly treated as a standalone challenge. I will present scTIE, a unified method that integrates temporal multimodal data and infers regulatory relationships predictive of cellular state changes. scTIE uses an autoencoder to embed cells from all time points into a common space using iterative optimal transport, followed by extracting interpretable information to predict cell trajectories. Using a variety of synthetic and real temporal multimodal datasets, we demonstrate scTIE achieves effective data integration while preserving more biological signals than existing methods, particularly in the presence

of batch effects and noise. Furthermore, on the exemplar multiome dataset we generated from differentiating mouse embryonic stem cells over time, we demonstrate scTIE captures regulatory elements highly predictive of cell transition probabilities, providing new potentials to understand the regulatory landscape driving developmental processes.

Dr Rachel Wang is a Senior Lecturer in the School of Mathematics and Statistics at the University of Sydney. Her research interests lie broadly in statistical network modelling, statistical machine learning, and applications in genomics data. She obtained her PhD in statistics from Berkeley before completing a Stein fellowship at Stanford. She was an ARC DECRA fellow, a Harrington faculty fellow at UT Austin, and a recipient of the Moran medal in statistical sciences.

Delineate the regulatory map in silico

Ge Gao
Peking University

Human individual cells, as the basic biological units of our bodies, carry out their functions through rigorous regulation of gene expression and exhibit heterogeneity among each other in every human tissue. In addition to identifying individual genes, one is often interested in how multiple genes interact to form regulatory circuits and carry out cellular functions. Combining massive omics data and leading-edge statistical modeling/machine learning approaches, we have developed a set of novel bioinformatic technologies to delineate the regulatory map and characterize the functional genome in action globally during past years. Here we will present our recent advances as well as their potential applications in clinical and translational study.

As a bioinformatician, Ge Gao is interested in developing novel computational technologies to analyze, integrate and visualize high-throughput biological data effectively and efficiently, with applications to decipher the function and evolution of gene regulatory systems. In particular, he has been working on delineating the cellular regulatory map and characterizing the functional genome in action globally during past decades, by combining massive (single-cell) omics data and leading-edge statistical modeling/machine learning approaches.

Round table discussion: Integration of GWAS with scATAC-seq data
Chaired by Hongyu Zhao

Modelling of cellular dynamics on differentiation and lineage

Yuanhua Huang
University of Hong Kong

The recent advances in single-cell RNA-seq technologies offer a promising way to dissect the cellular dynamics of both differentiation and lineages. However, various statistical and computational challenges exist in inferring these temporal latent variables or structures. In this talk, we will introduce the recent methodology progress in the single-cell RNA velocity field and discuss a few potential strategies

that may further enhance the robustness to be applicable for a broad range of biological systems. We will also introduce how lineage reconstruction techniques may elucidate the clonal preference in cell fate decisions.

Dr Huang is an assistant professor in the School of Biomedical Sciences and the Department of Statistics and Actuarial Science at the University of Hong Kong (HKU). His lab works on statistical machine-learning models for single-cell genomic data and is supported by NSFC Young Excellent Scientist Fund.

Joint tensor modeling of single cell 3D genome and epigenetic data with Muscle

Kwangmoon Park
University of Wisconsin-Madison, USA

Emerging single cell technologies that simultaneously capture long-range interactions of genomic loci together with their DNA methylation levels are advancing our understanding of three-dimensional genome structure and its interplay with the epigenome at the single cell level. While methods to analyze data from single cell high throughput chromatin conformation capture (scHi-C) experiments are maturing, methods that can jointly analyze multiple single cell modalities with scHi-C data are lacking. Here, we introduce Muscle, a semi-nonnegative joint decomposition of Multiple single cell tensors, to jointly analyze 3D conformation and DNA methylation data at the single cell level. Muscle takes advantage of the inherent tensor structure of the scHi-C data, and integrates this modality with DNA methylation. We developed an alternating least squares algorithm for estimating Muscle parameters and established its optimality properties. Parameters estimated by Muscle directly align with the key components of the downstream analysis of scHi-C data in a cell type specific manner. Evaluations with data-driven experiments and simulations demonstrate the advantages of the joint modeling framework of Muscle over single modality modeling or a baseline multi modality modeling for cell type delineation and elucidating associations between modalities.

Kwangmoon Park is a third-year Statistics Ph.D. student at University of Wisconsin-Madison. He is currently working on statistical genomics and high dimensional statistics with Professor Sündüz Keleş. Before joining UW-Madison, he earned a Master's degree in Statistics at the Yonsei University in 2020 and Professor Seung-Ho Kang was his academic advisor. He earned his B.A. in Economics and Statistics at Yonsei University in Korea. He is mainly interested in questions related to understanding how genes are regulated by distal regions in the genome, particularly by functional non-coding regions. For that purpose, he develops statistical tools for analyzing High-dimensional genomic data including Hi-C and HiCHIP, and devises methods to link diverse types of genomic data with better statistical interpretation.

Combinatorial regulons (cregulon): a novel optimization model for unraveling cellular identity and state transitions through single multi-omics data

Yong Wang
Chinese Academy of Sciences, China

We propose combinatorial regulon (cRegulon) to model the combinations among TFs, which can better characterize cell types and serves as the driving forces for cell state transitions. By leveraging rapidly accumulated single multi-omics data, we develop an optimization model to systematically infer cRegulons (i.e., the representative TF modules, their associated regulatory elements and target genes formed regulatory network). In our approach, cRegulon is jointly reconstructed from i) identifying TF modules from TF combinatorial network, ii) explaining gene expression of scRNA-seq data, and iii) explaining gene activity of scATAC-seq data. Therefore, the inferred cRegulons provide details of how TF combinations utilize specific regulations to characterize the identity and transition of cell type/states.

Yong Wang is a Professor in the Institute of Applied Mathematics at Academy of Mathematics and Systems Science at Chinese Academy of Sciences (CAS). He is affiliated as a Professor in the National Center for Mathematics and Interdisciplinary Sciences (NCMIS) at Chinese Academy of Sciences, in the Center for Excellence in Animal Evolution and Genetics (CEAEG) at Chinese Academy of Sciences, in the School of Mathematics, University of Chinese Academy of Sciences at Chinese Academy of Sciences, and in the Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences at Chinese Academy of Sciences.

Tuesday July 4

Advances in single-cell RNA-Seq data
Sara Mostafavi

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

Hongkai Ji
Johns Hopkins University, USA

Pseudotime analysis with single-cell RNA-sequencing data has been widely used to study dynamic gene regulatory programs along continuous biological processes. While many methods have been developed to infer the pseudotemporal trajectories of cells within a biological sample, it remains a challenge to compare pseudotemporal patterns with multiple samples (or replicates) across different experimental conditions. 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 while adjusting for batch effects, and to detect changes in gene expression, cell density, and topology of a pseudotemporal 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 real scRNA-seq and simulation 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.

Dr. Ji is a Professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. His research is focused on developing statistical and computational methods and software tools for analyzing high-throughput functional genomic and single cell genomic data. He applies these tools to study gene regulation, development, immuno-oncology and infectious disease.

The curses of performing differential expression analysis using single cell data

Mengjie Chen
University of Chicago, USA

Differential expression analysis in single-cell transcriptomics provides essential insights into cell-type-specific responses to internal and external stimuli. While many methods are available to identify differentially expressed genes from single-cell transcriptomics, recent studies raise important concerns about the performance of state-of-the-art methods. As single-cell studies are scaled up to population-level quickly, powerful and accurate methods will be essential for obtaining meaningful results. In this context, we highlight various limitations and conceptual flaws in the

current workflows for single-cell differential expression analysis. Furthermore, we present a new paradigm that offers a potential solution to these issues.

Mengjie Chen is an Associate Professor in Section of Genetic Medicine in the Department of Medicine and the Department of Human Genetics at the University of Chicago.

A unified, reference-free analytic method revealing unseen transcript diversification in single cells

Julia Salzman
Stanford University, USA

Myriad mechanisms diversify the sequence content of eukaryotic transcripts and are of great interest to single cell biology. Currently, these events are detected using fragmented bioinformatic tools that require a predefined form of transcript diversification and rely on alignment to an incomplete reference genome, filtering out unaligned sequences which can be among the most interesting. I will present collaborative work based on a new statistics-first analytic method that performs unified, reference-free inference directly on raw sequencing reads. This method discovers and novel examples of transcript diversification in single cells, bypassing genome alignment and cell type metadata, which is impossible with current algorithms. Applying to 10,326 primary human single cells in 19 tissues profiled with SmartSeq2, we discover a set of splicing and histone regulators with highly conserved intronic regions that are themselves targets of complex splicing regulation, unreported transcript diversity in the heat shock protein HSP90AA1, and diversification in centromeric RNA expression, V(D)J recombination, RNA editing, and repeat expansions missed by existing methods.

Julia Salzman is an Associate Professor in the Department of Biomedical Data Science, Biochemistry and Statistics (by Courtesy). She joined the Stanford faculty in 2013. As a postdoctoral scholar in Patrick Brown's lab, Salzman developed statistical algorithms that led to the discovery of a ubiquitous expression of circular RNA missed by other computational and experimental approaches for decades. Her research spans the interface of statistical methodology and genomics aiming to use data driven experiments to uncover organizing principles of biological regulation.

Similarity-assisted variational autoencoder for nonlinear dimension reduction with application to single-cell RNA sequencing data

Hyonho Chun
Korea Advanced Institute of Science & Technology (KAIST), South Korea

Deep generative models naturally become nonlinear dimension reduction tools to visualize large-scale datasets such as single-cell RNA sequencing datasets for revealing latent grouping patterns or identifying outliers. The Variational autoencoder (VAE) is a popular deep generative method equipped with encoder/decoder structures. The encoder and decoder are useful when a new sample is mapped to the latent space and a data point is generated from a point in a latent space. However, the VAE tends not to show grouping pattern clearly without additional

annotation information. On the other hand, similarity-based dimension reduction methods such as t-SNE or UMAP present clear grouping patterns even though these methods do not have encoder/decoder structures. To bridge this gap, we propose a new approach that adopts similarity information in the VAE framework. In addition, for biological applications, we extend our approach to a conditional VAE (CVAE) to account for covariate effects in the dimension reduction step.Our method is able to produce clearer grouping patterns than those of other regularized VAE methods by utilizing similarity information encoded in the data via the highly celebrated UMAP loss function.

Hyonho Chun is an Associate Professor with interests in nonlinear dimension reduction for single-cell data.

RUV-III-NB: A robust method for normalization of single cell RNA-seq data

Agus Salim
University of Melbourne, Australia

Normalization of single cell RNA-seq data remains a challenging task. The performance of different methods can vary greatly between datasets when unwanted factors and biology are associated. Most normalization methods also only remove the effects of unwanted variation for the cell embedding but not from gene-level data typically used for differential expression (DE) analysis to identify marker genes. We propose RUV-III-NB, a method that can be used to remove unwanted variation from both the cell embedding and gene-level counts. Using pseudo-replicates, RUV-III-NB explicitly takes into account potential association with biology when removing unwanted variation. The method can be used for both UMI or read counts and returns adjusted counts that can be used for downstream analyses such as clustering, DE and pseudotime analyses. Using published datasets with 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 DE analyses, and lead to results exhibiting greater concordance with independent datasets of the same kind. The performance of RUV-III-NB is consistent and is not sensitive to the number of factors assumed to contribute to the unwanted variation.

A/Prof Agus Salim is a Biostatistician with extensive experience in developing and applying novel statistical methods for "Big" biomedical Data, including single-cell omics, continuous glucose monitoring (CGM) and data from wearable activity trackers. Within single-cell omics, his main research interest lies in development and evaluation of computational methods for normalization and differential expression of single-cell RNA-seq data. He has worked mostly within the space of traditional statistical methods, most notably generalized linear models but more recently, he has also been interested in combining these traditional methods with variational autoencoders for better preservation of biological signals.

Modelling group heteroscedasticity in single-cell RNA-seq pseudo-bulk data

Matthew Ritchie
Walter and Eliza Hall Institute of Medical Research, WEHI, Australia

Group heteroscedasticity is commonly observed in pseudo-bulk single-cell RNA-seq datasets and its presence can hamper the detection of differentially expressed genes. Since most bulk RNA-seq methods assume equal group variances, we introduce two new approaches that account for heteroscedastic groups, namely voomByGroup and voomWithQualityWeights using a blocked design (voomQWB). Compared to current gold-standard methods that do not account for group heteroscedasticity, we show results from simulations and various experiments that demonstrate the superior performance of voomByGroup and voomQWB in terms of error control and power when group variances in pseudo-bulk single-cell RNA-seq data are unequal.

Professor Matt Ritchie is a Laboratory Head in the Epigenetics and Development Division at WEHI in Melbourne, Australia. He is an experienced computational researcher, with skills in analysing high-throughput sequencing data and developing statistical methods implemented in open-source R/Bioconductor software for Bioinformatics analysis. Through close collaboration with bench scientists and clinicians, his analyses have revealed new insights into epigenetic processes, hematopoiesis, apoptosis and mechanisms for treatment resistance in cancer.

One of these cells is not like the other - modelling variability of gene expression in single cell data

Jessica Mar
University of Queensland, Australia

Gene expression changes underpin the regulation of almost all cellular phenotypes in nature. While we typically focus on changes in average gene expression, we know that changes in gene expression variability can also impact regulation too. Single cell data has provided an incredible opportunity to study how the variability of gene expression impacts a cell population. But like any data science question, there are challenges in how to model variability in single cell data. This talks highlights studies from my group which have focused on how to model heterogeneity in single cell RNA-seq data and its role in regulating phenotypes like ageing and differentiation.

Associate Professor Jessica Mar is a Group Leader at the University of Queensland in Brisbane, Australia. She leads a computational biology group that focuses on developing new methods related to variability in gene expression and investigating its relationship to cellular regulation. A/Prof Mar received her PhD in Biostatistics from Harvard University in 2008. She was a postdoctoral fellow at the Dana-Farber Cancer Institute in Boston (2008-11), and an Assistant Professor at Albert Einstein College of Medicine in New York (2011-2018) before relocating back to Australia.

Round table discussion: Grand challenges in single cell data
Chaired by Sunduz Keles and Jean Yang

Robust normalization and integration of single-cell protein expression across CITE-seq datasets

Zheng Ye

Fred Hutchinson Cancer Center, USA

CITE-seq technology enables the direct measurement of protein expression, known as antibody-derived tags (ADT), in addition to RNA expression. The increase in the copy number of protein molecules leads to a more robust detection of protein features compared to RNA, providing a deep definition of cell types. However, due to added discrepancies of antibodies, such as the different types or concentrations of IgG antibodies, the batch effects of the ADT component of CITE-seq can dominate over biological variations, especially for the across-study integration. We present ADTnorm as a normalization and integration method designed explicitly for the ADT counts of CITE-seq data. Benchmarking with existing scaling and normalization methods, ADTnorm achieves a fast and accurate matching of the negative and positive peaks of the ADT counts across samples, efficiently removing technical variations across batches. Further quantitative evaluations confirm that ADTnorm achieves the best cell-type separation while maintaining the minimal batch effect. Therefore, ADTnorm facilitates the scalable ADT count integration of massive public CITE-seq datasets with distinguished experimental designs, which are essential for creating a corpus of well-annotated single-cell data with deep and standardized annotations.

Ye Zheng is currently a Postdoctoral Research Fellow at the Fred Hutchinson Cancer Center in Seattle. She is supervised by Drs. Raphael Gottardo and Steven Henikoff from both quantitative modeling and molecular biology perspectives. Her work involves statistical modeling of single-cell transcriptomics, proteomics and epigenomics data with application to immunology and immunotherapy. Before her postdoctoral training, Ye received the Ph.D. in statistics from the University of Wisconsin-Madison. Her thesis focuses on statistical modeling and computational tools development to investigate the three-dimensional chromatin structure and the gene cis-regulation mechanism.

Improving the Resolution of Single-Cell TCR-seq

Kelly Street
University of Southern California, USA

T-cell receptors (TCRs) are hypervariable protein complexes that recognize foreign antigens and play an important role in immune response. Modern sequencing technology allows for the full characterization of these complexes at single-cell resolution and has the potential to serve as a broadly applicable diagnostic tool. However, single-cell TCR sequencing data is often ambiguous, making it difficult to differentiate between cells with distinct clonotypes. Many modern analyses have focused solely on cells with complete information, discarding ambiguous cells and thereby losing data. We propose an expectation maximization (E-M) algorithm for clonotype assignment, which leverages data from ambiguous cells to provide superior repertoire characterization.

Kelly Street is a biostatistician with a focus on computational methods for single-cell genomics. He is an Assistant Professor of Population and Public Health Sciences in the Division of Biostatistics at the Keck School of Medicine of USC. He has worked on trajectory inference, immune repertoire profiling, and multi-omic integration.

Integrative analysis of scRNA-seq, scTCR-seq, and TCR-seq to identify and characterize antigen-specific T cells

David Shih
University of Hong Kong, Hong Kong

Investigating the phenotypic profiles of antigen-specific T cells is critical to understanding T cell responses against pathogens as well as improving the efficacy of therapeutics and vaccines. However, current methodologies for identifying antigen-reactive T cells are limited in scope, throughput, or specificity. We have therefore developed an integrative approach to identify antigen-specific T cells in blood samples and characterize their single-cell transcriptomes. Our approach involves first identifying pathogen-specific T cells by modeling the temporal expansion trajectories in longitudinal bulk TCR-seq data, followed by using TCR sequences as barcodes to label the identified antigen-specific T cells in matched scRNA-seq and scTCR-seq data. Applying our approach to a clinical study of an experimental vaccine against human cytomegalovirus, we are able to characterize the single-cell transcriptomes of vaccine-specific T cells and discover transcriptional signatures of transient and durable T cell response to cytomegalovirus. Our approach can thus facilitate the study of T cell responses to vaccines and pathogens. To develop our methodology further, we propose a new longitudinal clustering method using Bayesian nonparametrics.

David Shih completed his BSc, MSc, and PhD at the University of Toronto, and undertook postdoctoral training in the Department of Data Science at Dana-Farber Cancer Institute, with cross-appointments in the Department of Biostatistics at Harvard T.H. School of Public Health and in the Cancer Program at the Broad Institute. He also did a postdoctoral fellowship in Systems Biology at MD Anderson Cancer Institute. Prior to joining HKU, Dr. Shih was a Research Assistant Professor in the School of Biomedical Informatics at the University of Texas Health Science Center, while serving as Co-Director of the Data Science and Informatics Core for Cancer Research.

Wednesday 5 July

Advances in single-cell data
Hongyu Zhao

scNovel: a neural network framework for novel rare cell detection of single-cell transcriptome data

Yu Li
The Chinese University of Hong Kong (CUHK), Hong Kong

Since bulk RNA sequencing can only provide a holistic perspective on the differences between samples, researchers are eager to obtain single-cell resolution of cell types within diseased tissues to develop more precise therapies. Single-cell RNA-sequencing has become a powerful tool to study biologically significant characteristics at explicitly high resolution. With the unprecedented boom in cell atlases, auto-annotation tools have become more prevalent due to their speed, accuracy, and user-friendly features. However, these tools have mostly focused on general cell type annotation and have not adequately addressed the challenge of detecting novel rare cell types. In this work, we introduce scNovel, a powerful model that specifically focuses on novel rare cell detection. By testing our model on diverse datasets with different scales, protocols, and degrees of imbalance, we demonstrate that scNovel significantly outperforms previous state-of-the-art novel cell detection models, reaching the most AUROC performance. Furthermore, we validate scNovel's performance on a million-scale dataset and demonstrate its ability to detect novel cell clusters for biological discovery through the analysis of clinical data. We believe that scNovel will be an important tool for high-throughput clinical data in a wide range of applications. To be more specific, scNovel can help to predict cell-type-specific gene expression profiles with biological significance, which helps biologists and medical researchers to perform downstream analysis leading to enlightening biological results and precision medical diagnoses.

Yu Li is an assistant professor in the Department of Computer Science and Engineering at CUHK. He is also the Visiting Assistant Professor at MIT/Harvard, working with Prof. James Collins. He works at the intersection between machine learning, healthcare and bioinformatics, developing new machine learning methods to resolve the computational problems in biology and healthcare. He has published over 50 papers in top venues, such as Nature Communications, PNAS, and Nature Computational Science. In 2022, he was selected to the Forbes 30 Under 30 Asia list, Healthcare & Science. He obtained his Ph.D. in computer science from KAUST in Saudi Arabia in 2020, after which he was nominated KAUST Alumni Change Makers Awards in 2022. Before that, he got his Bachelor degree in Biosciences from University of Science and Technology of China (USTC).

Guided-topic modelling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes

Yue Li
McGill University, Canada

Cell-type composition is an important indicator of health. We present Guided Topic Model for deconvolution (GTM-decon) to automatically infer cell-type-specific gene topic distributions from single-cell RNA-seq data for deconvolving bulk transcriptomes. GTM-decon performs competitively on deconvolving simulated and real bulk data compared with the state-of-the-art methods. Moreover, as demonstrated in deconvolving disease transcriptomes, GTM-decon can infer multiple cell-type-specific gene topic distributions per cell type, which captures sub-cell-type variations. GTM-decon can also use phenotype labels as a guide to infer phenotype-specific gene distributions. In a nested-guided design, GTM-decon identified cell-type-specific differentially expressed genes from bulk breast cancer transcriptomes.

Yue Li obtained a PhD degree in Computer Science from the University of Toronto in 2014. Between 2015 and 2018, he was a postdoc at the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT. In 2019, he started his tenure-track assistant professor in the School of Computer Science, McGill University. He is also an associate member at Mila – Quebec AI institute. He holds a Canada Research Chair for Machine Learning (ML) in Genomics and Healthcare (2022-2027). His research program is focused on developing ML algorithms to detect network connections among genes, regulatory elements, and phenotypes that underlie complex human diseases.

Towards a more reliable single-cell RNA-seq clustering - new measure to preserve global cell type relationships

Haiyan Huang
UC Berkeley, USA

Unsupervised cell clustering based on meaningful biological variation in single-cell RNA sequencing (scRNA-seq) data has received significant attention, as it assists with identifying ontological subpopulations among the data. A key step in the clustering process is to compute distances between cells using a specified distance measure. Although certain distance measures may successfully separate cells into biologically relevant clusters, they may fail to retain the global structure of the data, such as the relative similarity between cell clusters. In this talk, I will introduce a new measure that can more consistently retain the global cell type relationships than commonly used distance measures for scRNA-seq clustering. We used this measure to uncover compositional differences between annotated leukocyte cell groups in a compendium of Mus musculus scRNA-seq assays comprising 12 tissues.

Haiyan Huang received her BS in Mathematics at Peking University in 1997, her PhD in Applied Mathematics at the University of Southern California in 2001. She did postdoc at Harvard University from 2001-2003. Currently, she is a Professor in the Department of Statistics at UC Berkeley. As an applied statistician, her research is at the interface between statistics and data-rich scientific disciplines such as biology.

Mosaic single cell data integration

Shila Ghazanfar

University of Sydney, Australia

Currently available single cell -omics technologies capture many unique features with different biological information content. Data integration aims to place cells, captured with different technologies, onto a common embedding to facilitate downstream analytical tasks. Current horizontal data integration techniques use a set of common features, thereby ignoring non-overlapping features and losing information. Here we introduce StabMap, a mosaic data integration technique that stabilises mapping of single cell data by exploiting the non-overlapping features. StabMap is a flexible approach that first infers a mosaic data topology based on shared features, then projects all cells onto supervised or unsupervised reference coordinates by traversing shortest paths along the topology. We show that StabMap performs well in various simulation contexts, facilitates “multi-hop” mosaic data integration, and enables the use of novel spatial gene expression features for mapping dissociated single cell data onto a spatial transcriptomic reference.

Dr Shila Ghazanfar is an Australian Research Council DECRA Fellow at the University of Sydney, and is an expert in statistical and computational analysis of spatial transcriptomics and single cell RNA-seq data. Dr Ghazanfar completed her undergraduate and PhD studies in statistics and statistical bioinformatics at The University of Sydney, before completing a Royal Society Newton International Fellowship at The University of Cambridge. Dr Ghazanfar's interests are in developing statistical bioinformatics and biomedical data science approaches for the meaningful integration of complex and high dimensional biological datasets, particularly across various technological or omics modalities, and using these statistical and computational techniques to extract novel biological insights. Her multidisciplinary knowledge and skills in statistics, statistical bioinformatics, and computational biology enable her to devise strategies to jointly model the processes generating diverse data sources.

Cell-type-specific co-expression inference from single cell RNA-sequencing data

Emma Zhang
Emory University, USA

The advancement of single cell RNA-sequencing (scRNA-seq) technology has enabled the direct inference of co-expressions in specific cell types, facilitating our understanding of cell-type-specific biological functions. For this task, the high sequencing depth variations and measurement errors in scRNA-seq data present two significant challenges, and they have not been adequately addressed by existing methods. We propose a statistical approach, CS-CORE, for estimating and testing cell-type-specific co-expressions, that explicitly models sequencing depth variations and measurement errors in scRNA-seq data. Systematic evaluations show that most existing methods suffer from inflated false positives as well as biased co-expression estimates and clustering analysis, whereas CS-CORE gave accurate estimates in these experiments. When applied to scRNA-seq data from postmortem brain samples from Alzheimer’s disease patients/controls and blood samples from COVID-19 patients/controls, CS-CORE identified cell-type-specific co-expressions

and differential co-expressions that were more reproducible and/or more enriched for relevant biological pathways than those inferred from existing methods.

Dr. Emma Jingfei Zhang is an Associate Professor of Information Systems & Operations Research at the Goizueta Business School of Emory University. Her research focuses on the developments of statistical methods and theory for networks, graphs, tensors, and point processes, with applications in biology and medicine.

ClusterDE: a post-clustering differentially expressed (DE) gene identification method robust to false-positive inflation caused by double-dipping

Jingyi Jessica Li
University of California, USA

In typical single-cell RNA-seq data analysis, first, a clustering algorithm is applied to cluster cells; then, a statistical method is used to identify the differentially expressed (DE) genes between the cell clusters. However, this common procedure uses the same data twice, an issue known as “double dipping”: the same gene expression data are used to define cell clusters and DE genes, leading to false-positive DE genes even when the cell clusters are spurious. To overcome this challenge, we propose ClusterDE, a post-clustering DE method for controlling the false discovery rate (FDR) regardless of clustering quality. The core idea of ClusterDE is to generate in silico negative control data with only one cluster, which can be used in contrast to real data for evaluating the whole clustering+DE procedure. Using comprehensive simulation and real data analysis, we show that ClusterDE not only has solid FDR control but also finds cell-type marker genes that are biologically meaningful. ClusterDE is fast, transparent, and adaptive to a wide range of clustering methods and statistical tests.

Jingyi Jessica Li is Professor in the Department of Statistics (primary) and the Departments of Biostatistics, Computational Medicine, and Human Genetics (secondary) at University of California, Los Angeles (UCLA). Jessica leads a research group, titled the Junction of Statistics and Biology, at UCLA since 2013 after she completed Ph.D. at UC Berkeley. Jessica’s research focuses on developing robust, interpretable, and efficient statistical methods to address cutting-edge questions in biomedical sciences, especially those related to high-throughput genomics data. Her group had multiple successes in developing bioinformatics tools and collaborating with biomedical scientists. A particular emphasis of Jessica’s research is the statistical rigor in biomedical data analysis, including the use of in silico negative controls to avoid excess false discoveries. Jessica’s work was recognized by multiple awards, including NSF CAREER Award, Sloan Research Fellowship, Johnson & Johnson WiSTEM2D Math Scholar Award, MIT Technology Review 35 Innovators Under 35 China, Harvard Radcliffe Fellowship, COPSS Emerging Leader Award, and ISCB Overton Prize.

Benchmarking computational methods for single cell and spatial transcriptomics data

Mark Robinson

University of Zurich, Switzerland

Computational methods represent the lifeblood of modern molecular biology. However, the field is experiencing a somewhat unprecedented explosion of computational tools, especially for the analysis of single cell and spatial transcriptomics data. I will motivate the situation with a couple examples of benchmarking tools and methods related to our own research, but also discuss the topic of benchmarking more generally by reporting on a meta-analysis of single-cell method benchmarks. I'll propose a new computational system for flexible continuous benchmarking that allows the community to be engaged at various levels.

Mark Robinson has been an Associate Professor since 2017 after joining the Department of Molecular Life Sciences of the University of Zurich (UZH) in 2011. He studied Applied Mathematics (BSc, Uni. Guelph) and Statistics (MSc, Uni. British Columbia), and did a PhD in Statistical bioinformatics at the University of Melbourne. He has predoctoral experience at the Banting and Best Department of Medical Research (Uni Toronto) and postdoctoral experience in Cancer Epigenomics at the Garvan Institute in Sydney. The Robinson group at UZH develops statistical methods for interpreting high-throughput sequencing and other genomics technologies in the context of genome sequencing, gene expression and regulation and analysis of epigenomes, with a current focus on the analysis of single-cell and spatial datasets.

Thursday, July 6

Challenges in spatially resolved single cell data
Rafael Irizarry

Statistical challenges in Single-Cell RNA-Seq and spatial transcriptomics

Rafael Irizarry
Dana-Farber Cancer Institute, USA

I will start the talk by describing general statistical challenges in high-throughput genomics related to batch effects and systematic errors. Then I will describe some of our recent work related to cell-type classification and clustering with single-cell RNA-Seq (scRNA-Seq) and spatial transcriptomics.

Rafael Irizarry is Professor and Chair of the Department of Data Science at the Dana-Farber Cancer Institute and Professor of Biostatistics at Harvard School of Public Health. He is an applied statistician generally interested in using data to help solve real-world problems. During his career, he has worked and co-authored papers on a variety of topics including musical sound signals, infectious diseases, circadian patterns in health, fetal health monitoring, estimating the effects of Hurricane María in Puerto Rico, and COVID19 vaccine effectiveness. His main focus during the last two decades has been in Genomics. Specifically, he has worked on the analysis and signal processing of data generated by emerging high-throughput technologies. He has distinguished himself by disseminating statistical methodology as open source software shared through the Bioconductor Project, a leading open source and open development software project for the analysis of high-throughput genomic data. His widely downloaded software tools have helped him become one of the most highly cited scientist in his field.

Spatial Deconvolution Method Considering Platform Effect Removal, Sparsity and Spatial Information

Xiting Yan
Yale University, USA

Spatial barcoding-based transcriptomic (ST) technologies unbiasedly measure mRNA expression of cells with physical locations in intact tissue. But the measured gene expression data lack single-cell resolution and require cell type deconvolution for cellular-level downstream analysis. We developed SDePER to deconvolve ST data using reference single-cell RNA sequencing (scRNA-seq) data from the same tissue type. A conditional variational autoencoder (CVAE) was used to remove platform effects, i.e., the systematic differences between reference scRNA-seq and ST data, and a graph Laplacian regularized model (GLRM) was developed to consider both spatial information and sparsity. Based on the estimated cell type compositions, a random walk was constructed to impute cell type compositions and gene expression at enhanced resolution. We compared the performance of SDePER and six existing methods using both simulated and real data. Results showed that SDePER was robust to platform effects and achieved the most accurate estimation.

Furthermore, applications to four different real ST datasets with histological staining images demonstrated that SDePER achieved results with the highest consistency with the staining images. SDePER also had the most accurate imputed gene expression of known marker genes. In summary, SDePER achieved significantly more accurate and robust results than the existing ST data deconvolution methods.

Dr. Yan is a computational biologist and bioinformatician, with extensive research experiences in genetic, genomics and transcriptomic data analyses. Her research interest is in developing novel statistical and computational methodologies to understand pathogenesis and heterogeneity of chronic lung diseases using large-scale omics data. She is specifically interested in development of novel analytical methods for single-cell RNA sequencing data, spatial transcriptomic data, drug perturbation data and integration of different omics data.

SpatialScope: A unified approach for integrating spatial and single-cell transcriptomics data using deep generative models

Can Yang
Hong Kong University of Science and Technology (HKUST), Hong Kong

The rapid emergence of spatial transcriptomics (ST) technologies is revolutionizing our understanding of tissue spatial architecture and their biology. Current ST technologies based on either next generation sequencing (seq-based approaches) or fluorescence in situ hybridization (image-based approaches), while providing hugely informative insights, remain unable to provide spatial characterization at transcriptome-wide single-cell resolution, limiting their usage in resolving detailed tissue structure and detecting cellular communications. To overcome these limitations, we developed SpatialScope, a unified approach to integrating scRNA-seq reference data and ST data that leverages deep generative models. With innovation in model and algorithm designs, SpatialScope not only enhances seq-based ST data to achieve single-cell resolution, but also accurately infers transcriptome-wide expression levels for image-based ST data. We demonstrate the utility of SpatialScope through comprehensive simulation studies and then apply it to real data from both seq-based and image-based ST approaches. SpatialScope provides a spatial characterization of tissue structures at transcriptome-wide single-cell resolution, greatly facilitating the downstream analysis of ST data, such as detection of cellular communication by identifying ligand-receptor interactions from seq-based ST data, localization of cellular subtypes, and detection of spatially differently expressed genes.

Prof. Yang Can is currently Dr Tai-chin Lo Associate Professor of Science, Department of Mathematics, The Hong Kong University of Science and Technology. His research focuses on data science with the development of novel statistical and computational methods for large-scale data analysis, including deep generative models, graph neural networks and adversarial domain translation. Prof. Yang has also established industrial collaboration supported by the Innovation and Technology Fund of the Hong Kong Government.

Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data

Jean Yee Hwa Yang
University of Sydney, Australia

Recent advances in subcellular imaging transcriptomics platforms have enabled high-resolution spatial mapping of gene expression, while also introducing significant analytical challenges in accurately identifying cells and assigning transcripts. Existing methods grapple with cell segmentation, frequently leading to fragmented cells or oversized cells that capture contaminated expression. To this end, we present BIDCell, a self-supervised deep learning-based framework with biologically-informed loss functions that learn relationships between spatially resolved gene expression and cell morphology. BIDCell incorporates cell-type data, including single-cell transcriptomics data from public repositories, with cell morphology information. Using a comprehensive evaluation framework consisting of metrics in five complementary categories for cell segmentation performance, we demonstrate that BIDCell outperforms other state-of-the-art methods according to many metrics across a variety of tissue types and technology platforms. Our findings underscore the potential of BIDCell to significantly enhance single-cell spatial expression analyses, including cell-cell interactions, enabling great potential in biological discovery.

Professor Jean Yang is an applied statistician with expertise in statistical bioinformatics. She was awarded the 2015 Moran Medal in Statistics from the Australian Academy of Science in recognition of her work on developing methods for molecular data arising in cutting-edge biomedical research. She has made contributions to the development of novel statistical methodology and software for the design and analysis of high-throughput biotechnological data. Recently, much of her focus has been on the integration of single-cell biotechnologies with clinical data to answer a variety of scientific questions. As a statistical data scientist who works in the interface of statistics, biomedicine, and health. She enjoys developing novel methods with translational potential in a collaborative environment, working closely with investigators from diverse backgrounds.

Rethinking assumptions in spatial molecular data analysis: the role and impact of library size normalisation

Melisa Davis
University of Adelaide

Spatial molecular technologies have revolutionised the study of disease microenvironments by providing spatial context to tissue heterogeneity. Recent spatial technologies are increasing the throughput and spatial resolution of measurements, resulting in larger datasets at single cell resolution. The added spatial dimension and volume of measurements poses an analytics challenge that has, in the short-term, been addressed by adopting methods designed for the analysis of single-cell RNA-seq data. Though these methods work well in some cases, they do not necessarily translate appropriately to spatial technologies. A common assumption is that total sequencing depth, also known as library size, represents technical variation in single-cell RNA-seq technologies, and this is often normalised out during analysis. Through analysis of several different spatial datasets, we noted that this assumption does not necessarily hold in spatial

molecular data. To formally assess this, we explore the relationship between library size and independently annotated spatial regions, across 23 samples from 4 different spatial technologies with varying throughput and spatial resolution. We found that library size confounded biology across all technologies, regardless of the tissue being investigated. Statistical modelling of binned total transcripts shows that tissue region is strongly associated with library size across all technologies, even after accounting for cell density of the bins. Through a benchmarking experiment, we show that normalising out library size leads to sub-optimal spatial domain identification using common graph-based clustering algorithms. On average, better clustering was achieved when library size effects were not normalised out explicitly, especially with data from the newer sub-cellular localised technologies. Taking these results into consideration, we recommend that spatial data should not be specifically corrected for library size prior to analysis unless strongly motivated. We also emphasise that spatial data are different to single-cell RNA-seq and care should be taken when adopting algorithms designed for single cell data.

Melissa Davis is a Professor of Computational Biology and leader of the Program in Computational Systems Oncology at the newly established South Australian ImmunoGenomics Cancer Institute (SAiGENCI), where she is part of the senior leadership team recruited to establish Australia's first new medical research institute in over a decade. She also holds appointments at the Walter and Eliza Hall Institute (WEHI), the University of Melbourne (Clinical Pathology) and the Frazer Institute. Over the last ten years, her highly translational research has contributed to three clinical trials. She is internationally recognised for her work using computational biology, network analysis and knowledge engineering in the analysis of cancer plasticity and progression. Her recent work developing methods for newly emerging spatial molecular measurement platforms has already resulted in substantial international recognition and collaborations with research labs and technology development companies worldwide.

Accurate and scalable spatial domain detection via integrated reference-informed segmentation for spatial transcriptomics

Xiang Zhou
University of Michigan, USA

Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and increasingly large, providing unprecedented opportunities for characterizing the spatial and functional organization of complex tissues. Here, we present a computational method, IRIS, that can characterize the spatial organization of complex tissues through accurate and efficient detection of spatial domains. IRIS is unique in its ability in leveraging single-cell RNA-seq data for reference-informed spatial domain detection, integrating multiple SRT tissue slices jointly while explicitly accounting for the correlation both within and across slices, and taking advantage of multiple algorithmic innovations for highly scalable computation. We demonstrate the advantages of IRIS through in-depth analysis of six SRT datasets from different technologies across distinct tissues, species, and spatial resolutions. In these applications, IRIS achieves 51% ~ 97% accuracy gain over existing methods in the dataset with known ground truth. In addition, IRIS is 4.6 ~ 134.7 times faster than existing methods in moderate-sized datasets and is the only method applicable to

large-scale SRT datasets including stereo-seq and 10x Xenium. As a result, IRIS captures the fine-scale structures of brain regions, reveals the spatial heterogeneity of tumor microenvironments, and characterizes the structural changes of the seminiferous tubes in the testis underlying diabetes, all at a speed and accuracy unattainable by existing approaches.

Dr. Xiang Zhou is an Associate Professor in the Department of Biostatistics at the University of Michigan's School of Public Health. Additionally, he serves as an Assistant Director at the University of Michigan Precision Health. Dr. Zhou has held the position of Associate Professor since September 2019. Prior to that, he joined the department as an Assistant Professor in 2014 and was subsequently appointed as the John G. Searle Assistant Professor from 2018 to 2019. Before joining the University of Michigan, Dr. Zhou worked as a Williams H. Kruskal Instructor in the Department of Statistics at the University of Chicago from 2013 to 2014. Dr. Zhou earned his MS degree in Statistics in 2009 under the guidance of Prof. Scott Schmidler and his PhD degree in Neurobiology in 2010 with Prof. Fan Wang as his advisor, both from Duke University. Following his doctoral studies, he was a postdoctoral scholar working alongside Prof. Matthew Stephens at the University of Chicago from 2010 to 2013.

Round table discussion: Gaps and Opportunities in Spatial Omics
Chaired by Can Yang

Gene set tests and cell-cell communication in scRNA-seq data

Di Wu
University of North Carolina, USA

Gene set tests and detection of cell-cell communications (CCC) in single cell RNAseq (scRNAseq) data are two key analysis methods to interpret the data for biological follow-up. Two-Sigma-G is a competitive test to test whether the gene in a prior defined gene set, e.g., from a pathway or other researchers' experiments, are more differentially expressed compared to the randomly selected gene sets. It employs the Two-Sigma framework based on zero-inflated negative binomial distribution and allowing random effects since many cells are from one biological sample and there may be multiple samples in a sample group. Simulations have been run for model fitting and the control of type I, and type II error in the tests. Methods are applied in well-designed HIV related scRNAseq datasets for biological discovery. We also have developed a statistical method to detect (CCC) mediated by ligand-receptor (LR) complexes, associated with the sample groups. We simultaneously model the data distribution that are featured with excess zeros, and performing the statistical test for differential CCC for a pair of LR and a pair of cell types, applied in scRNAseq data of humanized mouse spleen samples with or without the infection of acute human immunodeficiency virus (HIV).

Di Wu is a biostatistician working in the bioinformatics field. She is an associate professor in Biostatistics and jointly with the dental school at UNC Chapel Hill. She has developed novel statistical bioinformatics methods to handle biomedical and genomics data. She also has developed gene set testing methods with high citations, in the empirical Bayesian framework, to take care of small complex design

and genewise correlation structure. The focus of her research group includes microbiome/metabolomics data analysis, single cell RNAseq data, cancer genomics and electronic medical for precision medicine.

Inference of donor-specific co-expression networks across cohorts

Gerald Quon
University of California, Davis, USA

Gene co-expression networks are routinely inferred to identify co-expression modules and pathways active in diverse cell types. Their construction and inference is challenging due to their high dimensional nature and typically few samples available for inference. Here I discuss a multi-task framework for inferring and comparing multiple co-expression networks across individuals in a cohort. I will demonstrate that despite low information content in single cell transcriptome data, we are still able to parse out major differences in network structure that are correlated with phenotypes such as stem cell potential.

Gerald Quon is an Assistant Professor in the Department of Molecular and Cellular Biology at UC Davis. His work focuses on using computational methods to characterize the mechanisms by which inherited genetic variation leads to risk of different human diseases.

Identifying changes in cell states related to their spatial context in tissue microenvironment

Ellis Patrick
University of Sydney, Australia

The human body comprises over 37 trillion cells with diverse forms and functions, which can exhibit dynamic changes based on their environmental context. Understanding the spatial interactions between cells and changes in their state within the tissue microenvironment is crucial to comprehending the development of human diseases. State-of-the-art technologies such as PhenoCycler, IMC, CosMx, Xenium, and others can deeply phenotype cells in their native environment, providing a high-throughput means of identifying spatially related changes in cell state. The Statial Bioconductor package offers a suite of complementary approaches for identifying changes in cell state explained by changes in cell type localization. In this presentation, we introduce new functionality in the Statial package that can 1) identify changes in cell state between distinct tissue environments, 2) uncover changes in marker expression associated with cell proximities, and 3) model spatial relationships between cells in the context of hierarchical cell lineage structures. We provide context for these approaches and explain when and why modeling spatial relationships between cells in these ways is appropriate. Finally, we demonstrate how these approaches can be used in a classification setting to predict patient prognosis or treatment response.

Dr Ellis Patrick is an applied statistician and bioinformatician. He is currently a Senior Lecturer in the School of Mathematics and Statistics, a Faculty member at The Westmead Institute for Medical Research and the Cluster Lead of Bioinformatics in

the Sydney Precision Data Science Centre. He obtained his PhD at the University of Sydney, worked as a postdoc with joint appointments at Brigham and Women's hospital, Harvard Medical School and the Broad Institute, and recently completed his Australian Research Council DECRA fellowship. Dr Patrick's research interests lie at the interface of statistics and biomedical data science where he specialises in developing and implementing innovative bioinformatics tools to enhance our understanding and treatment of human diseases.

Friday 7 July

Scaling up for single cell data science
Joshua Ho

Scalable analysis methods for single cell omics data and lineage tracing

Joshua Ho
University of Hong Kong, Hong Kong

Single cell RNA-seq (scRNA-seq) analysis is gaining widespread adoption in many areas of biomedical research. A large amount of scRNA-seq data is being generated at a rapid rate. Nonetheless, it is challenging to quickly and efficiently process a large collection of scRNA-seq data. Extending our team's pioneering experience in developing scalable bioinformatics tools, we have developed a suite of cloud-accelerated scRNA-seq analysis tools that efficiently process read-level information at a highly scalable manner. In this talk, we will showcase a number of tools that we have developed to enable large-scale single cell RNA-seq analysis. Furthermore, we will discuss new methods that enable somatic genetic variants in mitochondria to be used as endogenous lineage tracing markers in scRNA-seq data. We will showcase the use of this type of novel lineage tracing technology to study cancer using human clinical tumour samples.

Dr Joshua Ho is an Associate Professor at the School of Biomedical Sciences at the University of Hong Kong. He completed a BSc and PhD in Bioinformatics at University of Sydney, then undertook a postdoctoral fellowship at Harvard Medical School. Joshua's research focuses on scalable analysis of large single cell omics data and using endogenous somatic mutations to support lineage tracing at the single cell level.

Graphical generative model for identification of disease associated perturbations to intercellular communications in single-cell RNA sequencing data

Zuoheng Wang
Yale University, USA

Diverse types of cells interact and communicate with each other to maintain tissue homeostasis and perform biological functions. Perturbations to these interactions can break the homeostasis of the tissue microenvironment, leading to disease. Understanding intercellular communication changes in disease is critical for therapeutic development. Cell-cell communication networks (CCCNs) inferred from single-cell RNA sequencing data are highly variable and only capture a snapshot of the dynamic intercellular communication system. We develop a graphical generative model to compare CCCNs between disease and control samples to identify disease associated perturbations to intercellular communications. The distribution of CCCNs is learned using variational graph autoencoder (VGAE) in disease and control groups separately. Then a large number of graphs is generated to assess the significance of the difference between the two distributions using different graph distance measures.

We demonstrate the advantage of this approach in improving the power of identifying disease associated perturbations to intercellular communications through both simulation studies and real scRNA-seq datasets.

Dr. Wang is Associate professor of Biostatistics at Yale School of Public Health. Her research focuses on combining genetics, genomics, immunology, and statistical modeling to answer biologically important questions in genetic epidemiological studies. Her statistical expertise lies in longitudinal data analysis, varying coefficient models, mixed effects models, kernel machine methods, mediation analysis, machine learning methods, and network analysis. She develops statistically innovative methods and computationally efficient tools in large-scale genetic and genomic studies to identify genetic susceptibility variants and advance the understanding of the etiology of complex diseases including breast cancer, alcohol and drug abuse, asthma, autism, obesity, lung and cardiovascular diseases. Current studies include using next-generation sequencing data to detect rare genetic variants in longitudinal genetic studies, combining knowledge in genomics and immunology to understand the risk of breast cancer survival, addressing statistical challenges in single-cell RNA sequencing data and spatial transcriptomics, and machine learning for risk prediction in electronic health records data.

Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2

Yingxin Lin
Yale University, USA

The recent emergence of multi-sample multi-condition single-cell multi-cohort studies allows researchers to investigate different cell states. The effective integration of multiple large-cohort studies promises biological insights into cells under different conditions that individual studies cannot provide. Here, we present scMerge2, a scalable algorithm that allows data integration of atlas-scale multi-sample multi-condition single-cell studies. We have generalized scMerge2 to enable the merging of millions of cells from single-cell studies generated by various single-cell technologies. Using a large COVID-19 data collection with over five million cells from 1000+ individuals, we demonstrate that scMerge2 enables multi-sample multi-condition scRNA-seq data integration from multiple cohorts and reveals signatures derived from cell-type expression that are more accurate in discriminating disease progression. Further, we demonstrate that scMerge2 can remove dataset variability in CyTOF, imaging mass cytometry and CITE-seq experiments, demonstrating its applicability to a broad spectrum of single-cell profiling technologies.

Yingxin Lin recently joined the Department of Biostatistics at the Yale School of Public Health as a Postdoctoral Associate advised by Prof. Hongyu Zhao. She completed her PhD in Statistics at the University of Sydney in 2022 in School of Mathematics and Statistics where she also obtained her Bachelor of Science in Statistics, with the University Medal in 2017. Her research interests lie broadly in statistical modelling and machine learning for various omics, biomedical, and clinical data, with a specific focus on methodological development and data analysis for single-cell multi-omics data.

Cross-species single-cell atlases: analysis and challenges

Angela Wu
Hong Kong University of Science and Technology (HKUST), Hong Kong

The rapid emergence of large-scale atlas-level single-cell RNA-seq (scRNA-seq) datasets presents remarkable opportunities for broad and deep biological investigations through integrative analyses. However, harmonizing such datasets requires integration approaches to be not only computationally scalable, but also capable of preserving a wide range of fine-grained cell populations. As part of the Tabula Microcebus Consortium whose mission is to create a single cell atlas of the grey mouse lemur, we faced such challenges during the integration of scRNA-seq data generated from multiple animals, tissue types, batches, and technologies. Manual annotation of each dataset, particularly the identification of rare cell-types, proved to be difficult and tedious. To address these challenges, we embarked on a detailed exploration of large-scale scRNA-seq data, uncovered underlying features of their data distributions, and created two tools for data integration: FIRM and Portal. These two algorithms were used to construct the Tabula Microcebus single cell atlas, and are suitable for scRNA-seq datasets with different characteristics. I will present the findings of the Tabula Microcebus, as well as present perspectives on our current work in cross-species analyses.

Angela Ruohao Wu is an associate professor in the Division of Life Science and the Department of Chemical and Biological Engineering at the Hong Kong University of Science and Technology. She completed her Ph.D. and post-doctoral training in Bioengineering at Stanford University, and soon after that co-founded Agenovir Corporation, a genome editing-based antiviral therapeutics company that was ultimately acquired by Vir Biotechnologies. Angela is one of the earliest scientists to work in single cell genomics, and she pioneered the field of microfluidic chromatin immunoprecipitation (ChIP). Her current research focuses on using genomics and microfluidics to address complex biological questions, as well as applying genomics in the clinic. Angela's interdisciplinary work has been recognized for bridging important gaps between microfluidics and biology; she was named one of MIT Technology Review Innovators under 35 Asia; a World Economic Forum Young Scientist, and an Outstanding Young Faculty by IEEE EMBS (Micro and Nanotechnology in Medicine).

An informatics framework for assembling human cell atlases as a digital life

Xuegong Zhang
Tsinghua University, China

Profiling molecular features of all cells is essential for understanding the human body in health and diseases. Scientists are enthusiastic in building such atlases of human cells using single-cell omics technologies. More and more single-cell studies have been conducted in the world with the rapid development and popularization of single-cell sequencing technologies, generating tremendous amount of single-cell data in the public domain. This suggests the possibility of building cell atlases by assembling data in scattered publications. However, the information complexity and

volume of cell atlas data are magnitudes larger than that of the human genome project. We proposed a unified information framework for assembling atlases from data of various sources and built the first prototype of human Ensemble Cell Atlas (hECA). We argued that the ideal cell atlas should be like a “digital life” or a “virtual human body” composed of virtual cells. We developed an “in data” cell experiment scheme that allows extracting cells from the atlas using a logic formula to investigate scientific questions such as drug side effects that may involve multiple organs and cell types.

Xuegong Zhang received his BS degree in Industry Automation in 1989 and his Ph.D. degree in Pattern recognition and Machine Intelligence in 1994, both from Tsinghua University, after which he joined the faculty of Tsinghua University. He visited Harvard School of Public Health in 2001-2002, and is now a Professor of Pattern Recognition and Bioinformatics in the Department of Automation, Tsinghua University, and Adjunct Professor of the School of Life Sciences and School of Medicine. He is an ISCB Fellow and CAAI Fellow. His major research field is machine learning, bioinformatics, human cell atlas, and intelligent precision medicine.