Distribution Agreement

Signature:

Xin Wei	Date

EDClust: An EM-MM hybrid method for cell clustering in population-level single cell RNA sequencing

By

Xin Wei
Master of Science in Public Health
Department of Biostatistics and Bioinformatics

Hao Wu, Ph.D.
Committee Chair

Zhaohui Qin, Ph.D.
Committee Member

EDClust: An EM-MM hybrid method for cell clustering in population-level single cell RNA sequencing

By

Xin Wei

B.S.
Southern University of Science and Technology
2019

Thesis Committee Chair: Hao Wu, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2021

Abstract

EDClust: An EM-MM hybrid method for cell clustering in
population-level single cell RNA sequencing

By Xin Wei

Single-cell RNA sequencing (scRNA-seq) technology has revolutionized the genomics research by enabling the measurement of the transcriptomic profile at the level of single cells. One of the most fundamental problems in scRNA-seq data analysis is cell clustering, for which a rather large number of methods have been developed. With the increasing application of scRNA-seq in larger scale studies, people face the problem of cell clustering when the scRNA-seq data are from more than one subject. One challenge in analyzing such data is the subject-specific systematic variations: heterogeneity from multiple subjects may have a significant impact on the clustering accuracy. However, existing methods addressing such effect suffered from several limitations. In this work, we develop a novel statistical method named 'EDClust' for scRNA-seq cell clustering when data are from multiple subjects. EDClust models the sequence read counts by a mixture of Dirichlet-Multinomial distributions, and explicitly accounts for the cell type heterogeneity, subject heterogeneity, and the clustering uncertainty. An EM-MM hybrid algorithm is derived for maximizing the data likelihood and clustering the cells. We perform a series of simulation studies to evaluate the proposed method and demonstrate the outstanding performance of EDClust. Comprehensive benchmarking on four real scRNA-seq datasets with various tissue types and species demonstrates the substantial accuracy improvement of EDClust compared to the existing methods.

EDClust: An EM-MM hybrid method for cell clustering in population-level single cell RNA sequencing

By

Xin Wei

B.S.
Southern University of Science and Technology
2019

Thesis Committee Chair: Hao Wu, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2021

Acknowledgement

First and foremost, I am deeply grateful to my advisor, Dr. Hao Wu, who is always generous in sharing his enormous knowledge and providing continuous support. His invaluable advice guides me to learn to identify problems, foster divergent thinking, and practice writing skills.

I would also like to offer my special thanks to Dr. Ziyi Li and Dr. Kenong Su for their assistance and insightful suggestions on my thesis. I also want to thank Dr. Zhaohui Qin as the second reader of this thesis, and I am grateful for his invaluable comments.

In addition, I would like to express my gratitude to all the teaching and research staff in the Biostatistics Department at Emory for the comprehensive curriculum and consistent support.

Finally, I would like to express my sincere gratitude to my parents and my fiancee for their unwavering support and encouragement throughout my years of study.

Contents

1 . Introduction

Single-cell RNA-sequencing (scRNA-seq) is a powerful technology for measuring the gene expression at the single cell level. It offers unprecedented opportunities to answer questions related to cell-specific changes in transcriptome, such as identification of rare cell types and heterogeneity of cell responses[1]. Several experimental protocols of scRNA-seq have been developed in the past few years, including SMART-seq2[2], CEL-seq2[3] and Drop-seq[4], providing additional choices to meet diverse research needs. Among all, droplet-based technologies encapsulate each individual cell in a nanoliter droplet together with a bead[4], and substantially reduced the experimental cost. Moreover, droplet-based methods utilize unique molecular identifiers (UMIs) to eliminate the effects of PCR amplification bias[5]. The good scalability, high efficiency and low cost make droplet-based method the top choice for scRNA-seq experiments in population-scale studies.

In scRNA-seq data analysis, the first step is usually cell clustering. The main purpose of clustering is to group cells by their transcriptomic similarity, and then annotate the groups to cell types based on existing biological knowledge. This is a fundamental step in scRNA-seq analysis, since many downstream analyses, including cellular composition estimation, cell type-specific differential expression, and rare cell type discovery, are all carried out based on the clustering results[6]. Though classic unsupervised clustering methods such as K-means and hierarchical clustering can be applied, in view of the sparse and noisy characteristics of scRNA-seq data, many unsupervised methods customized for scRNA-seq data have been developed and widely used. For example, SC3 combines feature selection and dimension reduction in a consensus clustering framework and has been proven to be a highly robust clustering method[7]. Seurat is another popular method that adopts community-detection to identify similar cells and shows good scalability for large datasets[8]. TSCAN fits a mixture of multivariate normal distributions and uses hierarchical clustering to identify cell clusters[9]. Lastly, observing the needs for clustering large-scale study with thousands to millions of cells, SHARP is developed recently for ultra-fast clustering through a divide-and-conquer strategy[10].

All the aforementioned clustering methods are developed without the consideration of systematic biases in the data, that is, the expressions for a gene from all cells in the same cell type are considered to be identically distributed. However, similar to many other high-throughput technologies, scRNA-seq data also suffers from a number of technical biases. One such bias in population level study is the subject effect: there could be systematic, subject-specific shift in the gene expression. Thus, the distributions of the gene expression can be different between subjects even within the same cell type. That shift can be induced by different characteristics of the subjects, such as demographics or clinical conditions. Or it can be the result of batch effect when different subjects are profiled at different time/location/lab. It is worth mentioning that the batch effects can be severe in scRNA-seq, since it is exacerbated by the fact that most scRNA-seq protocols require fresh tissue for experiments, thus a randomized experimental design for removing batch effect might become impossible in many cases. Nevertheless, most existing clustering methods do not explicitly address the heterogeneity among multiple subjects. Direct application of those methods on data from population studies can lead to inaccurate clustering results due to correlated measurement errors instead of biological similarities[11].

One possible remedy for the problem is to consider the subject effect as batch effect, and correct for that before cell clustering. Several computational methods have been developed for batch effect correction and can be applied before clustering. For example, ComBat and ComBat-seq[12] are developed originally for bulk sequencing data and use linear models to remove batch effects. Mutual-nearest-neighbor (MNN) corrects batch effects by constructing a shared space between datasets[13]. Harmony is another popular batch correction method and uses an iterative approach to eliminate batch effects for cells calculated in PCA space[14].

Though it is possible to cluster the cells after removing the subject effects, this “two-step” approach has some drawbacks. First, the batch effect correction procedure often produces negative values for gene expressions, which will generate errors in many cell clustering tools. Secondly, such approach is in general not efficient due to the transformation of data and alteration of data structure. For example, several clustering methods make distributional assumptions on the count data, while the data after batch effect correction is not counts anymore. Such discrepancy will lead to undesirable clustering performances for those methods.

In comparison, a more rigorous and potentially better approach is to design a clustering method that takes subject effects into consideration. Both BAMM-SC[15] and BUSseq[16] are tailored methods for addressing subject effect during clustering. BAMM-SC implements a Bayesian mixture model which utilizes information across genes and individuals to account for the heterogeneity. BUSseq adopts a more complicated hierarchical model that strictly follows the data generation process of scRNA-seq experiments to correct batch effects and cluster cells. Both methods use Markov Chain Monte Carlo (MCMC) to solve the model, which do not scale well for large datasets. To provide a complementary approach to address the cell clustering problem in population-scale scRNA-seq data, we design EDClust, which is an EM[17] and MM[18] hybrid method based on Dirichlet-Multinomial mixture model, for clustering. EDClust takes the raw count data from multiple subjects without transformation, avoiding the possible destruction of data structure and loss of information. Meanwhile, EDClust explicitly quantifies the effects of heterogeneity from different sources and provides posterior probabilities for cells being in each cluster. Through extensive simulation studies and four real datasets, we show EDClust has better clustering accuracy compared with existing methods. In the following sections, we first introduce the data model and derivation of the EM-MM method. The simulation design and results are presented in Section 3. Lastly, we showcase the performance and utility of EDClust using four real scRNA-seq datasets in Section 4.

2 . Methods

To cluster population-scale scRNA-seq data, we propose the following Dirichlet-Multinomial mixture model to capture the cell type and subject effects on gene expression. We assume the number of cell types in the data, K, is known, and all subjects share the same K cell types. Our aim is to cluster all the cells from these subjects simultaneously. Note that K can be specified by investigators based on biological knowledge, or can be determined by a number of software tools. Throughout this work, we will just assume K is known.

2.1 Data model

Let y_{lji} represent the sequence counts for gene j in cell i from subject l ($1 \leqslant i \leqslant I_l$, $1 \leqslant j \leqslant J$, $1 \leqslant l \leqslant L$), where I_l, J and L indicate the total number of cells (in subject l), genes and subjects, respectively. Based on the assumption that $\boldsymbol{Y}_{li} = (Y_{l1i}, Y_{l2i}, \ldots, Y_{lJi})$ follows a Dirichlet-Multinomial mixture distribution, $\boldsymbol{Y}_{li}$ can be viewed as generated in two steps. First, a cell type label $W_{li} \in \{1, 2, \ldots, K\}$ is assigned to cell i in subject l with probability $\Pr(W_{li} = k) = \pi_{lk}$. Second, given the cell label (i.e., $W_{li} = k$), $\boldsymbol{Y}_{li}$ will be generated from a Multinomial distribution by $\boldsymbol{Y}_{li} \sim \text{Multinomial}(T_{li}, p_{li})$. Here, $T_{li} = \sum_j^J Y_{lji}$ indicates total read counts, and the proportion p_{li} represents the relative gene expressions. We further assume that p_{li} follows a cell-type specific prior distribution Dirichlet$(\alpha_{lk}) =$ Dirichlet$(\alpha_{lk1}, \alpha_{lk2}, \ldots, \alpha_{lkJ})$. To simultaneously account for cell type and subject effects, we assume the overall effect α_{lk} can be expressed as the sum of cell type effect α_{0kj} and subject effect δ_{lkj}: $\alpha_{lkj} = \alpha_{0kj} + \delta_{lkj} > 0$. Finally, we assume that all cells in all L subjects are independent and treat cell type label $W_{li} = k$ as the missing data. Then the observed and complete data log likelihood can be written as:

$$\begin{aligned} l(\boldsymbol{Y};\boldsymbol{T},\boldsymbol{\Theta}) &= \sum_{l=1}^{L}\sum_{i=1}^{I_l} log[\sum_{k=1}^{K} \pi_{lk} P(Y_{li}|T_{li}, \alpha_{0k} + \delta_{lk})] && (2.1) \\ l_c(\boldsymbol{Y},\boldsymbol{W};\boldsymbol{T},\boldsymbol{\Theta}) &= \sum_{l=1}^{L}\sum_{i=1}^{I_l}\sum_{k=1}^{K} I(W_{li} = k)[log\pi_{lk} + logP(Y_{li}|T_{li}, \alpha_{0k} + \delta_{lk})] && (2.2) \end{aligned}$$

Here $\boldsymbol{W} = \{W_{li} : i = 1, \ldots, I_l, l = 1, \ldots, L\}$ includes the indicator of cell type labels, and $\boldsymbol{\Theta} = \{\pi_{lk}, \alpha_{0k}, \delta_{lk} : k = 1, \ldots, K, l = 1, \ldots, L\}$ contains all the model parameters. $P(Y_{li}|T_{li}, \alpha_{0k} + \delta_{lk})$ represents the Dirichlet-Multinomial probability density, which is

$$P(Y_{li}|T_{li}, \alpha_{0k} + \delta_{lk}) = \binom{T_i}{Y_{li}} \frac{\prod_{j=1}^{J}(\alpha_{0kj} + \delta_{lkj})(\alpha_{0kj} + \delta_{lkj} + 1)\cdots(\alpha_{0kj} + \delta_{lkj} + Y_{lij} - 1)}{\|\alpha_{0k} + \delta_{lk}\|_1 (\|\alpha_{0k} + \delta_{lk}\|_1 + 1)\cdots(\|\alpha_{0k} + \delta_{lk}\|_1 + T_{li} - 1)} \quad (2.3)$$

Here, $\|\alpha_{0k} + \delta_{lk}\|_1 = \sum_{j=1}^{J} |\alpha_{0kj} + \delta_{lkj}| = \sum_{j=1}^{J}(\alpha_{0kj} + \delta_{lkj})$.

2.2 The EM-MM hybrid algorithm for maximum likelihood

The introduction of the latent variable $\boldsymbol{W}$ allows one to implement the EM algorithm to maximize the observed data likelihood and obtain posterior probabilities for cell type assignment (W_{li}). An EM algorithm iterates between two steps: an expectation step (E-step) and a maximization step (M-step)[17]. Let $\boldsymbol{\Theta}^{(t)}$ be the parameter estimate in iteration t. In the E-step, we compute the conditional expectation of W_{li}:

$$\begin{aligned}\mu_{lik}^{(t)} &= E[I(W_{li}=k)|\boldsymbol{Y},\boldsymbol{\Theta}^{(t)}] = P(W_{li}=k|\boldsymbol{Y},\boldsymbol{\Theta}^{(t)}) \\ &= \frac{\pi_{lk}^{(t)}P(Y_{li}|T_{li},\alpha_{0k}^{(t)}+\delta_{lk}^{(t)})}{\sum_{k'=1}^{K}\pi_{lk'}^{(t)}P(Y_{li}|T_{li},\alpha_{0k'}^{(t)}+\delta_{lk'}^{(t)})}\end{aligned} \tag{2.4}$$

In the M-step, we maximize the “Q function” (the expected complete data log-likelihood with respect to $\boldsymbol{\Theta}$) to obtain $\boldsymbol{\Theta}^{(t+1)}$. The update for π_{lk} can be obtained by solving $\partial Q/\partial \pi_{lk}=0$.

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = E[l(\boldsymbol{\Theta})|\boldsymbol{Y},\boldsymbol{\Theta}^{(t)}] = \sum_{l=1}^{L}\sum_{i=1}^{I_l}\sum_{k=1}^{K}\mu_{lik}^{(t)}[log\pi_{lk}+logP(Y_{li}|T_{li},\alpha_{0k}+\delta_{lk})] \tag{2.5}$$

$$\pi_{lk}^{(t+1)} = \frac{\sum_{i=1}^{I_l}\mu_{lik}^{(t)}}{I_l} \tag{2.6}$$

The M-step derivation for α_{0k} and δ_{lk} is much more difficult, and there is not closed form solution. For that, we design the following Minorization-Maximization (MM) algorithm[18] for updating α_{0k} and δ_{lk}. Conceptually, to maximize an objective function $f(\theta)$, an MM algorithm iterates between two steps. In the first step, one uses the current parameter estimate $\theta^{(n)}$ to construct a surrogate function $g(\theta|\theta^{(n)})$ such that $g(\theta|\theta^{(n)})$ minorizes $f(\theta)$, i.e.,

$$\begin{aligned} f(\theta) &\geq g(\theta|\theta^{(n)}) \qquad \forall\, \theta \neq \theta^{(n)} \\ f(\theta^{(n)}) &= g(\theta^{(n)}|\theta^{(n)}) \end{aligned} \tag{2.7}$$

In the second step, one finds θ to maximize the surrogate function $g(\theta|\theta^{(n)})$, which gives a new parameter estimate $\theta^{(n+1)}$. Since

$$f(\theta^{(n+1)}) \geq g(\theta^{(n+1)}|\theta^{(n)}) \geq g(\theta^{(n)}|\theta^{(n)}) = f(\theta^{(n)}) \tag{2.8}$$

$f(\theta^{(n)})$ will never decrease as n increases. The algorithm will converge to a stationary point, usually a mode of the objective function.

Extending the work by Zhou and Lange[19], we rewrite the log-likelihood function in (2.2) as the following.

$$\begin{aligned}
l(\Theta) &= \sum_{l=1}^{L}\sum_{k=1}^{K}\left[\sum_{i=1}^{I_l} I(W_{li}=k)log\pi_{lk} - \sum_{c_{1l}} r_{lkc}log(\|\alpha_{0k}+\delta_{lk}\|_1 + c_{1l}) + \right.\\
&\left. \sum_{j=1}^{J}\sum_{c_{2lj}} s_{lkjc}log(\alpha_{0kj}+\delta_{lkj}+c_{2lj})\right] + const. \qquad (2.9)\\
r_{lkc} &= \sum_{i=1}^{I_l} I(W_{li}=k)I(T_{li}\geqslant c_{1l}+1), \quad s_{lkjc} = \sum_{i=1}^{I_l} I(W_{li}=k)I(Y_{lji}\geqslant c_{2lj}+1)
\end{aligned}$$

Here the index c_{1l} ranges from 0 to $\max_i(T_{li})-1$, and the index c_{2lj} runs from 0 to $\max_i(Y_{lij})-1$. In MM algorithm, we design a surrogate function that minorizes the log-likelihood function. Assuming that $\alpha_{0kj}>0$ and $\delta_{lkj}\geqslant 0$, we can utilize the following inequalities:

$$\begin{aligned}
-log(c+\|\alpha_{lk}\|_1) &\geqslant -\frac{1}{\left\|\alpha_{0k}^{(n)}+\delta_{lk}^{(n)}\right\|_1 + c}(\|\alpha_{0k}+\delta_{lk}\|_1) + const. \qquad (2.10)\\
log(\alpha_{0kj}+\delta_{lkj}+c) &\geqslant \frac{\alpha_{0kj}^{(n)}}{\alpha_{0kj}^{(n)}+\delta_{lkj}^{(n)}+c}log(\alpha_{0kj}) + \frac{\delta_{lkj}^{(n)}}{\alpha_{0kj}^{(n)}+\delta_{lkj}^{(n)}+c}log(\delta_{lkj}) + const. \ (2.11)
\end{aligned}$$

For them, the equality holds when $\alpha_{0kj}=\alpha_{0kj}^{(n)}$ and $\delta_{lkj}=\delta_{lkj}^{(n)}$. We construct the following surrogate function $g(\Theta|\Theta^{(t,n)})$ as:

$$\begin{aligned}
g(\Theta|\Theta^{(t,n)}) &= \sum_{l=1}^{L}\sum_{k=1}^{K}\left\{\sum_{i=1}^{I_l}\mu_{lik}^{(t)}log\pi_{lk} - \sum_{c_{1l}} r_{lkc}^{(t)}\frac{\|\alpha_{0k}+\delta_{lk}\|_1}{\left\|\alpha_{0k}^{(t,n)}+\delta_{lk}^{(t,n)}\right\|_1 + c_{1l}} + \right.\\
&\left. \sum_{j=1}^{J}\sum_{c_{2lj}} s_{lkjc}^{(t)}\left[\frac{\alpha_{0kj}^{(t,n)}log(\alpha_{0kj})}{\alpha_{0kj}^{(t,n)}+\delta_{lkj}^{(t,n)}+c_{2lj}} + \frac{\delta_{lkj}^{(t,n)}log(\delta_{lkj})}{\alpha_{0kj}^{(t,n)}+\delta_{lkj}^{(t,n)}+c_{2lj}}\right]\right\} + const. \ (2.12)\\
r_{lkc}^{(t)} &= \sum_{i=1}^{I_l}\mu_{lik}^{(t)}I(T_{li}\geqslant c_{1l}+1), \quad s_{lkjc}^{(t)} = \sum_{i=1}^{I_l}\mu_{lik}^{(t)}I(Y_{lji}\geqslant c_{2lj}+1)
\end{aligned}$$

By solving $\partial g(\Theta|\Theta^{(t,n)})/\partial\delta_{lkj}=0$ and $\partial g(\Theta|\Theta^{(t,n)})/\partial\alpha_{0kj}=0$, we obtain the MM updates for δ_{lkj} and α_{0kj} as:

$$\delta_{lkj}^{(t,n+1)} = \left(\sum_{c_{2lj}} \frac{s_{lkjc}^{(t)}\delta_{lkj}^{(t,n)}}{\alpha_{0kj}^{(t,n)}+\delta_{lkj}^{(t,n)}+c_{2lj}}\right)\Bigg/\left(\sum_{c_{1l}} \frac{r_{lkc}^{(t)}}{\left\|\alpha_{0k}^{(t,n)}+\delta_{lk}^{(t,n)}\right\|_1+c_{1l}}\right) \tag{2.13}$$

$$\alpha_{0kj}^{(t,n+1)} = \left(\sum_{l=1}^{L}\sum_{c_{2lj}} \frac{s_{lkjc}^{(t)}\alpha_{0kj}^{(t,n)}}{\alpha_{0kj}^{(t,n)}+\delta_{lkj}^{(t,n)}+c_{2lj}}\right)\Bigg/\left(\sum_{l=1}^{L}\sum_{c_{1l}} \frac{r_{lkc}^{(t)}}{\left\|\alpha_{0k}^{(t,n)}+\delta_{lk}^{(t,n)}\right\|_1+c_{1l}}\right) \tag{2.14}$$

Within the M-step in each EM iteration, EDClust runs multiple MM iterations to update α_0 and δ. To reduce the computational burden, we only run 3 MM iterations in each M-step. Real data analyses show that such procedure provides comparable performance as running more (such as 20) iterations.

2.3 Feature selection

Feature selection is one of the key steps before clustering. We aim to select a subset of informative genes that can identify the structure of data and thus improve the performance of clustering. A recently developed feature selection tool tailored to scRNA-seq, FEAture SelecTion (FEAST)[20], shows great potential for improving clustering accuracy. FEAST computes the F-statistics for each feature based on embedded consensus clustering results and provides a ranking list of feature significant. By default, EDClust applies FEAST to obtain the top 500 features for clustering. In the software implementation, users have option to specify the gene features.

2.4 Determine the initial values

It is known that EM algorithm often suffers from locally optimal solutions. Our problem, due to the high dimensionality and complex nature of the data, is particularly prone to such problem. Thus, it is crucial to provide good initial values for the parameters, especially α_0 and δ. For that, we run unsupervised clustering on a randomly chosen subject to obtain initial clusters, and then the initial value for α_0 can be computed from these initial clusters. To be specific, we set a randomly chosen subject as the baseline, and thus its overall effect is entirely contributed by the cell type effect α_0. Based on the initial clusters, we obtain a naive estimate $\hat{\alpha}_0$ according to the relative gene expression in each initial cluster and take it as the initial value. We set the selected subject with a subject effect of zero, and set initial values for the rest of δ's to be small positive numbers (10^{-5} by default).

We use SHARP[10] as the unsupervised clustering method to determine initial values due to its good computational performance.

2.5 Software implementation

Overall, the complete EDClust algorithm is summarized in Figure 1. EDClust is designed as an R package with core algorithm written in Julia for better computational efficiency. Scripts for EDClust are available at https://github.com/weix21/EDClust. We are working to develop a software package and will submit to Bioconductor soon.

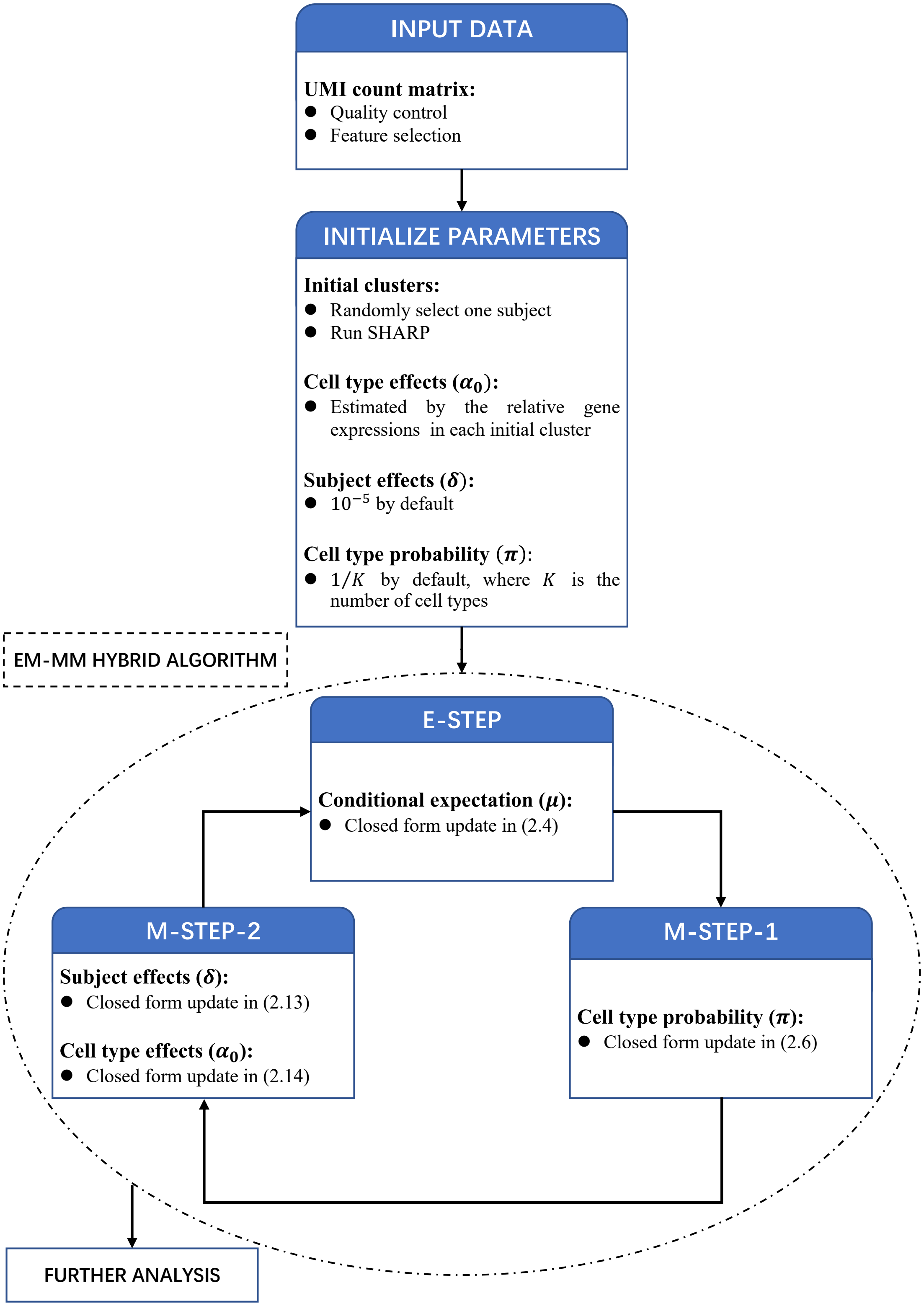
INPUT DATA
UMI count matrix:
● Quality control
● Feature selection
INITIALIZE PARAMETERS
Initial clusters:
● Randomly select one subject
● Run SHARP
Cell type effects (α_0):
● Estimated by the relative gene expressions in each initial cluster
Subject effects (δ):
● 10^{-5} by default
Cell type probability (π):
● $1/K$ by default, where K is the number of cell types
EM-MM HYBRID ALGORITHM
E-STEP
Conditional expectation (μ):
● Closed form update in (2.4)
M-STEP-2
Subject effects (δ):
● Closed form update in (2.13)
Cell type effects (α_0):
● Closed form update in (2.14)
M-STEP-1
Cell type probability (π):
● Closed form update in (2.6)
FURTHER ANALYSIS

Figure 1: Summary of the EDClust algorithm.

3 . Simulation studies

We design a series of simulation studies to comprehensively evaluate the performance of EDClust, and compare it to a number of competing methods. The simulations are based on a set of real scRNA-seq from human skin studies (described later). We evaluate the methods when data have different levels of subject specific effects (low, medium and high), and with different sample size selections (5, 10, 15). Specifically, for gene j, cell i and subject l, we generate observed single cell RNA-seq counts from a Dirichlet-Multinomial distribution by $\boldsymbol{Y}_{li} \sim \text{Multinomial}(T_{li}, \boldsymbol{p}_{li})$ where $\boldsymbol{p}_{li} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{lk})$. The total read counts T_{li} match the real data. We specify a log-normal prior distribution on $\boldsymbol{\alpha}_{lk}$ as $\boldsymbol{\alpha}_{lk} \sim \text{LogNormal}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$. The mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\sigma}_k^2$ are first predefined for each random sample and cell types by adopting hyper-prior $\boldsymbol{\mu}_k \sim \text{Uniform}(0,1)$ and $\boldsymbol{\sigma}_k^2 \sim \text{Gamma}(\beta, \tau)$. In all the simulations, we specify τ as 1 and changed the value of β to control the cross subject heterogeneity in the data. Larger heterogeneity indicates stronger subject-specific effects, and thus it is more difficult to cluster.

We compare EDClust with the other four unsupervised clustering methods (SC3, Seurat, SHARP and TSCAN) which embed Harmony for batch effect correction before clustering. We use adjusted Rand index (ARI)[21] as the evaluation criterion to benchmark the predicted cell type labels. All the simulation results are summarized over 20 Monte Carlo data sets.

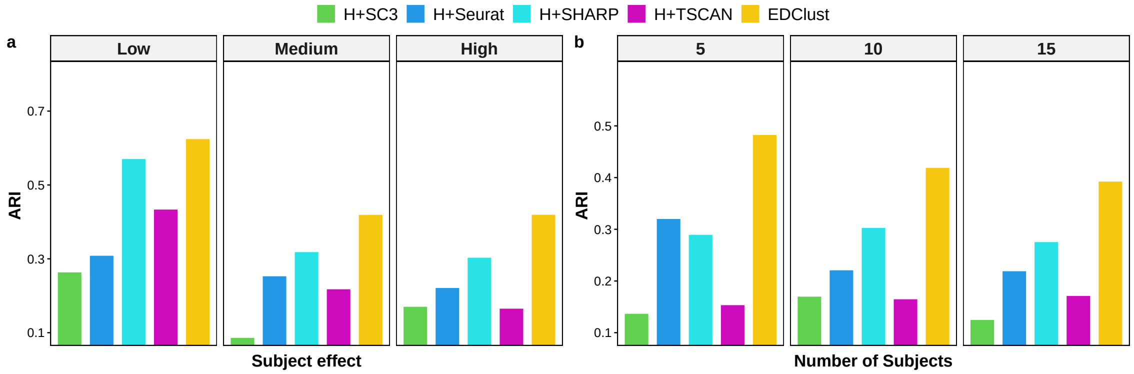

Figure 2: Barplots of average ARIs for 5 clustering methods across over 20 simulations, where "H +" indicates that the simulation data are processed by Harmony to remove the subject effects. **a** Influence of subject effect on clustering results. Simulation data is consisted of 10 subjects. **b** Influence of the number of subjects on clustering results with medium level of the subject effect.

As shown in Figure 2a, across three scenarios with different levels of cross subject heterogeneity, EDClust constantly achieves the highest average ARI. The performance of EDclust remains stable even as the subject effect varies from medium level to high level. Figure 2b presents the influence of number of subjects on clustering. As the number of subjects increased, so did the sources of heterogeneity. As expected, the average ARIs for most of the clustering methods decrease when the number of subjects increases. EDClust consistently outperforms other four methods in terms of ARI. The simulation studies showcase great potential of EDClust in accounting for subject-specific effects and clustering population-scale scRNA-seq data with outstanding performance.

4 . Real data analyses

We benchmark EDClust and other methods on four sets of real scRNA-seq with multiple subjects. More description of the datasets and data processing procedures are provided in each of the subsections below. Here in Table 1, we present the overall results for all four datasets, including the mean and standard deviation of ARIs from 50 runs in each datasets. In addition to the four clustering methods compared in the simulation study, we also compare EDClust with DIMM-SC and BAMM-SC, which have similar model assumptions. Since both TSCAN and Seurat are deterministic clustering methods, they don't have standard deviation in the results. The average ARIs are also displayed in Figure 3. These results show that for three out of the four dataset, EDClust has the best performance, and the performance improvement can be significant. For example in the Mouse Retina data, EDClust has mean ARI 0.87, while the second best performer (Harmony+SC3) only has ARI 0.70. In the Mouse Lung data, EDClust performs slightly worse than BAMM-SC and Harmony+SC3, but not very far off.

Table 1: The ARI of fifty times clustering analyses for each method on four real datasets

Method	Mouse Retina		Baron Pancreas		Human Skin		Mouse Lung	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BAMM-SC	0.4273	0.0058	0.6411	0.0686	0.7732	0.0688	0.7354	0.0323
DIMM-SC	0.4221	0.0065	0.6968	0.0692	0.7975	0.0839	0.7003	0.0643
Harmony+SC3	0.6972	0.0687	0.5590	0.1035	0.8260	0.0249	0.7652	0.0026
Harmony+Seurat	0.1024	-	0.5137	-	0.6520	-	0.5307	-
Harmony+SHARP	0.6572	0.0095	0.3115	0.0236	0.8369	0.0533	0.7029	0.0271
Harmony+TSCAN	0.2905	-	0.6392	-	0.6486	-	0.6354	-
EDClust	0.8735	0.0251	0.8047	0.0747	0.9191	0.0813	0.7227	0.0435

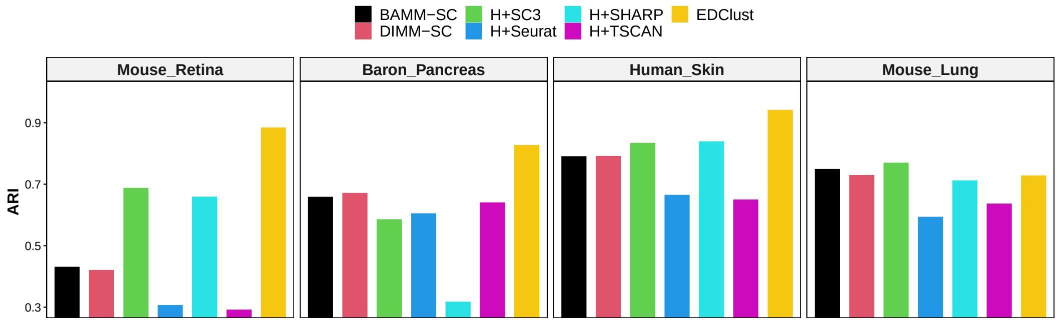

Figure 3: Bar plots of performance of EDClust and competing methods measured by average ARI across 50 clustering results. "H +" represents that the clustering methods are implemented on the real datasets which have been processed by Harmony to remove the batch effects.

4.1 Mouse Retina dataset

We first evaluate the clustering performance of EDClust in mouse tissues through a mouse retina dataset, which is collected from 14-day-old mice in seven batches[4]. Cells are first pooled together to filter out low-expression genes based on dropout rate. We apply FEAST to generate a ranking list of features, and select top 500 genes based on it. Five major cell types are retained and the number of cells is 43,603.

As shown in Table 1 and Figure 3, most of the methods struggle on this dataset with very low average ARI. Though the performances of Harmony+SC3 and Harmony+TSCAN are slightly better, their average ARIs are still below 0.70. EDClust achieves the highest ARI (0.8735), suggesting the excellent performance of EDClust. To visualize the clustering results, we generate some t-SNE plots as shown in Figure 4. The t-SNE plot generated based on the clustering result of EDClust is highly similar to the t-SNE plot with the true labels. We also show the t-SNE plot based on the clustering results from BAMM-SC and Harmony+SC3, where the circled regions highlight the incorrectly clustered cells. These plots provide clear visualization for demonstrating the improved performance of EDClust.

4.2 Baron Pancreas dataset

To evaluate the performance of EDClust in human tissue, we analyze a set of human pancreas data (named "Baron Pancreas" dataset). The original data includes over 12,000 pancreatic cells from four human donors and two mouse strains[22]. We extract cells from the human donors and filter out

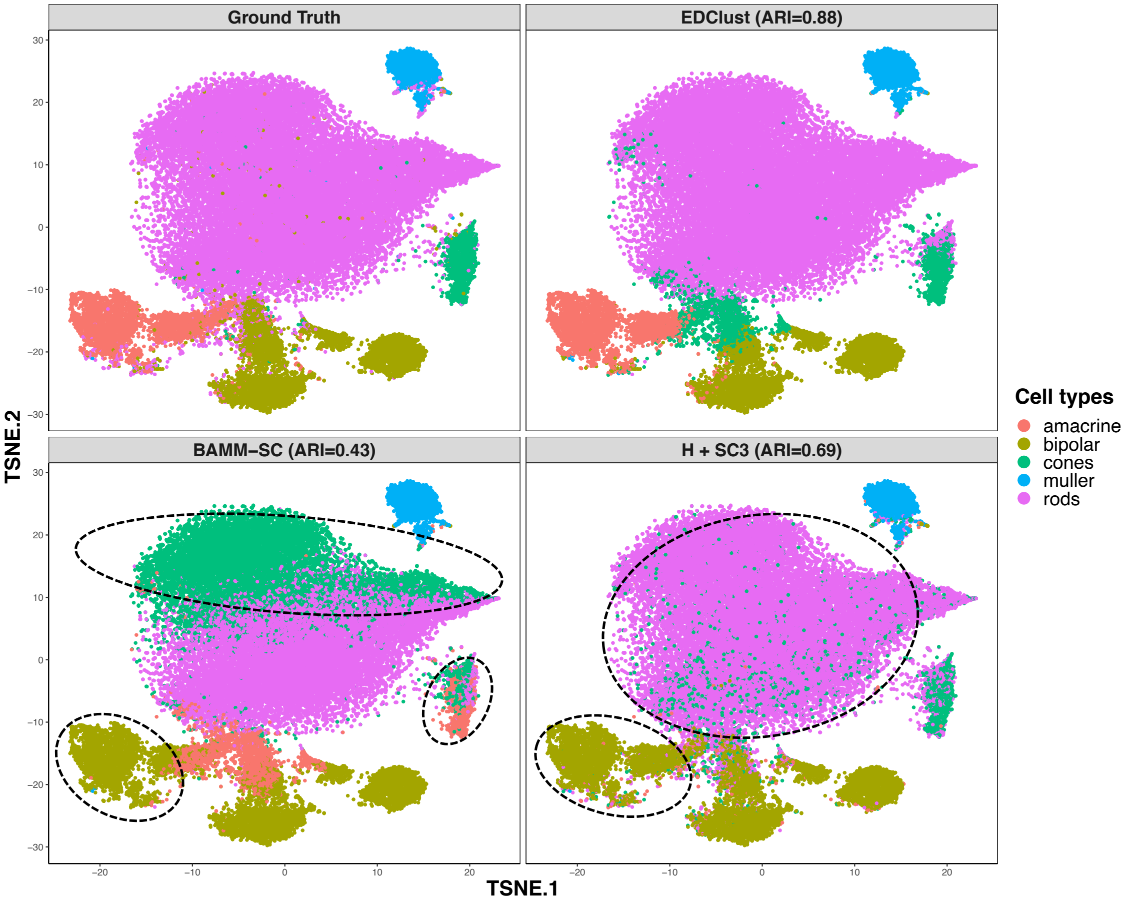

Figure 4: The t-SNE plots of cells in the mouse retina dataset[4]. Each plot is colored by the ground truth, labels inferred by EDClust, BAMM-SC, and Harmony+SC3 (H+SC3), respectively.

the lowly expressed genes. The processed data contains 500 genes a total of 8,506 cells. Some very rare cell types with only a few cells (such as T cells) are removed and 10 major cell types are kept for further analysis.

Both Table 1 and Figure 3 show that the average ARI of EDClust is up to 0.8047 while all other methods fail to achieve the average ARI of 0.70. Figure 5 elucidates that Harmony+SC3 mixed massive cells. Compared to BAMM-SC, EDClust correctly identifies beta cells. Based on EDClust, for most of the cells, we are able to assign labels that are close to the approximated truth. These results showcase the outstanding performance of EDClust in the Baron Pancreas dataset.

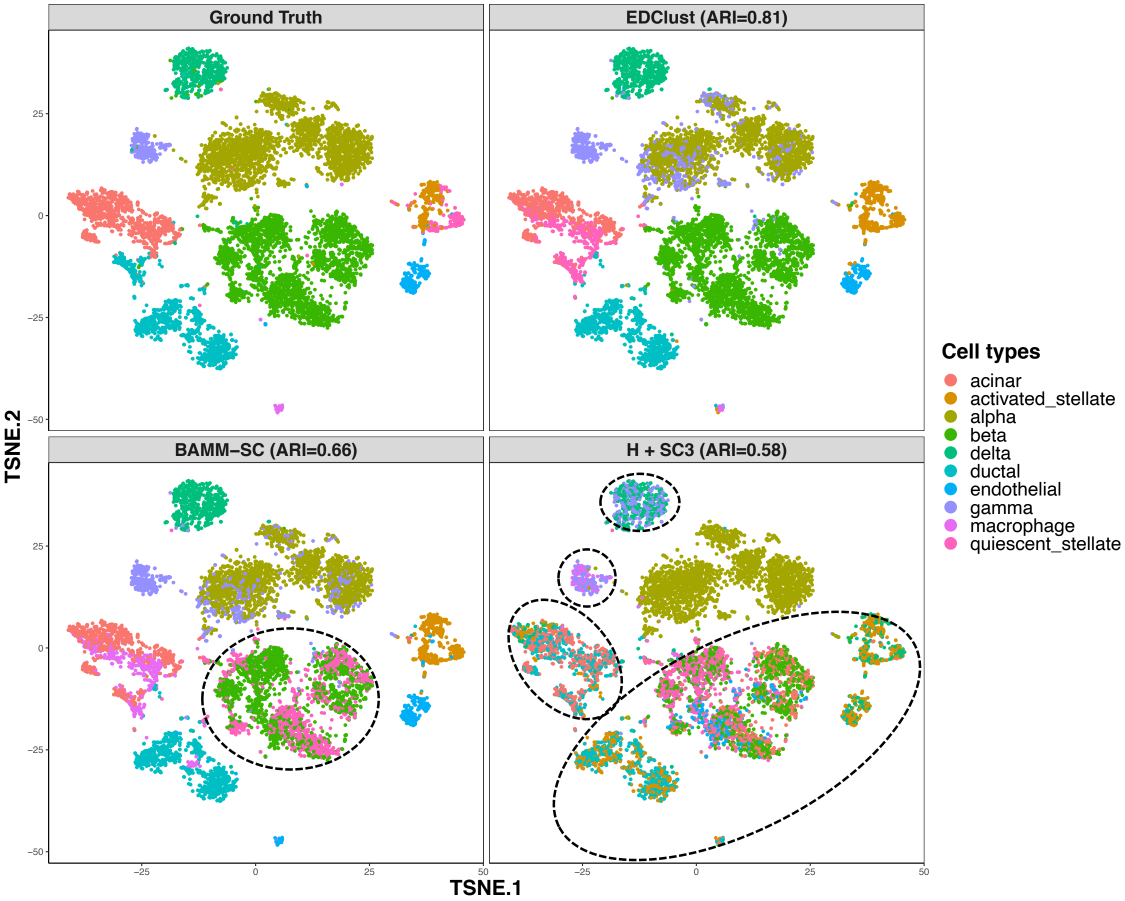

Figure 5: The t-SNE plots of cells in the Baron Pancreas dataset[22]. Each plot is colored by the ground truth, labels inferred by EDClust, BAMM-SC, and Harmony+SC3 (H+SC3), respectively.

4.3 Human Skin dataset

We further evaluate the clustering performance of EDClust in another human tissues through a human skin dataset, which includes skin samples collected from three health donors in a systemic sclerosis study[15]. In their study, Sun et al.[15] identified eight major types of cells. We use their results as the ground truth, but remove cells with uncertain cell type. After quality control and feature selection, 3,067 cells with 500 selected genes are used in the clustering analysis.

From Table 1 and Figure 3, we can find that EDClust has the most outstanding performance (average ARI = 0.9191) among all the methods. Although the average ARIs for most methods are close to 0.80, EDClust is more accurate in the clustering of several cell types. As shown in Figure 6, compared with BAMM-SC and Harmony+SC3, macrophages/DC, basal keratinocytes and suprabasal keratinocytes can all be classified by EDClust and each is assigned a specific cell type label, which points out the superior performance of EDClust on the Human Skin dataset.

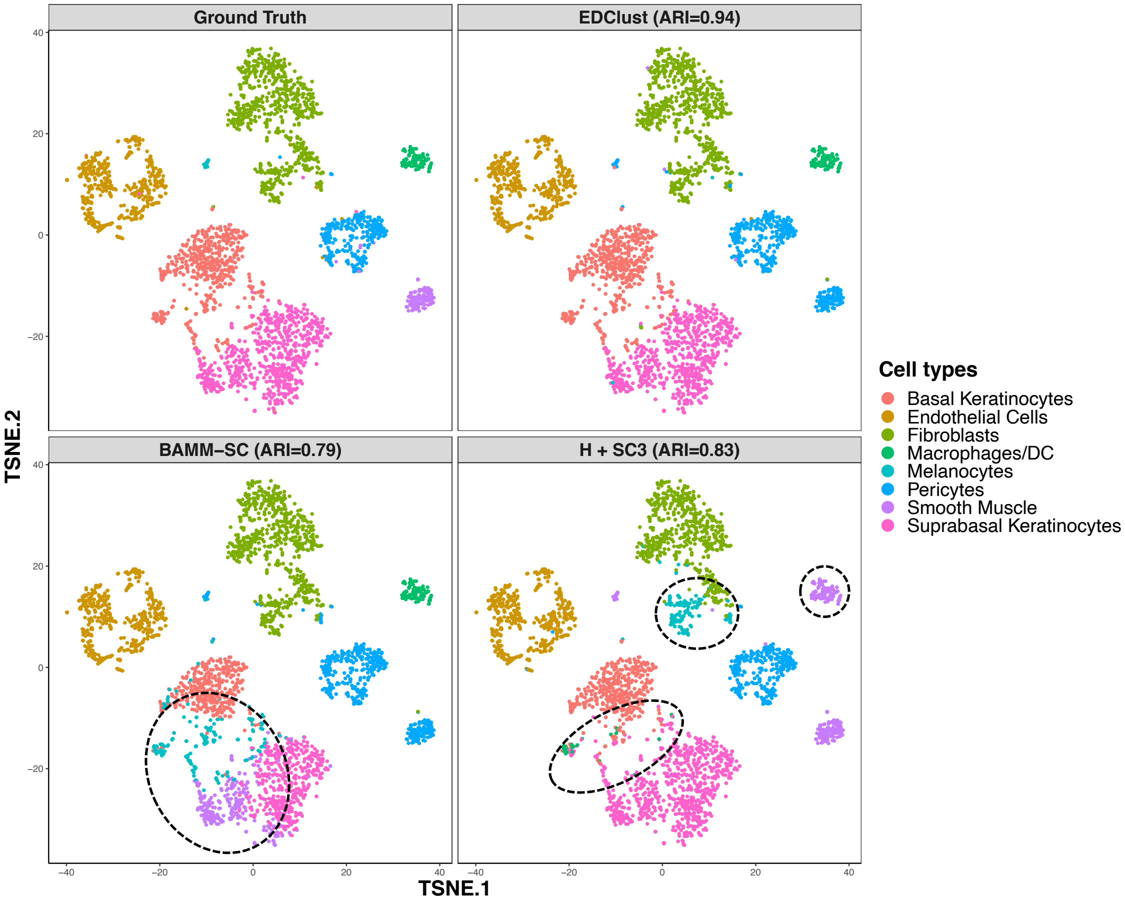

Figure 6: The t-SNE plots of cells in the Human Skin dataset[15]. Each plot is colored by the ground truth, labels inferred by EDClust, BAMM-SC, and Harmony+SC3 (H+SC3), respectively.

4.4 Mouse Lung dataset

At last, we evaluate the performance of EDClust in a real dataset with fewer cells. We mainly analyze a mouse lung dataset, which is obtained by collecting lung mononuclear cells from four mouse samples in Streptococcus pneumonia (SP) infected group and control group[15]. After data processing step, we obtain 500 top features provided by FEAST and a total of 1,756 cells. Each cell is assigned a cell type label according to previous study by Sun et al.[15] and the expected number of clusters is set as six.

All methods have similar performances on the Mouse Lung dataset (Table 1 and Figure 3). The performance of EDClust (average ARI = 0.7227) is slightly lower than Harmony+SC3 and BAMM-SC. Figure 7 presents consistent pattern. In general, despite some mixed cell types, EDClust has an excellent performance in characterization of endothelial cells and neutrophils.

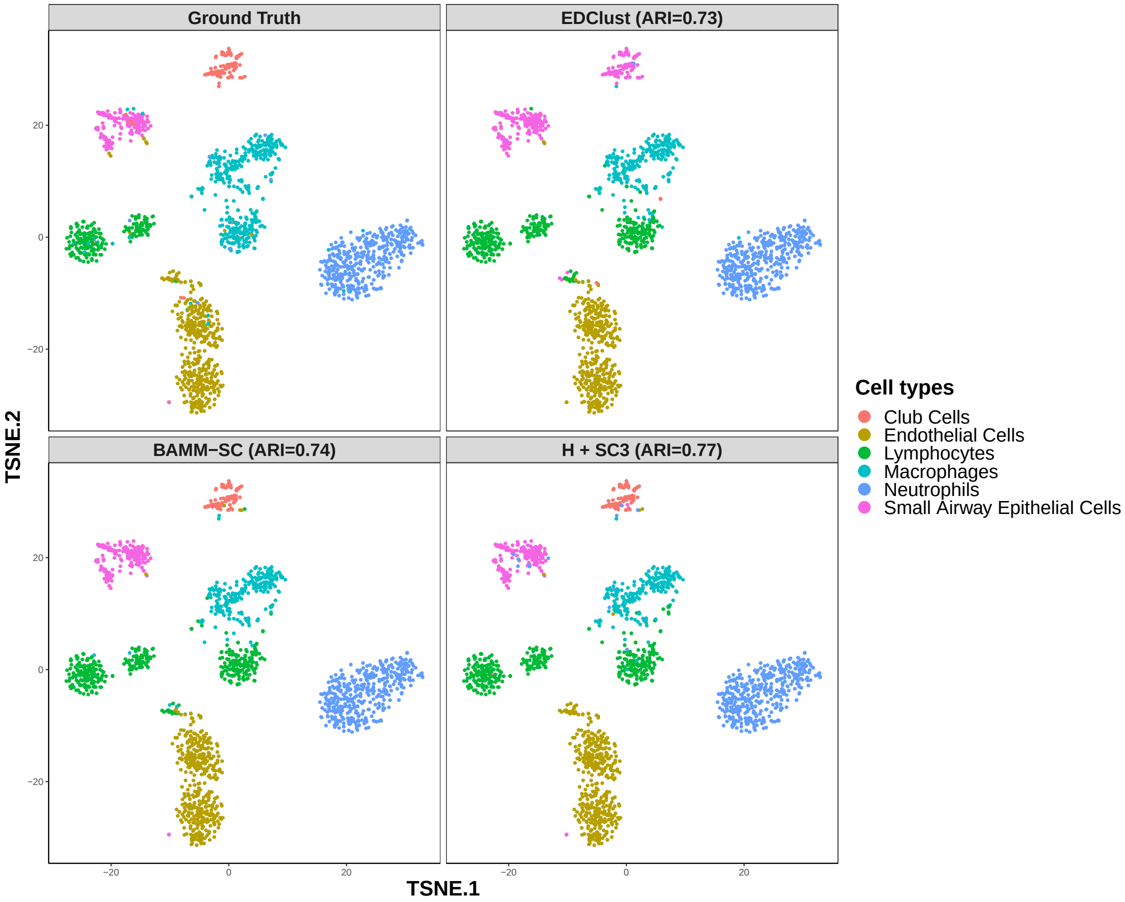

Figure 7: The t-SNE plots of cells in the Mouse Lung dataset[15]. Each plot is colored by the ground truth, labels inferred by EDClust, BAMM-SC, and Harmony+SC3 (H+SC3), respectively.

4.5 Computational performance

The EM algorithm usually converges slowly and has heavy computational burden. Our proposed method embeds a few MM iteration within each EM iteration, which brings higher computational cost. However, we implement the software in Julia, with an interface to R, and achieves reasonable computational performance. We benchmark the computational performances of all methods under comparison shown in Table 2. It shows that for the biggest dataset we have tried (Mouse Retina dataset with 7 batches and 43,603 cells), it takes about 35 minutes on a normal computer with a single process. This is considerably faster than BAMM-SC, which serves the same purpose but uses MCMC. Other methods either ignore the subject effects or perform a two-step approach (batch effect removal then cell clustering), which are not really comparable to EDClust. We will continue our development of EDClust and implement parallel computing to further improve the performance. Overall, EDClust provides satisfactory computational performance.

Table 2: The average computational time (in min) of each method on four real datasets

Method	Mouse Retina	Baron Pancreas	Human Skin	Mouse Lung
BAMM-SC	180	90	13	7
DIMM-SC	4	3	2	1
Harmony+SC3	57	24	6	8
Harmony+Seurat	34	4	1	<1
Harmony+SHARP	37	4	2	<1
Harmony+TSCAN	33	4	1	<1
EDClust	35	60	25	10

5 . Discussion

In this work, we develop a novel statistical method for cell clustering in multi-subject scRNA-seq data. We model the read counts by a Dirichlet-Multinomial mixture distribution, where the Dirichlet parameters contain subject and cell type effects. We develop an EM-MM hybrid algorithm for fitting the mixture model and performing model-based clustering.

Compared to existing clustering methods that ignore the subject specific effects, EDClust has the following advantages: (1) EDClust provides a tool to describe data heterogeneity among multiple subjects and more effectively identify subject-specific cell types. (2) Utilizing the shared information among subjects, EDClust clusters all the cells from all subjects at the same time, which improves the accuracy of cell clustering. (3) Most of the clustering methods can only be performed after several preprocessing approaches, e.g. normalization and batch effect removal, while EDClust offers a one-stop service which can be directly applied on raw count data. (4) EDClust quantifies cluster uncertainty with the probability that each cell belongs to a given cluster, contributing to further statistical inference and biological interpretation.

In our simulation studies, we investigate the influence of heterogeneity among multiple subjects on clustering results. We generate simulated data from Dirichlet-Multinomial distribution with specifying different subject effects and different sample size selections. In real data analyses, we compare the performance of EDClust with competing methods in four droplet-based scRNA-seq datasets with multiple subjects collected from different tissue types or species. In contrast to existing clustering methods that work on data with batch-effect removal through Harmony, EDClust considerably improves clustering accuracy under various experimental designs. Compared to existing clustering methods which account for subject variability, EDClust adopts a straightforward and highly explanatory model but at the same time outperforms those methods in terms of ARI on most of the datasets we tested.

For real data analysis, since the initial values of the cell type effects are set as the naive estimates based on the clustering results given by SHARP, we recommend running EDClust multiple times, each time using a different random seed, and select the one with the best likelihood as the final result. Estimation of $\sum_{j=1}^{J} \alpha_{0kj}$ provide by Ronning[23] or moment estimates proposed by Weir and Hill[24] can also be an appropriate choice for obtaining initial values. Moreover, determination of the number of clusters is a crucial step. We suggest predefining it based on prior biological knowledge or model selection criteria such as AIC[25] and BIC[26].

There are several limitations of EDClust. First, similar to all the methods using EM algorithm, EDClust could be sensitive to the initial values. The current algorithm computes initial values based on existing unsupervised clustering method. How to better determine initial values is our research direction in the near future. For example, we could use a small subset of cells with strong confidence in cell types, or apply prior biological knowledge. Secondly, EDClust is computationally heavy, especially when the dataset contains some cells with a large total count. We will investigate whether we can build a different surrogate function in order to improve the computational performance.

References

[1] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[2] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

[3] Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron De Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, et al. Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome biology*, 17(1):1–7, 2016.

[4] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[5] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[6] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317, 2019.

[7] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[8] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

[9] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[10] Shibiao Wan, Junil Kim, and Kyoung Jae Won. Sharp: hyperfast and accurate processing of single-cell rna-seq data via ensemble random projection. *Genome research*, 30(2):205–213, 2020.

[11] Yifan Tang. Cluster analysis with batch effect. 2015.

[12] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *NAR genomics and bioinformatics*, 2(3):lqaa078, 2020.

[13] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

[14] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

[15] Zhe Sun, Li Chen, Hongyi Xin, Yale Jiang, Qianhui Huang, Anthony R Cillo, Tracy Tabib, Jay K Kolls, Tullia C Bruno, Robert Lafyatis, et al. A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nature communications*, 10(1):1–10, 2019.

[16] Fangda Song, Ga Ming Angus Chan, and Yingying Wei. Flexible experimental designs for valid single-cell rna-sequencing experiments allowing batch effects correction. *Nature communications*, 11(1):1–15, 2020.

[17] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[18] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[19] H. Zhou and K. Lange. Mm algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19(3):645–665, 2010.

[20] Kenong Su, Tianwei Yu, and Hao Wu. Accurate feature selection improves single-cell rna-seq cell clustering. *Briefings in Bioinformatics*, 2021.

[21] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[22] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[23] Gerd Ronning. Maximum likelihood estimation of dirichlet distributions. *Journal of statistical computation and simulation*, 32(4):215–221, 1989.

[24] Bruce S Weir and William G Hill. Estimating f-statistics. *Annual review of genetics*, 36(1):721–750, 2002.

[25] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[26] Gideon Schwarz et al. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464, 1978.

[27] Zhe Sun, Ting Wang, Ke Deng, Xiao-Feng Wang, Robert Lafyatis, Ying Ding, Ming Hu, and Wei Chen. Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, 34(1):139–146, 2018.

[28] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21(1):1–32, 2020.