Frontiers in Single-cell Technology, Application and Data Analysis BIRS Workshop 19w5032, Feb 24, 2019 – March 1, 2019

Quan Long [1], Jie Peng[2], Pei Wang [3]

[1] University of Calgary, quan.long@ucalgary.ca
[2]University of California, Davis, jiepeng@ucdavis.edu
[3]Icahn School of Medicine at Mount Sinai, pei.wang@mssm.edu

1. Overview

The last few years have seen an explosion of high throughput single-cell technologies for quantifying genetic, epigenetic, and RNA expression levels within individual cells, with the technologies for single-cell RNA sequencing comparatively the most mature. These breakthroughs pave the way for exploring biological systems at an unprecedented level of details. They allow us to look into inter-cellular variations and interactions and intra-tissue heterogeneity in biological samples, thus enabling the investigation of many fundamental biological questions far beyond those that could be tackled by traditional bulk tissue experiments. Single-cell technologies have led to the developments of novel computational and statistical methods that encompass data preprocessing, modeling and inference. Despite the progress, there is still much work to be done to meet the challenges and make use of the opportunities posed by the new data type.

Although there are sessions in statistical and computational biology conferences focusing on single-cell data analysis, researchers from these fields as well practitioners of these technologies would greatly benefit from a focused workshop, where it will be possible to exchange ideas, raise new questions and build future collaborations. Through this workshop, we aim to disseminate cutting-edge technological and computational advancements, to identify new challenges in data analysis and modeling, to provide a platform for interdisciplinary dialogue and to help shape future directions for this burgeoning field. The workshop was organized around the following topics:

- Single-cell technology: Experimental techniques and challenges.
- Biology applications based on single-cell technologies.
- Single-cell data analysis.
  - Computational and bioinformatics tools for data processing, integration, and visualization.
  - Statistical modeling to handle confounding effects, cell-to-cell variations, and intra-tissue heterogeneity.

During the workshop, we brought together both researchers who developed computational and statistical tools, as well as those who applied computational tools within their own domain, to identify key remaining challenges in these tools and develop collaborations to solve them. The workshop provided a unique opportunity for biologists, biotechnologists, statisticians and bioinformaticians to communicate, collaborate and work together to further advance this important research field.

2. Single-Cell Technologies

In the past decade, diverse technologies of single-cell experiments have been invented, including single-cell sequencing, single-cell transcriptomics, single-cell epigenomics and single-cell proteomics. Different types of experiments involve different technologies/platforms. Specifically, single-cell sequencing and transcriptomics are both based on the latest high-throughput sequencing technologies. In single-cell epigenomics experiments, sequencing techniques are further integrated with other physical/chemical/molecular approaches to monitor cytosine modification, protein-DNA interaction, chromatin structure, and three dimensional DNA organization. Moreover, single-cell mass cytometry instruments integrate mass spectrometry and cytometry technologies and enable real time monitoring of multi-protein targets.

3. Broad Biological Applications Based on Single-Cell Technologies

These diverse single-cell technologies have profoundly advanced single-cell biology research, from bench to clinic. In basic biology, all events at the single-cell level are stochastic, and single-cell experiments allow scientists to quantify and model this stochasticity, and from these models, draw inferences on the relationships between genes as well as the relationship between a gene and its epigenetic environment. Single-cell experiments allow much higher resolution in the study of transcriptional regulation. At the tissue level, single-cell experiments allow the discovery of new cell types and the characterization of known cell types and their relationships to each other. In clinical practice, single-cell experiments have direct applications, for example, to the quantification of intra-tumor heterogeneity in oncology and to the tracking of immune cell development in immunology.

3.1. Germline genomics and single-cell analysis of brain tumors

Brain cancer is one of the deadliest cancers, for example, GBM has poor clinic outcomes (i.e., with a survival of 1-2 years), low-grade glioma (LGG) with IHD1 wild-type (wt) has poor clinic outcomes as well, however, LGG with IDH1 mutations (mut) has good survival (i.e., with a survival of 5-6 years). It has been debated for years whether the LGG is the precursor of the GMB. Dr. Edwin Wang from University of Calgary conducted single-cell genomic analysis of LGG-IDH (wt), LGG-IDH (mut) tumors and GBM and developed methods to conduct the analysis. At the same time, Dr. Wang and team members analyzed the germline genomes of the patients. Their analysis showed that GMB with IDH (wt) and LGG with IDH (wt) have independent origins. Furthermore, they showed that nature killer (NK) cells play an important role in cancer progression and metastasis. The number of germline inherited variants affecting NK cell defects are negatively correlated with patient survival. A person who has high mummer of NK cell genetic defects has much higher chance to get brain tumors.

3.2. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression

Genetic variation affects human traits and disease risk via myriad pathways, including through changes in gene expression. Expression quantitative trait locus (eQTL) mapping is a widely-used approach to elucidate such effects. Dr. Daniel Seaton from The European Bioinformatics Institute leveraged a population-scale human induced pluripotent stem cell (iPSC) bank to study, in vitro, the effect of genetic variation on gene expression during human development. By combining single-cell transcriptome sequencing with a pooled experimental design of 125 cell lines, they assay gene expression variability across iPSC differentiation to two different fates: definitive endoderm and dopaminergic neurons. This allowed discovery of 1,000s of eQTLs in distinct developmental stages and cell types. Dr. Seaton and collaborators developed an allele-specific expression-based approach to identify eQTLs that are sensitive to differentiation and other cellular processes. For example, 800 eQTL effects were dynamic across differentiation to endoderm, and these dynamics were generally uncorrelated with overall changes in gene expression. In sum, their data and methods illustrate the power of combining pooled iPSC lines with scRNA-seq to simultaneously discover and characterise genetic variants affecting gene expression in differentiating cells.

3.3. Effectively comparing publicly available single cell datasets: a case study in glioblastoma multiforme

Glioblastoma multiforme (GBM) is an aggressive form of brain cancer, accounting for 17% of all brain tumours and it has a poor prognosis. Despite this, the minority of GBMs with isocitrate dehydrogenase gene mutations have relatively good prognostic outcomes. Microglia and macrophage content of GBMs is widely described as up to one-third of all the total cells within the tumor. Flow cytometry of IDH-mutant and wild type show more pro-inflammatory markers in IDH-mutant tumours, but macrophages and microglia cannot be reliably distinguished with protein markers. To determine the distinct roles of microglia and macrophages, GBM and GBM-IDH mutant single cell datasets from publicly available on GEO were analyzed, revealing that the immune phenotype of the good prognosis IDH-mutant GBMs is driven by pro-inflammatory microglia. Challenges in quality control, normalizing, clustering, cell-type labelling and visualizing disparate datasets were addressed in this analysis, as well the labelling of spectra of behaviour like pro- vs anti-inflammatory.

3.4. Understanding gene regulation using single cell RNA-seq data

Single-cell analytics offers tremendous opportunity for studying different levels of gene regulation at single-cell resolution. Dr. Liu briefly introduced some of their preliminary results for designing computational methods in understanding cell type-specific polyadenylation, miRNA regulation, as well as cancer drug response prediction.

3.5. Single cell transcriptomics and fate mapping of ependymal cells reveals an absence of neural stem cell function

Ependymal cells are multi-ciliated cells that form the brains ventricular epithelium and a niche for neural stem cells (NSCs) in the ventricular-subventricular zone (V-SVZ). In addition, ependymal cells are suggested to be latent NSCs with a capacity to acquire neurogenic function. This remains highly controversial due to a lack of prospective in vivo labeling techniques that can effectively distinguish ependymal cells from neighboring V-SVZ NSCs. Dr. Stratton described a transgenic system that allows for targeted labeling of ependymal cells within the V-SVZ. Single-cell RNA-seq revealed that ependymal cells are enriched for cilia-related genes and share several stem-cell-associated genes with neural stem or progenitors. Under in vivo and in vitroneural-stem- or progenitor-stimulating environments, ependymal cells failed to demonstrate any suggestion of latent neural-stem-cell function. These findings suggest remarkable stability of ependymal cell function and provide fundamental insights into the molecular signature of the V-SVZ niche.

3.6. Characterizing cell type-specific responses to stimuli using single cell RNA sequencing

Single cell RNA sequencing (scRNA-seq) technologies are quickly advancing our ability to characterize the transcriptional heterogeneity of biological samples, given their ability to identify novel cell types and characterize precise transcriptional changes during previously difficult-to-observe processes such as differentiation and cellular reprogramming. An emerging challenge in scRNA-seq analysis is the characterization of cell type-specific transcriptional responses to stimuli, when the similar collections of cells are assayed under two or more conditions, such as in control/treatment or cross-organism studies.

Dr. Quon has presented a novel computational strategy for identifying cell type specific responses using deep neural networks to perform unsupervised domain adaptation. Compared to other existing approaches, ours does not require identification of all cell types before alignment, and can align more than two conditions simultaneously. He has discussed on-going applications oftheirmodel to two problem domains: characterizing hematopoietic progenitor populations and their response to inflammatory challenges (LPS), in which their have identified putative subpopulations of long term HSCs that differentially respond to the challenge, and characterizing the malaria cell cycle process, in which they identify transcriptional changes associated with sexual commitment.

3.7. Single Cell Assessment of Tumor Heterogeneity

Each tumor is composed of multiple cell types, characterized by different genomic and transcriptomic profiles. Many methods have been proposed for the classification of tumor samples into different subtypes based on bulk tumor data. For this classification, a common practice is to utilize existing gene signatures that are expected to be upregulated in a particular subtype. A challenge when dealing with single cell transcriptomic data is due the high frequency of missing values. In fact, key markers specific to a particular subtype might be missing in most of the cells. Another challenge is that most of the existing gene signatures were experimentally validated using bulk tumor data; and therefore might not be appropriate for single cell transcriptomic data analysis. Dr. Petralia reviewed current methods utilized to classify single cells into different subtypes. In addition, Dr. Petralia propose a new method, which can classify cells into different subtypes while dealing with the sparse nature of the data. Different methods are compared based on single cell sequencing transcriptomic profiles of breast cancer data.

4. Single-Cell Data Analysis: Challenges and Opportunities

Single-cell data is complex and noisy, and presents new challenges arising from both the technical noise in the experiments as well the stochastic nature of single-cell biology. Some of the main technical issues with single-cell RNA sequencing arise from the experimental biases introduced within each cell in the RNA extraction, reverse transcription, and amplification steps. Cell size and cell cycle differences are also new sources of biological variation that needs to be modeled and accounted for. Moreover, the data is very sparse, with many zeros, due both to the technical issue of experimental dropout as well as the biological phenomenon of transcriptional bursting. How these various sources of noise impact downstream analyses, and how best to remove them, still remains under much debate. Yet, addressing these technical issues is necessary for reliable conclusions to be drawn from single-cell experiments.

The computational tools for analysis of single-cell profiles are still in their infancy. Both the sparsity and lower total read counts characteristic of single-cell profiles make tools previously developed for even the most basic analyses ill-suited for direct application to single-cells. As a result, in the past few years there has been an explosion in terms of new bioinformatics tools for performing tasks in common with bulk sample analysis and furthermore tools to analyze data for entirely new problems are now being developed (e.g. for ordering of single-cells along a differentiation trajectory, detecting bifurcating points in those trajectories, or identifying new cell types in collections of single-cells).

4.1. Transfer Learning in Single Cell Transcriptomics

Cells are the basic biological units of multicellular organisms. The development of single-cell RNA sequencing (scRNA-seq) technologies have enabled us to study the diversity of cell types in tissue and to elucidate the roles of individual cell types in disease. Yet, scRNA-seq data are noisy and sparse, with only a small proportion of the transcripts that are present in each cell represented in the final data matrix. Dr. Zhang proposed a transfer learning framework to borrow information across related single cell data sets for de-noising and expression recovery. The goal is to leverage the expanding resources of publicly available scRNA-seq data, for example, the Human Cell Atlas which aims to be a comprehensive map of cell types in the human body. Dr. Zhang's method is based on a Bayesian hierarchical model coupled to a deep autoencoder, the latter trained to extract transferable gene expression features across studies coming from different labs, generated by different technologies, and/or obtained from different species. Through this framework, Dr. Zhang and collaborators explore the limits of data sharing: How much can be learned across cell types, tissues, and species? How useful are data from other technologies and labs in improving the estimates from your own study? She has also discuss the implications of technical batch artifacts in the joint analysis of multiple data sets, and propose strategies for alignment of data across batch.

4.2. Impact of Misspecified Dependence on Clustering of RNA-seq Gene Expression Profiles

Clustering RNA-seq data is used to characterize environment-induced (e.g., treatment) differences in gene expression profiles by separating genes into clusters based on their expression patterns. Wang et al. (2013, Briefings in Bioinformatics) recently adopted the bi-Poisson distribution, obtained via the trivariate reduction method, as a model for clustering bivariate RNA-seq data. Dr. Leon discussed discuss the inadequacy of the bi-Poisson distribution in modelling the correlation between dependent Poisson counts, and its impact on clustering such data. Dr. Leon introduced the bi-Poisson Gaussian copula distribution as an alternative copula-based model that incorporates a flexible dependence structure for the counts. Dr. Leon then reported simulation results to investigate the impact on clustering of Poisson counts of misspecified dependence structures. Their simulations indicate that the clustering performance of the bi-Poisson distribution suffers when the cluster-specific correlations are negative, as the bi-Poisson distribution allows only positive correlations. Dr. Leon also found that although large positive values are also not admissible under the bi-Poisson distribution, their effect is minimal, especially when clusters are well separated. Dr. Leon illustrate their methodology on a lung cancer RNA-seq data.

4.3. Penalized Latent Dirichlet Allocation Model in Single Cell RNA Sequencing

Single cell RNA sequencing (scRNA-seq) data are counts of RNA transcripts of all genes in species genome. Viewing the genes as building blocks of the genetic language, Dr. Wu and collaborators adapt the Latent Dirichlet Allocation (LDA) model, a generative probabilistic model originated in natural language processing(NLP), to scRNA-seq experiments. Dr. Wu and collaborators considered the DNA as natures language using a four-letter alphabet, and the genome of a species defines its dictionary. The active transcriptome of a single cell is a document composed of different copies of various words, and the analogy of topics are biological functions a cell is performing. The observed transcript counts are a result of transcripts generated from a mixture of biological processes, each with a different gene usage frequency. She proposed a penalized version of LDA to reflect the sparsity expected in biological data. Dr. Wu and collaborators demonstrate that inferred biological topic frequency is a meaningful dimension reduced representation of the single cell transcriptomes and delivers improved accuracy in cell type clustering/classification.

4.4. Integrative Differential Expression Analysis and Gene Set Enrichment Analysis in Single Cell RNAseq Studies

Single cell RNA sequencing (scRNAseq) has been widely applied for transcriptomics analysis. One important analytic task in scRNAseq is to identify genes that are differentially expression (DE) between different cell types or cellular states and to perform subsequent gene set enrichment analysis (GSEA) to detect biological pathways that are enriched in the identified DE genes. These two types of analytic tasks – DE analysis and GSEA – are often treated as two sequential steps in commonly used analytic pipelines. However, these two tasks are intermingled with each other: while DE results are indispensable for detecting enriched gene sets and pathways, the detected enriched gene sets and pathways also contain invaluable information that can in turn improve the power of DE analysis. Therefore, integrating GSEA and DE analysis into a joint statistical framework can potentially improve the power of both. In the workshop, Dr. Zhou described a Bayesian hierarchical model (iDEA) to integrate GSEA and DE analysis. With simulations, Dr. Zhou show that, by integrating GSEA with DE, their method dramatically improves the power of DE analysis and the accuracy of GSEA over commonly used existing approaches. Dr. Zhou also illustrate the benefits of their new method with applications to two published scRNAseq data sets.

4.5. Fast and accurate alignment of single-cell RNA-seq samples using kernel density matching

With technologies improved dramatically over recent years, single cell RNA-seq (scRNA-seq) has been transformative in studies of gene regulation, cellular differentiation, and cellular diversity. As the number of scRNA-seq datasets increases, a major challenge will be the standardization of measurements from multiple different scRNA-seq experiments enabling integrative and comparative analyses. However, scRNA-seq data can be confounded by severe batch effects and technical artifact. In addition, scRNA-seq experiments generally capture multiple cell-types with only partial overlaps across experiments making comparison and integration particularly challenging. To overcome these problems, Dr. Chen and collaborators have developed a method, dmatch, which can both remove unwanted technical variation and assign the same cell(s) from one scRNA-seq dataset to their corresponding cell(s) in another dataset. By design, their approach can overcome compositional heterogeneity and partial overlap of cell types in scRNA-seq data. Dr. Chen further showed that this method can align scRNA-seq data accurately across tissues biopsies.

4.6. A statistical simulator scDesign for rational scRNA-seq experimental design

Single-cell RNA-sequencing (scRNA-seq) has revolutionized biological sciences by revealing genome-wide gene expression levels within an individual cell. However, a critical challenge faced by researchers is how to optimize the choices of sequencing platforms, sequencing depths, and cell numbers in designing scRNA-seq experiments, so as to balance the exploration of the depth and breadth of transcriptome information. In the workshop, Dr. Li present a flexible and robust simulator, scDesign, the first statistical framework for researchers to quantitatively assess practical scRNA-seq experimental design in the context of differential gene expression analysis. In addition to experimental design, scDesign also assists computational method development by generating high-quality synthetic scRNA-seq datasets under customized experimental settings. In an evaluation based on 17 cell types and six different protocols, scDesign outperformed four state-of-the-art scRNA-seq simulation methods and led to rational experimental design.

4.7. Reconstructing gene regulatory dynamics along pseudotemporal trajectories using single-cell RNA-seq

Single-cell RNA-seq (scRNA-seq) provides a powerful technology for analyzing gene expression landscape of individual cells in a heterogeneous cell population. Ordering cells along a pseudotemporal trajectory based on cells progressively changing transcriptome is a useful way to elucidate cells developmental lineages and decode dynamic gene expression programs along developmental processes. Today, scRNA-seq is the most widely used high-throughput single-cell functional genomic technology. However, this technology only measures transcriptome and does not directly provide information on cis-regulatory element (CRE) activities. Building upon previous work on predicting chromatin accessibility using RNA-seq in bulk samples, Dr. Ji developed a new method for predicting CRE activities in single cells using scRNA-seq. In the workshop, Dr. Ji introduced their new tool that uses scRNA-seq to construct cells pseudotemporal trajectories and infer CREs dynamic activities along pseudotime. Using this method, one can conduct pseudotime analysis of transcriptome and regulome simultaneously using only scRNA-seq data. Analyses of the Human Cell Atlas data demonstrate that this method is capable of reconstructing cells gene regulatory programs along developmental processes.

4.8. Reconstructing haplotypes from bulk-sequencing data

Pooled sequencing (Pool-seq) is a next-generation sequencing (NGS) strategy where the genomes of several individuals from a population are grouped together and bulk-sequenced. Pool-seq provides an efficient and cost-effective alternative to genome sequencing of individuals or single cells, especially in contexts where pathogen genomes are inherently mixed. To determine the frequencies of individual-level polymorphisms and linkage disequilibrium (LD) from a population, the aggregated variation data must be deconvoluted in silico, an even more difficult task when haplotypes are not previously known and must be assembled de novo. Dr. Long has proposed a program, PoolHap, approximates the genotypic resolution of single-cell sequencing using only Pool-seq data by integrating population genetics models with genomics algorithms to reconstruct haplotypes.

4.9. Missing Imputation in Single-Cell RNA Sequencing Data

Single-cell RNA-sequencing (ScRNA-seq) technology is widely used to obtain genome-wide gene expression data at single-cell level. Often ScRNA-seq data contains large number of missing values or zero gene expression levels, which could be either biologically driven or technically driven due to the low capture efficiency of the sequencing technology. This imposes a great challenge to the downstream analysis as many (advanced) data analysis tools/models can not deal with missing values. Dr. Chowdhury from Icahn School of Medicine at Mount Sinai and team members developed a novel imputation method: DreamAI which is a consensus imputation algorithm based on multiple imputation strategies involving prediction-based imputation algorithms, machine learning algorithms, nearest neighbor clustering and low rank matrix approximation algorithms. Dr. Chowdhury apply DreamAI on ScRNA-seq data to impute the missing values and compare its performance with some existing methods.

Acknowledgements

We thank Dr. Nancy Zhang and Dr. Yunlong Liu for chairing the two discussion sections during the workshop. We thank all the workshop participates for all the great talks and discussions during the workshop period.