Novel Statistical Methods in Analyzing Single Cell Sequencing Data

by

Zhe Sun

BS in Biological Sciences, Fudan University, China, 2012

MS in Biostatistics, Emory University, 2014

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Zhe Sun

on

July 25, 2019

Dissertation Advisor:

Ying Ding, PhD
Assistant Professor, Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Dissertation Co-Advisor:

Wei Chen, PhD
Associate Professor, Departments of Pediatrics, School of Medicine
University of Pittsburgh

Committee Members:

Ming Hu, PhD
Assistant Staff, Department of Quantitative Health Sciences, Lerner Research Institute
Cleveland Clinic Foundation, Cleveland, Ohio

Yong Seok Park, PhD
Assistant Professor, Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Kong Chen, PhD
Assistant Professor, Department of Medicine, School of Medicine
University of Pittsburgh

Ying Ding, PhD
Wei Chen, PhD

Novel Statistical Methods in Analyzing Single Cell Sequencing Data

Zhe Sun

University of Pittsburgh, 2019

Abstract

Understanding biological systems requires the knowledge of their individual components. Single cell RNA sequencing (scRNA-Seq) becomes a revolutionary tool to investigate cell-to-cell transcriptomic heterogeneity, which cannot be obtained in population-averaged measurements such as the bulk RNA-Seq. This dissertation focuses on developing novel statistical methods for analyzing droplet-based single cell data, which includes clustering methods to identify cell types from single or multiple individuals, and a joint clustering approach to analyze paired data from Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-Seq), a new state-of-art technology that allows the detection of cell surface proteins and transcriptome profiling within the same cell simultaneously.

In the first part of this dissertation, I developed DIMM-SC, a Dirichlet mixture model which explicitly models the raw UMI count for clustering droplet-based scRNA-Seq data and produces cluster membership with uncertainties. Both simulation studies and real data applications demonstrated that overall, DIMM-SC achieves substantially improved clustering accuracy and much lower clustering variability compared to other clustering methods.

In the second part, I developed BAMM-SC, a novel Bayesian hierarchical Dirichlet mixture model to cluster droplet-based scRNA-Seq data from population studies. BAMM-SC takes raw count data as input and accounts for data heterogeneity and batch effect among multiple individuals in a unified Bayesian hierarchical model framework. Extensive simulation studies and applications to multiple in house scRNA-Seq datasets demonstrated that BAMM-SC outperformed existing clustering methods with improved clustering accuracy.

In the third part, I developed BREM-SC, a novel random effects model that jointly cluster the paired data from CITE-Seq simultaneously. Simulations and analysis of in-house real data sets were performed, which successfully demonstrated the validity and advantages of our method in understanding the heterogeneity and dynamics of various cell populations.

Contribution to public health:

Recent droplet-based single cell sequencing technology and its extensions have brought revolutionary insights to the understanding of cell heterogeneity and molecular processes at single cell resolution. I believe the proposed statistical approaches in this dissertation for single cell data will help us fully understand cell identity and function. This will promote the innovation for the traditional public health and medical research.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation covers the major methodology work during my five years' PhD studies. First of all, I would like to express the greatest gratitude to my dissertation advisor, Dr. Wei Chen and Dr. Ying Ding whose detailed guidance, patience and continued motivation during this entire journey has not only helped in earning this degree but also increased my understanding of statistical theory. Both of them provide me with a great amount of help and advices on my research, and serve as role models for students like me. Having the chance to work with them is my great honor and I will benefit from it for my entire life.

I want to thank my committee members Dr. Ming Hu, Dr. Kong Chen and Dr. Yong Seok Park for serving on my dissertation committee and each providing me with their expert guidance on this dissertation. I would like to give special thanks to Dr. Ming Hu for his continued support and detailed instructions for the first two papers, when the Bayesian statistics was very new to me. I really learnt a lot from him. I also want to thank all my former and current lab mates for their support and help in both research and daily life.

Finally, I am extremely grateful for my supportive parents who always stand as pillars of strength, providing continuous emotional and mental support during this long but exciting journey. I dedicate this work to them.

1.0 Introduction

1.1 Transcriptomic Data

RNA is essential in coding, regulation and expression of genes. For a certain gene to be expressed and coded into protein, messenger RNA (mRNA) is necessary, which is the RNA that conveys information from the genome to ribosome to instruct the coding of proteins. This step is called transcription. The transcriptome is the set of all RNA molecules in one cell or a population of cells and reflects the genes that are being actively expressed at any given time. Transcript profiling (also known as transcriptomics) is defined as the simultaneous quantitation of multiple mRNAs in a biological sample. The study of transcriptomics examines the expression level of RNAs in a given cell population. As many thousands of gene transcripts can be quantified, transcriptomics provides a way of gaining experimental information on a biological system relatively quickly.

RNA sequencing (RNA-Seq) is one of the transcriptomics techniques. It uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment (Wang, et al., 2009). Expression is quantified by counting the number of reads that mapped to each locus in the step of transcriptome assembly. Expression is quantified to study cellular changes in response to external stimuli, differences between healthy and diseased states, and other research questions. A number of organism-specific transcriptome databases have been constructed and annotated to aid in the identification of genes that are differentially expressed in distinct cell populations.

1.2 Overview of Single Cell RNA Sequencing

Cells are the basic biological units of multicellular organisms. Within a cell population, individual cells vary in their gene expression levels, reflecting the dynamics of transcription across cells (Shalek, et al., 2014; Spencer, et al., 2009). Traditional bulk RNA-Seq technologies profile the average gene expression level of all cells in the population. In contrast, recent single cell RNA sequencing (scRNA-Seq) technology has the advantage in generating expression measurement for each individual cell. It can be used to dissect transcriptomic heterogeneity, which cannot be obtained in population-averaged measurements such as the bulk RNA-Seq. scRNA-Seq studies have led to the discovery of novel cell types and provided insights into regulatory networks during development.

Commercially available, microfluidic-based scRNA-Seq approaches have limited throughput (Pollen, et al., 2014). Plate-based methods often require time-consuming fluorescence-activated cell sorting (FACS) into many plates that must be processed separately (duVerle, et al., 2016; Jaitin, et al., 2014). To overcome these challenges, the newly developed a droplet-based system enables parallel processing with digital counting of thousands of single cells in a short period of time (Macosko, et al., 2015; Zheng, et al., 2017).

More recently, 10X Genomics has released a commercialized droplet-based Chromium system, which is a microfluidics platform based on Gel bead in Emulsion (GEM) technology. GEM generation takes place in a multiple-channel microfluidic chip that encapsulates single gel beads. It is efficient and cost-effective in isolating thousands of single cells with an average running time of ten minutes. Reverse transcription takes place inside each droplet. Unique Molecular Identifiers (UMI) was introduced as a barcoding technique to reduce amplification noise. In the parallelized droplet based systems, samples are processed in parallel, which allow for the analysis of a much larger number of cells. It performs direct counting of molecule copies using UMI and the detection result of UMI is minimally affected by gene length, resulting in low transcript bias.

1.3 Clustering Analysis

Clustering analysis is an unsupervised study where data of similar types are put into one cluster while data of another types are put into different cluster. This is the process of dividing data elements into different groups in such a way that the elements within a group possess high similarity while they differ from the elements in a different group. Broadly speaking, clustering can be divided into two subgroups: hard clustering and soft clustering. In hard clustering, each data point either belongs to a cluster completely or not. On the contrary, instead of putting each data point into a separate cluster, soft clustering provides a probability or likelihood of that data point to be in those clusters is assigned.

Cell clustering based on transcriptomic profiles plays an important role in single cell analysis. Different types of cells have different gene expression profiles (Silbereis, et al., 2016). Thus, they can be identified by these profiles, especially by expression of certain genes that tend to have cell-specific expression (marker genes). Characterization of these profiles has recently been facilitated by scRNA-Seq techniques (Tang, et al., 2009). Clustering scRNA-Seq data identifies and characterizes cell subtypes from heterogeneous tissues and enhances understanding of cell identity and functionality. The identification of cell types from a mass of heterogeneous cells can be used as covariates in downstream differential expression analysis. However, the intrinsic features of droplet-based scRNA-Seq data pose great statistical and computational challenges, particularly in handling the large number of single cells, data sparseness of UMI count, and multiple levels of uncertainties in a nested experiment design.

To cluster cells for scRNA-Seq data, unsupervised clustering methods such as K-means clustering, hierarchical clustering, and density-based clustering approach (Rodriguez, et al., 2014) have been used. Among them, a commonly used method for single cell clustering is K-means clustering on dimensional reduced data. The K-means (Gawad, et al., 2016) is one of the famous hard clustering algorithms. It takes the number of clusters as input parameter, and partitions a set of objects into clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Although K-means algorithm is significantly sensitive to the initial randomly selected cluster centers, it has been adapted to many scientific fields.

Choosing a particular clustering algorithm is solely dependent on the type of the data to be clustered and the purpose of the clustering applications. Hard clustering algorithm like K-means algorithm is suitable for exclusive clustering task. In some situations, we cannot directly consider that data belongs to only one cluster. It may be possible that some data's properties contribute to more than one cluster. For scRNA-Seq data, it is more appropriate to use soft clustering approaches, since a particular cell may be categorized into multiple different categories. For example, development often involves pluripotent cells transitioning into other cell types or in a series of different stages.

1.4 Overview of Cellular Indexing of Transcriptomes and Epitopes by Sequencing

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq) is a recently developed cutting-edge technology. It is the first technique that can measure single cell surface protein and mRNA expression level simultaneously in the same cell (Stoeckius, et al., 2017), and fully compatible with droplet-based single cell transcriptome sequencing (scRNA-Seq) technology by 10X Genomics Chromium.

Droplet-based single cell sequencing technology and its extension has brought revolutionary insights to our understanding of cell heterogeneity and molecular processes. Embedded in the scRNA-Seq experiment, the recently developed CITE-Seq (Stoeckius, et al., 2017) and cell hashing technologies (Stoeckius, et al., 2018) allow for immunophenotyping of single cells based on the expression of specific cell surface proteins together with simultaneous transcriptome profiling and sample origin detection within a cell. These promising and popular technologies provide the unprecedented opportunity for jointly analyzing transcriptome and cell-surface proteins, and examining the complex relationship between mRNA and surface protein at single cell level and in a cost-effective way.

CITE-Seq measures cellular surface protein abundance by counting the sequencing of oligos conjugated to antibodies directed against different cell surface proteins. Traditionally, transcriptomic and proteomic studies utilize different technologies. Previously, scRNA-Seq technologies are developed to obtain cellular transcriptomic expression level, while flow cytometry techniques are for cellular protein level measurements. At that time, there is no “bridge” to connect scRNA-Seq and flow cytometry, leading to the fact that there is no cell-matched RNA and protein data. However, the development of CITE-Seq technology, which uses DNA-barcoded antibodies to quantitatively measure protein levels through sequenceable readout, can match mRNA and surface protein data with unique cell barcode for each individual cell.

The rapid advances in single cell technologies help researchers better understand cell heterogeneity and identify cell types, which is a crucial step in single cell analyses, and it leads to the high demand of novel statistical methods and tools to analyze data with different characteristics. In companion with the great advances of these new technologies and their important applications in biomedical research, statistical and computational methods are emerging. Numerous methods have been recently proposed to address different aspects of single cell data such as clustering, differential gene analysis, and trajectory analysis. However, current statistical methods for jointly analyzing data from scRNA-Seq and CITE-Seq are still immature, which motivates us to develop some novel statistical approaches that fully utilizes the advantages and unique features of these single cell multi-omics data.

2.0 DIMM-SC: A Dirichlet Mixture Model for Clustering Droplet-Based Single Cell Transcriptomic Data

Single cell transcriptome sequencing (scRNA-Seq) has become a revolutionary tool to study cellular and molecular processes at single cell resolution. Among existing technologies, the recently developed droplet-based platform enables efficient parallel processing of thousands of single cells with direct counting of transcript copies using Unique Molecular Identifier (UMI). Despite the technology advances, statistical methods and computational tools are still lacking for analyzing droplet-based scRNA-Seq data. Particularly, model-based approaches for clustering large-scale single cell transcriptomic data are still under-explored.

I developed DIMM-SC, a **Di**richlet **m**ixture **m**odel for clustering droplet-based **s**ingle **c**ell transcriptomic data. This approach explicitly models UMI count data from scRNA-Seq experiments and characterizes variations across different cell clusters via a Dirichlet mixture prior. I performed comprehensive simulations to evaluate DIMM-SC and compared it with existing clustering methods such as K-means, CellTree and Seurat. In addition, I analyzed public scRNA-Seq datasets with known cluster labels and in-house scRNA-Seq datasets from a study of systemic sclerosis with prior biological knowledge to benchmark and validate DIMM-SC. Both simulation studies and real data applications demonstrated that overall, DIMM-SC achieves substantially improved clustering accuracy and much lower clustering variability compared to other existing clustering methods. More importantly, as a model-based approach, DIMM-SC is able to quantify the clustering uncertainty for each single cell, facilitating rigorous statistical inference and biological interpretations, which are typically unavailable from existing clustering methods.

2.1 Introduction

Single cell RNA sequencing (scRNA-Seq) technologies have advanced rapidly in recent years (Gawad, et al., 2016). Among them, the newly developed droplet-based technologies have generated great interests (Macosko, et al., 2015; Zheng, et al., 2017). They are able to measure the transcriptome of thousands of single cells simultaneously in a short time period and at a relatively low cost (Macosko, et al., 2015; Zheng, et al., 2016). More attractively, droplet-based technologies utilize Unique Molecular Identifier (UMI) to annotate the 3' end of each transcript in order to reduced PCR amplification bias, increase transcript capture efficiency, and substantially minimize batch effect (Islam, et al., 2014; Kivioja, et al., 2012). More recently, 10X Genomics has released a commercialized droplet-based Chromium system, which is efficient and cost-effective in isolating thousands of single cells with average running time of ten minutes based on the Gel bead in EMulsion (GEM) technology. They used this platform to comprehensively characterize and profile peripheral blood mononuclear cells (PBMC) (Zheng, et al., 2017). Harnessing the power of these exciting new technologies, droplet-based scRNA-Seq has brought revolutionary insights to understand cellular and molecular processes at single cell resolution.

One important question in the analysis of scRNA-Seq data is to identify and characterize cell subtypes from heterogeneous tissues, which is essential to fully understand cell identity and cell function. Clustering methods have been extensively studied for many areas in the past decades. For example, unsupervised clustering methods such as K-means clustering, hierarchical clustering, and Adaptive Density Peak (ADP) clustering (Rodriguez and Laio, 2014; Wang and Xu, 2015), can be applied to droplet-based scRNA-Seq data after certain data transformation. In addition, tailored methods such CellTree and Seurat have been proposed to analyze scRNA-Seq data with the motivation from early generation platforms (duVerle, et al., 2016; Jaitin, et al., 2014).

However, clustering methods tailored to droplet-based scRNA-Seq data are largely lagging behind. Although existing clustering methods can be adapted, there are at least three key limitations of using those methods to cluster droplet-based scRNA-Seq data. First of all, most existing methods are developed for continuous data (e.g. Fragments Per Kilobase of transcript per Million (FPKM) or log-transformed count data) while droplet-based scRNA-Seq data consist of the discrete count of the unique UMIs, which are direct measurements of transcript copies from each gene. Converting UMI counts into continuous measure will alter the straightforward interpretation of UMI, thus it is more appealing and reasonable to directly model the count data. Second, most existing methods are designed for the early generation of scRNA-Seq technologies that measure transcriptome across a relatively small number of single cells. It is unclear how these methods can be scaled up to cluster droplet-based scRNA-Seq data, which usually contain thousands of single cells. Last but not the least, most existing methods only provide a "hard" cluster membership for each cell without statistical uncertainty quantification. In order to conduct rigorous statistical inference and achieve reliable data interpretation, different sources of uncertainties in droplet-based scRNA-Seq data need to be explicitly taken into consideration in the clustering analysis.

To fill in these gaps, I proposed DIMM-SC, a **Di**richlet **m**ixture **m**odel for clustering droplet-based **sc**RNA-Seq data. DIMM-SC explicitly models both the within-cluster and between-cluster variability of the UMI count data, leading to rigorous quantification of clustering uncertainty for each single cell. I also implemented an efficient expectation-maximization (E-M) algorithm (Dempster, et al., 1977) for fast convergence. Furthermore, I proposed different strategies for initial value selection to ensure algorithm robustness. In the following sections, I first introduce the unique features of droplet-based scRNA-Seq data, as well as the details of the DIMM-SC method. Next, I compare the performance of DIMM-SC with three popular clustering methods, including K-means clustering, CellTree and Seurat, in both simulation studies and real data applications. K-means is one of the most popular clustering methods and has been used in the first 10X genomics publication (Zheng, et al., 2017). CellTree has been recently developed to cluster scRNA-Seq data based on Latent Dirichlet Allocation (LDA) (duVerle, et al., 2016). Seurat is a deterministic approach which relies on a graph-based clustering approach (Satija, et al., 2015).

2.2 Methods

2.2.1 Data description

The droplet-based scRNA-Seq data can be summarized into a UMI count matrix (Table 1), in which each row represents one gene and each column represents one single cell. Each entry in the UMI count matrix is the number of transcripts (unique UMIs) for one gene in one single cell. Compared to the data generated from early generation of scRNA-Seq technologies, droplet-based scRNA-Seq data have three important features (Gawad, et al., 2016; Stegle, et al., 2015; Zheng, et al., 2017). First, each experiment can generate thousands of cells, which dramatically increase the data dimension and computational burden. Second, the use of UMI can reduce PCR amplification bias and quantify the copies of captured molecules. Droplet-based sequencing protocol amplifies the 3' end of the transcript, so the number of UMI is independent of the total transcript length. The normalization method used in RPKM and FPKM, which adjusts for total transcript length, is invalid for analyzing droplet-based scRNA-Seq data. Therefore, the raw count data should be directly modeled to retain their biological interpretations. Third, the UMI count matrix is extremely sparse, and thus violates the statistical assumption of many existing clustering methods. Figure 18

(Appendix A) lists the empirical distribution of the UMI counts for a few representative genes, demonstrating the non-ignorable proportion of zeroes for different levels of expression. Pre-selection of informative single cells and informative genes are necessary before the downstream clustering analysis. After clustering analysis, the results are usually visualized by a t-distributed stochastic neighbor embedding (t-SNE) approach (van der Maaten and Hinton, 2008), which embeds high-dimensional transcriptome data into a two-dimensional scatter plot. Note that t-SNE is a visualization tool, and it is not intended to be used for clustering scRNA-Seq data.

Table 1. An example of the raw UMI count table from droplet-based scRNA-Seq data

	Cell 1	**Cell 2**	**Cell 3**	**…**	**Cell 2,000**
Gene1	0	0	0	…	0
Gene2	1	0	1	…	0
Gene3	23	12	9	…	3
…	…	…	…	…	…
Gene 10,000	22	6	7	9	3

2.2.2 Statistical model

I start with a matrix $\boldsymbol{X}$, of which the element X_{ij} represents the number of unique UMIs for gene i in cell j where i runs from 1 to the total number of genes G, and j runs from 1 to the total number of cells C (as showed in Table 1). X_{ij} is the count for the absolute number of transcripts. I denote the j th column of this matrix, which gives the number of unique UMIs in the j th single cell, by a vector $\boldsymbol{x_j} = \left(x_{1j}, x_{2j}, \dots, x_{Gj}\right)$, where $j = 1, \dots, C$. I assume that $\boldsymbol{x_j}$ is generated from a multinomial distribution with parameter vector $\boldsymbol{p_j} = (p_{1j}, p_{2j}, \dots, p_{Gj})$. The element of $\boldsymbol{p_j}$, p_{ij}, is the probability that a unique UMI count taken from cell j belongs to gene i. This gives a likelihood for each cell:

$$P(\boldsymbol{x_j}|\boldsymbol{p_j}) = \frac{T_j!}{\prod_{i=1}^{G} x_{ij}!} p_{1j}^{x_{1j}} p_{2j}^{x_{2j}} \dots p_{Gj}^{x_{Gj}},$$

where $T_j = \sum_i x_{ij}$ is the total number of unique UMIs for the j th cell. The joint likelihood of all C cells is the product of the likelihood for each cell: $\prod_{j=1}^{C} P(\boldsymbol{x_j}|\boldsymbol{p_j})$.

In a Bayesian framework, I need to define a prior distribution for the multinomial parameter probability vector $\boldsymbol{p_j}$. For multinomial distribution, a commonly used conjugate prior is the Dirichlet distribution. Specifically, I assume that the proportion $\boldsymbol{p_j} = (p_{1j}, p_{2j}, \dots, p_{Gj})$ follows a Dirichlet prior distribution $Dir(\boldsymbol{\alpha}) = Dir(\alpha_1, \alpha_2, \dots, \alpha_G)$:

$$P(\boldsymbol{p_j}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} p_{1j}^{\alpha_1 - 1} p_{2j}^{\alpha_2 - 1} \dots p_{Gj}^{\alpha_G - 1},$$

where $B(\boldsymbol{\alpha})$ is Beta function with parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_G)$. All the elements in $\boldsymbol{\alpha}$ are strictly positive ($\alpha_i > 0$). The mean and variance of p_{ij} are $\alpha_i/|\boldsymbol{\alpha}|$ and $\alpha_i(|\boldsymbol{\alpha}| - \alpha_i)/(|\boldsymbol{\alpha}|^2(|\boldsymbol{\alpha}| + 1))$, respectively, where $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2 + \cdots + \alpha_G$. A large $|\boldsymbol{\alpha}|$ gives small variance about the proportions $\boldsymbol{p_j}$, while a small $|\boldsymbol{\alpha}|$ leads to widely spread $\boldsymbol{p_j}$'s. When the cell population is homogeneous, I assume that $\boldsymbol{p_j}$'s all follow the same prior distribution $Dir(\boldsymbol{\alpha})$, and the full likelihood function is as follows:

$$P(\boldsymbol{x_j}|\boldsymbol{\alpha}) = \int P(\boldsymbol{x_j}|\boldsymbol{p_j}) P(\boldsymbol{p_j}|\boldsymbol{\alpha})\, d\boldsymbol{p_j} = \frac{T_j!}{\prod_{i=1}^{G} x_{ij}!} \left(\prod_{i=1}^{G} \frac{\Gamma(x_{ij} + \alpha_i)}{\Gamma(\alpha_i)} \right) \frac{\Gamma(|\boldsymbol{\alpha}|)}{\Gamma(T_j + |\boldsymbol{\alpha}|)}$$

I then assume that the cell population consists of K distinct cell types, where K can be pre-defined according to prior biological knowledge or can be estimated through model fitting. To provide a more flexible modeling framework and allow for unsupervised clustering, I extend the aforementioned single Dirichlet prior to a mixture of K Dirichlet distributions, indexed with $k =$

$1, \dots, K$, each with parameter $\boldsymbol{\alpha}_{(k)}$. If cell j belongs to the k th cell type, its gene expression profile $\boldsymbol{p_j}$ follows a cell-type-specific prior distribution $Dir(\boldsymbol{\alpha}_{(k)})$. The full likelihood function is then obtained by multiplying the Dirichlet mixture prior by the multinomial likelihood.

2.2.3 E-M algorithm for fitting the mixture of Dirichlet prior

I now use a latent variable vector $\boldsymbol{Z}$ with elements z_j to represent the cell type label for the cell j. This allows me to maximize the log posterior distribution using the E-M algorithm (Dempster, 1977). I have:

$$P(\boldsymbol{x_j}|z_j = k, \boldsymbol{\alpha}_{(k)}) \propto \left(\prod_{i=1}^{G} \frac{\Gamma(x_{ij} + \alpha_{ik})}{\Gamma(\alpha_{ik})}\right) \frac{\Gamma(|\boldsymbol{\alpha}_{(k)}|)}{\Gamma(T_j + |\boldsymbol{\alpha}_{(k)}|)}$$

and $P(z_j = k) = \pi_k$, where π_k is the proportion of the k th cell type among all cells. I can treat z_j as missing data, and use the E-M algorithm to estimate $\alpha_{1k}, \alpha_{2k}, \dots, \alpha_{Gk}$ and π_k. The complete log likelihood is

$$\log \prod_{j=1}^{C} P(\boldsymbol{x_j}, z_j = k) = \sum_{j=1}^{C} I(z_j = k) \log\left\{\left(\prod_{i=1}^{G} \frac{\Gamma(x_{ij} + \alpha_{ik})}{\Gamma(\alpha_{ik})}\right) \frac{\Gamma(|\boldsymbol{\alpha}_{(k)}|)}{\Gamma(T_j + |\boldsymbol{\alpha}_{(k)}|)}\right\}$$

The formula for updating $\alpha_{1k}, \alpha_{2k}, \dots, \alpha_{Gk}$ is derived from the Minka's fixed-point iteration for the leaving-one-out likelihood (Minka, 2000):

$$\hat{\alpha}_{ik}^{(t+1)} = \alpha_{ik}^{(t)} \frac{\sum_{j=1}^{C} \delta_{jk}\left\{x_{ij}/(x_{ij} - 1 + \alpha_{ik}^{(t)})\right\}}{\sum_{j=1}^{C} \delta_{jk}\left\{T_j/(T_j - 1 + |\boldsymbol{\alpha}_{(k)}^{(t)}|)\right\}}.$$

We repeat the above steps until the convergence of log likelihood or a maximum number of iterations is reached (see detailed algorithm in Appendix A).

2.2.4 Selection of the number of clusters and initial values

To implement DIMM-SC, it is critical to select the total number of clusters and the initial values for the E-M algorithm. Specially, the number of clusters K can be defined with prior knowledge or can be selected from model selection criteria such as AIC or BIC (Akaike, 1974; Schwarz, 1978). Meanwhile, there are many methods to determine the initial values of $\alpha_1, \alpha_2, \dots, \alpha_G$ in the E-M algorithm for fitting the Dirichlet mixture model. For example, Ronning (1989) suggests to estimate $\sum_{i=1}^{G} \alpha_i$ by

$$log \sum_{i=1}^{G} \alpha_i = \frac{1}{G-1} \sum_{i=1}^{G-1} log \left(\frac{E(p_i)\left(1 - E(p_i)\right)}{var\ (p_i)} - 1 \right)$$

where $E(p_i)$ can be approximated by $(\sum_{j=1}^{C} x_{ij} / \ T_j)/C)$ (Ronning, 1989). An alternative approach is to estimate the initial values using a method of moment estimates proposed by Weir and Hill (Weir and Hill, 2002). In this paper, I applied the K-means clustering to obtain the initial clustering results, and then used either the Ronning's method or the Weir and Hill's method to obtain the initial estimates of Dirichlet parameter α.

2.3 Simulation Studies

I performed comprehensive simulation studies to compare DIMM-SC with three existing clustering methods, including K-means clustering, Seurat and CellTree. The first two are deterministic approaches and the third one is a probabilistic approach.

In the simulation set-up, the UMI count matrix was sampled from the proposed Dirichlet mixture model. Specially, for a fixed total number of cell clusters K, I first pre-defined the values of $\alpha_{(k)} = \left(\alpha_{1(k)}, \alpha_{2(k)}, \dots, \alpha_{G(k)}\right)$ for the k th cell cluster, and then sampled the proportion $p_j = (p_{1j}, p_{2j}, \dots, p_{Gj})$ from a Dirichlet distribution $Dir\left(\alpha_{(k)}\right)$. Next, I sampled the UMI count vector $\boldsymbol{x_j}$ for the j th cell from the multinomial distribution $Multinomial(T_j, \boldsymbol{p_j})$. I fixed T_j as a constant across all cells.

In the simulation studies, I considered the following seven clustering methods. (1) DIMM-SC + K-means + Ronning (hereafter referred as DIMM-SC-KR), in which I used the K-means clustering to obtain the initial values of clustering labels and then used the Ronning's method to estimate initial values of $\boldsymbol{\alpha}$; (2) DIMM-SC + K-means + Weir (hereafter referred as DIMM-SC-KW), in which I used the K-means clustering to obtain the initial values of clustering labels and used the Weir and Hill's method to estimate initial values of $\boldsymbol{\alpha}$; (3) DIMM-SC + random + Ronning (hereafter referred as DIMM-SC-RR), in which I randomly selected the initial values of clustering labels and used the Ronning's method to estimate initial values of $\boldsymbol{\alpha}$; (4) DIMM-SC + random + Weir (hereafter referred as DIMM-SC-RW), in which I randomly selected the initial values of clustering labels and used the Weir and Hill's method to estimate initial values of $\boldsymbol{\alpha}$**;** (5) K-meaning clustering; (6) CellTree, a LDA-based approach to cluster scRNA-Seq data; and (7) Seurat. To perform the simulation analysis using Seurat, I followed the tutorial instructions from the Seurat website and used all genes as input to perform Principal Component Analysis (PCA). After that, I followed the "jackstraw" procedure implemented in Seurat, and identified first ten PCs for their downstream algorithm. I fixed the number of PCs in all the simulation runs under each scenario. Since Seurat requires users to self-specify a resolution parameter with increased values leading to a greater number of clusters, the clustering results are very sensitive to this resolution parameter. Seurat suggests that setting this resolution parameter between 0.6-1.2 typically returns good results for datasets of around 3,000 cells, so I ran Seurat using resolution parameter with 0.6, 0.8, 1.0 and 1.2, and chose the one with the highest adjusted rand index (ARI) value in each simulation setting. Note that ARI is a commonly-used metric of the similarity between the estimated clustering labels and the true clustering labels (Rand, 1971).

I used the signal-to-noise ratio (SNR) to measure the magnitude of difference among different cell clusters. When $K = 2$, SNR is defined as:

$$SNR = \frac{\left|\boldsymbol{\alpha}_{(1)} - \boldsymbol{\alpha}_{(2)}\right|_1}{G\sqrt{Var\left(\boldsymbol{\alpha}_{(1)}\right) + var\left(\boldsymbol{\alpha}_{(2)}\right)}},$$

where $|.|_1$ is the L_1 norm of a vector. I performed comprehensive simulations to investigate how different SNRs, different sequencing depths, different total numbers of cells/genes/clusters, and different proportions of noisy genes affect the clustering results. To evaluate the performance of DIMM-SC and other competing clustering methods, I used the following two metrics: (1) clustering accuracy measured by ARI and (2) Stability (the standard deviation of ARI). I expect a good clustering method should achieve both high accuracy and high stability.

2.4 Results

2.4.1 Simulation studies

Figure 1A shows the boxplots of ARI for seven clustering methods across 100 simulations at different SNRs. Four DIMM-SC based methods (KR, KW, RR, RW) achieved comparable performance, which produced higher accuracy and lower variability than K-means clustering, Seurat and CellTree. When SNR is high (i.e., substantial differences among cell clusters), all seven methods performed well. However, when SNR is low (i.e., different cell clusters are similar), K-means clustering, Seurat and CellTree produced less accurate and more variable clustering results, while four DIMM-SC based methods still performed well.

Figure 1B shows the boxplots of ARI for seven clustering methods across 100 simulations, when the total number of clusters is 2, 3, 4, 5 and 8 respectively. The four DIMM-SC based methods, especially the two methods with randomly selected initial cluster labels (RR and RW), achieved better clustering accuracy (i.e., higher ARI) with more number of clusters. K-means clustering has high variability for more clusters, since it is a deterministic procedure and is more likely to end at a local optimum when the total number of clusters increases. CellTree performed worse for more clusters, partially due to the over-parameterized LDA model and lack of fit to highly heterogeneous data. Seurat was run under different default recommended parameters and the performance varies with different parameters.

Figure 1C~F list the boxplots of ARI for seven clustering methods across 100 simulations, for different number of genes (Figure 1C), different number of cells (Figure 1D), different sequencing depths (Figure 1E) and different number of informative genes (Figure 1F) (i.e., differentially expressed genes among clusters), respectively. Consistent across all these four scenarios, more information (i.e., more genes, more cells, higher sequencing depths and more informative genes) lead to higher clustering accuracy and lower clustering variability. Four DIMM-SC based clustering methods consistently outperformed K-means clustering, Seurat and CellTree in all these simulation settings, suggesting the advantage of DIMM-SC.

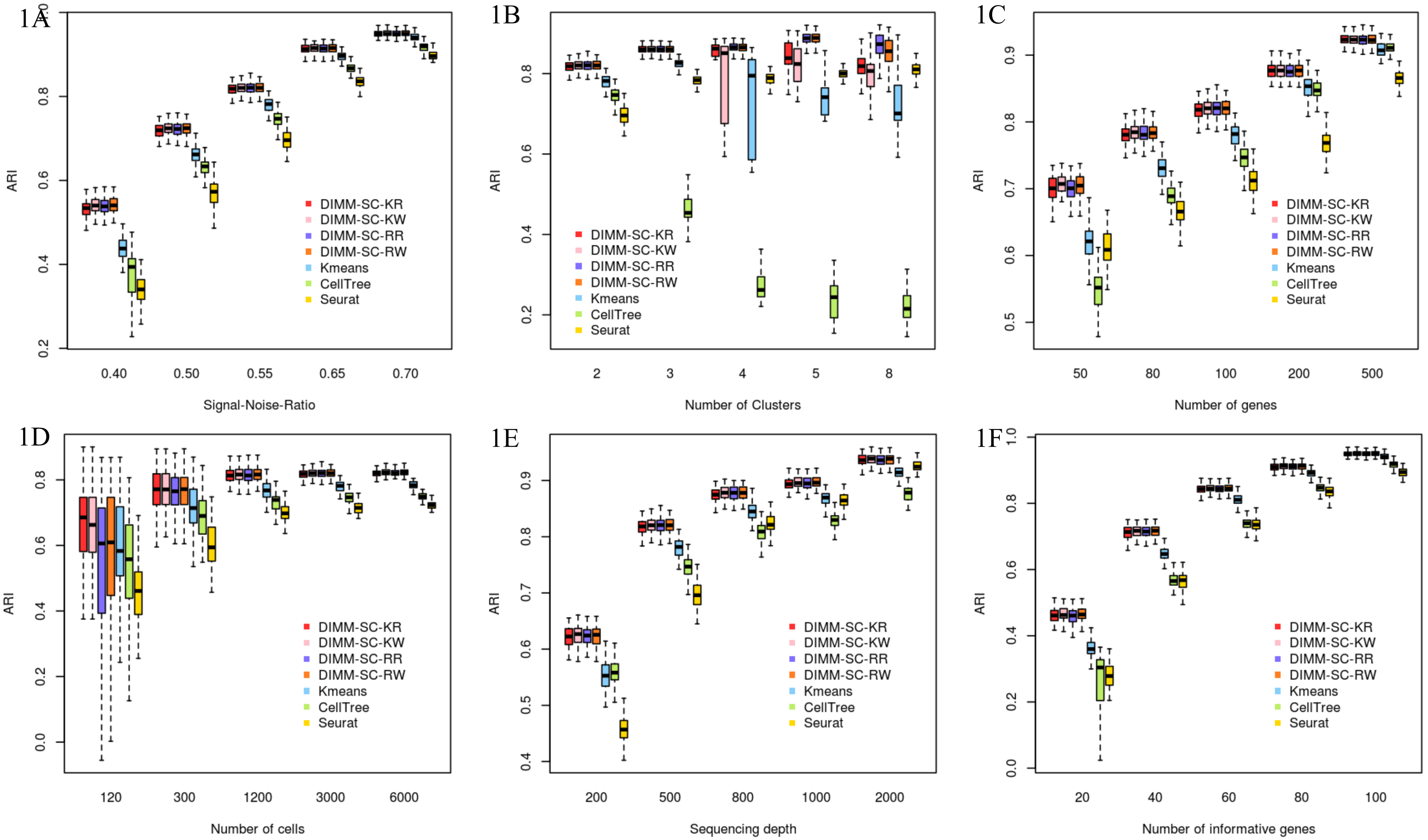


Figure 1. Boxplots of ARI for seven clustering methods in simulation studies

Boxplots of ARI investigating how different SNRs (A), number of clusters (B), number of genes (C), number of cells (D), sequencing depth (E) and the number of informative genes (F) affect clustering results.

2.4.2 Real data analysis: the publicly available 10X scRNA-Seq data

2.4.2.1 In silicon studies based on purified cell types from published scRNA-Seq data

To illustrate the application of DIMM-SC to real datasets, I first benchmarked our method against pre-defined measures in capturing true cell-to-cell similarities on published single-cell datasets. 10X Genomics has made eleven datasets from purified cell types available to public (Zheng, et al., 2017). Among which, over 10,000 cells were detected in most experiments. Here I considered two scenarios: (1) a simple case with cells from three highly distinct cell types (CD56+ NK cells, CD19+ B cells, and CD4+/CD25+ regulatory T cells); (2) a challenging case with cells from three similar cell types (CD8+/CD45RA+ naive cytotoxic T cells, CD4+/CD25+ regulatory T cells, and CD4+/CD45RA+/CD25- naive T cells) (Table 2). For visualization, I used the t-SNE algorithm to project the data into a two-dimensional space so that certain hidden structures in the data can be depicted intuitively (see the t-SNE visualization in Figure 19 and Figure 20 (Appendix A)).

I ran DIMM-SC, K-means clustering, CellTree and Seurat 50 times for both two scenarios. In the simple case, at each time, I randomly selected 1,000 CD56+ NK cells, 2,000 CD19+ B cells and 3,000 CD4+/CD25+ regulatory T cells from the 10X Genomics datasets, and combined them together. Thus the total number of cells for clustering is 6,000. Similarly, in the challenging case, 1,000 CD8+/CD45RA+ naive cytotoxic T cells, 2,000 CD4+/CD25+ regulatory T cells and 3,000 CD4+/CD45RA+/CD25- naive T cells were randomly selected at each time.

Cell types in each dataset were known as a priori and were further validated in the respective follow-up studies, providing a reliable gold standard to benchmark the clustering performance for each method. I compared the performance of four DIMM-SC methods with K-means clustering, CellTree and Seurat, in terms of clustering accuracy and stability.

In the simple case, I applied all these seven clustering methods on the top 100 variable genes ranked by their standard error among all cells. Table 3 shows that all methods provided good clustering results. Two DIMM-SC methods with randomly selected initial cluster labels (RR and RW) slightly outperformed K-means clustering in terms of accuracy and variability. For the challenging case, unlike what I did in the simple case, I chose different numbers of top variable genes. Table 3 and Figure 21 (Appendix A) show that the ARIs of CellTree and Seurat were lower than other methods when the total number of genes used for clustering was greater than 200. DIMM-SC slightly outperformed K-means clustering in terms of accuracy. K-means clustering made a great leap forward when the total number of genes increased to 300. However, there is no further improvement of ARI with K-means clustering when top 500 or more variable genes were used. Since in the challenging case, CD4+/CD25+ regulatory T cells and CD4+/CD45RA+/CD25- naive T cells were similar to each other, more and more noisy genes were included in the analysis when we increased the total number of genes, which undermined the performance of K-means clustering. Note that K-means clustering and Seurat were only able to provide a deterministic clustering label, while DIMM-SC and CellTree can additionally provide the probability that each cell belongs to each cluster.

Table 2. Total number of cells, genes and validated populations for two scenarios

Scenario	#Genes	#Cell	Cell type
Simple	32,738	8,385	CD56+ NK cells
		10,085	CD19+ B cells
		10,283	CD4+/CD25+ regulatory T cells
Challenging	32,738	11,953	CD8+/CD45RA+ naive cytotoxic T cells
		10,263	CD4+/CD25+ regulatory T cells
		10,479	CD4+/CD45RA+/CD25- naive T cells

Table 3. Performance of clustering in the simple case and the challenging case

#Genes	DIMM-SC-KR	DIMM-SC-KW	DIMM-SC-RR	DIMM-SC-RW	K-means clustering	CellTree	Seurat
The simple case							
100	0.952 (0.114)	0.951 (0.118)	0.982 (0.052)	**0.990 (0.002)**	0.951 (0.129)	0.983 (0.002)	0.983 (0.003)
The challenging case							
100	0.351 (0.140)	0.357 (0.140)	0.368 (0.140)	**0.408 (0.128)**	0.182 (0.012)	0.278 (0.018)	0.395 (0.027)
200	0.558 (0.014)	0.559 (0.014)	0.558 (0.014)	**0.559 (0.013)**	0.283 (0.050)	0.389 (0.022)	0.410 (0.017)
300	0.563 (0.013)	**0.564 (0.013)**	0.563 (0.013)	0.563 (0.013)	0.526 (0.063)	0.419 (0.023)	0.413 (0.022)
400	**0.571 (0.014)**	**0.571 (0.014)**	0.566 (0.040)	**0.571 (0.014)**	0.554 (0.014)	0.404 (0.050)	0.429 (0.012)
500	**0.572 (0.015)**	**0.572 (0.015)**	**0.572 (0.015)**	**0.572 (0.015)**	0.559 (0.014)	0.397 (0.067)	0.435 (0.011)
800	**0.562 (0.041)**	**0.562 (0.041)**	0.557 (0.057)	0.556 (0.056)	0.557 (0.041)	0.365 (0.078)	0.445 (0.011)

2.4.2.2 Real data analysis on the PMBC 68K dataset

To examine how DIMM-SC is applicable to large-scale dataset, I applied DIMM-SC-KR on the PBMC 68K dataset, which consists of >68,000 single cells. Among all 32,738 genes, I selected the top 1,000 genes with the highest variations. Figure 2A shows a clear separation of cell types as I expected. 11 purified sub-populations of PBMCs were used as the reference to identify the cell type of each single cell from the PBMC 68K dataset. I used the labels from cell classification analysis as the approximated truth. In this analysis, each cell was assigned to the purified population which has the highest correlation with its gene expression profile. I calculated ARIs between the true labels and inferred ones obtained from K-means clustering, CellTree, Seurat and DIMM-SC. The ARIs of K-means clustering, CellTree and DIMM-SC are 0.32, 0.28 and 0.41, respectively. To perform the analysis using Seurat, I used the default setting of Seurat to select the top 1,657 variable genes, and picked the first 22 PCs for the clustering analysis. The ARI of Seurat is 0.31, suggesting that DIMM-SC performed the best in the PMBC 68K dataset. Additionally, I highlighted vague cells in the t-SNE projection (Figure 2B), where vague cells are defined as cells with the largest posterior cluster-specific probability < 0.95. As shown in Figure 2B, most of vague cells are located at the boundary of different clusters, which reassuring the validity of the clustering results.

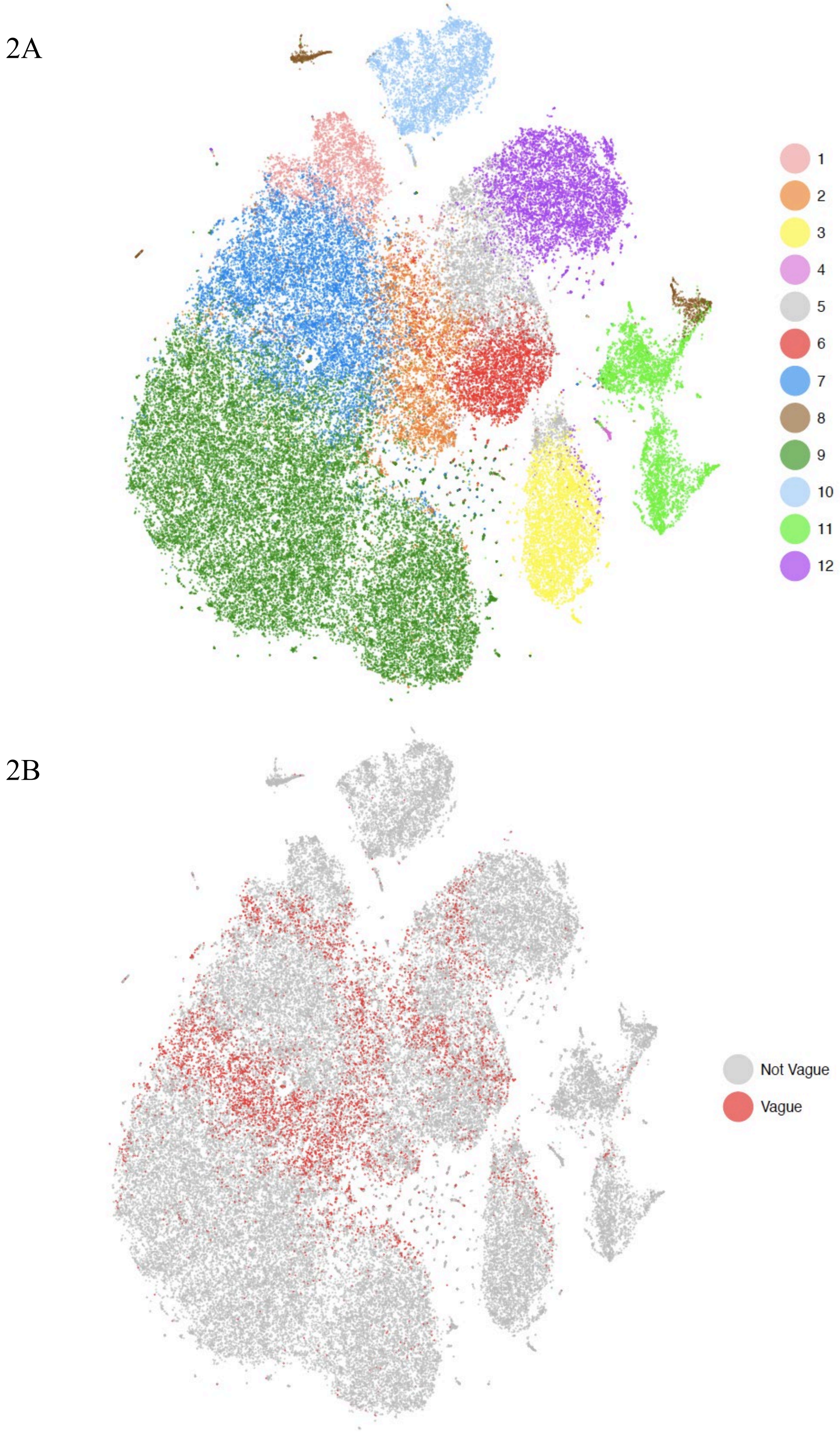


Figure 2. The t-SNE projection of 68K PBMCs, colored by the DIMM-SC clustering assignment and the illustration of vague cells with the largest posterior probability < 0.95

2.4.2.3 Analysis of the in-house scRNA-Seq data from systemic sclerosis study

Collaborating with investigators at the University of Pittsburgh, I am in the first place to use the 10X Chromium system to generate scRNA-Seq data in order to study systemic sclerosis. I applied DIMM-SC to the scRNA-Seq data of skin tissue collected from a systemic sclerosis patient. Starting from a UMI count matrix for 1,180 cells generated by the 10X genomics Cellranger pipeline, I first removed cells that had less than 300 genes expressed and filtered noisy genes that were expressed in less than five cells, then we extracted the top 1,000 highly variable genes based on their standard deviations. I set the total number of clusters to be six based on our prior knowledge and utilized the KR method to generate the initial cluster labels and the initial values for the parameter $\boldsymbol{\alpha}$. The six cell clusters from DIMM-SC included 92, 89, 45, 156, 469 and 271 cells, respectively. Figure 3 shows the t-SNE projection of the skin cells, colored by cluster labels inferred by DIMM-SC, and the dashed circles represent potential subtypes of skin cells according to the expressions of cell type specific markers. It is interesting that fibroblast cells exhibit two clusters, suggesting possible subtypes. For each cell cluster, I identified top marker genes that were differentially expressed between the specified cluster and all the other clusters. I recognized some subtypes of skin cells for the identified clusters based on the biological knowledge of cell specific markers, such as pericyte cells specifically expressed gene *RGS5*, T cells specifically expressed gene *IL32*, endothelial cells specifically expressed gene *VWF*, fibroblast cells specifically expressed gene *COL1A1*, basal keratinocyte cells specifically expressed gene *KRT14* and gene *KRT5*, and suprabasal keratinocyte cells specifically expressed gene *KRT1* and gene *KRT10*.

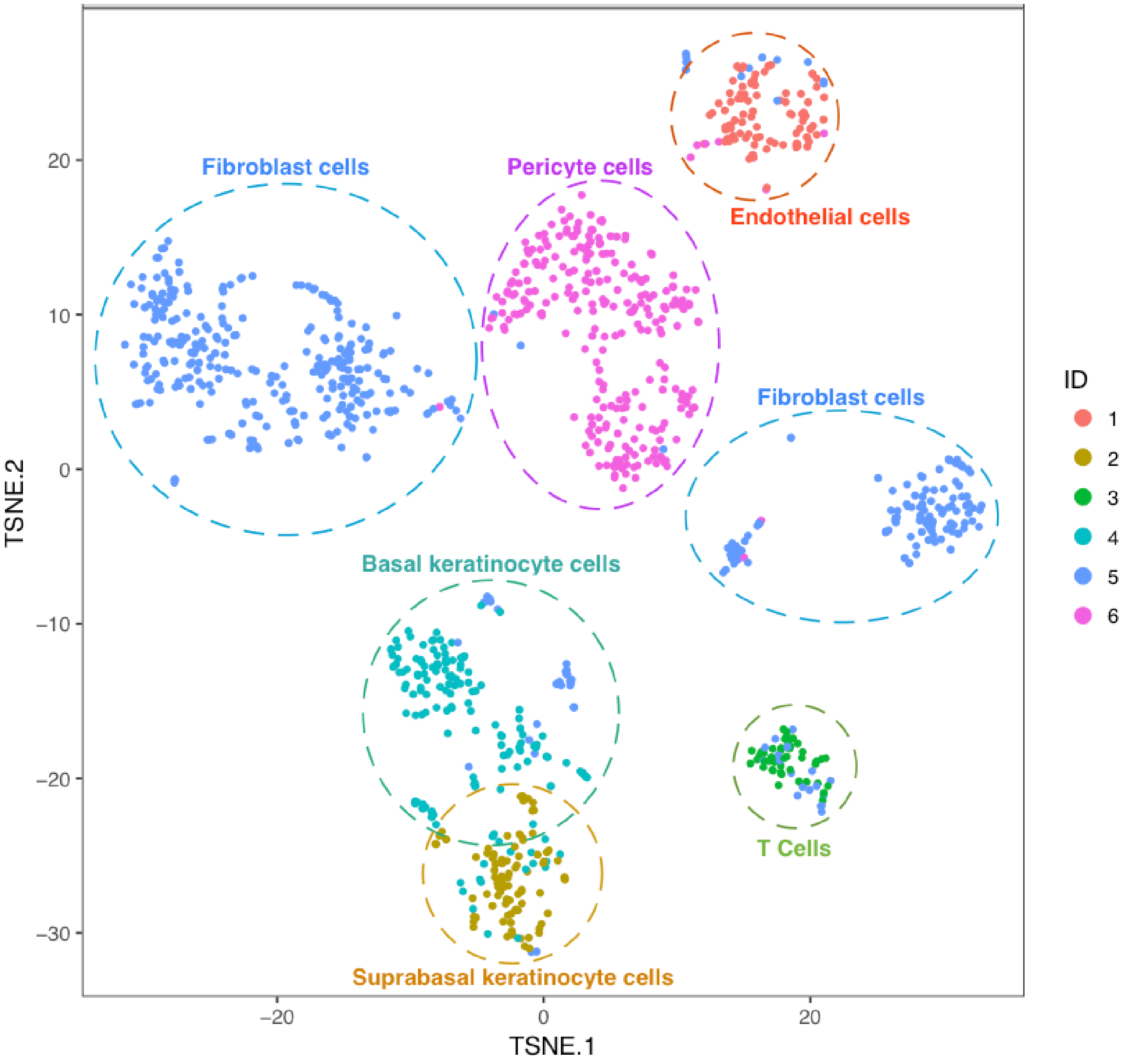


Figure 3. The t-SNE projection of cells from systemic sclerosis skin tissue, colored by the DIMM-SC clustering assignment

2.4.3 Model fitting diagnosis

An important step in applying model-based approach is to examine whether the proposed statistical model fits the real data well. In Dirichlet distribution, the marginal distribution of $\boldsymbol{p}$ is a Beta distribution. In addition, the mean of p_i, $\alpha_{ik}/|\boldsymbol{\alpha}_{(\boldsymbol{k})}|$, is approximately proportional to its variation $\alpha_{ik}(|\boldsymbol{\alpha}_{(\boldsymbol{k})}| - \alpha_{ik})/(|\boldsymbol{\alpha}_{(\boldsymbol{k})}|^2(|\boldsymbol{\alpha}_{(\boldsymbol{k})}| + 1))$. After applying DIMM-SC to the PBMC 68K scRNA-Seq dataset, I performed the following two analyses to evaluate the goodness of fit of the model. I first collected cells that belong to the same cell type using datasets of purified sub-populations of PBMCs from 10X Genomics, and then plotted the empirical marginal distribution of proportion p_i for top variable genes. I compared such empirical distribution with the marginal distribution $Beta(\alpha_{ik}, |\boldsymbol{\alpha}_{(k)}| - \alpha_{ik})$ at $\alpha_{(k)} = \hat{\alpha}_{(k)}$, where $\hat{\alpha}_{(k)}$ was estimated from the real scRNA-Seq data. Figure 22 (Appendix A) shows that the fitted distributions for top variable genes aligned very well with the empirical distributions, suggesting that DIMM-SC achieved good fit in real scRNA-Seq data.

Moreover, I explored the relationship between the mean and variance of p_i's, as commonly used in count data analysis, to evaluate whether any over-dispersion pattern exists. Similar to the previous analysis, I also collected cells from the same cell type, and calculated the mean and variation of p_i for each gene. The scatter plot of the log mean of p_i versus the log variance of p_i (Figure 23 (Appendix A)) shows a clear linear relationship between mean and variance. Derived from Dirichlet distribution, the expected intercept and slope can be approximated by 1 and $\log(|\boldsymbol{\alpha}|)$, respectively, where $\log(|\hat{\boldsymbol{\alpha}}|)$ was estimated from the real scRNA-Seq data. In CD56+ Natural Killer cells and CD19+ B cells, $\log(|\hat{\boldsymbol{\alpha}}|)$ equals to 6.60 and 6.67, respectively. As shown in Figure 22B, the intercept and slope of the fitted line (red line) are close to the expected values, indicating a good model fitting in this real scRNA-Seq data. I noticed that, due to both technical and biological uncertainties, a few genes exhibit extra variation, which cannot be fully explained by the mean-variation relationship posited by the Dirichlet distribution. I will pursue to extend DIMM-SC to account for such additional variation in the near future.

2.5 Discussion

Compared with the early generation scRNA-Seq technologies, the intrinsic characteristics of droplet-based scRNA-Seq data, including a much larger number of cells and direct counting of molecule copies using UMI, pose great challenges on statistical analysis and require new methodological development. In this study, I developed a model-based clustering method DIMM-SC for analyzing droplet-based scRNA-Seq data. DIMM-SC directly models UMI counts from scRNA-Seq data using a multinomial distribution with Dirichlet mixture priors. I demonstrated that DIMM-SC has achieved substantial improvements in clustering accuracy and stability compared to existing clustering methods such as K-means clustering, Seurat and CellTree. More importantly, my probabilistic model provides clustering uncertainty for each cell (how likely each cell belongs to each cluster), thus can benefit rigorous statistical inference and straightforward biological interpretations. In addition, DIMM-SC can be used to detect differentially expressed gene markers among different cell types, which is under our further investigation.

My probabilistic model coupled with a computationally efficient E-M algorithm is able to cluster large-scale droplet-based scRNA-Seq data. For example, it takes around 3 hours to cluster 68,000 cells using top 1,000 highly variable genes. In the analysis of scRNA-Seq data, both gene level filtering and cell level filtering are critical for clustering regardless of which clustering method to use. I recommend ranking genes by their variations among all cells and choosing top 500-1,000 highly variable genes. In addition, I also recommend running DIMM-SC 5~10 times, each with different random seeds, and choosing the one with the largest likelihood as the final results. For the number of clusters, I can pre-define it based on prior knowledge on the tissue or determine it using some model checking criteria such as AIC or BIC (Akaike, 1974; Schwarz, 1978). As shown in Figure 24 (Appendix A), AIC and BIC work well in the analysis of simulated datasets, the performance in real data needs further exploration. Alternatively, it can be determined using the procedure described in ADPclust (Wang and Xu, 2015) or the Dirichlet process (Teh, 2011). DIMM-SC is currently implemented in R/ Rcpp with satisfactory computing efficiency for most needs so far. Further improvement (e.g. parallel computing) can be made to accommodate larger-scaled data.

There are several noticeable limitations of my method. First, DIMM-SC only models variations among different cells from one single individual. To jointly model scRNA-Seq data from multiple individuals, a hierarchical structure can be posed in the current method to account for the individual level heterogeneity, but a more sophisticated numerical algorithm will be needed to reduce the computational cost. Second, DIMM-SC is an unsupervised method that infers structures from all data. Prior knowledge on cell-type-specific biomarkers may further improve the clustering accuracy. To use such prior information, a semi-supervised approach is needed to guide cluster inference. Furthermore, existing scRNA-Seq data from purified cells (e.g. via flow cytometry) can serve as external reference panels or training datasets to reduce experimental biases, remove outliers, and improve clustering reliability. Last but not least, my DIMM-SC model ignores the measurement errors and uncertainties buried in the UMI count matrix. Multiple factors such as dropout event, mapping percentage, sequencing depth, and PCR efficiency are not considered in the current model. These limitations can be largely overcome by extending my method. I will explore these directions in the near future.

I noticed that similar models have been proposed in the field of text-mining (Yamamoto and Sadamitsu, 2005) and microbiome (Holmes, et al., 2012), where word, article, and topic or taxa, individual, and meta-community are studied. However, in those applications, the clusters are not well defined and require a careful interpretation. On the contrary, scRNA-Seq data usually consist of a set of known cell types from prior knowledge and have a much larger signal-to-noise ratio for the clustering analysis. Although sharing the common types of data structure, these fields have different fundamental questions, so existing methods proposed from other fields need to be tailored or extended to incorporate intrinsic characteristics of scRNA-Seq data. For example, CellTree adapts the LDA approach from the text-mining field. Although LDA is more flexible and more widely used in text-mining field than the Dirichlet mixture model based methods, I have showed that DIMM-SC is more accurate, stable and efficient than CellTree in both simulation studies and real data applications in the context of scRNA-Seq clustering analysis.

In summary, I provide a novel statistical method and an efficient computational tool DIMM-SC for clustering droplet-based single cell transcriptomic data, which facilitates rigorous statistical inference of cell population heterogeneity. I am confident that DIMM-SC will be highly useful for the fast-growing community of large-scale single cell transcriptome analysis.

3.0 BAMM-SC: A Dirichlet Mixture Model for Clustering Droplet-based Single Cell Transcriptomic Data from Multiple Individuals

The recently developed droplet-based single cell transcriptome sequencing (scRNA-Seq) technology makes it feasible to perform a population-scale scRNA-Seq study, in which the transcriptome is measured for tens of thousands of single cells from multiple individuals. Despite the advances of many clustering methods, there are few tailored methods for population-scale scRNA-Seq studies. Here, I developed a **BA**yesian **M**ixture **M**odel for **S**ingle **C**ell sequencing (BAMM-SC) method to cluster scRNA-Seq data from multiple individuals simultaneously. BAMM-SC takes raw count data as input and accounts for data heterogeneity and batch effect among multiple individuals in a unified Bayesian hierarchical model framework. Results from extensive simulation studies and applications of BAMM-SC to in-house experimental scRNA-Seq datasets using blood, lung and skin cells from humans or mice demonstrated that BAMM-SC outperformed existing clustering methods with considerable improved clustering accuracy, particularly in the presence of heterogeneity among individuals.

3.1 Introduction

Single cell RNA sequencing (scRNA-Seq) technologies have been widely used to measure gene expression for each individual cell, facilitating a deeper understanding of cell heterogeneity and better characterization of rare cell types (Gawad, et al., 2016). Compared to early generation scRNA-Seq technologies, the recently developed droplet-based technology, largely represented by the 10X Genomics Chromium system, has quickly gained popularity because of its high-throughput (tens of thousands of single cells per run), efficiency (a couple of days), and relatively lower cost (< $1 per cell) (Macosko, et al., 2015; Zheng, et al., 2017). It is now feasible to conduct population-scale single cell transcriptomic profiling studies, where several to tens or even hundreds of individuals are sequenced.

A major task of analyzing droplet-based scRNA-Seq data is to identify clusters of single cells with similar transcriptomic profiles. To achieve this goal, classic unsupervised clustering methods such as K-means clustering, hierarchical clustering, and density-based clustering approaches can be applied after some normalization steps. Recently, scRNA-Seq tailored unsupervised methods, such as SIMLR (Wang, et al., 2018), CellTree (duVerle, et al., 2016), SC3 (Kiselev, et al., 2017), TSCAN (Ji and Ji, 2016) and DIMM-SC (Sun, et al., 2018), have been designed and proposed for clustering scRNA-Seq data. Supervised methods, such as MetaNeighbor, have been proposed to assess how well cell-type-specific transcriptional profiles replicate across different datasets (Crow, et al., 2018). However, none of these methods explicitly considers the heterogeneity among multiple individuals from population studies. In a typical analysis of population-scale scRNA-Seq data, reads from each individual are processed separately and then merged together for the downstream analysis. For example, in the 10X Genomics Cell Ranger pipeline, to aggregate multiple libraries, reads from different libraries are down-sampled such that all libraries have the same sequencing depth, leading to substantial information loss for individuals with higher sequencing depth. Alternatively, reads can be naively merged across all individuals without any library adjustment, leading to batch effects and unreliable clustering results.

Similar to the analysis of other omics data, several computational approaches have been proposed to correct batch effects for scRNA-Seq data. For example, Spitzer et al. adapted the concept of force-directed graph to visualize complex cellular samples via Scaffold (Single-Cell Analysis by Fixed Force- and Landmark-Directed) maps (Spitzer, et al., 2015), which can overlay data from multiple samples onto a reference sample(s). Recently, two new methods: Mutual Nearest Neighbors (Haghverdi, et al., 2018) (MNN) (implemented in scran) and Canonical Correlation Analysis (CCA) (Satija, et al., 2015) (implemented in Seurat) were published for batch correction of scRNA-Seq data. All these methods require the raw counts to be transformed to continuous values under different assumptions, which may alter the data structure in some cell types and lead to difficulty of biological interpretation.

To fill in this gap, I propose BAMM-SC, an extension of my previous DIMM-SC approach, to simultaneously cluster droplet-based scRNA-Seq data from multiple individuals. Specifically, BAMM-SC represents a Bayesian hierarchical Dirichlet multinomial mixture model, which explicitly characterizes three sources of heterogeneity (i.e., genes, cell types and individuals). Figure 4 provides an overview of the model structure in BAMM-SC, which directly models cell-type specific genes' UMI counts and their heterogeneity among different individuals through a hierarchical distribution structure in a Bayesian framework. This method has the following three key realistic assumptions. First, cell type clusters are discrete, and each cell belongs to one specific type exclusively. Second, heterogeneity exists among different individuals and across different cell types. The heterogeneity of the same cell type among different individuals is smaller than the heterogeneity across different cell types within the same individual. Third, cells of the same cell type share a similar gene expression pattern. That is, the underlying statistical distributions for cells within the same cell type are assumed to be the same. Compared to other clustering methods which ignore individual level variability, BAMM-SC has the following four key advantages: (1) BAMM-SC accounts for data heterogeneity among multiple individuals, such as unbalanced sequencing depths and technical biases in library preparation, and thus reduces the false positives of detecting individual-specific cell types. (2) BAMM-SC borrows information across different individuals, leading to improved power for detecting individual-shared cell types and higher reproducibility as well as stability of the clustering results. (3) BAMM-SC performs one-step clustering on raw UMI count matrix without any prior batch-correction step, which is required for most clustering methods in the presence of batch effect. (4) BAMM-SC provides a statistical framework to quantify the clustering uncertainty for each cell in the form of posterior probability for each cell type.

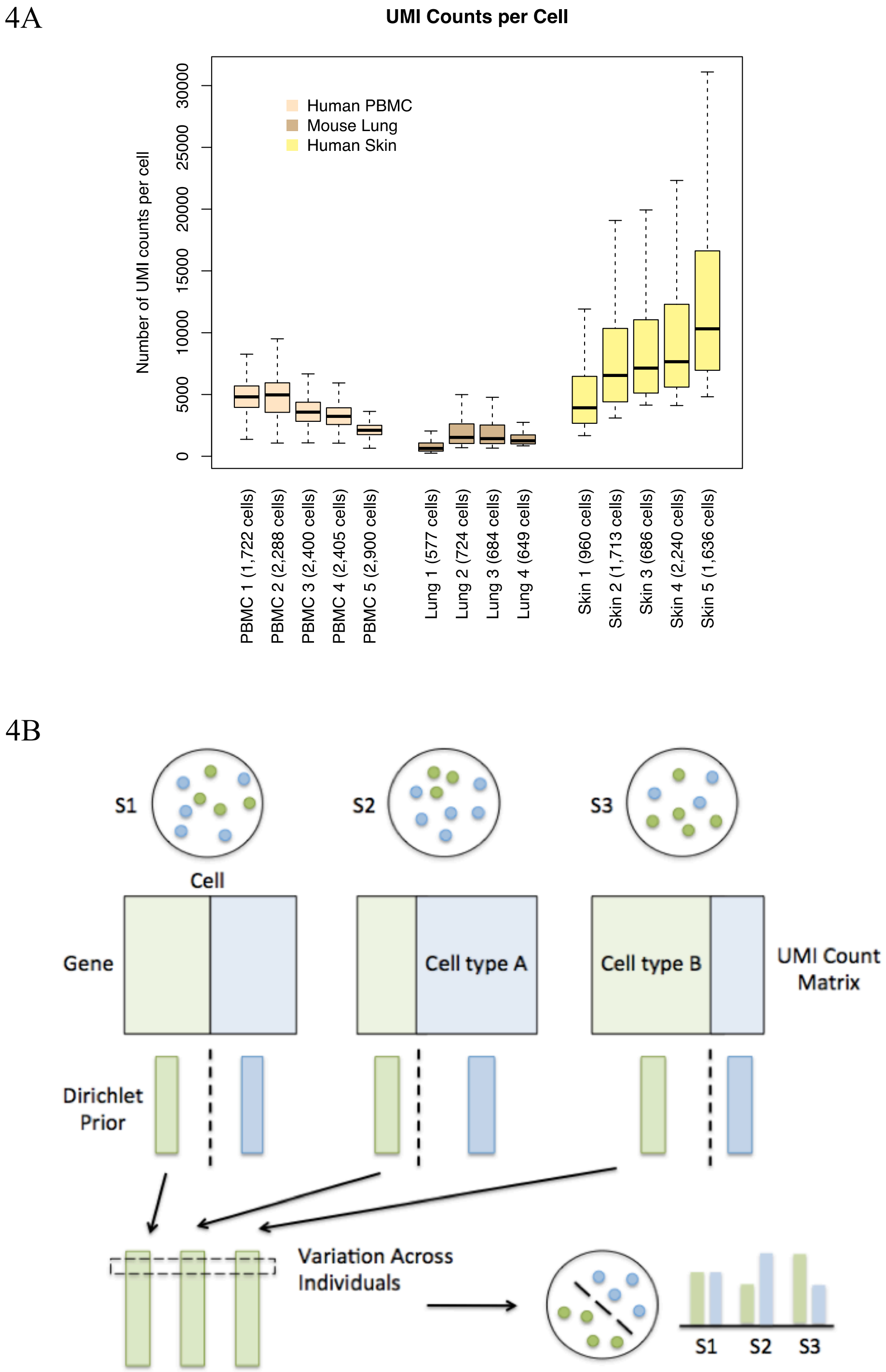


Figure 4. (A) UMI counts per cell of three droplet-based scRNA-Seq datasets. (B) An overall workflow of BAMM-SC

3.2 Motivating Example

I first conducted an exploratory data analysis using in-house droplet-based scRNA-Seq mouse data. For simplicity, I illustrated with three mice. Figure 25A (Appendix B) shows the t-SNE plot, suggesting that heterogeneity among single cells of different cell types (i.e., B cell vs. T cell) is greater than heterogeneity among three different mice (i.e., dots with three different colors). In addition, gene *Ftl1* shows higher expression in B cell than that in T cell (Figure 25B (Appendix B)). Noticeably, the variation of *Ftl1* gene expression between two cell types is larger than the variation of *Ftl1* gene expression across single cells within the same cell type among three mice. This motivating example demonstrates the importance of charactering different levels of variability among droplet-based scRNA-Seq data collected from multiple individuals.

To demonstrate the existence of batch effect in multiple individuals, I used both publicly available and three in-house synthetic droplet-based scRNA-Seq datasets including human peripheral blood mononuclear cells (PBMC), mouse lung and human skin tissues. Detailed sample information was summarized in Figure 4A and Table 4. Here, I use human PBMC as an example. Isolated from the whole blood obtained from 4 healthy donors, I used the 10X Chromium system to generate scRNA-Seq data. I also included one additional healthy donor from a published PBMC scRNA-Seq data (Zheng, et al., 2017) to mimic the scenario where we combine the local dataset with the public datasets. In this cohort, sample 1 and sample 2 were sequenced in one batch; sample 3 and sample 4 were sequenced in another batch; sample 5 was downloaded from the original study conducted by 10X Genomics (Zheng, et al., 2017). As an exploratory analysis, I produced a t-SNE plot based on the first 50 Principal Components (PCs) (Figure 5) of all cells from these 5 donors and observed a clear batch effect: samples from the same batch tend to cluster together.

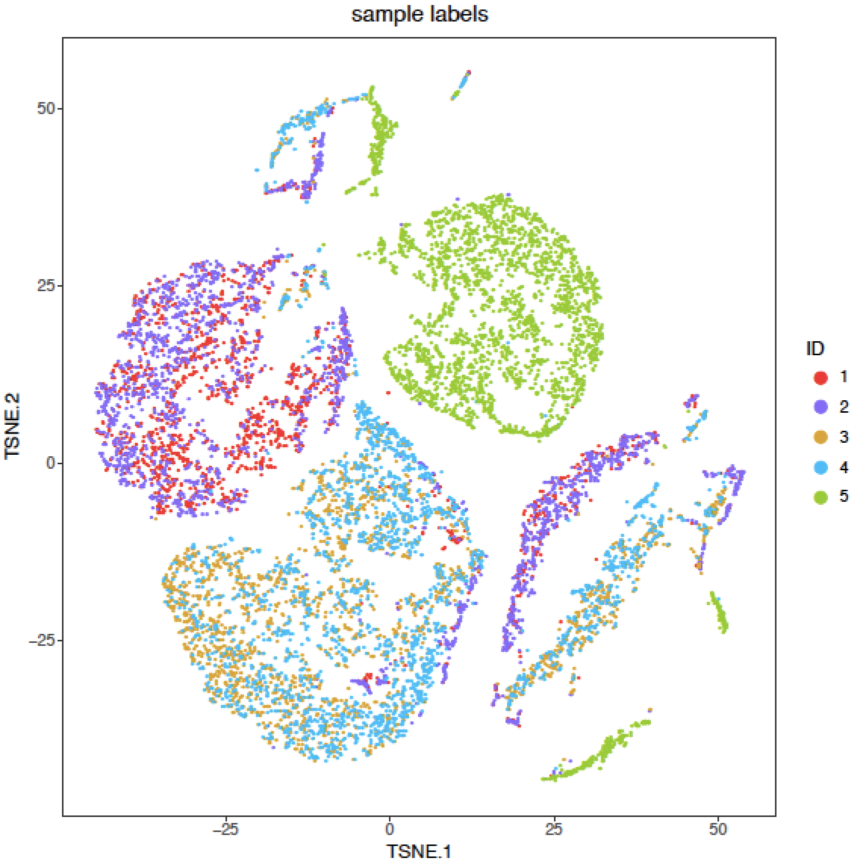


Figure 5. The t-SNE projection of PBMCs from 5 human samples

Table 4. Sample information of three droplet-based scRNA-Seq datasets

Dataset	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
			Version of Cell Ranger		
Human PBMC	2.0.0	2.0.0	1.2.0	1.2.0	1.1.0
Mouse Lung	1.2.0	1.2.0	1.2.0	1.2.0	
Human Skin	1.3.0	1.3.0	2.0.0	2.0.0	2.0.0
			Mean Reads per Cell		
Human PBMC	143,286	100,847	96,324	92,650	24,722
Mouse Lung	86,040	234,182	243,611	267,401	
Human Skin	215,718	168,951	157,677	51,669	107,424
			Number of Cells		
Human PBMC	1,722	2,288	2,400	2,405	2,900
Mouse Lung	577	724	684	649	
Human Skin	960	1,713	686	2,240	1,636

This illustrative example demonstrates the importance and urgent need for well characterizing different sources of variability and correcting potential batch effects among droplet-based scRNA-Seq datasets collected from multiple individuals. In addition, due to the computational burden, many methods cannot be scaled up to analyze population-scale droplet-based scRNA-Seq data with tens of thousands of cells collected from many individuals under various conditions. In this study, I propose a **BA**yesian **M**ixture **M**odel for **S**ingle **C**ell sequencing (BAMM-SC) to simultaneously cluster large-scale droplet-based scRNA-Seq data from multiple individuals. BAMM-SC directly works on the raw counts without any data transformation and models the heterogeneity from multiple sources by learning the distributions of signature genes in a Bayesian hierarchical model framework. In the following sections, I will describe this method, benchmark its performance against existing clustering methods in simulation studies, and evaluate our method for its accuracy, stability, and efficiency in three in-house synthetic scRNA-Seq datasets including PBMCs, skin, and lung tissues from humans or mice.

3.3 Statistical Model

I propose a Bayesian hierarchical Dirichlet multinomial mixture model to explicitly characterize different sources of variability in population scale scRNA-Seq data. Specifically, let x_{ijl} represent the number of unique UMIs for gene i in cell j from individual l ($1 \leq i \leq G$, $1 \leq j \leq C_l$, $1 \leq l \leq L$). Here G, C_l and L denote the total number of genes, cells (in individual l) and individuals, respectively. My goal is to perform simultaneous clustering for cells from all L individuals. I assume that within each individual, all single cells consist of K distinct cell types. Cell type clusters are discrete, and each cell belongs to one cell type exclusively. Here K is pre-defined according to prior biological knowledge, or will be estimated from the data, and K is the same among all L individuals.

3.3.1 Statistical model

Assume that $\boldsymbol{x}_{\cdot\boldsymbol{jl}} = (x_{1jl}, x_{2jl}, \dots, x_{Gjl})$, the gene expression for cell j in individual l, follows a multinomial distribution $Multi(T_{jl}, \boldsymbol{p}_{\cdot\boldsymbol{jl}})$. Here $T_{jl} = \sum_{i=1}^{G} x_{ijl}$ is the total number of UMIs, $\boldsymbol{p}_{\cdot\boldsymbol{jl}} = (p_{1jl}, p_{2jl}, \dots, p_{Gjl})$ is the probability vector for gene expression with $\sum_{i=1}^{G} p_{ijl} = 1$ (where larger p_{ijl} is associated with more UMI counts x_{ijl}). In addition, let $z_{jl} \in \{1,2, \dots, K\}$ represent the cell type label for cell j in individual l, where $z_{jl} = k$ indicates that cell j in individual l belongs to cell type k. Cells of the same cell type share a similar gene expression pattern. If cell j in individual l belongs to cell type k ($z_{jl} = k$), I assume that $\boldsymbol{p}_{\cdot\boldsymbol{jl}}$ follows a cell-type specific Dirichlet prior $DIR(\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}})$, where $\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}} = (\alpha_{1lk}, \alpha_{2lk}, \dots, \alpha_{Glk})$ is the Dirichlet prior parameter for cell type k in individual l.

$$P(\boldsymbol{p}_{\cdot\boldsymbol{jl}}|z_{jl} = k, \boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}) = \frac{1}{B(\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}})} p_{1jl}^{\alpha_{1lk}-1} p_{2jl}^{\alpha_{2lk}-1} \dots p_{Gjl}^{\alpha_{Glk}-1},$$

where $B(\alpha_{\cdot lk})$ is Beta function with parameter $\alpha_{\cdot lk} = (\alpha_{1lk}, \alpha_{2lk}, \dots, \alpha_{Glk})$. Then after integrating $p_{\cdot jl}$ out, we have:

$$P(\boldsymbol{x}_{\cdot\boldsymbol{jl}}|z_{jl} = k, \boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}) = \frac{T_{jl}!}{\prod_{i=1}^{G} x_{ijl}!} \left(\prod_{i=1}^{G} \frac{\Gamma(x_{ijl} + \alpha_{ilk})}{\Gamma(\alpha_{ilk})} \right) \frac{\Gamma(|\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}|)}{\Gamma(T_{jl} + |\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}|)}, \qquad |\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}| = \sum_{i=1}^{G} \alpha_{ilk}.$$

The joint distribution of $\boldsymbol{x}_{\cdot\boldsymbol{jl}}$ and z_{jl} is:

$$P(\boldsymbol{x}_{\cdot\boldsymbol{jl}}, z_{jl}|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}\cdot}) = \frac{T_{jl}!}{\prod_{i=1}^{G} x_{ijl}!} \sum_{k=1}^{K} I(z_{jl} = k) \left(\prod_{i=1}^{G} \frac{\Gamma(x_{ijl} + \alpha_{ilk})}{\Gamma(\alpha_{ilk})} \right) \frac{\Gamma(|\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}|)}{\Gamma(T_{jl} + |\boldsymbol{\alpha}_{\cdot\cdot\boldsymbol{k}}|)}.$$

I further assume that all C_l cells in individual l are independent, then the joint distribution for all cells in individual l is:

$$P(\boldsymbol{x}_{\cdot\cdot\boldsymbol{l}}, \boldsymbol{z}_{\cdot\boldsymbol{l}}|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}\cdot}) = \prod_{j=1}^{C_l} P(\boldsymbol{x}_{\cdot\boldsymbol{jl}}, z_{jl}|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}\cdot}).$$

Finally, I assume that all L individuals are independent, then the overall joint distribution for all cells across all individuals becomes:

$$P(\boldsymbol{x}_{\dots}, \boldsymbol{z}_{\cdot\cdot}|\boldsymbol{\alpha}_{\dots}) = \prod_{l=1}^{L} P(\boldsymbol{x}_{\cdot\cdot\boldsymbol{l}}, \boldsymbol{z}_{\cdot\boldsymbol{l}}|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}\cdot}) \propto \prod_{l=1}^{L} \prod_{j=1}^{C_l} \left\{ \sum_{k=1}^{K} I(z_{jl} = k) \left(\prod_{i=1}^{G} \frac{\Gamma(x_{ijl} + \alpha_{ilk})}{\Gamma(\alpha_{ilk})} \right) \frac{\Gamma(|\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}|)}{\Gamma(T_{jl} + |\boldsymbol{\alpha}_{\cdot\boldsymbol{lk}}|)} \right\}.$$

In this model, the two sets of parameters of interest are $\boldsymbol{z}_{\cdot\cdot} = \{z_{jl}\}_{1\le j\le C_l, 1\le l\le L}$, the cell type label for cell j in individual l, and $\boldsymbol{\alpha}_{\dots} = \{\alpha_{ilk}\}_{1\le i\le G, 1\le l\le L, 1\le k\le K}$, the Dirichlet parameters for gene i in cell type k in individual l. I adopt a full Bayesian approach and use Gibbs sampler to estimate the posterior distributions. Specifically, the joint posterior distribution for $\boldsymbol{z}_{\cdot\cdot}$ and $\boldsymbol{\alpha}_{\dots}$ are:

$$P(\boldsymbol{z}_{\cdot\cdot}, \boldsymbol{\alpha}_{\dots}|\boldsymbol{x}_{\dots}) \propto P(\boldsymbol{x}_{\dots}, \boldsymbol{z}_{\cdot\cdot}|\boldsymbol{\alpha}_{\dots}) \times Prior(\boldsymbol{\alpha}_{\dots}).$$

Since all α's are strictly positive, I propose a log-normal distribution as the prior distribution for α_{ilk}. I assume that for gene i in cell type k, α_{ilk} from all L individuals share the same prior distribution $LN(\mu_{ik}, \sigma_{ik}^2)$, that is,

$$Prior(\boldsymbol{\alpha_{i\cdot k}}) = \prod_{l=1}^{L} \frac{1}{\alpha_{ilk}\sqrt{2\pi\sigma_{ik}^2}} \exp\left\{-\frac{(\log \alpha_{ilk} - \mu_{ik})^2}{2\sigma_{ik}^2}\right\}.$$

Here μ_{ik} can be estimated by the mean of α_{ilk}'s: $\hat{\mu}_{ik} = \frac{1}{L}\sum_{l=1}^{L} \log(\alpha_{ilk})$. Estimation of σ_{ik}^2 can be challenging due to limited number of individuals. I can assume all σ_{ik}^2's follow a hyper-prior: Gamma distribution $Gamma(a_k, b_k)$, and use information across all genes to estimate variance. In addition, I assume a non-informative prior for μ_{ik}'s. Taken all together, I have the full posterior distribution as follows:

$$P(\boldsymbol{z}_{\cdot\cdot}, \boldsymbol{\alpha}_{\cdots}|\boldsymbol{x}_{\cdots}) \propto P(\boldsymbol{x}_{\cdots}, \boldsymbol{z}_{\cdot\cdot}|\boldsymbol{\alpha}_{\cdots}) \times \prod_{k=1}^{K}\prod_{i=1}^{G} Prior(\boldsymbol{\alpha_{i\cdot k}}) \times \prod_{k=1}^{K} Prior(\boldsymbol{\mu_{\cdot k}}) \times \prod_{k=1}^{K} Prior\left(\boldsymbol{\sigma^2_{\cdot k}}\right).$$

I use Gibbs sample to iteratively update α_{ilk} and z_{jl}. Details can be found in Appendix B.

3.4 Simulation Studies

I have conducted comprehensive simulation studies to benchmark the performance of BAMM-SC. Specifically, I simulated droplet-based scRNA-Seq data collected from multiple individuals from the posited Bayesian hierarchal Dirichlet multinomial mixture model. I considered different experimental designs, including different heterogeneities among multiple individuals and different number of individuals. In the posited hierarchical model, the log normal prior distribution $LN(\mu_{ik}, \sigma_{ik}^2)$ measures the heterogeneity of gene i in cell type k among multiple individuals, where μ_{ik} and σ_{ik}^2 are related to the mean and variation of gene expression. Without loss of generality, I used the mean of σ_{ik}^2 across all genes and all cell types to quantify the overall individual level heterogeneity. I applied BAMM-SC to each synthetic dataset, and compared the inferred cell type label of each single cell with the ground truth, measured by adjusted Rand index (ARI) (Rand, 1971). I compared BAMM-SC with other competing clustering methods (K-means, TSCAN, SC3 and Seurat), which are either methods from different clustering categories or recommended by recent reviews on clustering methods for single-cell data (Duo, et al., 2018; Freytag, et al., 2018). Since none of methods model batch effects and therefore each needs to be combined with a batch correction method as a pre-processing step in data analysis. I applied two recently published and prevalent methods srcan Mutual Nearest Neighbors (MNN) (Haghverdi, et al., 2018) and Seurat Canonical Correlation Analysis (CCA) (Satija, et al., 2015) prior to these clustering methods so that each combination can be a fair comparison with BAMM-SC, which does not need a separate batch correction step.

Specifically, I compared BAMM-SC with the other nine competing methods (MNN+K-means, MNN+TSCAN, MNN+SC3, MNN+Seurat, CCA+K-means, CCA+TSCAN, CCA+SC3, CCA+Seurat and DIMM-SC) in the simulation studies. Noticeably, DIMM-SC, my previously developed method for clustering scRNA-Seq data from a single individual, also takes the raw UMI count matrix as the input without any batch effect correction or data transformation. I pooled single cells from different individuals together while ignoring each individual label, and then applied DIMM-SC to the pooled data. I simulated 100 datasets and summarized the corresponding ARIs for each method.

3.5 Results

3.5.1 Simulation studies

As shown in Figure 6A, BAMM-SC consistently outperformed the other nine competing methods across a variety of individual level heterogeneities by achieving higher average ARI and lower variation of ARI among 100 simulations. As expected, the performance of all ten clustering approaches decreases as the among individual heterogeneity increases, measured by the mean σ_{ik}^2 values. In Figure 6B, with the increase of number of individuals, BAMM-SC achieved higher ARI, while ARIs of other methods either remained stable or decreased.

Furthermore, I performed comprehensive simulation studies by generating simulated scRNA-Seq datasets from different number of cell type clusters (Figure 26A (Appendix B)), different overall sequencing depths (Figure 26B), and different cell-type-specific heterogeneities (i.e., the mean difference of gene expression profiles between two distinct cell types) (Figure 26C). BAMM-SC consistently outperformed other methods in terms of accuracy and robustness in all these scenarios. Taken together, my comprehensive simulation studies have demonstrated that, when data are generated from the true model, BAMM-SC is able to appropriately borrow information across multiple individuals, account for unbalanced sequencing depths, and provide more accurate and robust clustering results than other competing methods.

To evaluate the robustness of BAMM-SC when data generation model is mis-specified, I simulated additional datasets using R package Splatter (Zappia, et al., 2017), a commonly used tool for scRNA-Seq data simulation using a completely different model. To make my simulated data a good approximation to the real data, I first downloaded the raw UMI count matrix of a purified B cell scRNA-Seq dataset from the 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/b_cells), and used the function "splatEstimate" to estimate the parameters related to mean of gene, library size, expression outlier, dispersion across genes and dropout rate. I assumed cell types are shared across multiple individuals, where each individual is treated as one batch with the same number of cells and genes. I further specified batch parameters and differential expression parameters to generate scenarios with different amount of group effect (i.e., cell type differences) and batch effect. As shown in Figure 7, BAMM-SC still outperformed most other competing methods in terms of clustering accuracy in all scenarios, although the improvement is less substantial than my own model simulations, which is expected.

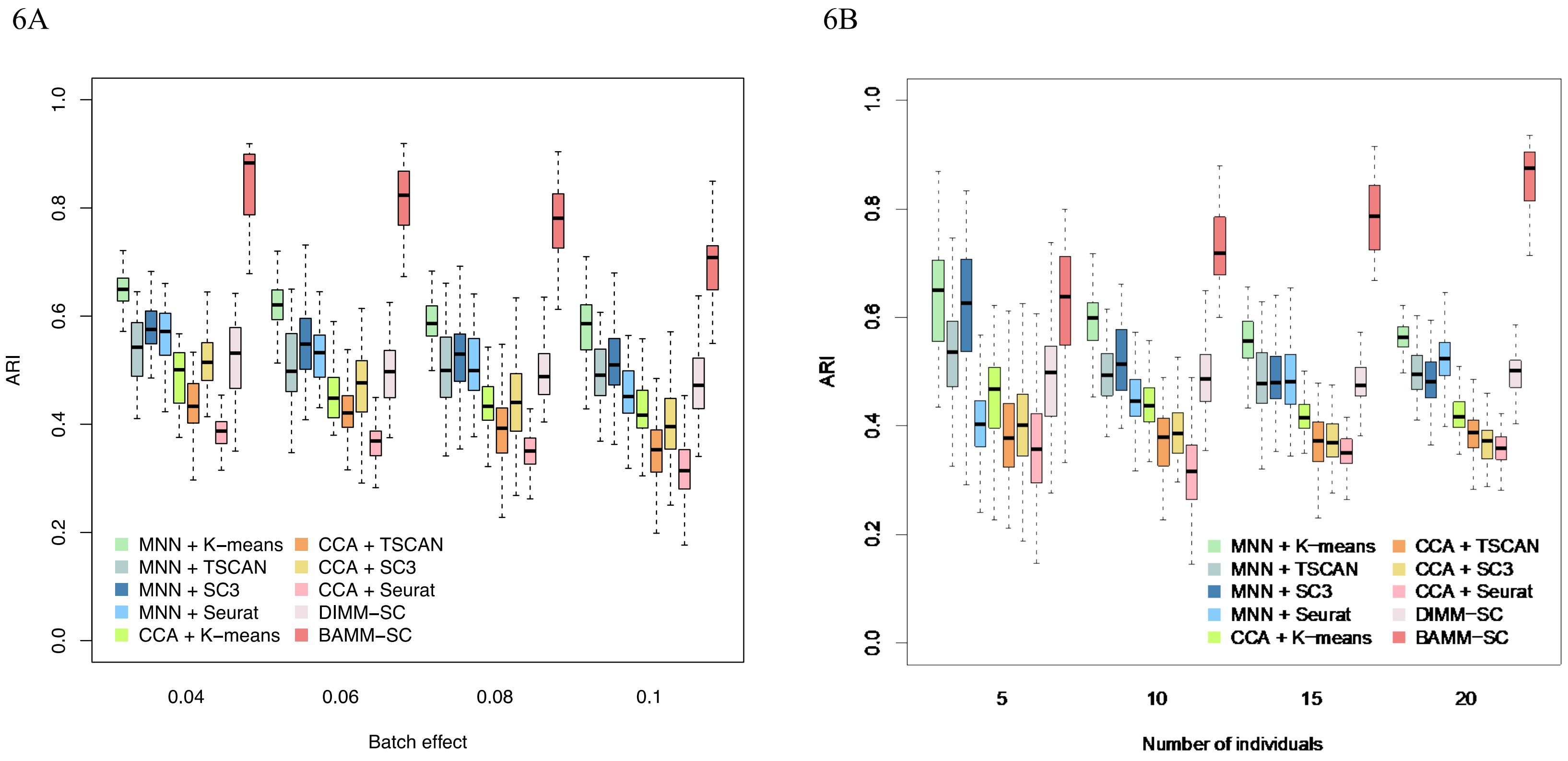


Figure 6. Boxplots of ARIs for ten clustering methods in simulation studies

Boxplots of ARIs for ten clustering methods across 100 simulations, investigating how (a) different heterogeneity among multiple individuals (measured by mean σ_{ik}^2 values) and (b) number of individuals affect clustering results. In (a), the simulated dataset consists of 10 individuals with 400 cells for each. In (b), we set the level of heterogeneity (mean of σ_{ik}^2) among individuals as 0.1.

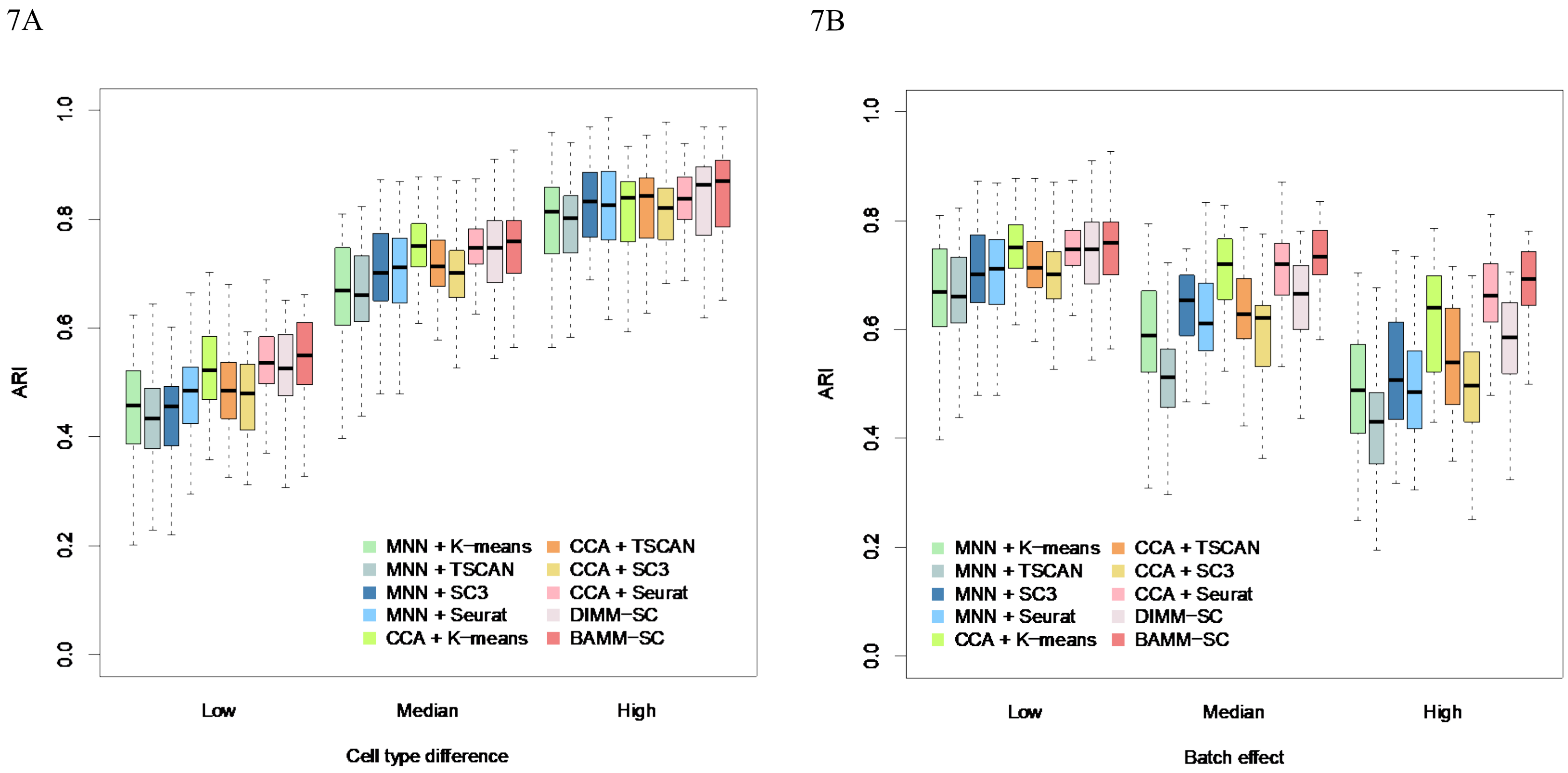


Figure 7. Boxplots of ARI for ten clustering methods across 100 simulations using Splatter

Boxplots of ARI for ten clustering methods across 100 simulations using Splatter, investigating how different levels of (a) group effect and (b) batch effect affect clustering results. In (a), we set the mean parameters of three cell types as (0.20, 0.21, 0.22), (0.20, 0.22, 0.24) and (0.20, 0.24, 0.28) to represent three levels (low, medium, high) of group difference. In (b), we set the mean parameters of the five individuals as (0.1, 0.1, 0.1, 0.1, 0.1), (0.12, 0.12, 0.12, 0.12, 0.12) and (0.14, 0.14, 0.14, 0.14, 0.14) to represent three levels (low, medium, high) of batch effects.

3.5.2 Real data analysis

I evaluated the performance of BAMM-SC together with other methods in three in-house synthetic scRNA-Seq datasets including human PBMC, mouse lung and human skin tissues generated using 10X Chromium system at the University of Pittsburgh.

3.5.2.1 Analysis of the in-house scRNA-Seq data from human PBMC samples

For aforementioned human PBMC samples, I first pooled cells from 5 donors together, filtered lowly expressed genes that were expressed in less than 1% cells. I then extracted the top 1,000 highly variable genes based on their standard deviations. As shown in Figure 27 (Appendix B), I identified 7 types of PBMCs based on the biological knowledge of cell-type-specific gene markers (Table 5). Using these gene markers, >70% single cells can be assigned to a specific cell type. Since there is no gold standard for clustering analysis in this real dataset, I used the labels of these cells as the approximated ground truth to benchmark the clustering performance for different clustering methods. Cells with uncertain cell types were removed when calculating ARIs.

Similar to the simulation studies, I applied ten clustering methods on these samples and repeated each method 10 times to evaluate the stability of its performance (Table 6). The total number of clusters was set as 7 based on the biological knowledge from cell-type-specific gene markers. Both TSCAN and Seurat are deterministic clustering methods and therefore they generate identical results for 10 analyses. As shown in Table 6, BAMM-SC achieved the highest ARI for human PBMC samples compared to all other competing methods.

Table 5. Gene markers used to specify cell types in human PBMC samples

Cell Types	Genes
CD8+ T cells	CD3+CD8A+CD4-
CD4+ T cells	CD3+CD8-CD4+IL2RA-IL7R+
B cells	CD3-CD19+MS4A1+
NK cells	NCAM1+NKG7+CD3-
CD14+ monocytes	CD3-CD19-CD14+HLA-
CD16+ monocytes	CD3-CD19-FCGR3A+
Dendritic cells	CD1C+CD14-HLA-FCER1A+

Table 6. Performance of clustering across 10 times analyses for human PBMC, mouse lung and human skin samples

	Human PBMC			Mouse Lung			Human Skin		
Method	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
MNN + K-means	0.379	0.083	(0.283, 0.485)	0.662	0.066	(0.596, 0.815)	0.597	0.075	(0.461, 0.676)
MNN + TSCAN	0.373	NA	NA	0.720	NA	NA	0.553	NA	NA
MNN + SC3	0.348	0.084	(0.266, 0.511)	0.640	0.061	(0.556, 0.687)	0.517	0.034	(0.436, 0.557)
MNN + Seurat	0.325	NA	NA	0.749	NA	NA	0.647	NA	NA
CCA + K-means	0.414	0.056	(0.307, 0.464)	0.695	0.114	(0.505, 0.883)	0.619	0.129	(0.424, 0.737)
CCA + TSCAN	0.210	NA	NA	0.611	NA	NA	0.398	NA	NA
CCA + SC3	0.145	0.052	(0.051, 0.215)	0.610	0.068	(0.531, 0.708)	0.369	0.071	(0.277, 0.488)
CCA + Seurat	0.468	NA	NA	0.729	NA	NA	0.702	NA	NA
DIMM-SC	0.333	0.071	(0.302, 0.529)	0.809	0.030	(0.742, 0.868)	0.715	0.045	(0.671, 0.779)
BAMM-SC	0.487	0.056	(0.362, 0.532)	0.882	0.042	(0.764, 0.910)	0.762	0.032	(0.717, 0.843)

I further generated t-SNE plots with each cell colored by their cell-type classification based on specific gene markers (i.e., the approximated truth) (Figure 8A (left figure)) and cluster labels inferred by BAMM-SC (Figure 8A (middle figure)), respectively. Despite some dendritic cells were wrongly identified as CD16+ Monocytes, these two plots are similar to each other (ARI=0.532), suggesting that BAMM-SC performed well in human PBMC samples compared with other clustering methods.

Moreover, I calculated the averaged cell proportions of each cell type inferred from BAMM-SC among 10 runs for 5 PBMC samples, compared with cell proportions calculated from the approximated truth based on gene markers. Figure 8A (right figure) shows that the proportions inferred from BAMM-SC are close to the truth, suggesting that BAMM-SC can adequately account for batch effect when clustering cells from multiple individuals. I also generated t-SNE projection plots colored by cluster labels inferred by other methods: MNN+K-means clustering (Figure 28A (Appendix B)), MNN+TSCAN (Figure 28B), MNN+SC3 (Figure 28C**)**, MNN+Seurat (Figure 28D), CCA+K-means (Figure 28E), CCA+TSCAN (Figure 28F), CCA+SC3 (Figure 28G), CCA+Seurat (Figure 28H) and DIMM-SC (Figure 28I).

3.5.2.2 Analysis of the in-house scRNA-Seq data from mouse lung samples

In addition to human PBMC samples, I also collected lung mononuclear cells from 4 mouse samples under 2 conditions: streptococcus pneumonia (SP) infected (sample 1 and 2) and naive (sample 3 and 4). Figure 9 shows the t-SNE plot of lung mononuclear cells from 4 mouse samples. Similar to the analysis of PBMC samples, after filtering lowly expressed genes, I pooled cells from 4 mice together and extracted the top 1,000 highly variable genes. As shown in Figure 29 (Appendix B), I identified 6 types of cells based on the biological knowledge of cell-type specific gene markers (Table 7). Taken together, > 66% of single cells can be assigned to a specific cell type. Therefore, I used the labels of these cells as the approximated truth and removed cells with uncertain cell types from the downstream analysis.

Figure 8B (left figure) and Figure 8B (middle figure) show the t-SNE plots with each cell colored by their cluster label based on cell-type-specific gene markers and cluster labels inferred by BAMM-SC, respectively. These two are highly similar (ARI=0.910), indicating the outstanding performance of BAMM-SC. Table 6 shows that BAMM-SC considerably outperformed other nine clustering methods in terms of ARI. I also generated t-SNE plots colored by cluster labels inferred by other competing clustering methods (Figure 30 (Appendix B)). As shown in Figure 31 (Appendix B), the proportions of neutrophils in SP infected samples (sample 1 and sample 2) are much higher than the proportions in naïve samples (sample 3 and sample 4). This is consistent with the fact that infections by bacteria and viruses may increase the number of neutrophils, which is a necessary reaction by the body (Chen and Kolls, 2013; Weiser, 2010). Interestingly, the proportion of cell types in naïve sample 3 is different from others, which may due to unsatisfactory sample quality or unexpected bacterial infections.

Table 7. Gene markers used to specify cell types in mouse lung cell samples

Cell Types	Genes
Macrophages	Ctss+Chil3+
Neutrophils	S100a8+S100a9+Il1b+
Endothelial	Lyve1+Egfl7+
Small airway Epithelial	Sftpc+Sftpd+Lyz1+
Club Cells	Scgb1a1+Scgb3a1+
Lymphocytes	Cd79b+Ms4a1+ / Gzma+Nkg7+

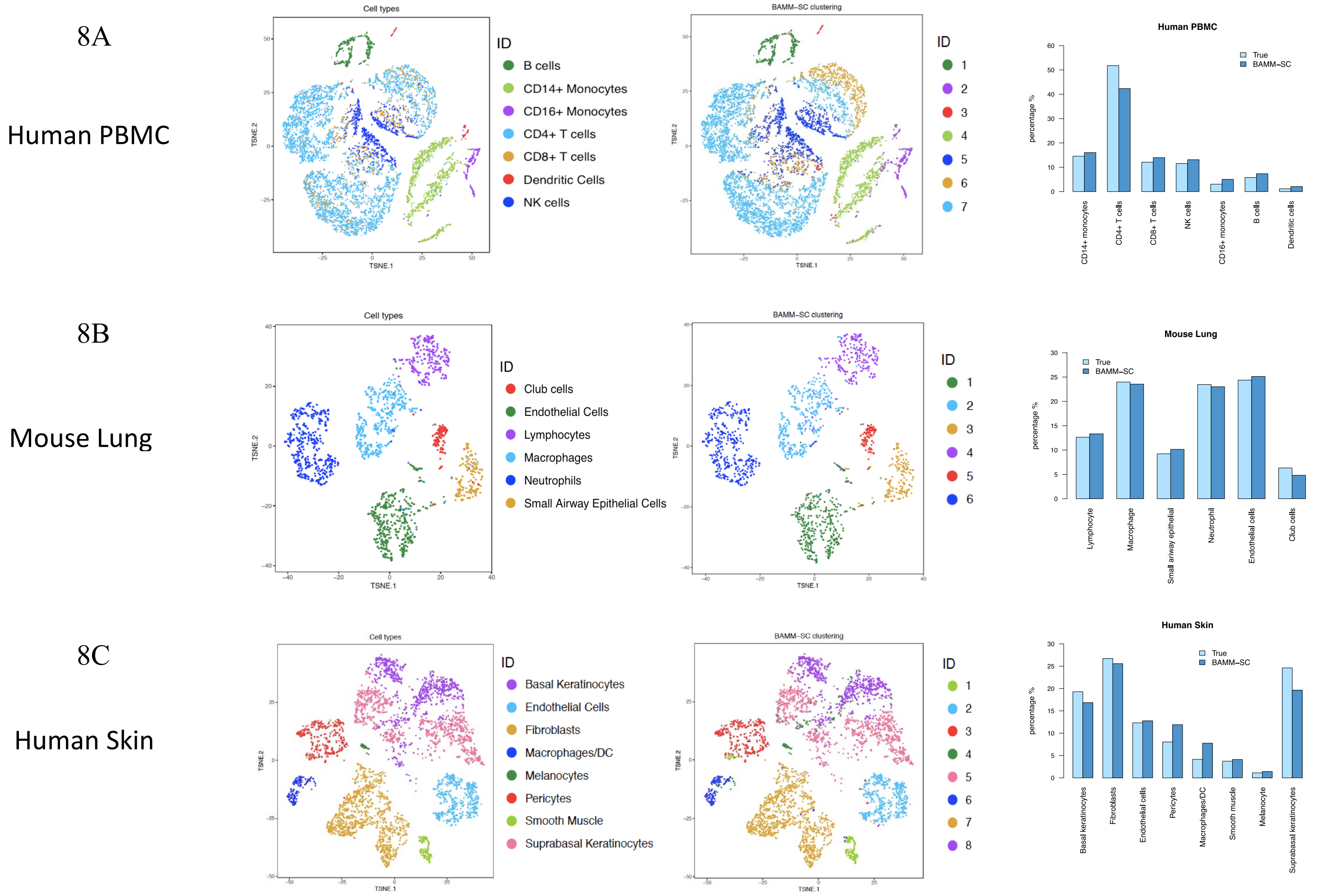


Figure 8. The t-SNE projection of cells and bar plots of proportions of cell types among all individuals for three real datasets

The t-SNE projection of cells (colored by the approximated truth and BAMM-SC clustering results) and bar plots of proportions of cell types among all individuals for human PBMC (A), mouse lung (B) and human skin (C) tissues, separately. BAMM-SC clustering labels are from the result with the highest ARI among 10 times analysis.

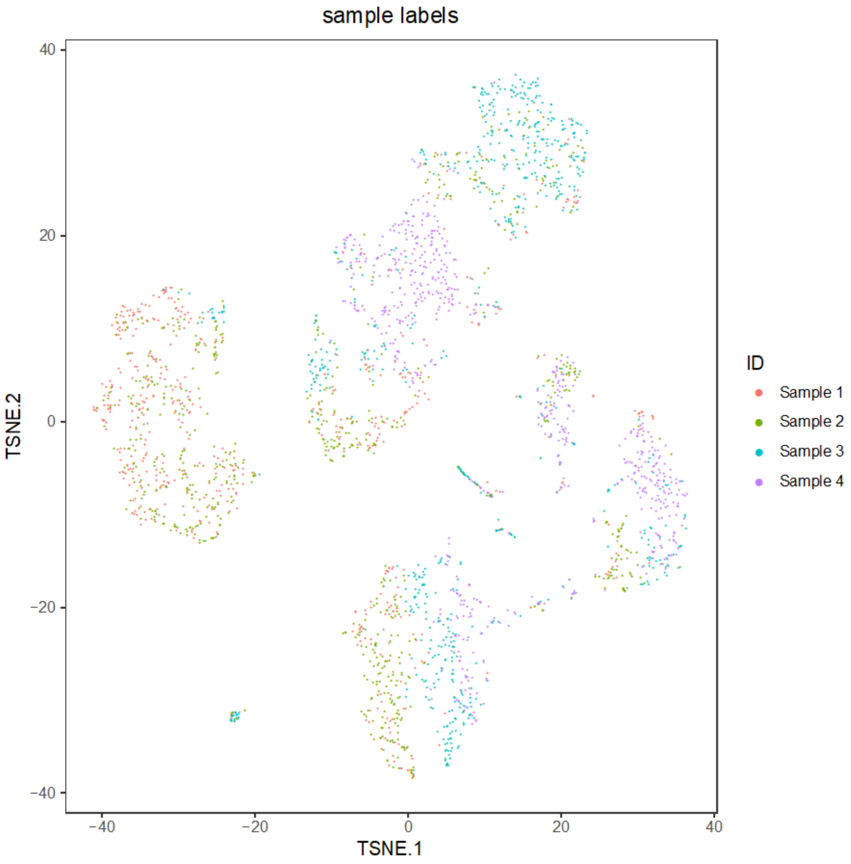


Figure 9. The t-SNE projection of lung mononuclear cells from 4 mouse samples

3.5.2.3 Analysis of the in-house scRNA-Seq data from human skin samples

To evaluate the clustering performance of BAMM-SC in solid human tissues, I collected skin samples from 5 healthy donors that are part of a systemic sclerosis study (Tabib, et al., 2018). Figure 4A and Table 4 list the detailed sample information and Figure 10 shows the t-SNE plot of cells from 5 human skin samples after the data processing similar as previous analyses. As shown in Figure 32 (Appendix B), I identified 8 major types of cells based on the biological knowledge of cell-type-specific gene markers (Table 8). Taken together, >67% of single cells can be assigned to a specific cell type. Like the other two real data analyses, I used the labels of these cells as the approximated truth and removed cells with uncertain cell types from the downstream analysis.

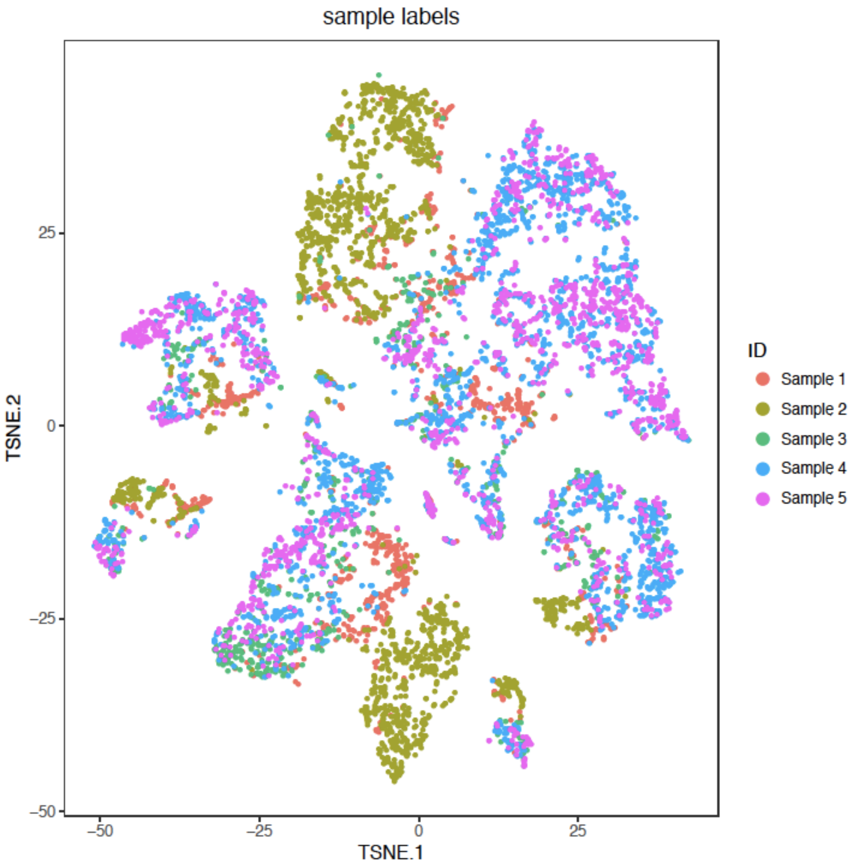


Figure 10. The t-SNE projection of cells from human skin dataset

As shown in Figure 8C, BAMM-SC performed well in human skin samples, since the t-SNE plot with each cell colored by their cell type label based on gene markers is highly similar to the plot generated from the clustering result of BAMM-SC (ARI=0.843). Also, BAMM-SC achieved higher ARI compared with all the other clustering methods (Table 6). As comparisons, I generated t-SNE plots colored by cluster labels inferred by different clustering (Figure 33 (Appendix B)).

Table 8. Gene markers used to specify cell types in human skin samples

Cell Types	Genes
Smooth muscle cells	DES+
Suprabasal keratinocytes	KRT1+KRT10+
Basal keratinocytes	KRT14+KRT5+
Endothelial cells	VWF+
Fibroblasts	COL1A1+
Pericytes	RGS5+VWF-
Melanocytes	PMEL+
Mecrophages/Dendritic cells	AIF1+

To further demonstrate the validity of BAMM-SC, I calculated the confusion matrix for three real datasets and reported the clustering accuracy (defined as the proportion of cells being classified into the correct cell-type cluster) (Table 9). In Table 9, the clustering results are selected based on the highest ARI among 10 times analysis. BAMM-SC outperformed other competing methods in all three datasets. Additionally, I performed a flow cytometry experiment, a gold standard method for quantifying cell population through cell surface markers, on the sample 3 from the human PBMC dataset, which has an additional aliquot from the same pool of cells. I used FlowJo software to gate each cell population through specific antibodies and calculated the percentage of each cell type. Then, I compared the proportions of different cell types from flow cytometry and the clustering result of BAMM-SC from scRNA-seq. Figure 34 (Appendix B) showed that the proportion of cells in each cell type classified by BAMM-SC is consistent with that being estimated by flow cytometry. I also calculated the Pearson's correlation coefficient of cell proportions for each clustering method (Table 10). Similarly, the clustering results are selected based on the highest ARI among 10 times analysis. Despite the different technology, the high correlation (Pearson correlation coefficient is 0.98) suggests that BAMM-SC is able to adequately account for heterogeneity among multiple individuals and provide reliable clustering results. To be noted, unlike other clustering methods we considered, Seurat cannot directly pre-specify the number of clusters K. Rather it needs to set a resolution parameter that indirectly controls the cluster number. In all three real data sets, after an extensive grid search, I found the resolution parameter that yields the same number of clusters as the one based on the biological knowledge. Therefore, for the two Seurat clustering methods, instead of using the clustering assignments that produced the highest ARI among 10 times analysis, I computed the confusion matrix and the proportions of different cell types based on this specific resolution parameter.

Table 9. Accuracy of the confusion matrix generated from different clustering methods for human PBMC, mouse lung and human skin samples

	Human PBMC	Mouse lung	Human skin
MNN + K-means	0.669	0.908	0.801
MNN + TSCAN	0.560	0.849	0.634
MNN + SC3	0.687	0.861	0.676
MNN + Seurat	0.511	0.598	0.775
CCA + K-means	0.673	0.947	0.797
CCA + TSCAN	0.452	0.827	0.638
CCA + SC3	0.528	0.811	0.591
CCA + Seurat	0.754	0.825	0.748
DIMM-SC	0.643	0.944	0.793
BAMM-SC	0.734	0.960	0.859

It is challenging to evaluate clustering algorithms in experimental data since the ground truth of cell type label is generally unknown. Other than using ARI based on cell-type-specific gene markers as approximated ground truth, I also used cluster stability and tightness to evaluate the clustering performance. Specifically, I calculated the average proportion of non-overlap (APN) (Datta, 2003) clustered cells and silhouette width (Rousseeuw, 1987) in three real datasets, respectively. APN is a cluster stability measurement which evaluates the stability of a clustering result by comparing it with the clusters obtained by removing one feature (i.e., one gene in our study) at a time. It measures the average proportion of observations not placed in the same cluster under both cases. To make computation affordable in our real data analysis, after extracting the top 1,000 highly variable genes, I compared the clustering results based on the full data (1,000 genes) to the clustering results based on a subset of data with 100 genes randomly removed. I repeated this step 10 times to calculate the APN. For cluster tightness, the silhouette width ranges from −1 to 1, where a higher value indicates that the observation is better matched to its own cluster and worse matched to other clusters. Here, the distance metric is Morisita dissimilarity. For both measurements, BAMM-SC achieved high cluster stability and high cluster tightness in most scenarios, compared with all other competing methods (Table 11, Table 12).

Table 10. The correlation of estimated proportions of cells in each cell type between different clustering methods and flow cytometry in human PBMC sample 3

	Human PBMC
MNN + K-means	0.95
MNN + TSCAN	0.94
MNN + SC3	0.92
MNN + Seurat	0.76
CCA + K-means	0.97
CCA + TSCAN	0.60
CCA + SC3	0.88
CCA + Seurat	0.99
DIMM-SC	0.97
BAMM-SC	0.98

Table 11. Performance of cluster stability measured by APN for human PBMC, mouse lung and human skin samples, respectively

Method	Human PBMC	Mouse Lung	Human Skin
MNN + K-means	0.24	0.21	0.25
MNN + TSCAN	0.16	0.11	0.29
MNN + SC3	0.43	0.44	0.56
MNN + Seurat	0.14	0.20	0.24
CCA + K-means	0.29	0.16	0.28
CCA + TSCAN	0.60	0.37	0.67
CCA + SC3	0.69	0.23	0.64
CCA + Seurat	0.11	0.16	0.19
DIMM-SC	0.23	0.14	0.17

Table 11 Continued

BAMM-SC	0.23	0.07	0.16

Table 12. Performance of cluster tightness measured by silhouette width for human PBMC, mouse lung and human skin samples, respectively

	Human PBMC	Mouse Lung	Human Skin
MNN + K-means	0.40	0.33	0.16
MNN + TSCAN	0.18	0.34	0.16
MNN + SC3	0.14	0.32	0.11
MNN + Seurat	0.34	0.33	0.20
CCA + K-means	0.13	0.34	0.11
CCA + TSCAN	-0.03	0.23	0.03
CCA + SC3	-0.12	0.33	-0.02
CCA + Seurat	0.03	0.29	0.11
DIMM-SC	0.21	0.34	0.12
BAMM-SC	0.35	0.35	0.17

Different from other deterministic methods, BAMM-SC has the ability to assess clustering uncertainty through the posterior probability for each cell to belong to each cell-type cluster. As shown in Figure 11, I highlighted vague cells in the t-SNE projection plot, where vague cells are defined as cells with the largest posterior cluster-specific probability < 0.95. In the human PBMC samples, most of the vague cells (colored in red) are located at the boundary of different clusters, which reassuring the validity of the clustering results. In real data analysis, users can decide to remove vague cells under a user-specified criterion (based on the posterior probability) for the downstream analysis such as differential gene expression analysis within each cell type.

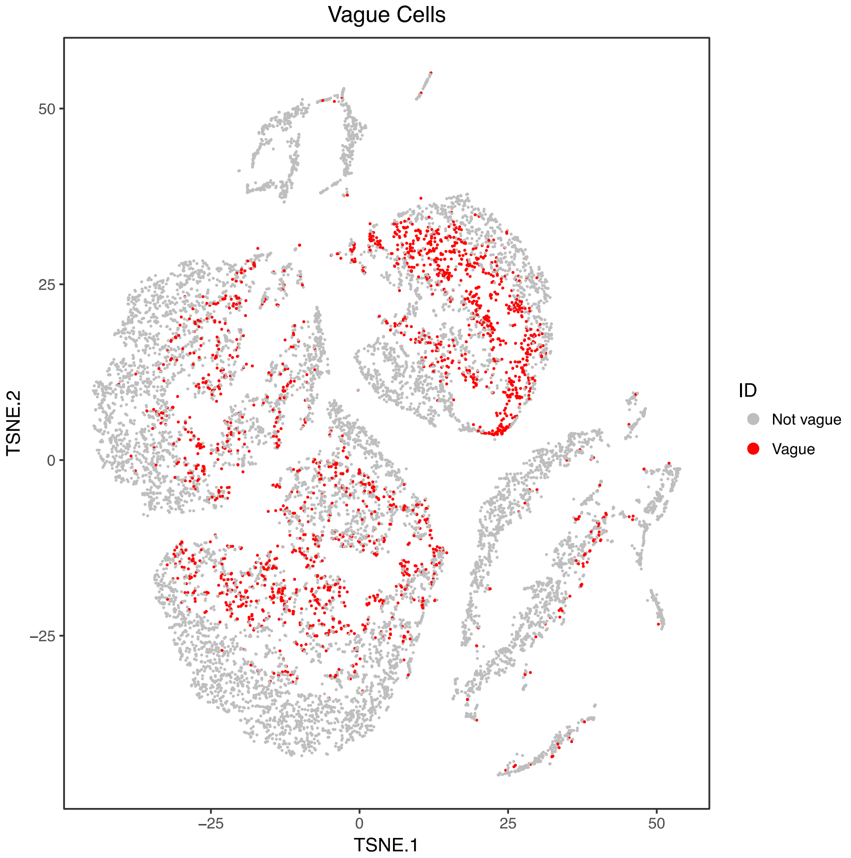


Figure 11. The t-SNE projection of human PBMC samples for the illustration of vague cells with the largest posterior probability < 0.95

3.6 Discussion

In summary, I have developed a novel Bayesian framework for clustering population-scale scRNA-Seq data. BAMM-SC retains the raw data information by directly modeling UMI counts without data transformation or normalization, facilitating straightforward biological interpretation. The Bayesian hierarchical model enables the joint characterization of multiple sources of uncertainty, including single cell level heterogeneity and individual level heterogeneity. Furthermore, BAMM-SC can borrow information across different individuals through its mixture hierarchical model structure and Bayesian computational techniques, leading to improved clustering accuracy. BAMM-SC is based on probabilistic models, thus providing the quantification of clustering uncertainty for each single cell.

My model coupled with a computationally efficient MCMC algorithm, which is able to cluster large-scale droplet-based scRNA-Seq data with feasible computational cost. For example, using 1, 000 highly variable genes, it takes about 1.5, 2.5 and 4.5 hours when analyzing the three real datasets (human PBMC, mouse lung and human skin), respectively. For the simulated dataset consist of 10 individuals with 4,000 cells each, the computational time for clustering is about 30 minutes. Figure 35 (Appendix B) demonstrates that the computational time of BAMM-SC increases approximately linearly with the increase of the number of cells in each individual, the number of individuals and the number of clusters, respectively. To further improve the computational efficiency, I provided a "supervised" clustering option in BAMM-SC for very large-scale datasets. Specifically, users can first apply BAMM-SC on a small subset of single cells in each individual, and save predicted cluster labels as well as other informative parameters such as

$\boldsymbol{\alpha}_{\cdot lk}$. Then for the remaining single cells, users can perform "supervised" classification via BAMM-SC instead of "unsupervised" clustering (Appendix B.2). By clustering a small number of single cells, this procedure will substantially reduce the computational cost. I used the simulated dataset of 10 individuals to demonstrate the effectiveness of this supervised option in Figure 12**.** I simulated two datasets: one dataset consists of 10 individuals with 400 cells each and the other dataset consists of 10 individuals with 4,000 cells each. I selected a subset of cells in each individual as the training set and treated the remaining cells as the test set. I set the proportion of cells in the training set from 10% to 100% and reported the ARIs for both training and test sets. When the proportion equals 100%, there is no test data set, thus only ARI for the training set is reported. I repeated this simulation procedure 100 times and reported ARIs in Figure 12 below. When the total number of cells in the training set is large enough (4,000 in total or more), the prediction performance (measured by ARI) in the test set is saturated. For the dataset consists of 10 individuals with 4,000 cells each, when I used 10% cells for training, it only takes ~90 seconds to obtain the clustering labels for all cells in both training and test sets with the similar performance from the full dataset. Therefore, for large datasets (e.g. > 100K cells), users can apply BAMM-SC to a smaller subset of cells in each individual to cluster distinct cell types, and then classify the remaining cells according to the predicted cell types. BAMM-SC is currently implemented in R/Rcpp with satisfactory computing efficiency to accommodate population scale scRNA-seq data. Further speed-up can be made through parallel computing or GPU.

Additionally, I can pre-define the number of clusters based on prior knowledge on the tissue or determine it using some standard model checking criterion such as AIC or BIC. As shown in Figure 36 (Appendix B), AIC and BIC work as expected in the analysis of simulated datasets and provide a reliable range of cluster numbers to guide real data analysis based on prior knowledge. However, in biological study, the number of clusters is often considered as a continuum because of the nature of cell growth, so I recommend trying a range of cluster numbers in practice. BAMM-SC is shown to be robust against model mis-specification. In my simulation studies, I applied Splatter to simulate scRNA-Seq data in which the data generation mechanism is different from my proposed BAMM-SC model. BAMM-SC still achieved higher clustering accuracy than other competing methods. In addition, I compared BAMM-SC with other clustering methods when the number of clusters is different from the true number of cell types. Figure 37 (Appendix B) shows that BAMM-SC still achieved the highest ARI in most scenarios.

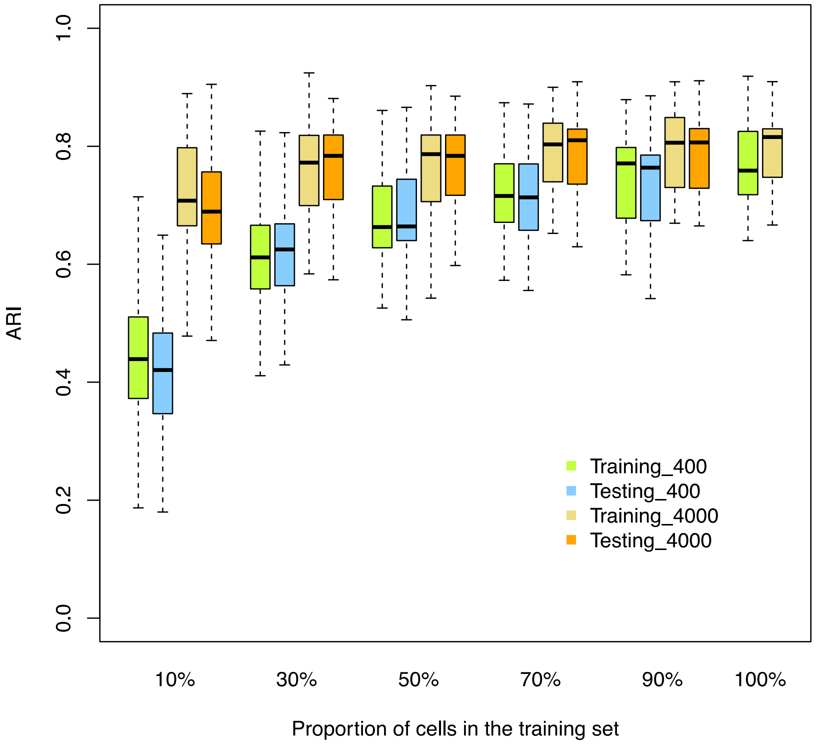


Figure 12. The Boxplots of ARI for BAMM-SC across 100 simulations, demonstrating the clustering accuracy under different proportions of cells being selected in the training set

Other than MNN and CCA, several other approaches have been proposed to correct batch effect across multiple individuals. One straightforward approach is taking one individual as the reference, producing a low-dimensional embedding of it and then projecting the other individuals onto that embedding. To perform low-dimensional embedding, diffusion map (Coifman, et al., 2005) is a tool for non-linear dimension reduction and has recently been adapted for the visualization of single cell gene expression data. Additionally, single-cell Variational Inference (scVI) is a scalable framework for batch correction based on variational inference and stochastic optimization of deep neural networks (Lopez, 2018). The performance of diffusion map and scVI combined with other clustering method was examined, which is worse than MNN and CCA in the three synthetic datasets (possibly due to unmet model underlying assumptions). I will explore more emerging methods in our future work.

There are several limitations of BAMM-SC. First, I filtered out genes with excessive zeros from the analysis under the assumption that lowly-expressed genes do not contribute much to clustering. This may be problematic for rare cell type identification. Second, I do not explicitly model a zero-inflation pattern, which may or may not affect clustering accuracy. A refined model that can handle inflated zeros can be further developed with a balance between computational complexity and model flexibility. Third, in my model, I assume that each cell belongs to one distinct cluster. The posterior probability measures the clustering uncertainty, which cannot be directly interpreted as a quantification of cell cycle or developmental stage. Finally, although the supervised strategy is proven to work for large datasets efficiently, it may potentially miss some rare clusters.

My method has the potential to be extended to perform trajectory analysis (Trapnell, 2015; Trapnell, et al., 2014), and accounts for both individual and batch level heterogeneity (e.g. two individuals spread evenly across two 10X chips in a properly blocked design) by adding another level of structure. In addition, the model parameters can be used for downstream differential gene expression analysis or construct cell-type specific biomarker panels. These interesting directions are beyond the scope of this dissertation and will be studied in future papers. Additionally, unlike the traditional way of analyzing scRNA-Seq data, BAMM-SC can be also used with batch effect correction. As shown in Figure 38 (Appendix B**),** I ran BAMM-SC on the mouse lung dataset first and extracted cells in cluster 4. Then I applied CCA (implemented in Seurat) on this specific cluster of cells and replotted the t-SNE plot. From Figure 38E, cells from different samples are superimposed on each other, suggesting that most batch effect has been removed. In practice, I recommend using BAMM-SC for clustering raw count data and then use other methods, such as MNN and CCA, to remove batch effect for each individual cell type if needed.

I have applied BAMM-SC to simulated datasets and three in-house synthetic datasets to showcase its performance on different tissue types and species. With the increased popularity of population-based scRNA-Seq studies, BAMM-SC will become a powerful tool for elucidating single cell level transcriptomic heterogeneity from population-bases studies and a complementary approach to existing clustering methods.

4.0 BREM-SC: A Random Effect Bayesian Mixture Model for Joint Clustering Single Cell CITE-Seq Data

Besides the single cell transcriptome sequencing (scRNA-Seq) technology, another revolutionary technology named Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq) came out more recently. Coupling with droplet-based scRNA-Seq, it allows the detection of cell surface proteins and transcriptome profiling within the same cell simultaneously. Despite the rapid advances in technologies, novel statistical methods and computation tools for analyzing CITE-Seq data are lacking. In this study, I developed BREM-SC, a novel random effects model that jointly cluster the paired data from CITE-Seq simultaneously. Simulations and analysis of in-house real data sets were performed, which successfully demonstrated the validity and advantages of this method in fully utilizing both types of data to identify cell clusters. We expect this new method will greatly help researchers better understand immune diseases as well as facilitate novel biological discoveries.

4.1 Introduction

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq) is a recently developed revolutionary tool, which is the first technique that can measure single cell surface protein and mRNA expression level simultaneously in the same cell (Stoeckius, et al., 2017). Oligonucleotide-labeled antibodies are used to integrate cellular protein and transcriptome measurements. It combines highly multiplexed protein marker detection with transcriptome profiling for thousands of single cells. CITE-Seq allows for immunophenotyping of cells using existing single cell sequencing approaches (Stoeckius, et al., 2017). It is fully compatible with droplet-based single cell RNA sequencing (scRNA-Seq) technology (by 10X Genomics Chromium (Zheng, et al., 2017)). These promising and popular technologies provide the opportunity for jointly analyzing transcriptome and surface proteins at single cell level in a cost-effective way.

The rapid advances in single cell technologies help researchers better understand cell heterogeneity and identify cell types, which leads to the high demand of novel statistical methods and tools to analyze data with different characteristics. Current statistical methods for jointly analyzing data from CITE-Seq are still immature. A novel joint clustering approach that fully utilizes the advantages and unique features of these single cell multi-omics data will lead to a more powerful tool in identifying rare cell types.

In CITE-Seq analysis, RNA measurements and the expression levels of Antibody-Derived Tags (ADT) are collected for a common set of cells. These two data sources represent different but highly related and dependent biological components. Traditional cell type identification relies on cell surface protein abundance, which can be measured with flow cytometry. Moreover, researchers can also use scRNA-Seq data to classify cell types, especially based on genes that are differentially expressed between different cell types. Both data sources have their unique characteristics and can provide complementary information. For example, the use of cell surface proteins for cell gating is advantageous in identifying common cell types but may not successfully identify some rare cell types due to its low dimensionality. On the other hand, cell clustering based on scRNA-Seq could identify more cell types because of its higher dimensionality. However, it may not be able to distinguish highly similar cell types, such as CD4+ T cells and CD8+ T cells, due to a poor observed correlation between an mRNA and its translated protein expression in single cell (Chen, et al., 2002; Haider and Pal, 2013).

Different clustering methods have been proposed for clustering gene expression data only. Recently, single cell interpretation via multi-kernel learning (SIMLR) (Wang, et al., 2018), Seurat (Satija, et al., 2015), SC3 (Kiselev, et al., 2017) and DIMM-SC (Sun, et al., 2018) have been proposed for clustering scRNA-Seq data. Among these methods, DIMM-SC directly models UMI counts using a multinomial distribution with Dirichlet mixture priors, and provides clustering uncertainty for each cell (i.e., how likely each cell belongs to each cluster). Therefore, it can benefit rigorous statistical inference and straightforward biological interpretations. In contrast, few methods are designed for directly clustering surface protein levels generated from CITE-Seq. Separate analyses of each data source may lack power and will not capture the associations between transcriptomes and expression of surface proteins. Multimodal data analysis can achieve a more detailed characterization of cellular phenotypes than using transcriptome measurements alone.

In this study, I proposed BREM-SC, a Bayesian mixture Random Effect Model for joint clustering single cell CITE-Seq data. In the following sections, I first introduce the BREM-SC method. Next, I compare the performance of BREM-SC with four popular clustering methods, including K-means clustering, SC3 (Kiselev, et al., 2017), TSCAN (Ji and Ji, 2016) and DIMM-SC (Sun, et al., 2018), in both simulation studies and real data applications. K-means is one of the most popular clustering methods and has been used in the first 10X Genomics publication (Zheng, et al., 2017). SC3, TSCAN and DIMM-SC have been proposed for clustering scRNA-Seq data. They are from different clustering categories. For example, SC3 is a single cell consensus clustering method. The consensus matrix is calculated using the Cluster-based Similarity

Partitioning Algorithm (CSPA). Unlike SC3, TSCAN performs model-based clustering on the transformed expression values.

4.2 Methods

4.2.1 Statistical model

Suppose there are C cells generated from CITE-Seq, denote by the transcriptomic data a matrix $X^{(1)}$ and its ADT levels $X^{(2)}$. I use a latent variable vector Z with elements z_j to represent the cell type label for the cell j, where $j = 1, \dots, C$.

For transcriptomic data, each element $x_{ij}^{(1)}$ represents the number of unique UMIs for gene i in cell j, where i runs from 1 to the total number of genes G, and j runs from 1 to the total number of cells C. Then I denote the number of unique UMIs in the j th single cell by a vector $x_j^{(1)} = \left(x_{1j}^{(1)}, x_{2j}^{(1)}, \dots, x_{Gj}^{(1)}\right)$. I assume that $x_j^{(1)}$ is generated from a multinomial distribution with parameter vector $p_j^{(1)} = \left(p_{1j}^{(1)}, p_{2j}^{(1)}, \dots, p_{Gj}^{(1)}\right)$. For multinomial distribution, I further assume that the proportion $p_j^{(1)} = \left(p_{1j}^{(1)}, p_{2j}^{(1)}, \dots, p_{Gj}^{(1)}\right)$ follows a Dirichlet prior distribution $Dir\left(\alpha^{(1)}\right) = Dir\ (\alpha_1^{(1)}, \alpha_2^{(1)}, \dots, \alpha_G^{(1)})$, with all the elements in $\alpha^{(1)}$ being strictly positive. Next, I assume that the cell population consists of K distinct cell types. To provide a more flexible modeling framework and allow for unsupervised clustering, I extend the aforementioned single Dirichlet prior to a mixture of K Dirichlet distributions, indexed by $k = 1, \dots, K$ and each with parameter $\alpha_{(k)}^{(1)}$. If cell j belongs to the k th cell type, its gene expression profile $p_j^{(1)}$ follows a cell-type-specific prior distribution $Dir(\alpha_{(k)}^{(1)})$. The full likelihood is then obtained by multiplying the Dirichlet mixture prior by the multinomial likelihood. The RNA data source-specific log likelihood is $log\prod_{j=1}^{C} P(x_j^{(1)}, z_j = k) = \sum_{j=1}^{C} I(z_j = k)\, log\left\{\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ij}^{(1)}+\alpha_{i(k)}^{(1)}\right)}{\Gamma\left(\alpha_{i(k)}^{(1)}\right)}\right)\frac{\Gamma(|\alpha_{(k)}^{(1)}|)}{\Gamma(T_j^{(1)}+|\alpha_{(k)}^{(1)}|)}\right\}$, where $T_j^{(1)} = \sum_i x_{ij}^{(1)}$ is the total number of unique UMIs for the j th cell. Similarly, I use the Dirichlet multinomial distribution to model ADT data. Suppose there are total D ADT markers, the density of Dirichlet multinomial is $P\left(x_j^{(2)}\middle|\alpha^{(2)}\right) = \frac{T_j^{(2)}!}{\prod_{d=1}^{D} x_{dj}^{(2)}!}\left(\prod_{d=1}^{D}\frac{\Gamma\left(x_{dj}^{(2)}+\alpha_d^{(2)}\right)}{\Gamma\left(\alpha_d^{(2)}\right)}\right)\frac{\Gamma(|\alpha^{(2)}|)}{\Gamma(T_j^{(2)}+|\alpha^{(2)}|)}$, where $d = 1, \dots, D$ and $T_j^{(2)} = \sum_d x_{dj}^{(2)}$. Here, $T_j^{(2)}$ is the total counts of ADT markers for the j th cell. Then the joint distribution for all cells is

$$\sum_{j=1}^{C}\sum_{k=1}^{K} I\left(z_j = k\right) log\left\{\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ij}^{(1)} + \alpha_{i(k)}^{(1)}\right)}{\Gamma\left(\alpha_{i(k)}^{(1)}\right)}\right)\frac{\Gamma(|\alpha_{(k)}^{(1)}|)}{\Gamma(T_j^{(1)} + |\alpha_{(k)}^{(1)}|)}\left(\prod_{d=1}^{D}\frac{\Gamma\left(x_{dj}^{(2)} + \alpha_{d(k)}^{(2)}\right)}{\Gamma\left(\alpha_{d(k)}^{(2)}\right)}\right)\frac{\Gamma(|\alpha_{(k)}^{(2)}|)}{\Gamma(T_j^{(2)} + |\alpha_{(k)}^{(2)}|)}\right\}.$$

To consider the correlation between cells from the same cell type, I add cell specific random effects into my framework. Given random effect $b_j \sim LogNormal\ (0, \sigma_b^2)$, $Prior(b_j) = \prod_{j=1}^{C}\frac{1}{b_j\sqrt{2\pi\sigma_b^2}} exp\left\{-\frac{(logb_j-0)^2}{2\sigma_b^2}\right\}$, I have $\alpha_{j(k)}^{(1)} = \alpha_{(k)}^{(1)} b_j$ and $\alpha_{j(k)}^{(2)} = \alpha_{(k)}^{(2)} b_j$, where $\alpha_{(k)}^{(1)}$ and $\alpha_{(k)}^{(2)}$ are the Dirichlet parameters of cell type k for RNA and ADT data, respectively. Then, I can have the complete log likelihood, which is

$$log\, P\left(\alpha^{(1)}, \alpha^{(2)}, Z, b_j \middle| X^{(1)}, X^{(2)}\right) \propto \sum_{j=1}^{C}\sum_{k=1}^{K} I\left(z_j = k\right)$$

$$log\left\{\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ij}^{(1)}+\alpha_{i(k)}^{(1)} b_j\right)}{\Gamma\left(\alpha_{i(k)}^{(1)} b_j\right)}\right)\frac{\Gamma\left(\left|\alpha_{(k)}^{(1)} b_j\right|\right)}{\Gamma\left(T_j^{(1)}+\left|\alpha_{(k)}^{(1)} b_j\right|\right)}\left(\prod_{d=1}^{D}\frac{\Gamma\left(x_{dj}^{(2)}+\alpha_{d(k)}^{(2)} b_j\right)}{\Gamma\left(\alpha_{d(k)}^{(2)} b_j\right)}\right)\frac{\Gamma(|\alpha_{(k)}^{(2)} b_j|)}{\Gamma(T_j^{(2)}+|\alpha_{(k)}^{(2)} b_j|)}\right\} +$$

$$\sum_{j=1}^{C}(-logb_j - \frac{(logb_j)^2}{2\sigma_b^2}) + \sum_{j=1}^{C}(-\frac{1}{2} log\sigma_b^2).$$

I use Gibbs sample to iteratively update z_j, $\alpha_{i(k)}^{(1)}$, $\alpha_{d(k)}^{(2)}$ and b_j. Details can be found in Appendix C.1.

4.2.2 Selection of the number of clusters and initial values

To implement BREM-SC, it is critical to select the total number of clusters and the initial values for MCMC. Specifically, the number of cluster K can be defined with prior knowledge or standard model checking criterion such as Akaike's Information Criteria (AIC) or Bayesian Information Criteria (BIC). Meanwhile, there are many methods to determine the initial values of $\alpha_1, \alpha_2, \ldots, \alpha_G$. As described in Chapter 2, I applied K-means clustering to get a preliminary clustering result, and then used Ronning's method to estimate initial values of $\boldsymbol{\alpha}$, similar to the estimation procedure in DIMM-SC.

4.2.3 Simulation studies

I performed comprehensive simulation studies to compare BREM-SC with three existing clustering methods, including K-means clustering, SC3 and TSCAN. They are hard clustering approaches, which assign each cell to an exclusive cluster. Based on different Dirichlet multinomial models, I simulated RNA expression and ADT measurements for each single cell. In the simulation set-up, the two count matrices were sampled from the proposed Dirichlet mixture models. Specifically, for a fixed number of cell clusters K, I first pre-defined the values of $\alpha_{(k)}^{(1)}$ and $\alpha_{(k)}^{(2)}$ for the k th cell cluster. The random effects b_j are generated from log-normal distribution with pre-specified value σ_b^2. Then I can get the transcriptomic profile for each single cell by multiplying $\alpha_{(k)}^{(1)}$ with b_j. Similarly, for cellular protein expression profile, I multiplied $\alpha_{(k)}^{(2)}$ with b_j to calculate $\boldsymbol{\alpha}_{\boldsymbol{j(k)}}^{(\mathbf{2})}$ for each cell. Next, I sampled the proportion $\boldsymbol{p}_{\boldsymbol{j}}^{(\mathbf{1})}$ (or $\boldsymbol{p}_{\boldsymbol{j}}^{(\mathbf{2})}$) from a Dirichlet distribution $Dir\left(\boldsymbol{\alpha}_{\boldsymbol{j(k)}}^{(\mathbf{1})}\right)$ (or $Dir\left(\boldsymbol{\alpha}_{\boldsymbol{j(k)}}^{(\mathbf{2})}\right)$). Lastly, I sampled the UMI count vector $\boldsymbol{x}_{\boldsymbol{j}}^{(\mathbf{1})}$ for the j th cell from the multinomial distribution $Multinomial(T_j^{(1)}, \boldsymbol{p}_{\boldsymbol{j}}^{(\mathbf{1})})$ and sampled the levels of ADT markers $\boldsymbol{x}_{\boldsymbol{j}}^{(\mathbf{2})}$ from another multinomial distribution $Multinomial(T_j^{(2)}, \boldsymbol{p}_{\boldsymbol{j}}^{(\mathbf{2})})$. In my simulation studies, I considered different experimental designs, including different number of cells in each cluster, different number of clusters, different cell-type-specific heterogeneity (i.e., the magnitude of difference among different clusters), and different heterogeneities among cells. All clustering methods were run under default parameters. For K-means, SC3 and TSCAN, I pooled data from RNA expression and ADT together while ignoring data source label, and then applied each clustering method on the pooled data. I simulated 100 datasets and summarized the corresponding adjusted rand index (ARIs) for each method.

4.3 Results

4.3.1 Results of simulation studies

As expected, the performance of all four clustering approaches decreases as the among cell heterogeneity increases, measured by the value of σ_b^2. As shown in Figure 13, BREM-SC outperformed the other three competing methods by achieving higher average ARI among 100 simulations when the level of cell heterogeneities is median or large. However, when σ_b^2 is very small, all b_j's are close to 0, indicating that there is little correlation between two data sources, and in this case, BREM-SC demonstrated larger variability compared with other methods.

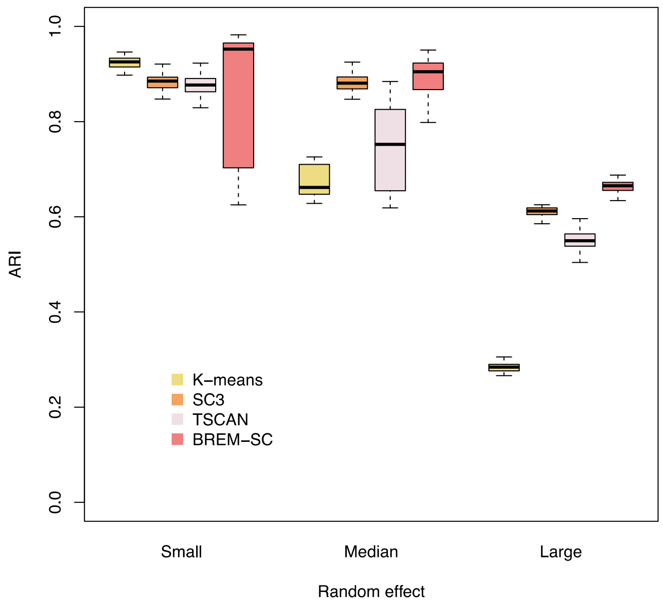


Figure 13. Boxplots of ARI for four clustering methods across 100 simulations, investigating how different levels of heterogeneities among cells affect clustering results

Figure 14 lists the boxplots of ARI for different cell-type-specific heterogeneity (Figure 14A) and different number of cells in each cluster (Figure 14B), respectively. In Figure 14A, I considered 4 scenarios in terms of signal strength from two data sources. To illustrate the advantage of joint clustering, I also applied K-means, SC3, TSCAN and DIMM-SC on ADT data alone. When the clustering signal is strong (i.e., difference among cell clusters is large) in both RNA expression and ADT data, all methods performed well. However, when cell clusters are similar in either proteomics or transcriptomics data, K-means and TSCAN produced less accurate clustering results, while BREM-SC and SC3 still performed well. Strong clustering signal leads to higher clustering accuracy and lower clustering variability. If the data of transcriptome as well as proteome from single cells are similar across different cell types, ARIs of all methods decreased but BREM-SC still performed better than the other methods. In Figure 14B, more cells can provide more accurate and robust clustering results, and BREM-SC achieved the highest ARI across a variety of number of cells. Consistent across these two scenarios, when data are generated from the true model, BREM-SC outperformed K-means clustering, SC3, TSCAN and DIMM-SC, suggesting its advantage in fully utilizing both types of data simultaneously. Furthermore, I performed simulation analysis by generating simulated scRNA-Seq datasets from different number of cell type clusters (Figure 39 (Appendix C)). BREM-SC still provided more accurate clustering results than other competing methods in this scenario.

To evaluate the robustness of BREM-SC when the data generation model is mis-specified, I simulated additional datasets using R package Splatter (Zappia, et al., 2017), a commonly used tool for scRNA-Seq data simulation. In Splatter, the final data matrix is a synthetic dataset consisting of counts from a Gamma-Poisson (or negative-binomial) distribution. Since there is no existing method for generating surface protein expression levels from CITE-Seq, in this work, I also used Splatter to generate ADT data. To make my simulated gene expression data a good approximation to the real data, I used the same approach as I did in Chapter 3, with model parameters (in Splatter) estimated from the real data downloaded from the 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/b_cells). For ADT data, I modified the Splatter parameters (such as dropout rate, library size, expression outlier, and dispersion across features) to make the simulated data similar to real observed ADT data. I assumed all cell types are shared between gene expression and ADT data, and further specified differential expression parameters to generate scenarios with different amount of cell type differences. As shown in Figure 15, BREM-SC still outperformed other competing methods in terms of clustering accuracy in all scenarios.

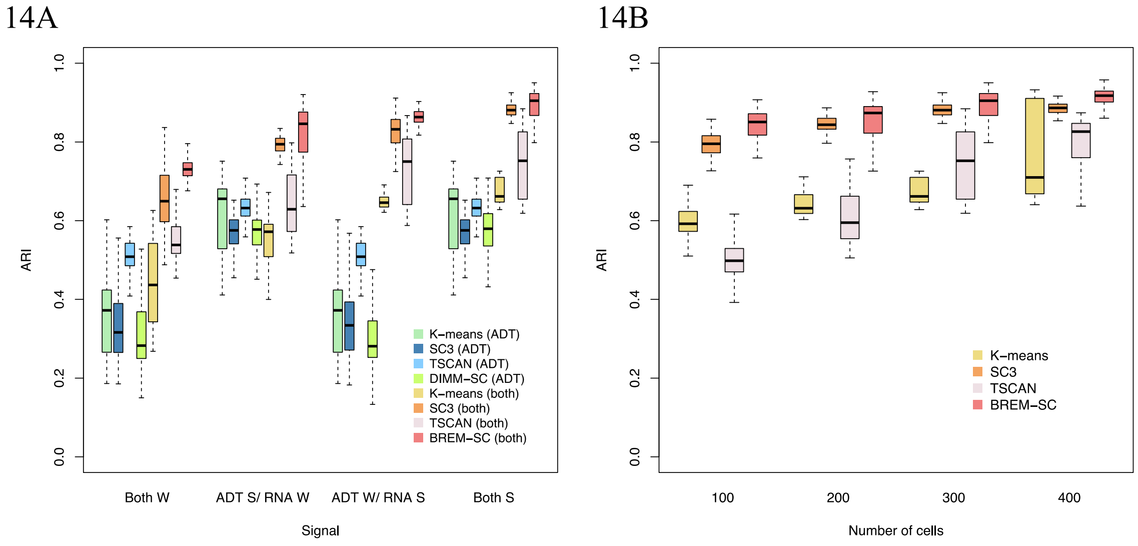


Figure 14. Boxplots of ARI for four clustering methods across 100 simulations, investigating how different cell-type-specific heterogeneity (14A) and number of cells (14B) affect clustering results

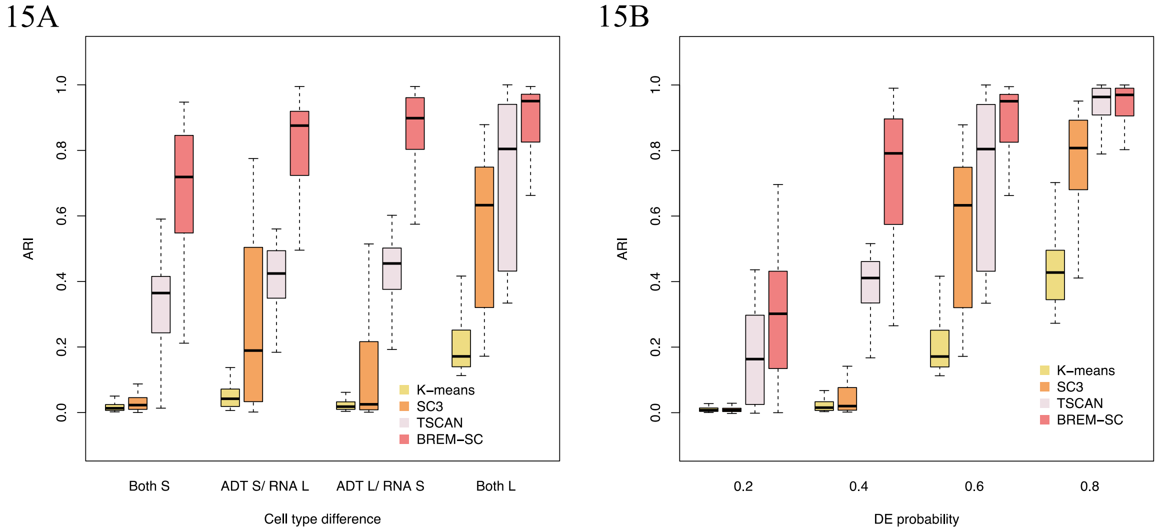


Figure 15. Boxplots of ARI for four clustering methods across 100 simulations using Splatter

We investigated how different levels of cell-type-specific heterogeneity affect clustering results. In Figure 15A, we set the mean parameters of three cell types as (0.15, 0.151, 0.152) and (0.15, 0.2, 0.25) to represent two levels (small and large) of cell type difference. In Figure 15B, we set the probability that a gene will be selected to be differentially expressed as different values.

4.3.2 Real data analysis

4.3.2.1 Analysis of the public human PBMC sample

To evaluate the clustering performance of BREM-SC, I downloaded a published human PBMC CITE-Seq dataset from the website of 10X Genomics. A total of 7,865 cells and 14 ADT markers are included in this dataset. Similar to the analysis of in-house human PBMC dataset, I extracted the top 100 highly variable genes based on their standard deviations, and identified seven cell types based on the biological knowledge of both ADT and gene markers (Figure 40 (Appendix C)). Taken together, more than 80% of single cells can be assigned to a specific cell type. I applied five clustering methods (K-means clustering, TSCAN, SC3, DIMM-SC and BREM-SC) on this dataset. As shown in Table 13, BREM-SC outperformed other methods in terms of ARI. Figure 16 shows the t-SNE plots with each cell colored by their label based on cell-type-specific markers and cluster labels inferred by BREM-SC, respectively. These two are highly similar (ARI = 0.868), indicating the outstanding performance of BREM-SC.

Table 13. Performance of clustering for the public human PBMC real dataset

	Mean (SD)	Median (Range)
K-means (ADT)	0.653 (0.093)	0.648 (0.436, 0.760)
TSCAN (ADT)	0.389 (NA)	0.389 (0.389, 0.389)
SC3 (ADT)	0.649 (0.045)	0.668 (0.568, 0.692)
DIMM-SC (ADT)	0.673 (0.048)	0.679 (0.618, 0.737)
K-means	0.645 (0.158)	0. 689 (0.383, 0.889)
TSCAN	0.472 (NA)	0.472 (0.472, 0.472)
SC3	0.543 (0.130)	0.534 (0.395, 0.853)
BREM-SC	0.737 (0.125)	0.713 (0.556, 0.868)

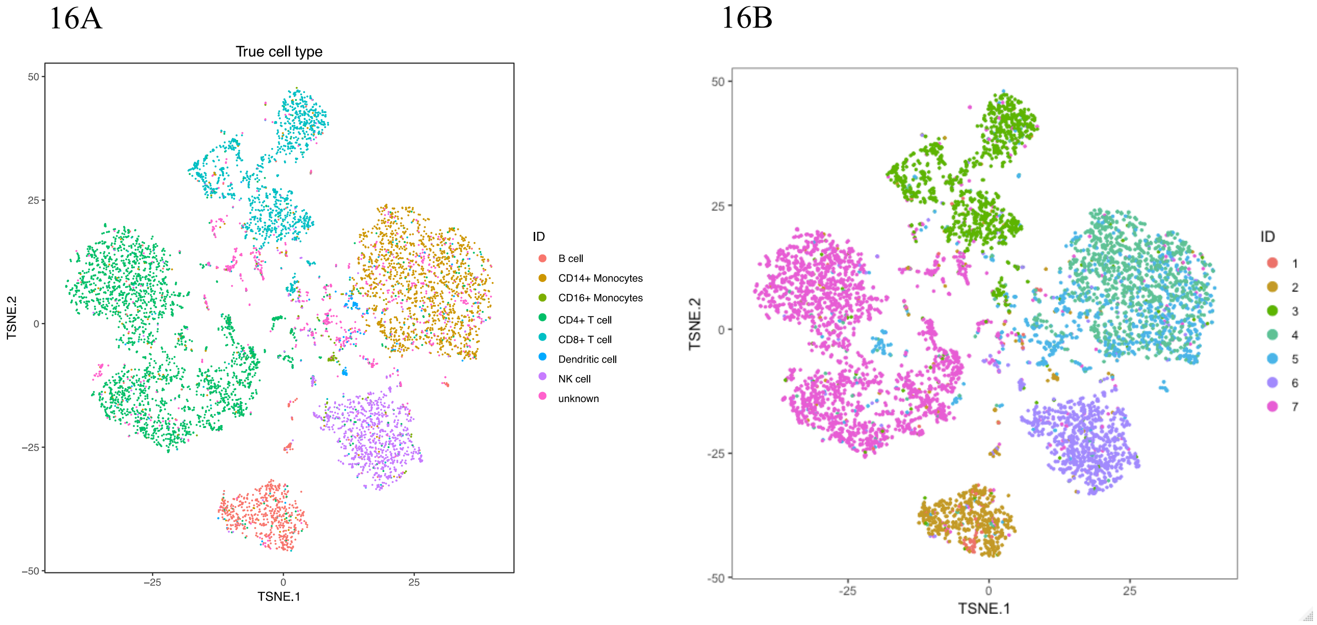


Figure 16. The performance of BREM-SC for public human PBMC CITE-Seq dataset

The t-SNE projection of cells are colored by the ground truth and BREM-SC clustering results.

4.3.2.2 Analysis of the in-house human PBMC sample

To evaluate the clustering performance of BREM-SC in solid human tissues, isolated human peripheral blood mononuclear cells (PBMCs) from the whole blood were obtained from a healthy donor and the 10X Chromium system was used to generate CITE-Seq data, which yielded a total of 1,388 cells. There were 10 cell surface markers designed in the experiment. Similarly, I extracted the top 100 highly variable genes based on their standard deviations. As shown in Figure 17, I identified six subtypes of PBMCs based on the biological knowledge of cell-type-specific ADT markers. Using these markers, >85% single cells can be assigned to a specific cell type (Figure 41 (Appendix C)). I used the labels of these cells as the ground truth to benchmark the clustering performance for different clustering methods. Cells with uncertain cell types were removed when calculating ARIs.

Similar to the simulation studies, I applied five clustering methods (K-means clustering, TSCAN, SC3, DIMM-SC and BREM-SC) on the PBMC dataset and repeated each method ten times to evaluate the stability of its performance (Table 14). The total number of clusters was set as six. In Table 14, BREM-SC achieved the highest ARI for human PBMC sample compared to all other competing methods. Note that TSCAN is a deterministic clustering method and therefore it generated identical results for ten analyses. As shown in Figure 17, BREM-SC performed well in the human PBMC samples, since the t-SNE plot with each cell colored by their cell-type label based on ADT markers is highly similar to the plot generated from the clustering result of BREM-SC (ARI = 0.895).

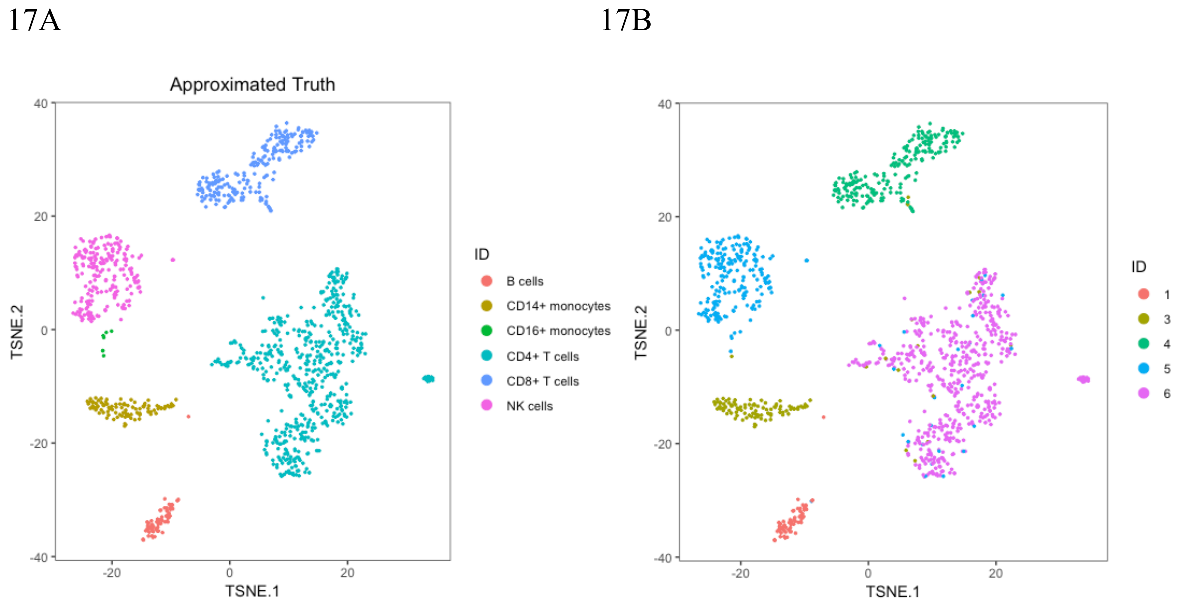


Figure 17. The performance of BREM-SC for in-house human PBMC CITE-Seq dataset

The t-SNE projection of cells are colored by the ground truth and BREM-SC clustering results.

Table 14. Performance of clustering across ten times analyses for human PBMC real dataset

	Mean (SD)	Median (Range)
K-means (ADT)	0.930 (0.120)	0.986 (0.702, 0.989)
TSCAN (ADT)	0.877 (NA)	0.877 (0.877, 0.877)
SC3 (ADT)	0.921 (0.039)	0.931 (0.816, 0.951)
DIMM-SC (ADT)	0.959 (0.037)	0.985 (0.916, 0.992)
K-means	0.596 (0.142)	0.613 (0.434, 0.803)
TSCAN	0.443 (NA)	0.443 (0.443, 0.443)
SC3	0.777 (0.047)	0.788 (0.679, 0.848)
BREM-SC	0.857 (0.048)	0.874 (0.749, 0.895)

In real world, researchers may not have any biological information of a specific cell type when designing ADT markers. In this case, we will not be able to identify all the cell types in the data only based on ADT markers. To mimic the situation where the pre-designed ADT markers cannot capture the characteristics of all cell types, I randomly removed three ADT markers (CD8A, CD16, CD127) in this human PBMC dataset, and applied five clustering methods (K-means clustering, TSCAN, SC3, DIMM-SC and BREM-SC) on the subset dataset and repeated this process ten times to evaluate the stability of all approaches (Table 15). The total number of clusters was still set as six. In Table 15, BREM-SC achieved the highest ARI compared with all other clustering methods.

Table 15. Performance of clustering for the subset of human PBMC real dataset with three ADT markers (CD8A, CD16, CD127) removed

	Mean (SD)	Median (Range)
K-means (ADT)	0.752 (0.126)	0.707 (0.666, 0.991)
TSCAN (ADT)	0.662 (NA)	0.662 (0.662,0.662)
SC3 (ADT)	0.757 (0.106)	0.818 (0.552, 0.840)
DIMM-SC (ADT)	0.782 (0.124)	0.758 (0.647, 0.983)
K-means	0.466 (0.023)	0.478 (0.437, 0.485)
TSCAN	0.371 (NA)	0.371(0.371, 0.371)
SC3	0.711 (0.095)	0.688 (0.614, 0.831)
BREM-SC	0.839 (0.017)	0.845 (0.805, 0.851)

4.4 Discussion

BREM-SC directly models count data from CITE-Seq using two multinomial distributions (one for each data type) with cells being treated as random effects. Unlike many other clustering methods which typically convert the counts into continuous measures, BREM-SC can work on the full matrix files compiled by users from the 10X Genomics Cellranger pipeline directly, to preserve the straightforward interpretation of count data. BREM-SC is also the first statistical approach to jointly cluster the paired data (scRNA and ADT) from CITE-Seq simultaneously. I demonstrated that BREM-SC has achieved substantial improvements in clustering accuracy compared to applying existing scRNA data clustering methods (e.g., K-means, TSCAN, SC3 and DIMM-SC) on CITE-Seq data. This probabilistic model provides clustering uncertainty for each cell (i.e., how likely each cell belongs to each cluster), which can enjoy the advantage of rigorous statistical inference and straightforward biological interpretations. Unlike DIMM-SC, BREM-SC considers the correlation between different data sources for the same cell. The random effect part further accounts for data heterogeneity among cells and therefore reduces the false positives of detecting rare cell types.

When analyzing data from different data sources, ensemble clustering may be considered to integrate the separate clusterings and determine an overall partition of cells that agrees the most with the source-specific clusterings. However, most of ensemble clustering methods assume that the separate clusterings are known in advance and do not inherently model the uncertainty (Wang, 2011). At the other extreme, a joint analysis that ignores the heterogeneity of the data may not capture important features that are specific to each data source. A fully integrative clustering approach is necessary to effectively combine the discriminatory power from transcriptome and protein measurements.

However, there are several noticeable limitations of this method. First, BREM-SC uses a computationally intensive MCMC algorithm which may cluster large datasets (e.g., >10 K cells) with a high computational cost. BREM-SC is currently implemented in R/Rcpp to accommodate large scale CITE-Seq data. Further speed-up can be made through block-wise MH within Gibbs sampling approach or graphics processing unit. Second, BREM-SC model ignores the measurement errors and uncertainties buried in count matrices. Multiple factors such as drop-out event, mapping percentage and PCR efficiency are not considered in the current model. These limitations can be largely overcome by extending the method. I will explore these directions in the near future.

In summary, I provide a novel statistical method BREM-SC for clustering CITE-Seq data, which facilitates rigorous statistical inference of cell population heterogeneity. I am confident that BREM-SC will be highly useful for the fast-growing community of large-scale single cell analysis.

5.0 Discussion and Future Work

The research work comprising this dissertation focuses on developing statistical methods for clustering the count data generated from scRNA-Seq and CITE-Seq technologies. In the first part, I developed DIMM-SC, a Dirichlet mixture model for clustering droplet-based scRNA-Seq data. I performed comprehensive simulations and real data applications to evaluate DIMM-SC and compared it with existing clustering methods. Both simulation studies and real data applications demonstrated that overall, DIMM-SC achieves substantially improved clustering accuracy and much lower clustering variability compared to other existing clustering methods. In the second part, I developed BAMM-SC, a novel Bayesian hierarchical Dirichlet mixture model to cluster droplet-based scRNA-Seq data from population studies. To be noted, BAMM-SC is able to account for data heterogeneity among multiple individuals such as unbalanced sequencing depths, read length and technical bias. I demonstrated that BAMM-SC achieves substantially improved clustering accuracy compared to other existing clustering methods. I applied this method to both human and mouse datasets. In the third part, I developed BREM-SC, a novel statistical method of joint clustering for data from CITE-Seq. Analysis of simulations and in-house real data were performed. It has been demonstrated that BREM-SC can account for the correlation between transcriptomes and cell surface proteins measurements within each single cell, and it is a powerful tool to jointly analyze RNA and surface protein data at single cell level.

However, there are some challenges that have not been fully addressed in this dissertation. For example, all the models proposed in this dissertation directly models UMI counts from scRNA-Seq data using a multinomial distribution with Dirichlet mixture priors. These models ignore the drop-out event, meaning that a gene which is expressed even at a relatively high level may be undetected due to technical limitations such as the inefficiency of reverse transcription. Such errors are distinct from random sampling and can often lead to significant error in cell-type identification and downstream analyses. Other than that, DIMM-SC, BAMM-SC and BREM-SC are unsupervised clustering method that infers structures from all data. Prior knowledge on cell-type-specific biomarkers may further improve the clustering accuracy. To use such prior information, a semi-supervised approach is needed to guide cluster inference.

The development of machine learning methods (especially deep learning) is also crucial in the analysis of single cell data. Machine learning methods are well-known to their high prediction efficiency which is based on prior knowledge from a training process. For prediction with omics data whose number of features is usually large, machine learning methods are expected to have better performance compared to traditional statistical tools. We expect that the utilization of deep learning methods on single cell data from the revolutionary scRNA-Seq and CITE-Seq technology will greatly advance our understanding of cell biology, tissue heterogeneity, and disease pathogenesis.

Appendix A (For DIMM-SC)

A.1 The E-M Algorithm

I used the E-M algorithm to maximize the log posterior distribution. Specifically, I first denoted $P\left(z_j = k\right) = \pi_k$, where π_k is the proportion of the k th cell type among all cells. I then treated z_j as missing data and used the E-M algorithm to estimate $\alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{Gk}$ and π_k. The complete data likelihood is:

$$\prod_{j=1}^{C} P(x_j, z_j) = \prod_{j=1}^{C} \left\{ \left(\prod_{i=1}^{G} \frac{\Gamma(x_i + \alpha_{ik})}{\Gamma(\alpha_{ik})} \right) \frac{\Gamma(|\alpha_{(k)}|)}{\Gamma(T_j + |\alpha_{(k)}|)} \right\}^{I(z_j=k)},$$

where $\left|\alpha_{(k)}\right| = \alpha_{1k} + \alpha_{2k} + \cdots + \alpha_{Gk}$ and the log likelihood is:

$$log \prod_{j=1}^{C} P(x_j, z_j) = \sum_{j=1}^{C} I(z_j = k)\, log \left\{ \left(\prod_{i=1}^{G} \frac{\Gamma\left(x_{ij} + \alpha_{ik}\right)}{\Gamma(\alpha_{ik})} \right) \frac{\Gamma(|\alpha_{(k)}|)}{\Gamma(T_j + |\alpha_{(k)}|)} \right\}.$$

E-step:

At the t th iteration, with the current realization of parameters $\Theta^{(t)} = (\alpha_{1k}^{(t)}, \alpha_{2k}^{(t)}, \ldots, \alpha_{Gk}^{(t)},\ \pi_k^{(t)})$, the conditional expectation is:

$$E_{z_j|x_j,\Theta^{(t)}}\, log\, P\left(x_j, z_j\right) = log \left\{ \left(\prod_{i=1}^{G} \frac{\Gamma\left(x_{ij} + \alpha_{ik}\right)}{\Gamma(\alpha_{ik})} \right) \frac{\Gamma\left(\left|\alpha_{(k)}\right|\right)}{\Gamma\left(T_j + \left|\alpha_{(k)}\right|\right)} \right\} * P\left(z_j = k \middle| x_j, \Theta^{(t)}\right),$$

where

$$P\left(z_j = k \middle| x_j, \Theta^{(t)}\right) = \frac{\left(\prod_{i=1}^{G} \frac{\Gamma\left(x_{ij} + \alpha_{ik}^{(t)}\right)}{\Gamma\left(\alpha_{ik}^{(T)}\right)} \right) \frac{\Gamma\left(\left|\alpha_{(k)}^{(t)}\right|\right)}{\Gamma\left(T_j + \left|\alpha_{(k)}^{(t)}\right|\right)} \pi_k^{(t)}}{\sum_{k=1}^{K} \left(\prod_{i=1}^{G} \frac{\Gamma\left(x_{ij} + \alpha_{ik}^{(t)}\right)}{\Gamma\left(\alpha_{ik}^{(T)}\right)} \right) \frac{\Gamma\left(\left|\alpha_{(k)}^{(t)}\right|\right)}{\Gamma\left(T_j + \left|\alpha_{(k)}^{(t)}\right|\right)} \pi_k^{(t)}} = \delta_{jk}.$$

Here δ_{jk} represents the probability that the j th cell belongs to the k th cluster. We calculated δ_{jk} in the E-step at each iteration.

M-step:

At the t th iteration, the estimation of the proportion of the k th cell type is

$$\hat{\pi}_k^{(t+1)} = \sum_{j=1}^{C} \delta_{jk}{}^{(t)} / C.$$

The update formula for $\alpha_{1k}, \alpha_{2k}, \dots, \alpha_{Gk}$ is derived from the Minka's fixed-point iteration for the leaving-one-out (LOO) likelihood (Minka , 2000):

$$\hat{\alpha}_{ik}^{(t+1)} = \alpha_{ik}^{(t)} \frac{\sum_{j=1}^{C} \delta_{jk} \left\{ x_{ij} / (x_{ij} - 1 + \alpha_{ik}^{(t)}) \right\}}{\sum_{j=1}^{C} \delta_{jk} \left\{ T_j / (T_j - 1 + |\alpha_{(k)}^{(t)}|) \right\}}.$$

After the M-step, I calculated $\sum_{k=1}^{K} (\hat{\pi}_k^{(t+1)} - \hat{\pi}_k^{(t)})^2$ and the relative difference of log likelihood between two consecutive iterations. Convergence tolerances for difference between iterations are pre-defined. I repeated the above steps until the convergence of log likelihood and $\hat{\pi}_k^{(t)}$, or a maximum number of iterations was reached. The default maximum number of iterations is 200.

A.2 Additional Figures

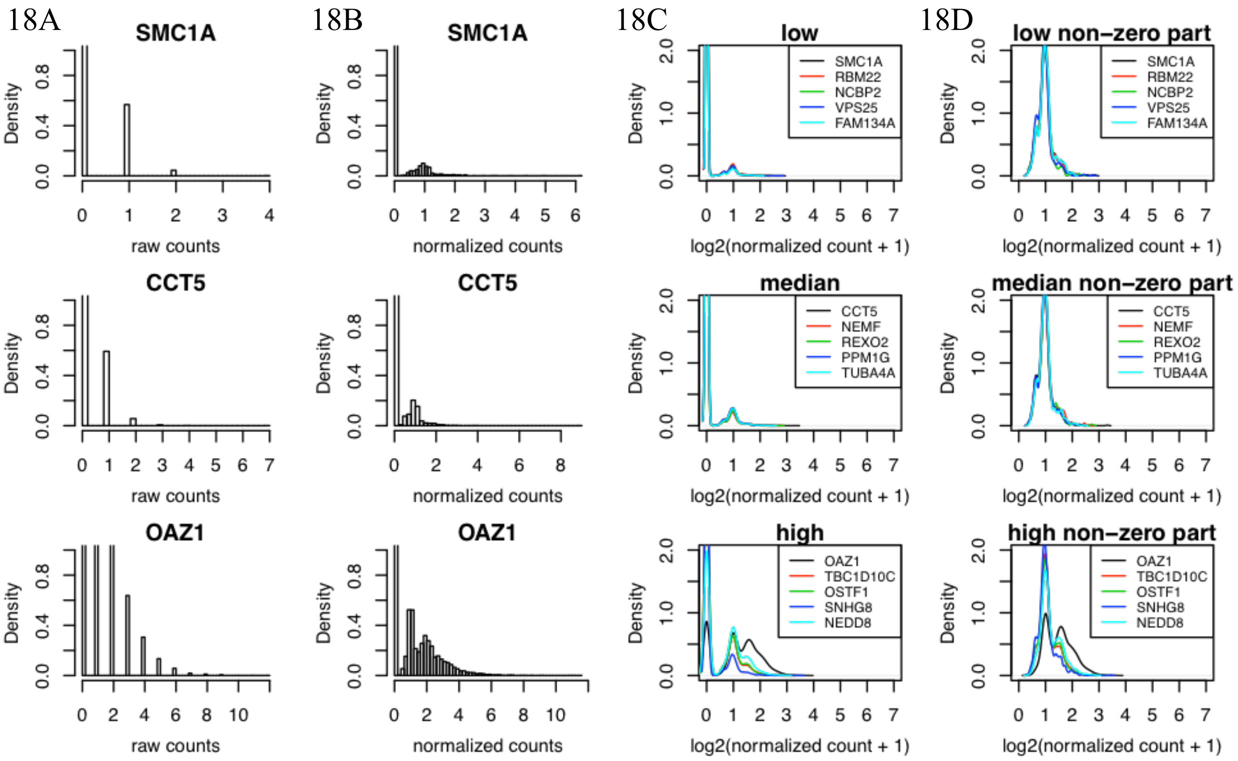


Figure 18. The empirical distribution of UMI counts for a few representative genes

We used the scRNA-Seq data from CD56+ NK cells. We first removed 29683 genes with zero count in more than 95% single cells and then performed normalization such that all single cells have the same total number of UMI counts. Normalization was performed by dividing UMI counts by the total number of UMI counts in each cell, and then multiplied by the median of the total UMI counts across all cells. We divided the remaining 3055 genes into three equal-sized groups based on their average gene expression levels. We randomly selected five genes from each group to generate the density plot of normalized count, for all data (Figure 18A) and non-zero part (Figure 18B), respectively. In addition, we provided the histograms of the raw counts (Figure 18C) and the untransformed normalized counts (Figure 18D) for gene SMC1A, gene CCT5 and gene OAZ1. All these empirical distribution plots of the UMI counts have demonstrated that drop-seq data contain extensive zeroes, for genes with different levels of expression.

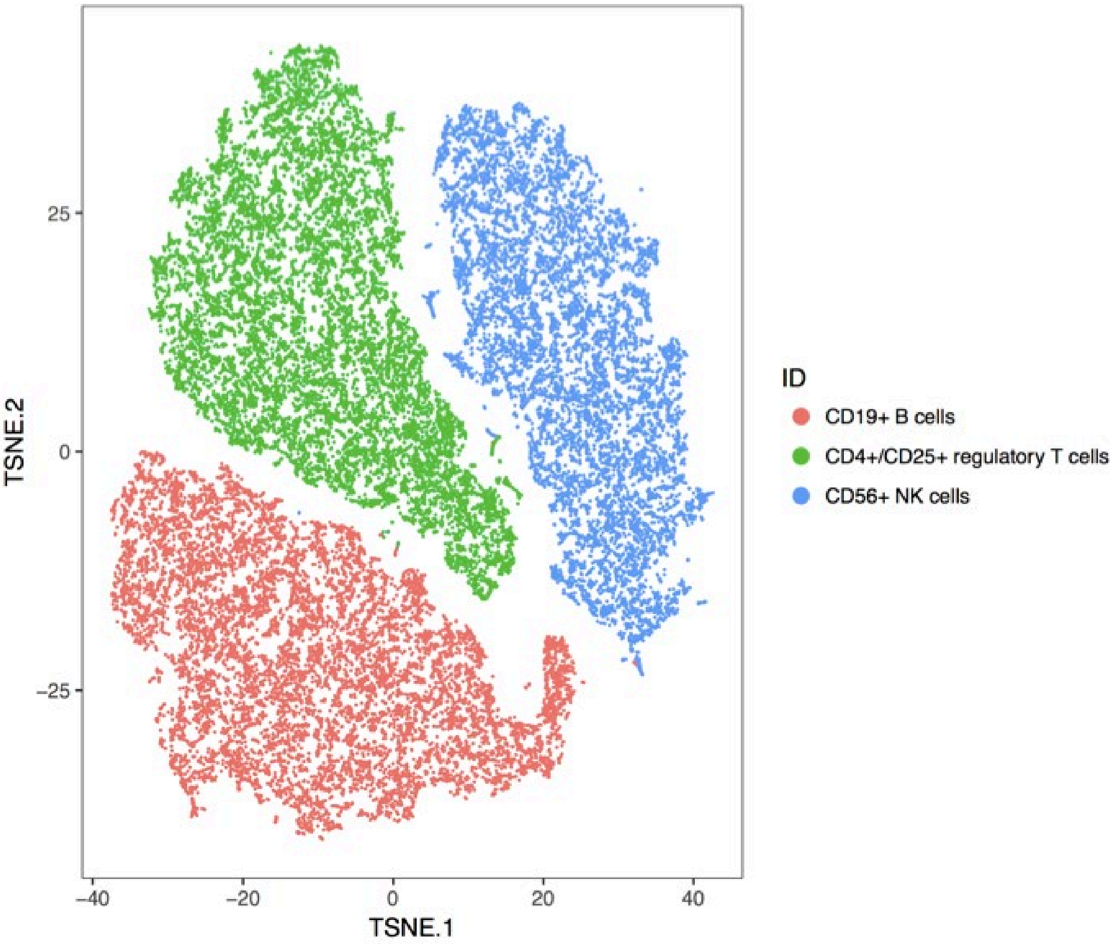


Figure 19. The t-SNE projection of the simple case

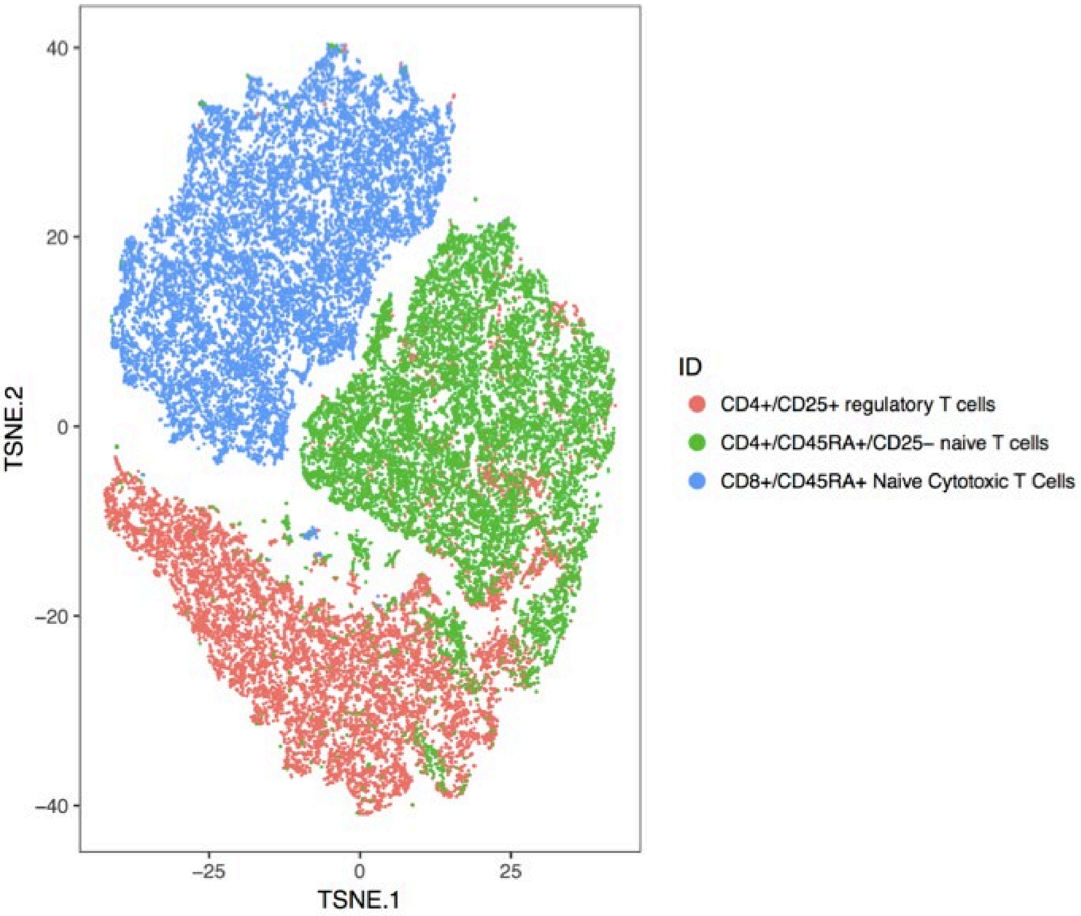


Figure 20. The t-SNE projection of the challenging case

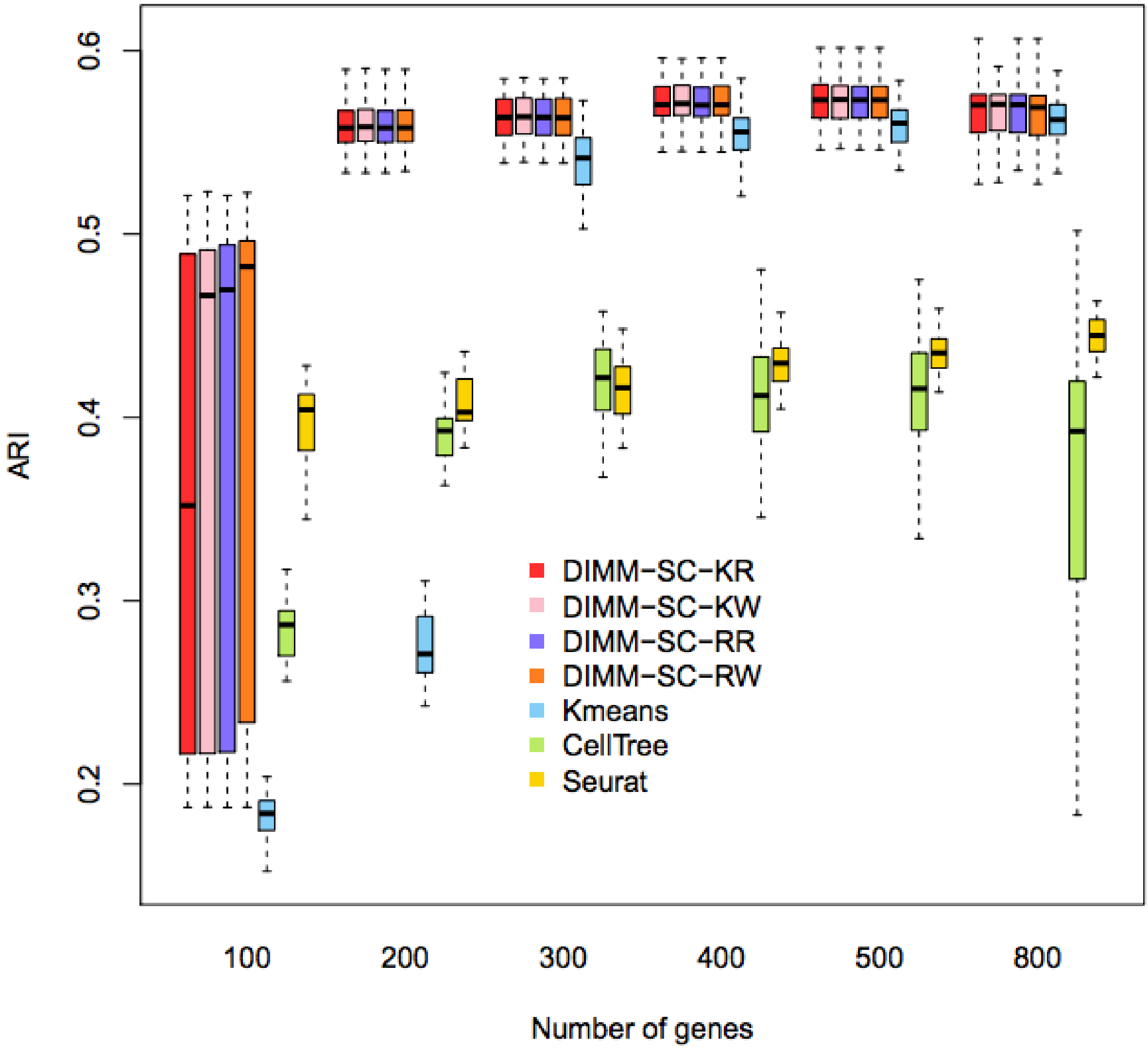


Figure 21. The boxplots of ARI for seven clustering methods across 50 simulations in the challenging case

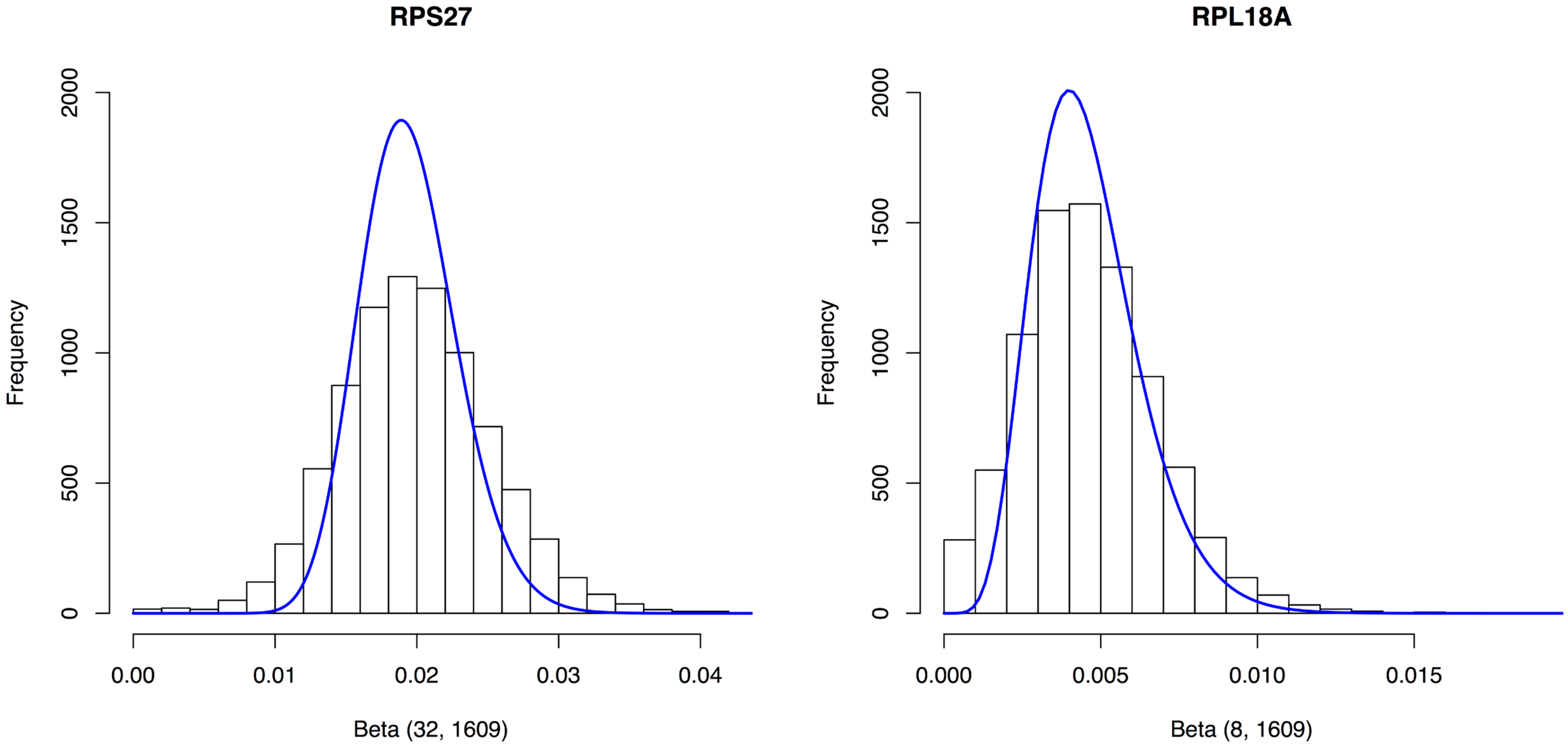


Figure 22. The histogram of proportion p_i for gene *RPS27* and gene *RPL18A* and the theoretical marginal beta distribution (solid blue line) in CD56+ NK cells

To obtain the theoretical marginal beta distribution, we used the top 1% (327) genes and calculate α_i of the Dirichlet distribution by the Ronning's method (Ronning, 1989).

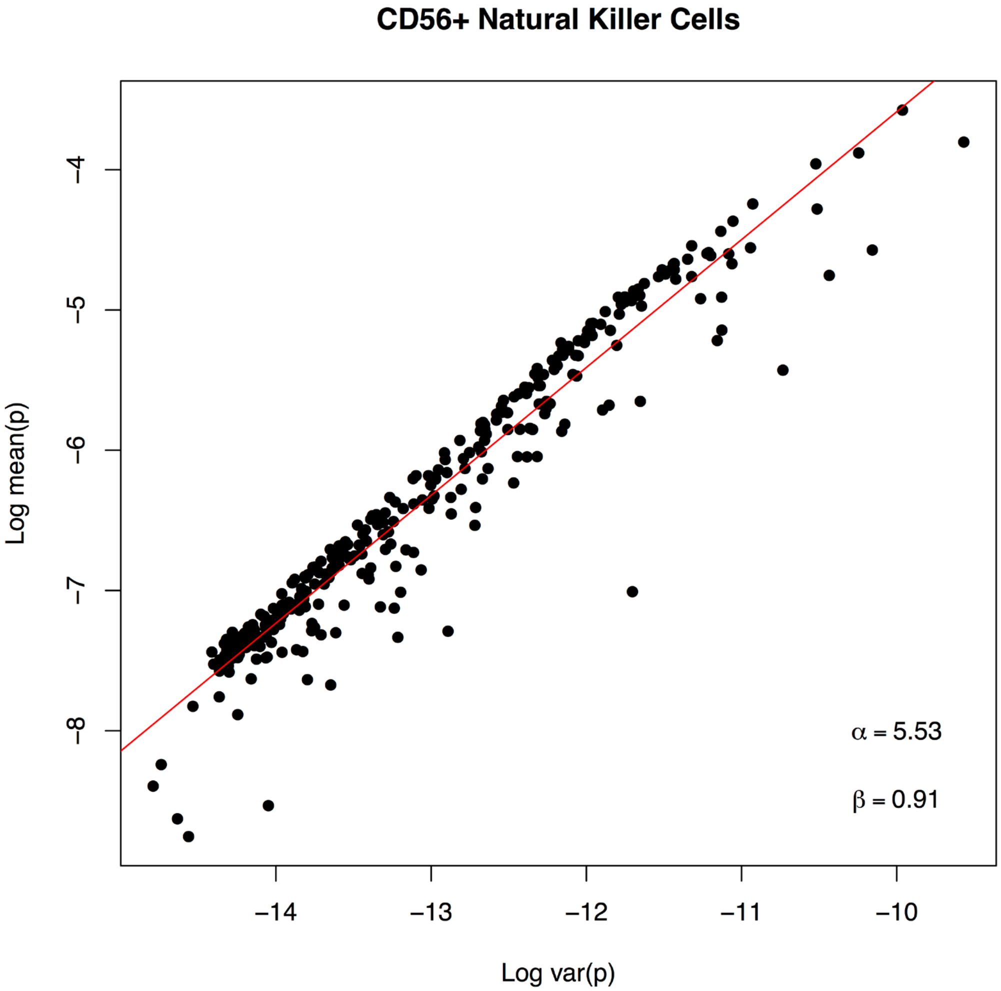


Figure 23. The scatter plot of the log mean of p_i versus the log variance of p_i in the CD56+ NK cells

α and β are linear regression intercept and slope, respectively. Each dot represents one gene. This figure includes the top 1% (327) highly variable genes.

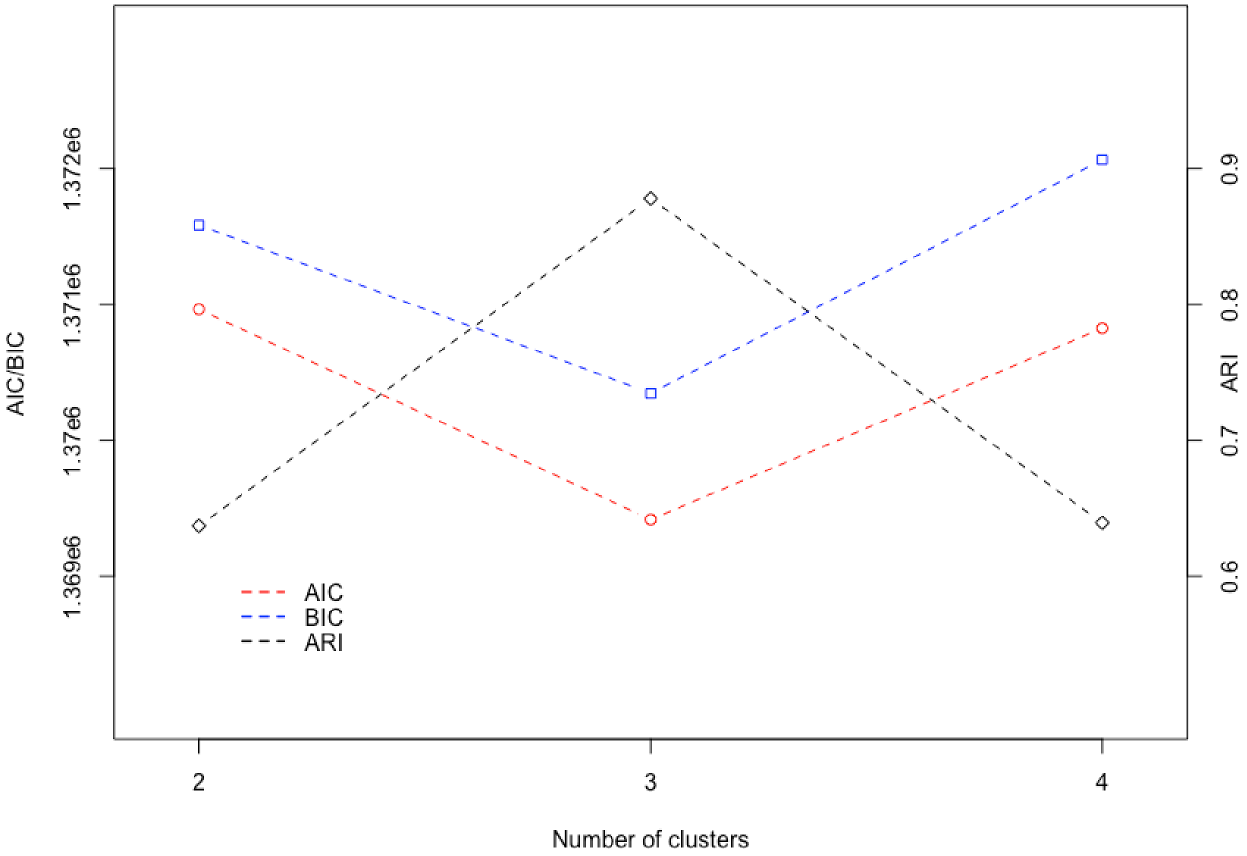


Figure 24. The performance of AIC/BIC criteria when selecting number of clusters

The dot plots of AIC and BIC for the final clustering results in the simulated dataset, where the true number of clusters is 3. Blue dots and red dots denote values of BIC and AIC, respectively. Black dots denote ARIs.

Appendix B (For BAMM-SC)

B.1 Details of Gibbs Sample

Based on Bayes formula, I have the full posterior distribution as follows:

$$P\big(\boldsymbol{z}_{..},\boldsymbol{\alpha}_{..(\cdot)}|\boldsymbol{x}_{...}\big)\propto P\big(\boldsymbol{x}_{...},\boldsymbol{z}_{..}|\boldsymbol{\alpha}_{..(\cdot)}\big)\times\prod_{k=1}^{K}\prod_{i=1}^{G}Prior(\boldsymbol{\alpha}_{\boldsymbol{i}\cdot\boldsymbol{k}})\times\prod_{k=1}^{K}Prior(\boldsymbol{\mu}_{\cdot\boldsymbol{k}})\times\prod_{k=1}^{K}Prior\big(\boldsymbol{\sigma}_{\cdot\boldsymbol{k}}^{\mathbf{2}}\big).$$

The complete log likelihood is:

$$\log P\big(\boldsymbol{z}_{..},\boldsymbol{\alpha}_{..(\cdot)}|\boldsymbol{x}_{...}\big)=\sum_{l=1}^{L}\sum_{j=1}^{C_l}\sum_{k=1}^{K}I\big(z_{jl}=k\big)*\log\left\{\left(\prod_{i=1}^{G}\frac{\Gamma(x_{ijl}+\alpha_{il(k)})}{\Gamma(\alpha_{il(k)})}\right)\frac{\Gamma(|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}(\boldsymbol{k})}|)}{\Gamma(T_{jl}+|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}(\boldsymbol{k})}|)}\right\}$$

$$+\sum_{k=1}^{K}\sum_{i=1}^{G}\sum_{l=1}^{L}\left\{-\log\alpha_{ilk}-\frac{(\log\alpha_{ilk}-\mu_{ik})^2}{2\sigma_{ik}^2}\right\}+\sum_{k=1}^{K}\sum_{i=1}^{G}\left\{-\frac{L}{2}\log\sigma_{ik}^2\right\}$$

$$+NonInformativePrior(\mu_{..})+\sum_{k=1}^{K}\sum_{i=1}^{G}logGammaPDF(\sigma_{ik}^2,a_k,b_k).$$

Here the hyper-prior parameters a_k and b_k can be pre-specified, or estimated from data via an empirical Bayes approach. I use Gibbs sample to iteratively update $\{z_{jl}\}_{1\le j\le C_l,1\le l\le L}$, $\{\alpha_{il(k)}\}_{1\le i\le G,1\le l\le L,1\le k\le K}$. For a given pair of l and j, the conditional distribution for z_{jl} is a multinomial distribution, where

$$P\big(z_{jl}=k\big)=\frac{1}{Constant}*\left(\prod_{i=1}^{G}\frac{\Gamma\big(x_{ijl}+\alpha_{il(k)}\big)}{\Gamma\big(\alpha_{il(k)}\big)}\right)\frac{\Gamma\big(|\alpha_{\cdot l(k)}|\big)}{\Gamma\big(T_{jl}+|\alpha_{\cdot l(k)}|\big)}.$$

Where the normalization constant is:

$$Constant=\sum_{k=1}^{K}\left(\prod_{i=1}^{G}\frac{\Gamma(x_{ijl}+\alpha_{il(k)})}{\Gamma(\alpha_{il(k)})}\right)\frac{\Gamma\big(|\alpha_{\cdot l(k)}|\big)}{\Gamma\big(T_{jl}+|\alpha_{\cdot l(k)}|\big)}.$$

I use random walk Metropolis within Gibbs to iteratively update $\alpha_{il(k)}$. For a given triple of i, l and k, the conditional log likelihood for $\alpha_{il(k)}$ is:

$$log\,P\left(\alpha_{il(k)}\middle|x_{\dots},z_{..}\right) \propto \sum_{j=1}^{C_l} I\left(z_{jl}=k\right) * log\left\{\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ijl}+\alpha_{il(k)}\right)}{\Gamma\left(\alpha_{il(k)}\right)}\right)\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}(\boldsymbol{k})}\right|\right)}{\Gamma\left(T_{jl}+\left|\boldsymbol{\alpha}_{\cdot\boldsymbol{l}(\boldsymbol{k})}\right|\right)}\right\}$$

$$-\,log\,\alpha_{ilk}-\frac{(log\,\alpha_{ilk}-\mu_{ik})^2}{2\sigma_{ik}^2}.$$

Similarly, I use random walk Metropolis within Gibbs to iteratively update σ_{ik}^2. For a given pair of i and k, the conditional log likelihood for σ_{ik}^2 is:

$$log\,P(\sigma_{ik}^2|\dots) \propto \sum_{l=1}^{L}\left\{-\frac{(log\,\alpha_{ilk}-\mu_{ik})^2}{2\sigma_{ik}^2}\right\}-\frac{L}{2}log\,\sigma_{ik}^2+logGammaPDF(\sigma_{ik}^2,a_k,b_k)$$

In random walk Metropolis algorithm, I adaptively select the step size of proposal distribution, to make sure that the acceptance rate is 20% ~ 30%.

B.2 Classification and Computational Acceleration

To further improve the computational efficiency, I provide a supervised option in BAMM-SC. Specifically, for very large-scale dataset, I use BAMM-SC to train a prediction model using a subset of cells from each individual and predict the clustering labels for the rest of cells. First, I randomly select a subset of cells from each individual and applied BAMM-SC on these selected cells. The estimate of α_{ilk} is computed as the average after deletion of the first 100 (default) iterations as burn-in. I then predict the cell type labels for other cells with realization of parameters: $\hat{\Theta} = (\hat{\boldsymbol{\alpha}}_{\cdot\boldsymbol{1}\cdot}, \dots, \hat{\boldsymbol{\alpha}}_{\cdot\boldsymbol{L}\cdot}, \hat{\boldsymbol{\pi}}_{\boldsymbol{1}}, \dots, \hat{\boldsymbol{\pi}}_{\boldsymbol{L}})$.

$$P\left(z_{jl} = k \middle| \boldsymbol{x_{jl}}, \hat{\Theta}\right) = \frac{\left(\prod_{i=1}^{G} \frac{\Gamma\left(x_{ijl} + \hat{\alpha}_{ilk}\right)}{\Gamma(\hat{\alpha}_{ilk})}\right) \frac{\Gamma(|\hat{\boldsymbol{\alpha}}_{\cdot\boldsymbol{lk}}|)}{\Gamma\left(T_j + |\hat{\boldsymbol{\alpha}}_{\cdot\boldsymbol{lk}}|\right)} \hat{\pi}_{lk}}{\sum_{k=1}^{K} \left(\prod_{i=1}^{G} \frac{\Gamma\left(x_{ij} + \hat{\alpha}_{ilk}\right)}{\Gamma(\hat{\alpha}_{ilk})}\right) \frac{\Gamma(|\hat{\boldsymbol{\alpha}}_{\cdot\boldsymbol{lk}}|)}{\Gamma\left(T_j + |\hat{\boldsymbol{\alpha}}_{\cdot\boldsymbol{lk}}|\right)} \hat{\pi}_{lk}}$$

This approach can substantially reduce the computational cost for very large-scale datasets while maintaining the accuracy as shown in Figure 35 (Appendix B).

B.3 Single Cell Sequencing Library Construction

10X Genomics Chromium system, which is a microfluidics platform based on Gel bead in EMulsion (GEM) technology, was used for generating real test datasets (Zheng et al., 2017). Cells mixed with reverse transcription reagents were loaded into the Chromium instrument. This instrument separated cells into mini-reaction "partitions" formed by oil micro-droplets, each containing a gel bead and a cell, known as GEMs. GEMs contain a gel bead, scaffold for an oligonucleotide that is composed of an oligo-dT section for priming reverse transcription, and barcodes for each cell and each transcript as described. GEM generation takes place in a multiple-channel microfluidic chip that encapsulates single gel beads. Reverse transcription takes place inside each droplet. Approximately 1,000-fold excess of partitions compared to cells ensured low capture of duplicate cells. The reaction mixture/emulsion was removed from the Chromium instrument, and reverse transcription was performed. The emulsion was then broken using a recovery agent, and following Dynabead and SPRI clean up cDNAs were amplified by PCR (C1000, Bio-Rad). cDNAs were sheared (Covaris) into ~200 bp length. DNA fragment ends were repaired, A-tailed and adaptors ligated. The library was quantified using KAPA Universal Library Quantification Kit KK4824 and further characterized for cDNA length on a Bioanalyzer using a High Sensitivity DNA kit. All sequencing experiments were conducted using Illumina NextSeq 500 in the Genomics Sequencing Core at the University of Pittsburgh.

Data description

Human PBMC dataset: Peripheral blood was obtained from healthy donors by venipuncture. Peripheral blood mononuclear cells (PBMC) were isolated from whole blood by density gradient centrifugation using Ficoll-Hypaque. PBMC were then counted and re-suspended in phosphate buffered saline with 0.04% bovinue serum albumin, and were processed through the Chromium 10X Controller according to the manufacturers' instructions, targeting a recovery of ~2,000 cells. The following steps were all performed under the aforementioned protocol developed by 10X Genomics.

Human skin dataset: Skin samples were obtained by performing 3 mm punch biopsies from the dorsal mid-forearm of healthy control subjects after informed consent under a protocol approved by the University of Pittsburgh Institutional Review Board. Skin for scRNA-seq was digested enzymatically (Miltenyi Biotec Whole Skin Dissociation Kit, human) for 2 hours and further dispersed using the Miltenyi gentleMACS Octo Dissociator. The resulting cell suspension was filtered through 70 micron cell strainers twice and re-suspended in PBS containing 0.04% BSA. Cells from biopsies were mixed with reverse transcription reagents then loaded into the Chromium instrument (10X Genomics). ~2,600-4,300 cells were loaded into the instrument to obtain data on ~1,100-1,800 cells, anticipating a multiplet rate of ~1.2% of partitions. The following steps were all performed under the aforementioned protocol developed by 10X Genomics.

Mouse lung dataset: Lung single cell suspension from naïve and infected C57BL/6 mice were subject to scRNA-seq library preparation protocol. Briefly, left lobs of both naïve and infected mice were removed and digested by Collagenase/DNase to obtain single cell suspension. Mononuclear cells after filtration with a 40□M cell strainer were separated into mini-reaction "partitions" or GEMs formed by oil micro-droplets, each containing a gel bead and a cell, by the Chromium instrument (10X Genomics). The reaction mixture/emulsion with captured and barcoded mRNAs were removed from the Chromium instrument followed by reverse transcription. The cDNA samples were fragmented and amplified using the Nextera XT kit (Illumina). The following steps were all performed under aforementioned the protocol developed by 10X Genomics.

B.4 Additional Figures

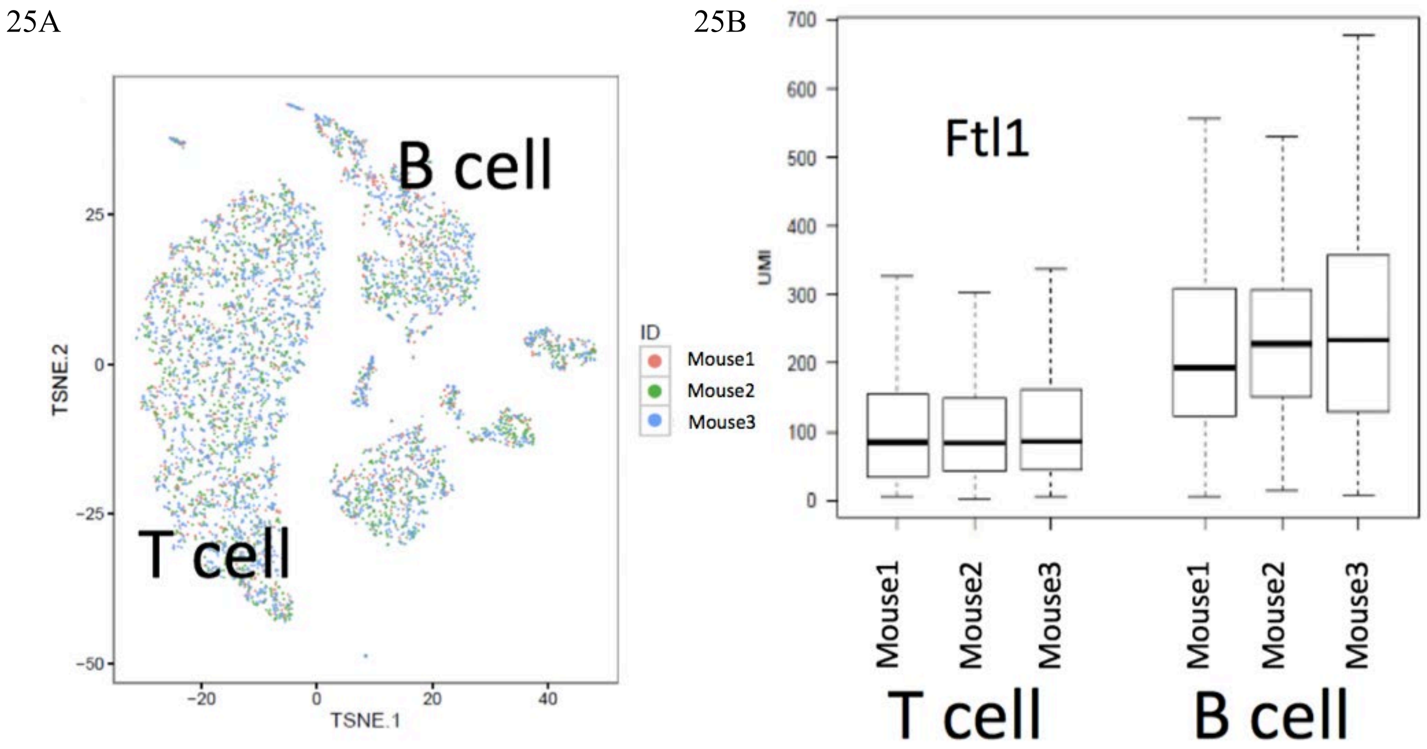


Figure 25. The t-SNE projection of cells from 3 mouse samples (colored by different sample labels) (A) and the boxplot of UMI counts for gene *Ftl1* in T cells and B cells, separately (B)

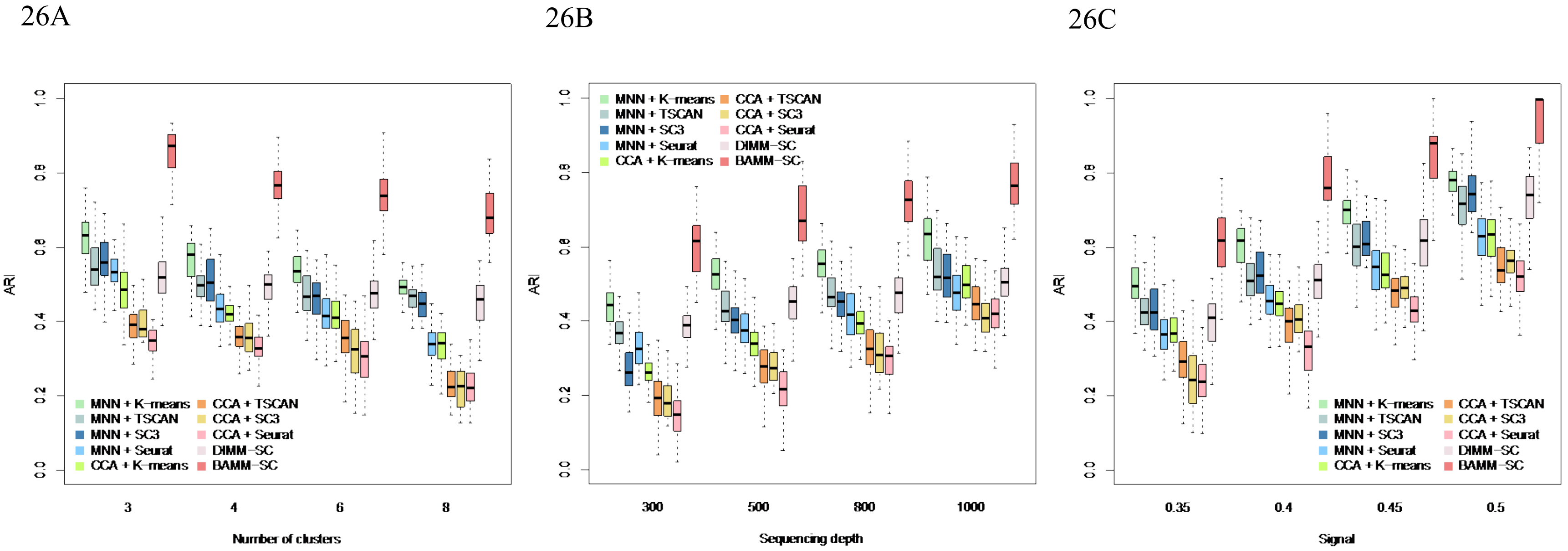


Figure 26. The Boxplots of ARI for ten clustering methods across 100 simulations, investigating how different number of clusters (A), sequencing depth (B) and cell-type-specific heterogeneities (C) affect clustering results

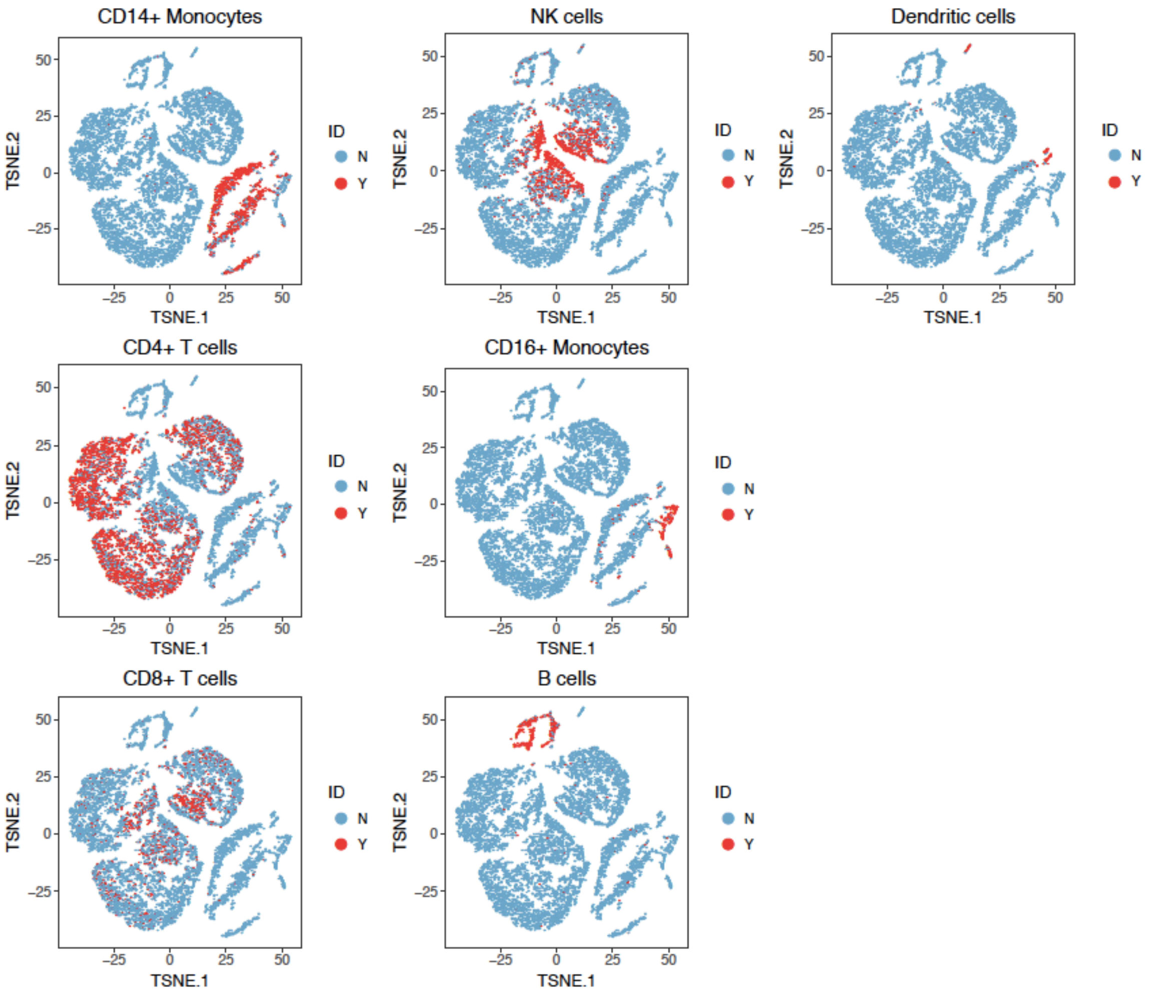


Figure 27. The t-SNE projection of PBMCs from 5 human samples illustrating different cell subtypes, with each cell is colored by their classification based on specific gene markers

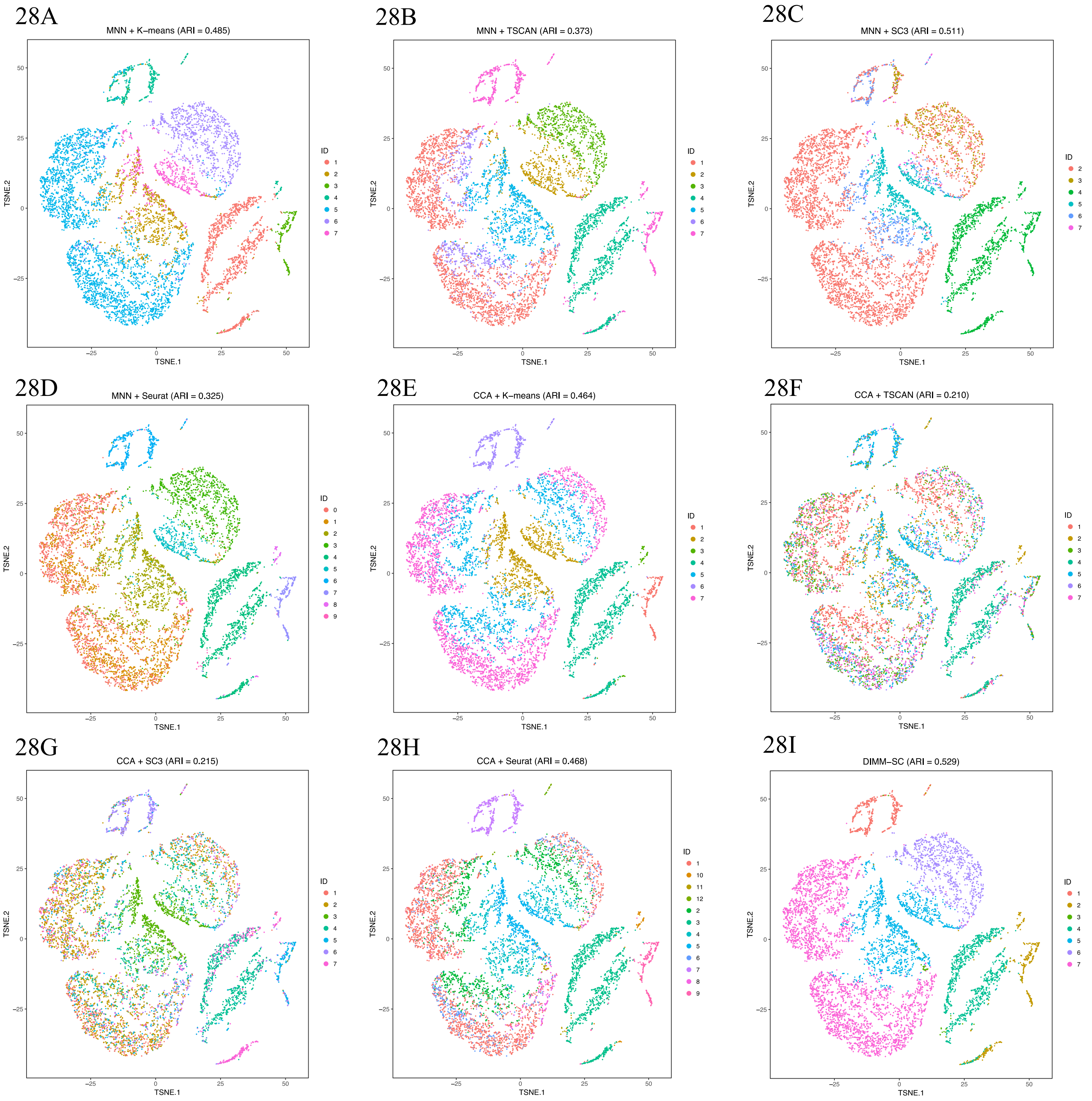


Figure 28. The t-SNE projection of cells from PBMC dataset, colored by different clustering assignments

The t-SNE projection of cells from human PBMC dataset, colored by the MNN + K-means clustering (A), MNN + TSCAN (B), MNN + SC3 (C), MNN + Seurat (D), CCA + K-means (E), CCA + TSCAN (F), CCA + SC3 (G), CCA + Seurat (H) and DIMM-SC (I) clustering assignment. All clustering labels are from the result with the highest ARI among 10 times analysis.

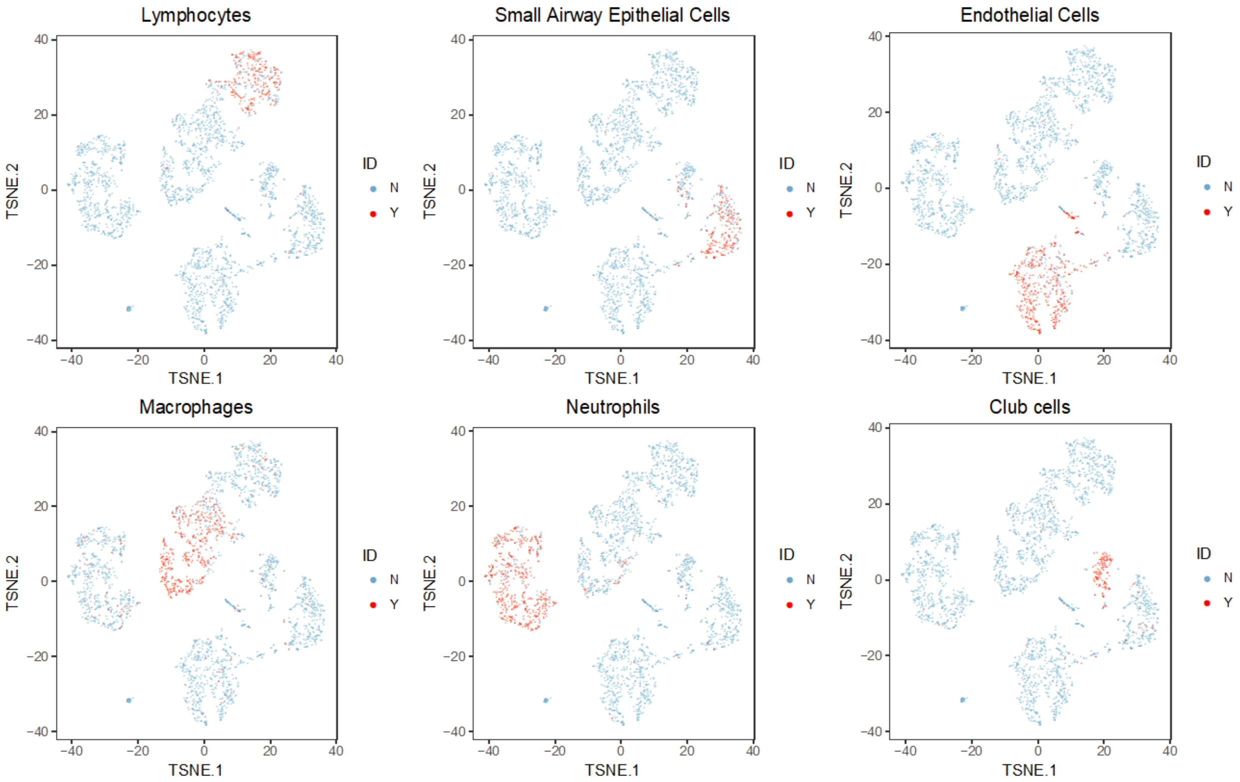


Figure 29. The t-SNE projection of lung mononuclear cells from 4 mouse samples illustrating different cell subtypes, with each cell is colored by their classification based on specific gene markers

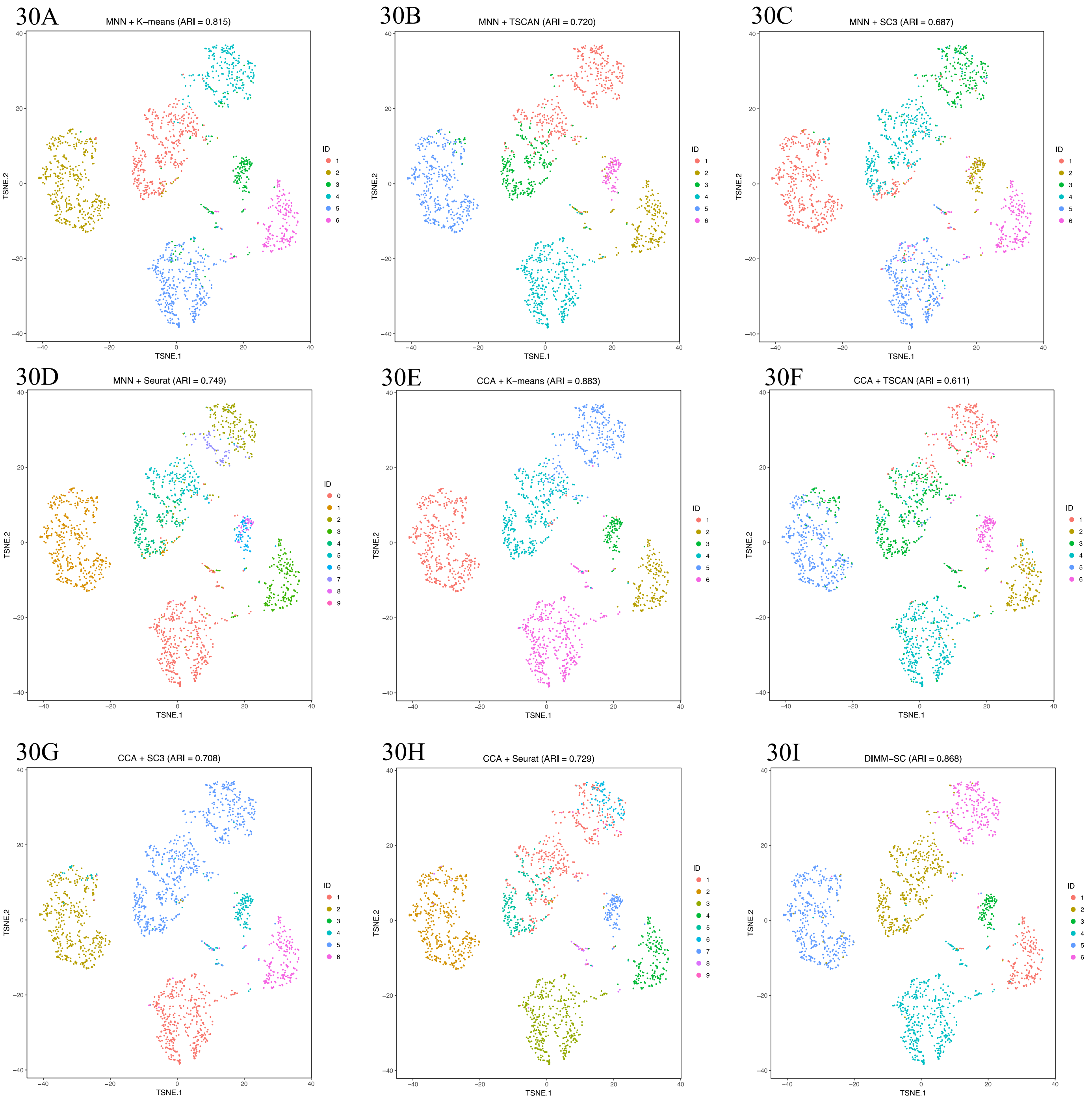


Figure 30. The t-SNE projection of cells from mouse lung dataset, colored by different clustering assignments

The t-SNE projection of cells from mouse lung dataset, colored by the MNN + K-means clustering (A), MNN + TSCAN (B), MNN + SC3 (C), MNN + Seurat (D), CCA + K-means (E), CCA + TSCAN (F), CCA + SC3 (G), CCA + Seurat (H) and DIMM-SC (I) clustering assignment. All clustering labels are from the result with the highest ARI among 10 times analysis.

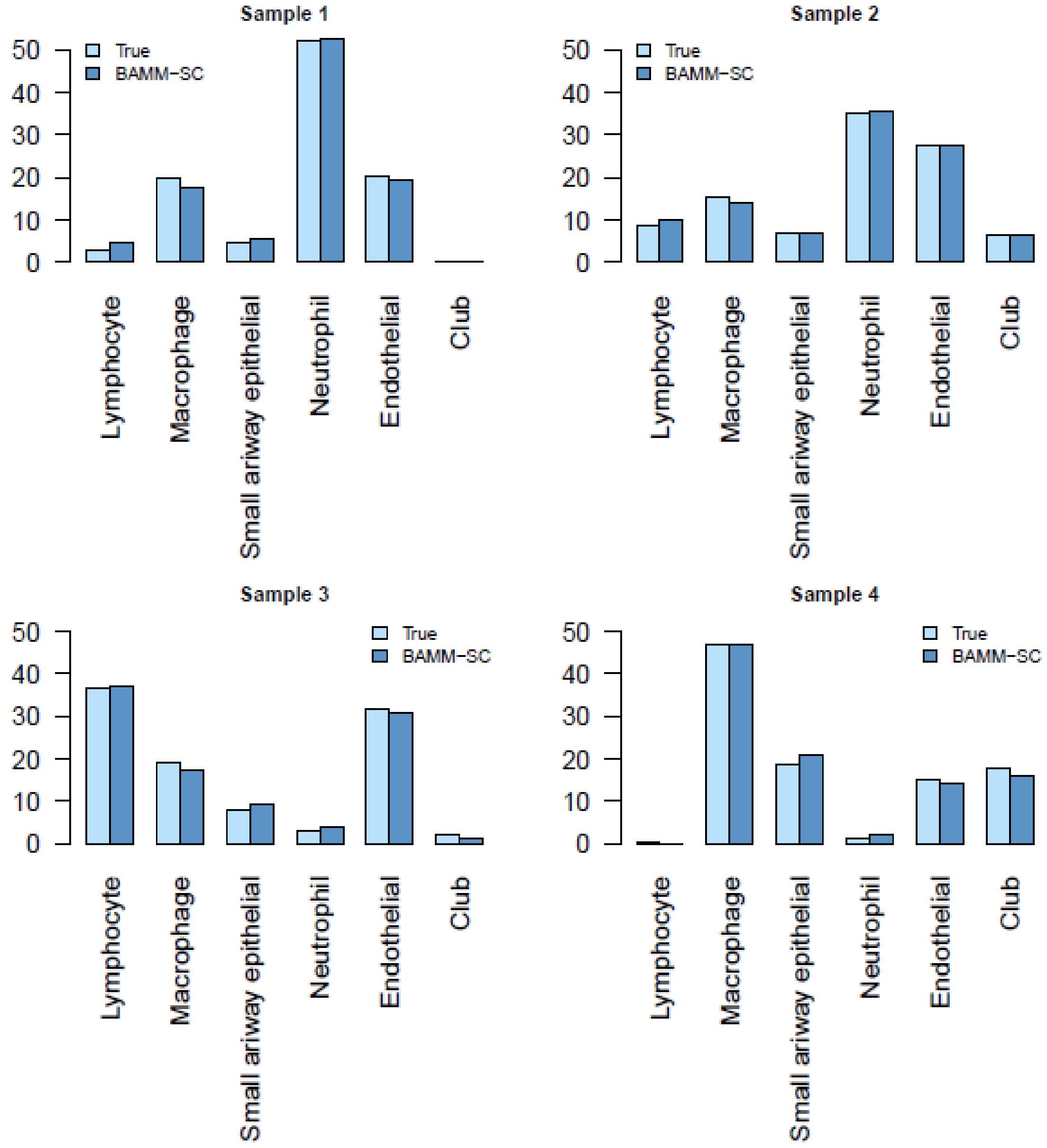


Figure 31. Bar plots of proportions of cell types for each individual in mouse lung dataset

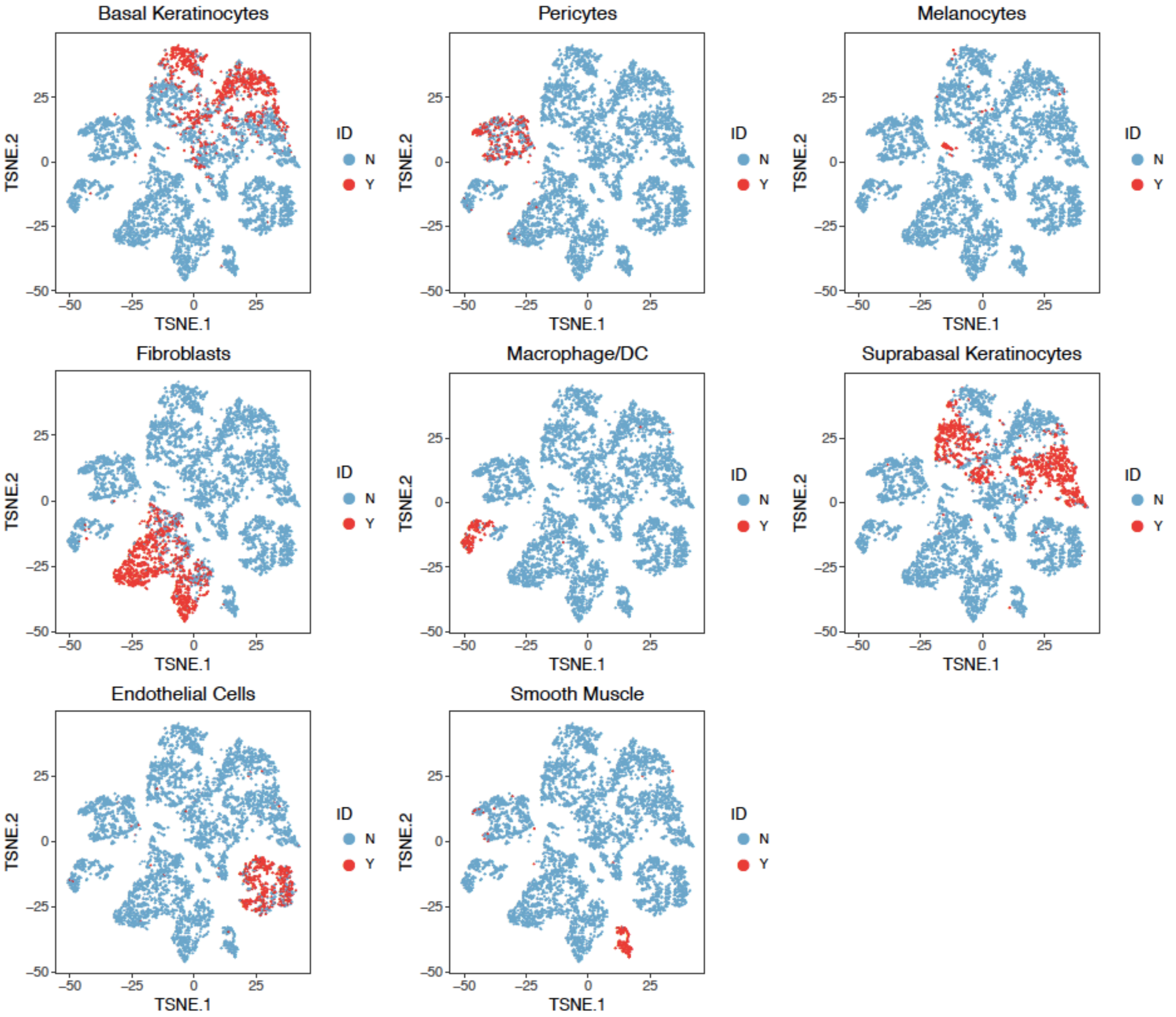
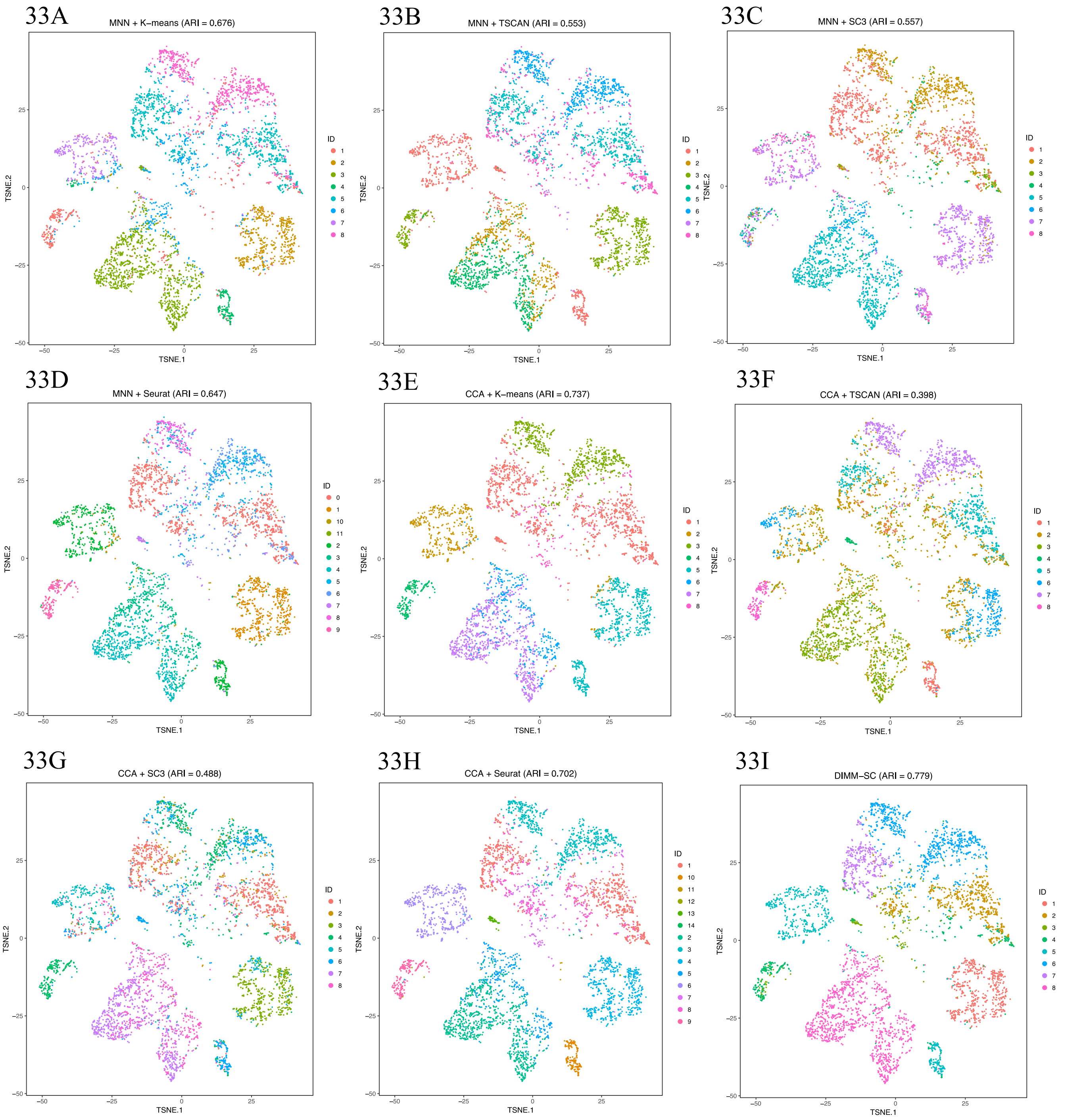


Figure 32. The t-SNE projection of cells from human skin dataset, colored by different types of PBMCs based on the biological knowledge of cell-type-specific gene markers



33A MNN + K–means (ARI = 0.676)

33B MNN + TSCAN (ARI = 0.553)

33C MNN + SC3 (ARI = 0.557)

33D MNN + Seurat (ARI = 0.647)

33E CCA + K–means (ARI = 0.737)

33F CCA + TSCAN (ARI = 0.398)

33G CCA + SC3 (ARI = 0.488)

33H CCA + Seurat (ARI = 0.702)

33I DIMM–SC (ARI = 0.779)

Figure 33. The t-SNE projection of cells from human skin dataset, colored by different clustering assignments

The t-SNE projection of cells from human skin dataset, colored by the MNN + K-means clustering (A), MNN + TSCAN (B), MNN + SC3 (C), MNN + Seurat (D), CCA + K-means (E), CCA + TSCAN (F), CCA + SC3 (G), CCA + Seurat (H) and DIMM-SC (I) clustering assignment. All clustering labels are from the result with the highest ARI among 10 times analysis.

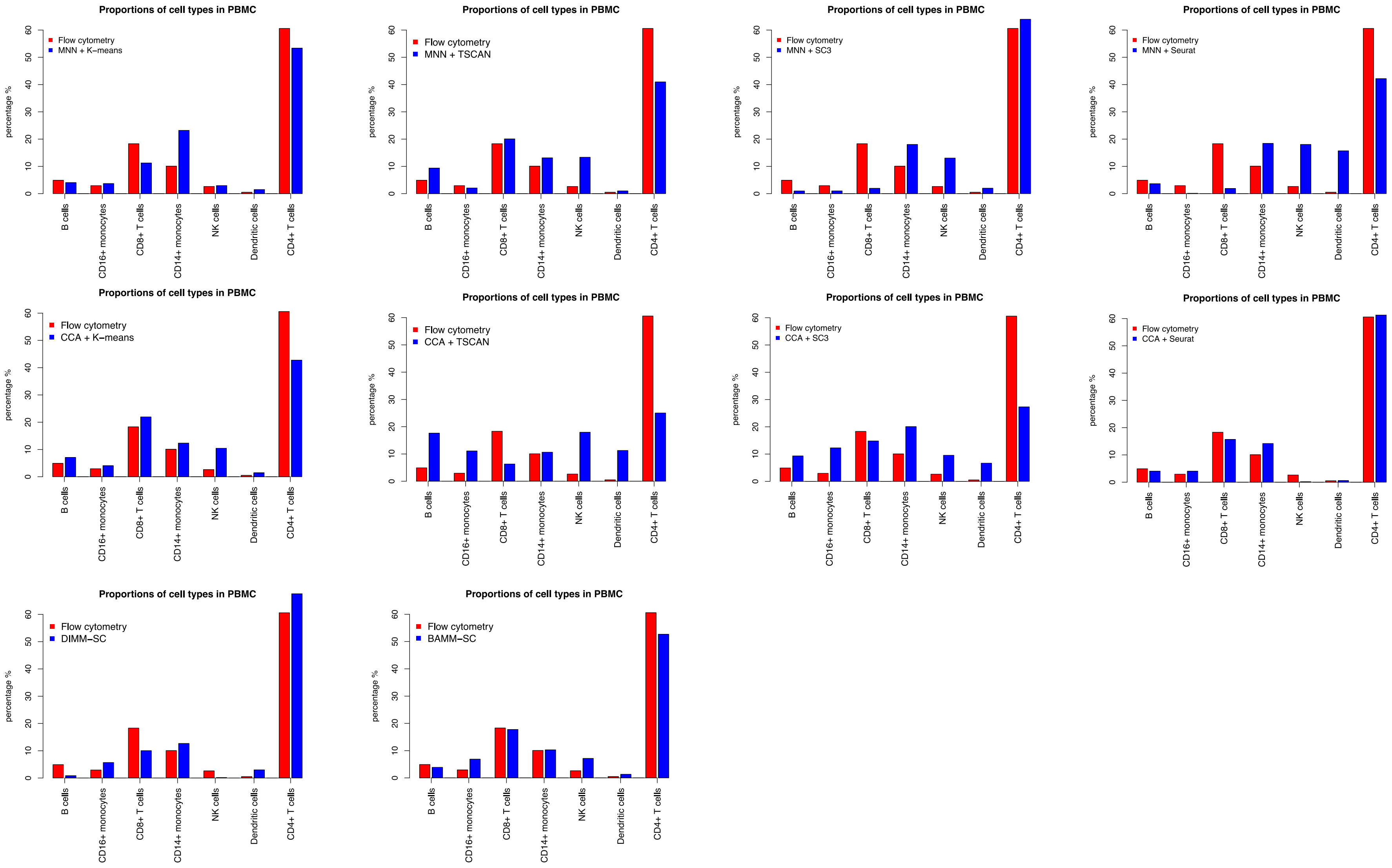


Figure 34. Bar plot of cell proportions from flow cytometry and different clustering methods in individual 3 from the human PBMC dataset

All clustering assignments are from the result with the highest ARI among 10 times analysis.

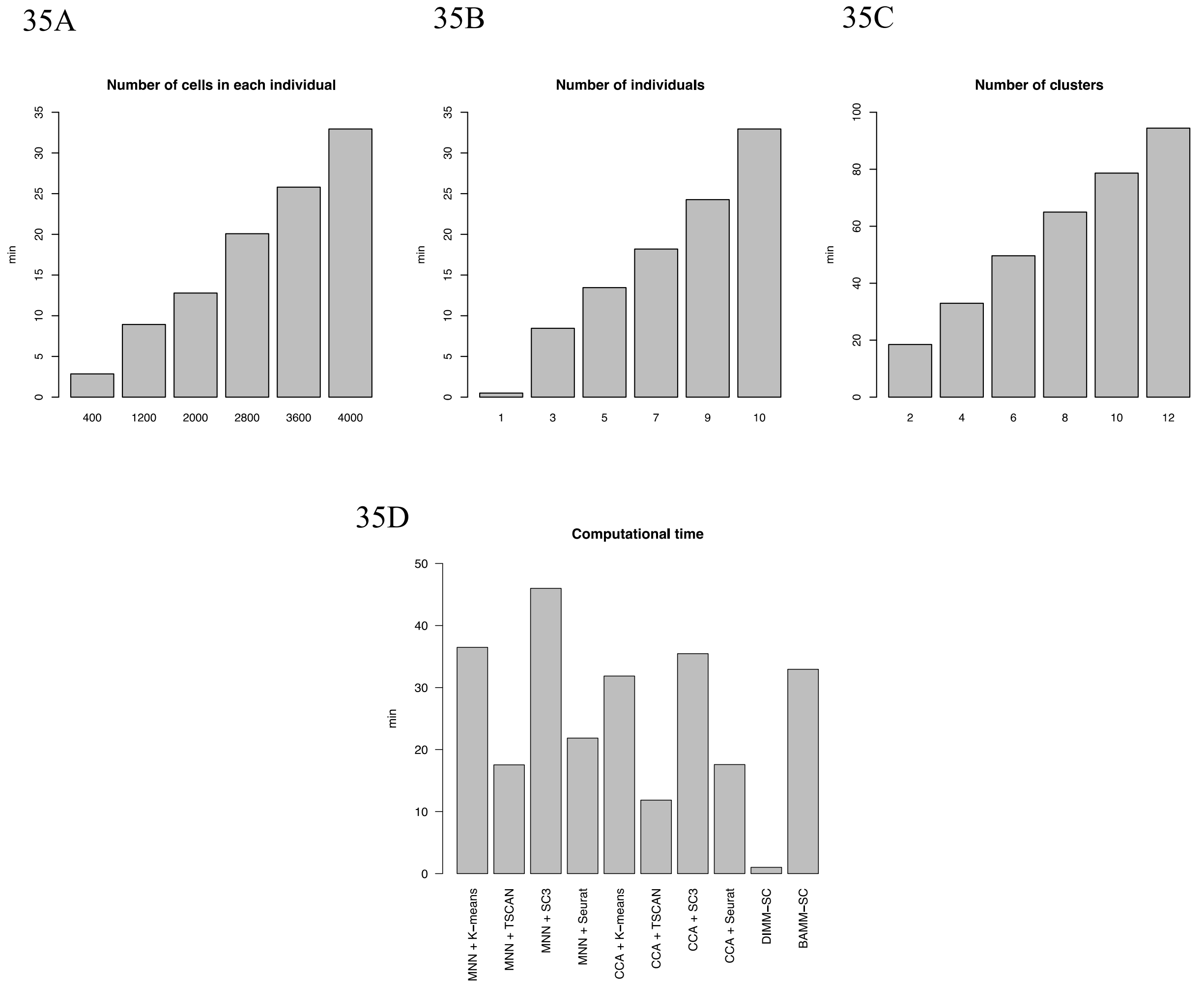


Figure 35. The bar plot of computational time for BAMM-SC in simulated dataset

The bar plot of computational time for BAMM-SC in simulated dataset with different number of cells in each individual (A), different number of individuals (B), different number of clusters (C), and the bar plot of computational time of different clustering methods in simulated dataset (D). In (D), we set the number of single cells in each individual as 4,000, the number of individuals as 10, and the number of clusters as 4, to benchmark the computational cost of different methods. To be noted, K-means clustering itself is very fast, the process of batch effect correction and calculating dimension reduction representations takes most of the computational time.

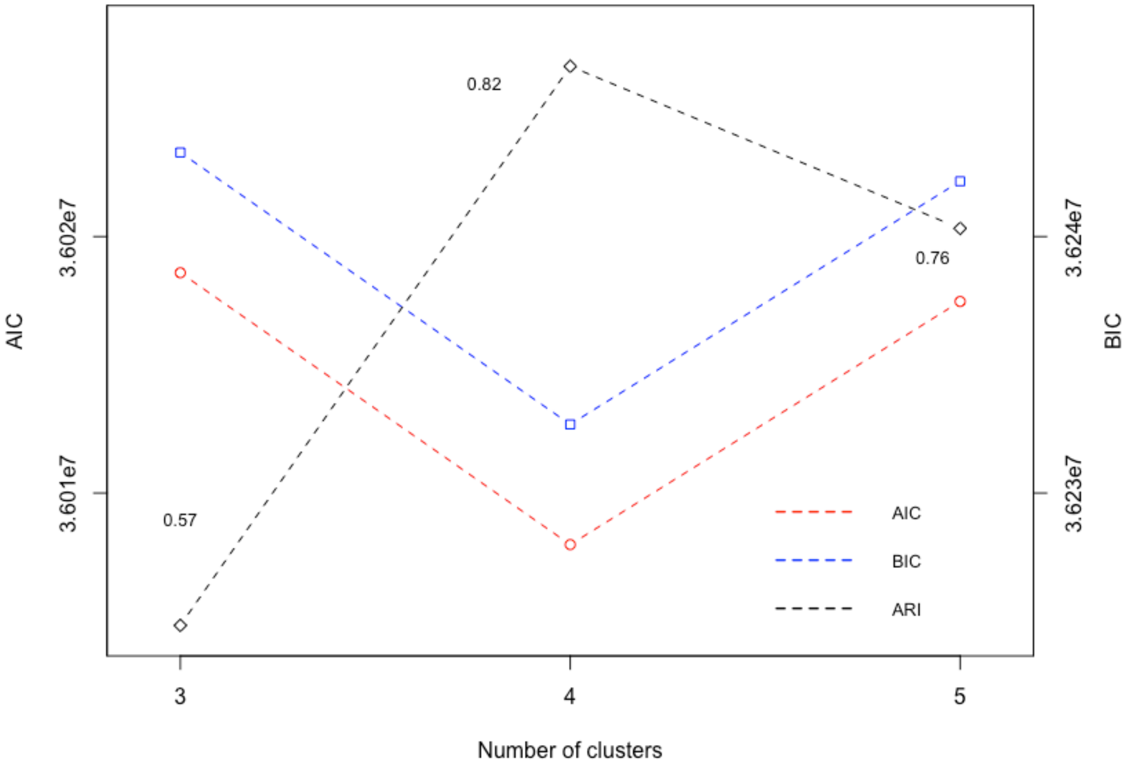


Figure 36. The dot plots of AIC and BIC for the final clustering results in the simulated dataset

The dot plots of AIC and BIC for the final clustering results in the simulated dataset, where the true number of clusters is 4 and the number of individuals is 10. Blue dots and red dots denote values of BIC and AIC, respectively. Black dots denote ARIs.

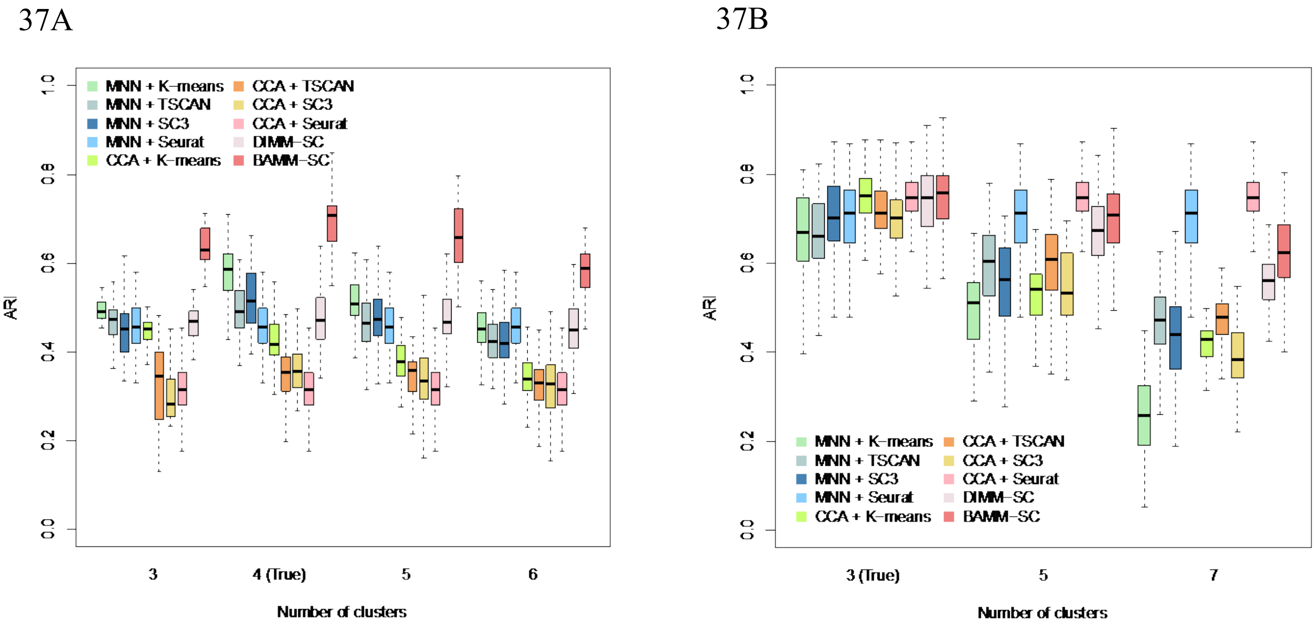


Figure 37. The Boxplots of ARI for ten clustering methods across 100 simulations when number of clusters is mis-specified using simulated data (A) and data generated from Splatter (B)

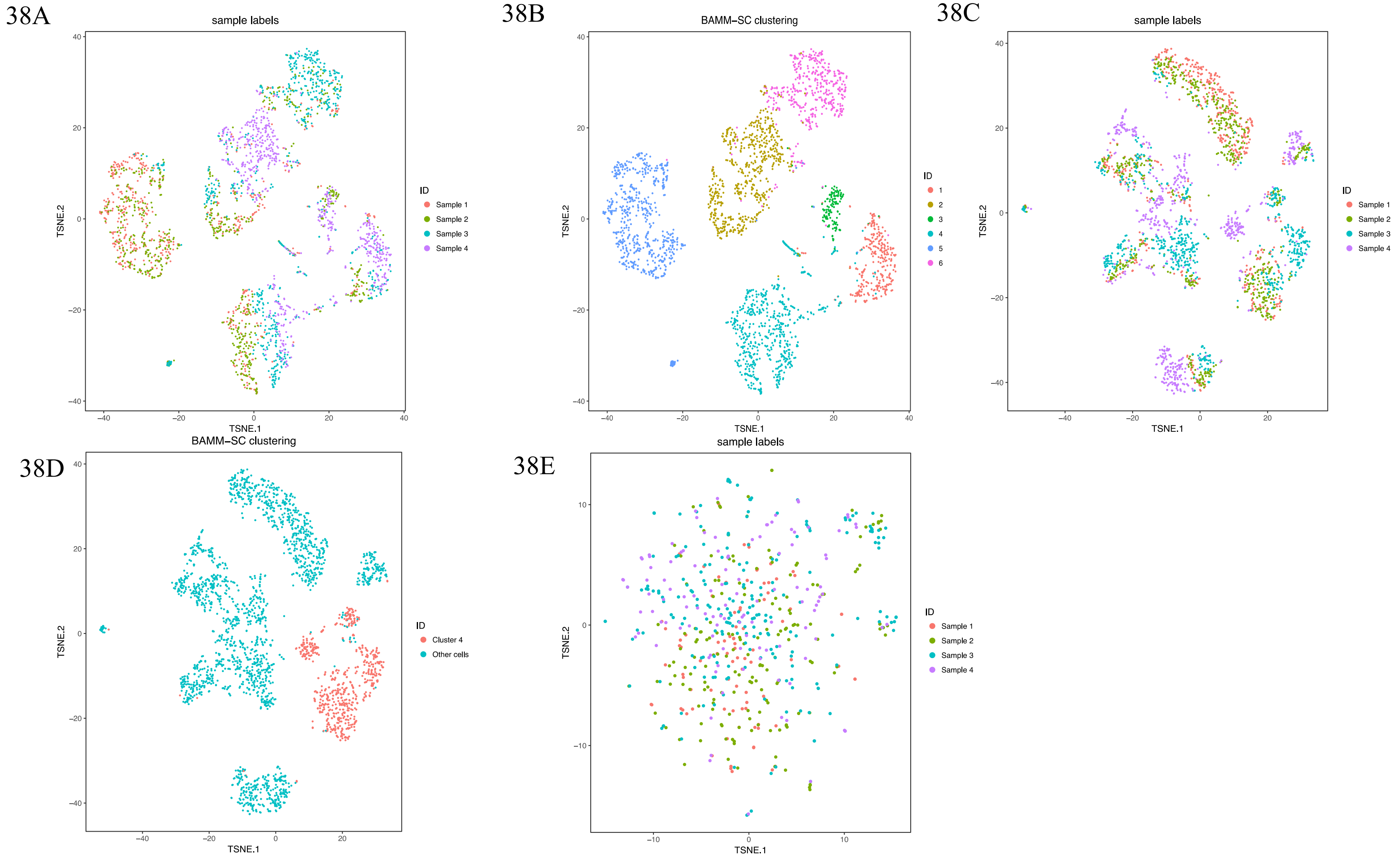


Figure 38. The t-SNE projection of cells from mouse lung dataset

The t-SNE projection of cells from mouse lung dataset without batch effect correction, colored by different sample IDs (A) and BAMM-SC clustering assignment (B), the t-SNE projection of cells after CCA batch effect correction, colored by different sample IDs (C) and the clustering assignment (based on the result of BAMM-SC in (B)) (D), and the t-SNE projection of cells in cluster 4 (based on the result of BAMM-SC in (B)) with CCA correction (E).

Appendix C (for BREM-SC)

C.1 Details of Gibbs Sample

I propose a general Bayesian framework for estimation. I use Gibbs sample to iteratively update z_j, $\alpha_{i(k)}^{(1)}$, $\alpha_{d(k)}^{(2)}$ and b_j. Specifically, I will use random walk Metropolis within Gibbs to iteratively update b_j, $\alpha_{i(k)}^{(1)}$ and $\alpha_{d(k)}^{(2)}$.

For a given cell j, the conditional distribution for z_j is a Multinomial distribution, where

$$P\left(z_j = k\right) = \frac{1}{constant}\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ij}^{(1)} + \alpha_{i(k)}^{(1)}b_j\right)}{\Gamma\left(\alpha_{i(k)}^{(1)}b_j\right)}\right)\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j\right|\right)}{\Gamma\left(T_j^{(1)} + \left|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j\right|\right)}$$

$$\left(\prod_{d=1}^{D}\frac{\Gamma\left(x_{dj}^{(2)} + \alpha_{d(k)}^{(2)}b_j\right)}{\Gamma\left(\alpha_{d(k)}^{(2)}b_j\right)}\right)\frac{\Gamma(|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j|)}{\Gamma(T_j^{(2)} + |\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j|)}\pi_k$$

where the normalization constant is:

$$\sum_{k=1}^{K}\left\{\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ij}^{(1)} + \alpha_{i(k)}^{(1)}b_j\right)}{\Gamma\left(\alpha_{i(k)}^{(1)}b_j\right)}\right)\frac{\Gamma(|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j|)}{\Gamma(T_j^{(1)} + |\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j|)}\left(\prod_{d=1}^{D}\frac{\Gamma\left(x_{dj}^{(2)} + \alpha_{d(k)}^{(2)}b_j\right)}{\Gamma\left(\alpha_{d(k)}^{(2)}b_j\right)}\right)\frac{\Gamma(|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j|)}{\Gamma(T_j^{(2)} + |\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j|)}\pi_k\right\}$$

For a given gene i and cell type k, the conditional log likelihood for $\alpha_{i(k)}^{(1)}$ is

$$logP\left(\alpha_{i(k)}^{(1)}\middle|\ldots\right) \propto \sum_{j=1}^{C} I(z_j = k)\log\{\left(\frac{\Gamma\left(x_{ij}^{(1)}+\alpha_{i(k)}^{(1)}b_j\right)}{\Gamma\left(\alpha_{i(k)}^{(1)}b_j\right)}\right)\frac{\Gamma(|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j|)}{\Gamma(T_j^{(1)}+|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j|)}\}.$$

Similarly, for a given ADT marker d and cell type k, the conditional log likelihood for $\alpha_{d(k)}^{(2)}$ is

$$logP\left(\alpha_{d(k)}^{(2)}\middle|\ldots\right) \propto \sum_{j=1}^{C} I(z_j = k)\log\{\left(\frac{\Gamma\left(x_{dj}^{(2)}+\alpha_{d(k)}^{(2)}b_j\right)}{\Gamma\left(\alpha_{d(k)}^{(2)}b_j\right)}\right)\frac{\Gamma(|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j|)}{\Gamma(T_j^{(2)}+|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j|)}\}.$$

For a given cell j, we can have the conditional log likelihood for b_j:

$$logP\left(b_j\middle|\ldots\right) \propto \sum_{k=1}^{K} I\left(z_j = k\right)\right)\ log$$

$$\left\{\left(\prod_{i=1}^{G}\frac{\Gamma\left(x_{ij}^{(1)}+\alpha_{i(k)}^{(1)}b_j\right)}{\Gamma\left(\alpha_{i(k)}^{(1)}b_j\right)}\right)\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j\right|\right)}{\Gamma\left(T_j^{(1)}+\left|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{1})}b_j\right|\right)}\left(\prod_{d=1}^{D}\frac{\Gamma\left(x_{dj}^{(2)}+\alpha_{d(k)}^{(2)}b_j\right)}{\Gamma\left(\alpha_{d(k)}^{(2)}b_j\right)}\right)\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j\right|\right)}{\Gamma\left(T_j^{(2)}+\left|\boldsymbol{\alpha}_{(\boldsymbol{k})}^{(\mathbf{2})}b_j\right|\right)}\right\}$$

$$-logb_j\ -\ \frac{(logb_j)^2}{2\sigma_b^2}.$$

C.2 Additional Figures

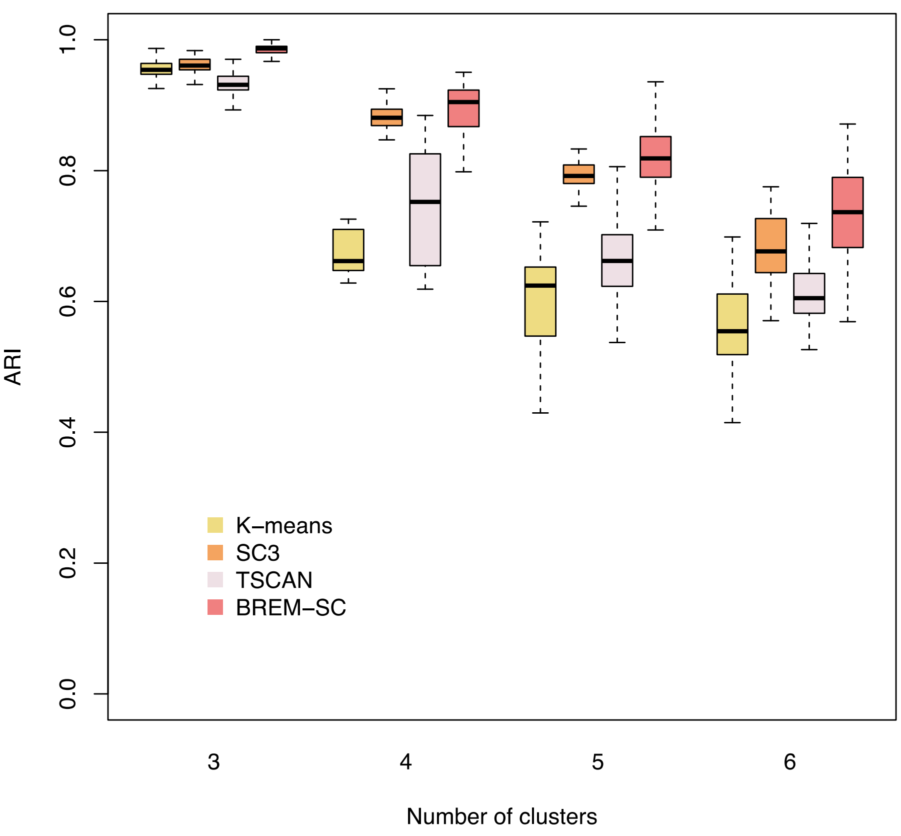


Figure 39. Boxplot of ARI for five clustering methods across 100 simulations, investigating how number of clusters affect clustering results

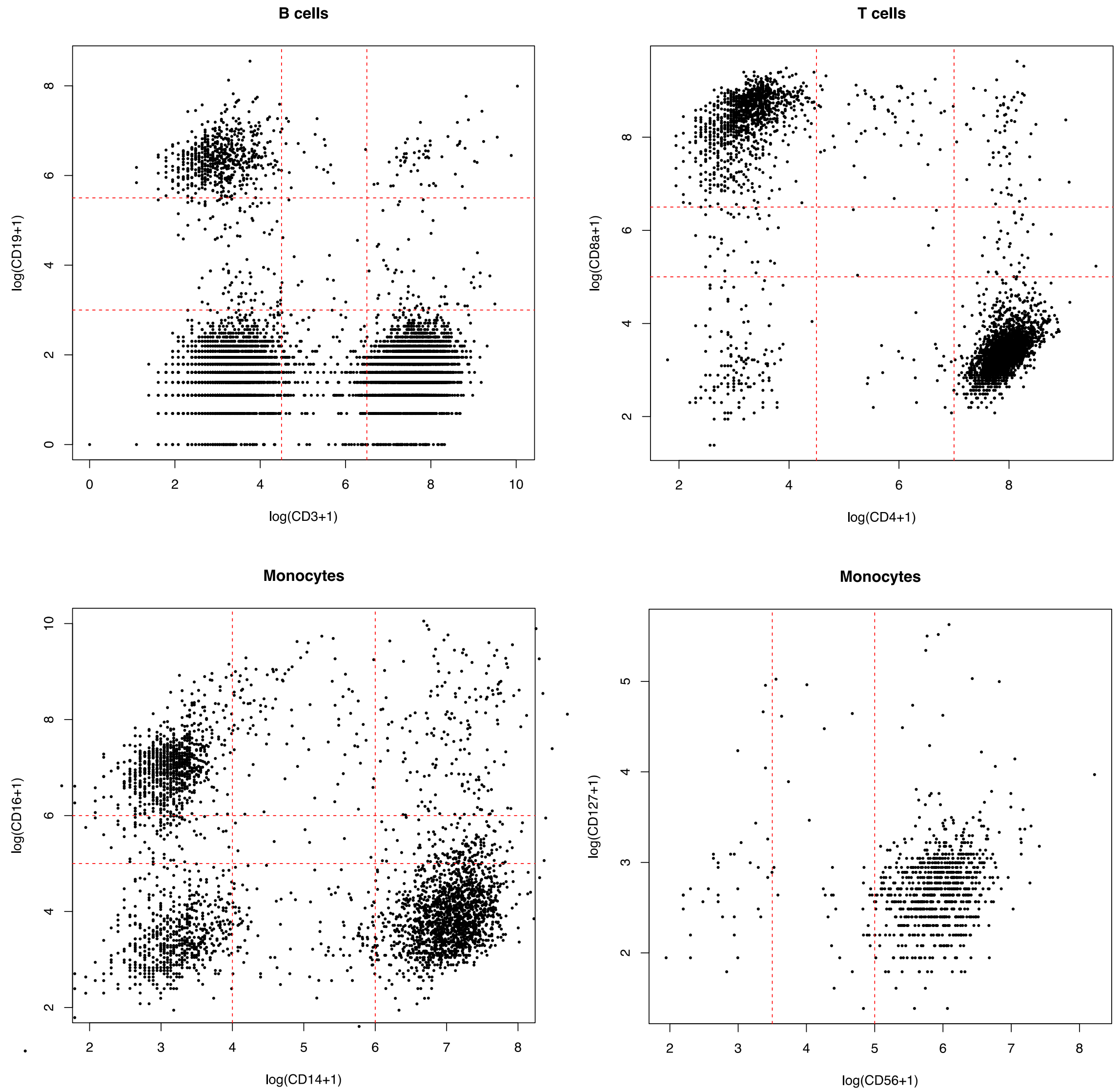


Figure 40. Scatter plot of cells illustrating how to get the approximated truth in 10X human PBMC dataset

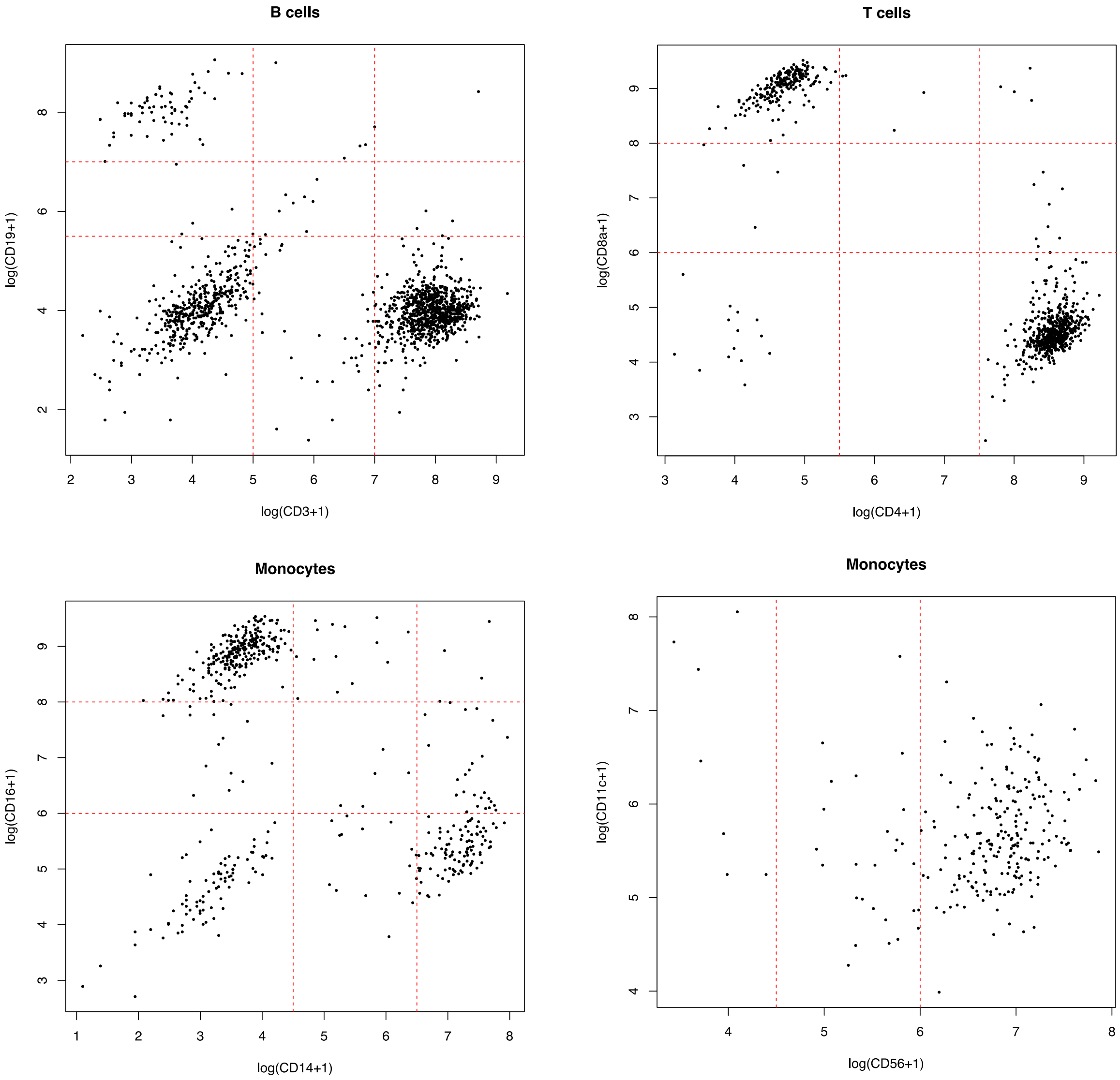


Figure 41. Scatter plot of cells illustrating how to get the approximated truth for in-house human PBMC dataset

Bibliography

Akaike, H. New Look at Statistical-Model Identification. *Ieee T Automat Contr* 1974;Ac19(6):716-723.

Chen, G., *et al.* Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics* 2002;1(4):304-313.

Chen, K. and Kolls, J.K. T cell-mediated host immune defenses in the lung. *Annu Rev Immunol* 2013;31:605-633.

Coifman, R.R., *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci U S A* 2005;102(21):7426-7431.

Crow, M., *et al.* Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat Commun* 2018;9(1):884.

Datta, S., & Datta, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 2003;19(4):459-466.

Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* 1977:1-38.

Duo, A., Robinson, M.D. and Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* 2018;7:1141.

duVerle, D.A., *et al.* CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics* 2016;17(1):363.

Freytag, S., *et al.* Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Res* 2018;7:1297.

Gawad, C., Koh, W. and Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016;17(3):175-188.

Haghverdi, L., *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;36(5):421-427.

Haider, S. and Pal, R. Integrated analysis of transcriptomic and proteomic data. *Curr Genomics* 2013;14(2):91-110.

Holmes, I., Harris, K. and Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 2012;7(2):e30126.

Islam, S., *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;11(2):163-166.

Jaitin, D.A., *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014;343(6172):776-779.

Ji, Z. and Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* 2016;44(13):e117.

Kiselev, V.Y., *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* 2017;14(5):483-486.

Kivioja, T., *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 2012;9(1):72-74.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. & Yosef, N. . Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing. *bioRxiv* 2018;292037.

Macosko, E.Z., *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015;161(5):1202-1214.

Minka, T. Estimating a Dirichlet distribution. http://www.msr-waypoint.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf 2000.

Pollen, A.A., *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* 2014;32(10):1053-1058.

Rand, W.M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 1971;66(336):846-850.

Rodriguez, A. and Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* 2014;344(6191):1492-1496.

Ronning, G. Maximum-likelihood estimation of dirichlet distributions. *Journal of Statistical Computation and Simulation* 1989;32:215-221.

Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 1987;20:53-65.

Satija, R., Farrell, J.A. and Gennert, D. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* 2015;33:495-502.

Satija, R., *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;33(5):495-502.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics* 1978;6(2):461-464.

Shalek, A.K., *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 2014;510(7505):363-369.

Silbereis, J.C., *et al.* The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron* 2016;89(2):248-268.

Spencer, S.L., *et al.* Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* 2009;459(7245):428-432.

Spitzer, M.H., *et al.* IMMUNOLOGY. An interactive reference framework for modeling a dynamic immune system. *Science* 2015;349(6244):1259425.

Stegle, O., Teichmann, S.A. and Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;16(3):133-145.

Stoeckius, M., *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;14(9):865-868.

Stoeckius, M., *et al.* Simultaneous epitope and transcriptome measurement in single cells. 2017;14(9):865.

Stoeckius, M., *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* 2018;19(1):224.

Sun, Z., *et al.* DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* 2018;34(1):139-146.

Tabib, T., *et al.* SFRP2/DPP4 and FMO1/LSP1 Define Major Fibroblast Populations in Human Skin. *J Invest Dermatol* 2018;138(4):802-810.

Tang, F., *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* 2009;6(5):377-382.

Teh, Y.W. Dirichlet process. *Encyclopedia of machine learning* 2011:Springer US, 280-287.

Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* 2015;25(10):1491-1498.

Trapnell, C., *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;32(4):381-386.

van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *J Mach Learn Res* 2008;9:2579-2605.

Wang, B., *et al.* SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning. *Proteomics* 2018;18(2).

Wang, H., et al. Bayesian cluster ensembles. *Stat. Anal. Data Mining* 2011;4:54–70.

Wang, X.F. and Xu, Y. Fast clustering using adaptive density peak detection. *Stat Methods Med Res* 2015.

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10(1):57-63.

Weir, B.S. and Hill, W.G. Estimating F-statistics. *Annual review of genetics* 2002;36:721-750.

Weiser, J.N. The pneumococcus: why a commensal misbehaves. *J Mol Med (Berl)* 2010;88(2):97-102.

Yamamoto, M. and Sadamitsu, K. Dirichlet mixtures in text modeling. *CS Technical report CS-TR-05-1, University of Tsukuba* 2005.

Zappia, L., Phipson, B. and Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;18(1):174.

Zheng, G.*, et al.* Massively parallel digital transcriptional profiling of single cells. *bioRxiv* 2016.

Zheng, G.X.*, et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;8:14049.