IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 10, 2020, accepted June 23, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3008681

LAK: Lasso and K-Means Based Single-Cell RNA-Seq Data Clustering Analysis

JIAO HUA[1], HONGKUN LIU[2], BOYANG ZHANG[1], AND SHUILIN JIN[1]
[1]School of Mathematics, Harbin Institute of Technology, Harbin 150001, China
[2]Network and Information Center, Ocean University of China, Qingdao 266100, China

Corresponding author: Shuilin Jin (jinsl@hit.edu.cn)

ABSTRACT The single-cell RNA sequencing provides a way to obtain marker genes of different cells, which lays the foundation for discovering new cell types. The general strategy of achieving this goal is to build a clustering pipeline and derive differentially expressed genes, followed by the cell type enrichment analysis and driving force analysis. Throughout the entire analysis process, clustering models and appropriate methods of dimension reduction are two vital and challenging tasks. In this study, we present a novel method LAK (a computational pipeline for single-cell RNA-seq data clustering analysis using Lasso and K-means based feature selection method) that can be applied to single-cell RNA-seq data by selecting the candidate genes. To deal with the sparse high-dimensional data, we integrated Lasso penalty into clustering method for single-cell RNA-seq data as the feature selection method, which extracts out the genes that have an actual effect on clustering. We also improved the parameter selection algorithm to search the appropriate parameters automatically by binary search according to the size of the data. Compared with other computational approaches, LAK obtains a better performance in reliability, stability, convenience and accuracy applied to the real datasets, the simulation data, and the datasets with a large number of dropout events.

INDEX TERMS Clustering analysis, Lasso, single-cell RNA-seq data.

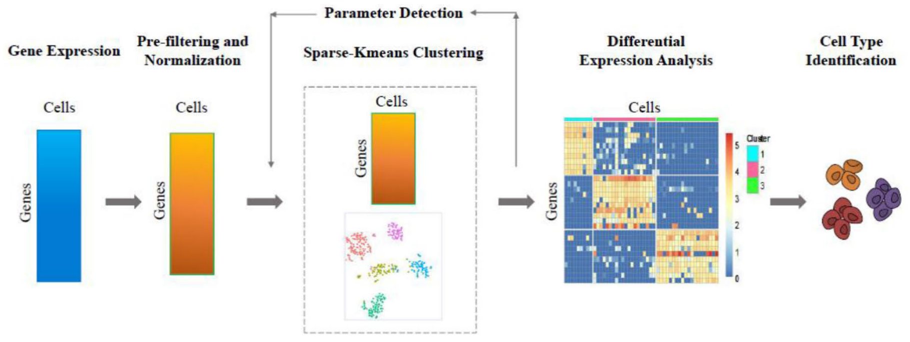
I. INTRODUCTION

The single-cell RNA sequencing (scRNA-seq) technology is now a powerful tool that demonstrates unprecedented precision in exploring biological processes and disease mechanisms. Recently, many works focus on the pathogenesis of the coronavirus disease 2019 (COVID-19) by the single-cell RNA sequencing technology [1]–[4]. By the single-cell RNA-seq analysis, somatic mutations at the individual cell levels and cell types in a sample are understood with high precision [5], [6]. The major advantage of scRNA-seq is that it enables unsupervised learning of population structure, and discovers the novel subtypes and rare cell species by dissecting complex and heterogeneous cell populations effectively [7]. Also, it facilitates a deeper understanding of cell heterogeneity [7].

In the single-cell RNA-seq data analysis, one of the relatively significant studies is unsupervised single-cell clustering analysis, which aims to cluster unknown cells of the sample into clusters using the cluster algorithm. The downstream analysis contains detection of the differentially expressed genes in each cluster and the enrichment analysis of cell types, from which samples are matched to real cell types and new cell subtypes discovered. Therefore, the clustering results are often applied to the downstream analysis, which will have a substantial impact on the final conclusion [8].

Considering the sparse and high-dimensional characteristics of the scRNA-seq data, many clustering methods are improved in terms of the execution time, the clustering accuracy, the detectability of small cell subtypes, and the data visualization compare with traditional cluster analysis methods [9]–[13]. However, the clusters of cells in scRNA-seq datasets still face statistical and computational challenges [14]. The main problem in the clustering analysis for scRNA-seq data is that they are so sparse and high-dimensional that most of the measurements are zero or near to zero. In addition, scRNA-seq data shows high differences in gene expression levels, even within the same cluster of cells. Based on these challenges, several clustering methods have been developed recently. For instance, the scDeep-Cluster applied the deep learning method of the model to

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.



(a)

Datasets	N	Units
Biase	49	FPKM
Treutlein	80	FPKM
Yan	90	RPKM
Goolam	124	CPM
Camp1	777	FPKM
Li	561	FPKM
Patel	430	TPM
Zeisel	3005	UMI

(b)

Datasets	REF	LAK	SC3	SINCERA
Biase	3	3	3	3
Treutlein	5	6	3	30
Yan	6	6	6	4
Goolam	5	6	6	4
Camp1	7	6	13	15
Li	9	7	11	7
Patel	5	5	17	10
Zeisel	9	3	30	8

(c)

FIGURE 1. Systemic pipeline of LAK. (a) Framework overview of the LAK. (b) Published datasets. N, the number of cells; RPKM, reads per kilo base of transcript per million mapped reads; RPM, reads per million mapped reads; FPKM, fragments per kilo base of transcript per million mapped reads; CPM, counts per million mapped reads. (c) The estimation of the cluster number *k*. REF, the number of clusters originally identified by the authors.

the cluster analysis of single-cell RNA-seq data [15]. The SC3 combines multiple clustering solutions through a consensus approach [16]. The BAMM-SC develops the Bayesian mixture model to cluster the scRNA-seq data [17]. The SOUP is a semisoft clustering method that classify both pure and intermediate clustering at the same time [18]. Besides, the SSCC is a clustering framework based on the random projection and the feature construction [19]. The pcaReduce based the hierarchical cluster analysis to generate a hierarchy of cell states where each cluster branch is associated with a changing principal component and is used to differentiate between two cell states [20]. The SINCERA is a complete single-cell clustering analysis pipeline that identify the cell types, identify the specific gene signatures of cell types and determine the driving forces of given cell types [21].

The Lasso [22] method was first used in the regression analysis. Recently, with the development of knowledge, the Lasso method has a wide range of applications, such as the gene set selection via lasso punishment regression (SLPR) method for the quantifying multiple linear regression [23], and the DropLasso method for learning a molecular signature from scRNA-seq data [24].

In this paper, we propose a computational pipeline to cluster single cell RNA-seq data (Fig. 1a). We introduce the Lasso penalty to the clustering process, which is suitable for the high dimension and sparse scRNA-seq data, improves the accuracy and high interpretive of the clustered results. We also improved the parameter selection algorithm to binary search the appropriate parameters automatically according to the size of the data. In addition, our cluster number estimation method is based on the Gap statistics [25], which is highly precise compared with other commonly used methods. We used eight published single cell datasets in the consolidation form [26] including the Biase [27], Treutlein [28], Yan [29], Goolam [30], Camp1 [31], Li [11], Patel [32], and Zeisel [33] datasets to demonstrate the higher accuracy and stability of the LAK method by comparing with the methods SC3, t-SNE [34] and hierarchical clustering, pcaReduce, and SINCERA. More detailed information about eight datasets are shown in Fig. 1b. In addition, we show that our method has higher stability than the SC3 using the simulation datasets with different dropout events.

II. METHODS

A. NORMALIZATION AND IDENTIFICATION OF THE CLUSTER NUMBER K

The LAK takes the Linnorm [35] as the default normalization method, and users can choose other normalization methods and pass the results to LAK as input. Although there are a number of methods to determine the number of clusters,

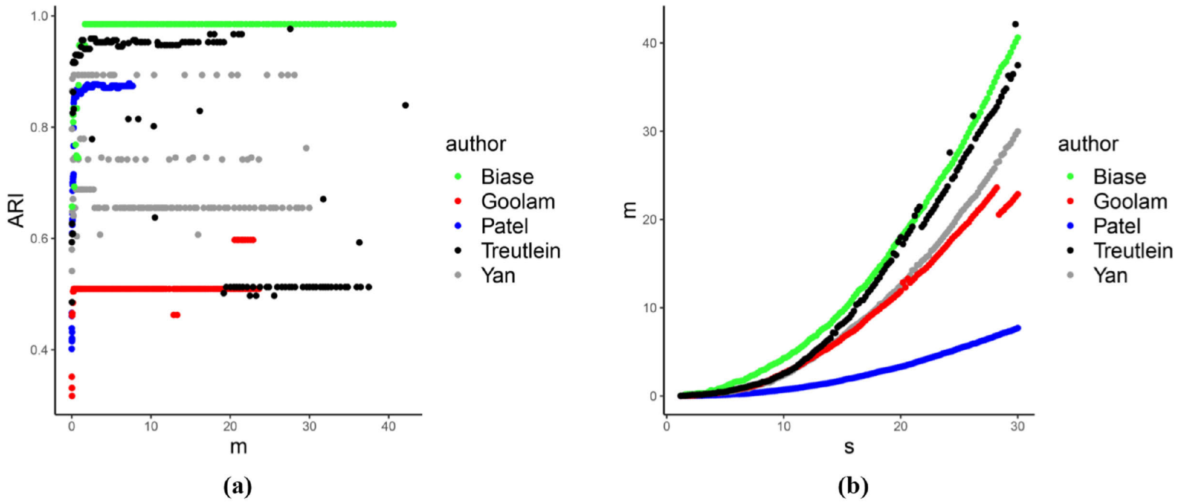


FIGURE 2. **Optimization of optimization parameter. *s* is the optimization parameter in our method, let *n* be the number of non-zero weighs genes, and *c* be the number of cells, then *m = n/c*. (a) ARI grow fast with the growth of *m*, but encountered a cliff-like fall on the dataset Yan and Treutlein. when *m* = 2, all datasets start to achieve high ARI values. (b) the *m* and *s* grow simultaneously in every dataset, demonstrating that the parameter *s* determines the number of non-zero weigh genes directly. We set the objective function *min(abs (m – 2))* to optimize the parameter *s*, because smaller value of the *m* means less non-zero genes, which can speed up the following analysis.**

there is no uniform standard to measure the unsupervised learning problem. We suggest that users should assist in subjective experience of their own data and draw up the preliminary category number range by themselves. Nevertheless, we incorporate the Gap statistics with the high precision and acceptable computing time into our pipeline. One of the advantages of Gap statistics is that in case of ambiguities, instead of just giving a simple arbitrary number, it can display graphics, and finally determine the clustering number by combining with subjective judgment. Let $\boldsymbol{M}$ be the result matrix of normalization, each row of $\boldsymbol{M}$ is a gene and each column is a sample.

B. FEATURE SELECTION AND CLUSTERING

For consistency consideration, we did not bring in any gene filtering methods to our pipeline nevertheless chose other ways to control quality. That is, the D.M's framework—sparse clustering [36] can effectively select out the genes that affect clustering, so no additional gene filtering method is needed, which is also a key advantage of LAK over other clustering methods.

The k-means clustering minimizes the within-cluster sum of squares (WCSS). Suppose that we wish to cluster n observations on p dimensions, i.e. genes. The k-means algorithm attempts to partition the n cells into K sets, or clusters, such that the WCSS is minimal, which is equal to maximize the between-cluster sum of squares (BCSS):

$$BCSS = \sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n} d_{i,l,j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,l\in C_k} d_{i,l,j}\right) \quad (1)$$

where p is the number of genes, n_k is the number of cells in cluster k, C_k contains the indices of the observations in cluster k, $d_{i,l,j}$ denotes the distance between cell i and cell l on gene j, this paper will take Euclidean distance.

The LAK takes the same Lasso and L_2 as penalty as the sparse k-means clustering algorithm [36], and the objective function is defined as follows:

$$\max_{w,C_1,\ldots,C_k}\left\{\sum_{j=1}^{p} w_j\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n} d_{i,l,j} - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,l\in C_k} d_{i,l,j}\right)\right\}$$
$$s.t.\ \|w\|^2 \leq 1,\quad \|w\|_1 \leq s,\ w_j \geq 0,\ \forall j \quad (2)$$

where s is an optimization parameter, and it directly determines the magnitude of dimension reduction and the validity of the final clustering results. The value of s plays a role in controlling the degree of feature selection, and the larger s tends to incorporate more features. From a mathematical point of view, when the sample data is much smaller than the variable dimension, the amount of information is far from enough to support effective analysis. Due to the limitation of technology and cost, the sample size of scRNA-seq data is often insufficient, which is also a key factor affecting the subsequent analysis of gene data.

Let n be the number of non-zero weighs genes, and c be the number of cells, then $m = n/c$. As shown in Fig. 2a, with the growth of m, ARI first grows fast and then encounter a cliff-like fall at some point. Fig. 2b shows the relationship between m and parameter s, which plays a decisive role in the degree of feature selection. Therefore, we designed an algorithm that selects optimization parameter s by binary search to make the number of non-zero genes closest to twice the number of cells, which automatically searches for the appropriate parameters according to the size of the data flexibly. Defined as follows:

```
Input:
  1, User-provided information: data matrix $\boldsymbol{M}$ (each row of $\boldsymbol{M}$ is a gene and each column is a cell); the number of cluster $k$.
  2, Default parameters: searching range $f = 5$, $u = 105$ (default $f$ means the floor bound, $u$ means the upper bound); default $e$ is the error limit of $s$ and set to be 0.1, and default threshold $t$ to stop searching is set to 0.1.
  3, Notation: the number of non-zero genes $g$; the number of cells $c$; the sparse k-means clustering algorithm $SK$.
  Algorithm to select the optimization parameter $s$
  $v = (f + u)/2$, run $SK$ with $s = v$, and get the number of non-zero weights, i.e. the number of genes $g$.
  While ($f \leq u$):
  If ($|g/c - 2| < t$): break.
  Else if ($g < 2c$): $f = v + e$, $v = (f + u)/2$, run $SK$ with $s = mid$, and get $g$.
  Else if ($g > 2c$): $u = v - e$, $v = (f + u)/2$, run $SK$ with s $= v$, and get $g$.
```

Considering extreme situations, if the number of cells is smaller than 50 or larger than 5000, we fixed the value of c in condition $|g/c - 2| < t$ to 50 or 5000, respectively.

C. ADJUSTED RAND INDEX

The adjusted rand index (ARI) [37] can be used to calculate the similarity between our method and the published clustering if cell-labels are available (for example, from a published dataset). ARI is defined as follows:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2}\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{n_{i\cdot}}{2} + \sum_{j}\binom{n_{\cdot j}}{2}\right] - \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2}\Big/\binom{n}{2}} \quad (3)$$

Given a set of n objects and two partitions of these objects, the overlap between the two partitions can be summarized in a contingency table, in which each n_{ij} denotes the number of objects in common between the two partitions.

D. DETECTION OF DIFFERENTIALLY EXPRESSED GENES

Firstly, the dataset was preprocessed, including normalizing the dataset with the Linnorm method in which the lines where all the genes express the value of 0 have been deleted. Secondly, our method is adopted for the clustering analysis (k refers to the clustering number that is produced by gap statistics). Meanwhile, the effective genes selected by the clustering method were used for the further analysis. Third, according to the clustering results, we calculated the p-value of the effective genes of each cluster through a one tailed Welch's t-test [38]. Fourth, we selected the differentially expressed genes according to the p-value of each group and the appropriate control conditions, and then we selected the first 10 highly differentially expressed genes in each cluster. Finally, the selected differentially expressed genes were visualized by the heatmap. With differentially expressed genes, we used some heatmaps to prove the validity of the LAK clustering results (Fig. 3d, Fig. S2 (Supplemental File 1)).

III. RESULTS

A. OVERVIEW OF LAK

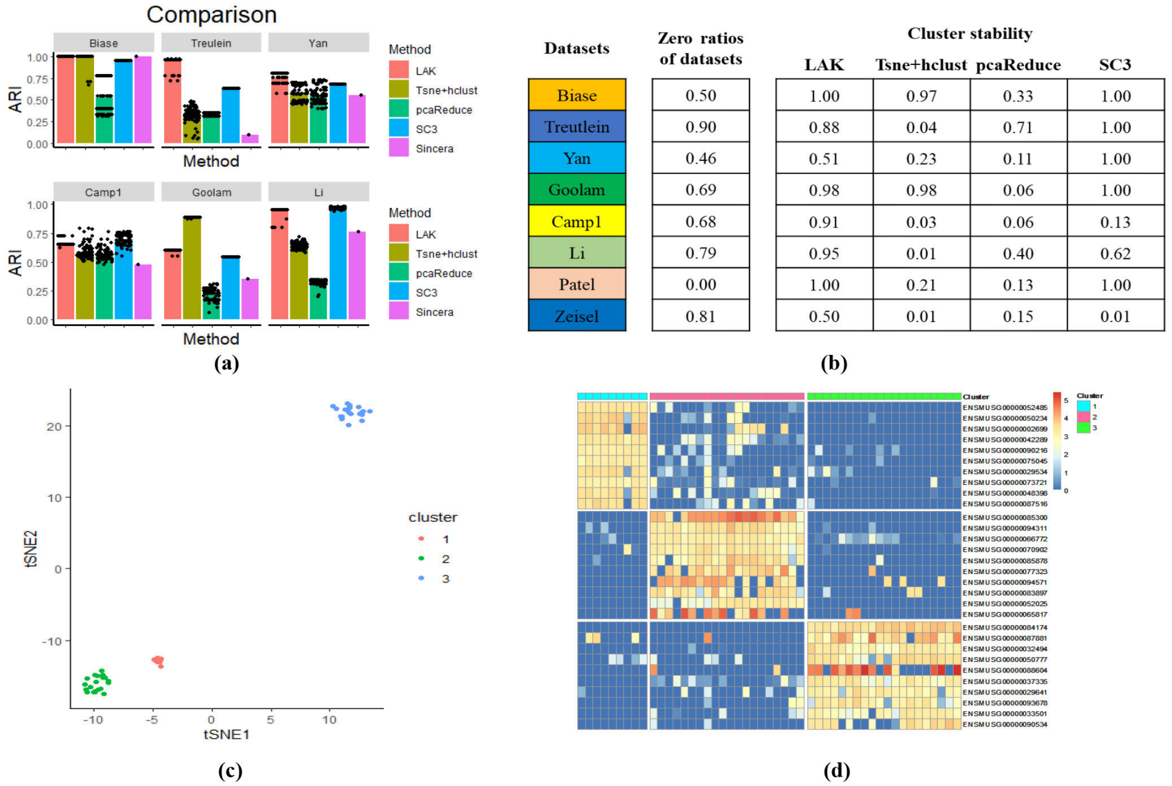
As the systemic pipeline indicated in Fig. 1a, for input single-cell RNA-seq data expression matrix, first, pre-filtering the genes by deleting the genes whose all expression value was 0 or close to 0, and the Linnorm normalizing transformation was used to transform the gene expression matrix into a linear model that does not need to go through the origin. Then, under an appropriate parameter of *clusGap* in R the Package cluster, the number of the clusters was determined, as well as the results of eight datasets in Fig. 1c. After that, in the clustering procedure, we chose to exploit a feature selection clustering algorithm based on the D.M's framework, which includes the Lasso penalty to the clustering process and is suited for high dimension data with high sparsity. Because the default parameter selection algorithm based on the original sparse k-means method tends to retain more unnecessary genes, we improved the parameter selection algorithm to binary search the appropriate parameters automatically according to the size of the data sets. Then, we implemented t-SNE as dimension reduction and visualizing technique used to visualize the clustering results. Finally, according to the clustering results, the differential expression analysis was performed by the one-tailed Welch's t-test or wilcoxon test and cell type identification.

B. BENCHMARKING

We applied the LAK method to the eight published single-cell RNA-seq datasets in consolidation form. Here, we also select four other methods – the SC3, t-SNE and hierarchical clustering, pcaReduce, and SINCERA, to benchmark our method. The LAK method performed better than the four tested methods across nearly all benchmark datasets with only a few exceptions. The results were evaluated by ARI and then visualized using the t-SNE 2D-plot.

As shown in Fig. 3a, the stochastic methods (the LAK method, t-SNE and hierarchical clustering, pcaReduce, and SC3) were applied 100 times to each dataset. SINCERA is deterministic and was run only once. Dots represent ARI values in each run and bars correspond to median ARI. The higher ARI value means the more accurate clustering results, where ARI value of 1 means that all cells are classified correctly. As shown in Fig. 3a, LAK method got higher ARI value in Biase, Treutlein, Yan and Goolam datasets than four other methods. Therefore, we concluded that LAK method performs better than the SC3, t-SNE and hierarchical clustering, pcaReduce, and SINCERA methods in the dataset Biase, Treutlein, Yan and Goolam.

Fig. 3b shows the cluster stability and the zero ratios of datasets. The results indicate that the zero ratios of the



Datasets	Zero ratios of datasets	Cluster stability			
		LAK	Tsne+hclust	pcaReduce	SC3
Biase	0.50	1.00	0.97	0.33	1.00
Treutlein	0.90	0.88	0.04	0.71	1.00
Yan	0.46	0.51	0.23	0.11	1.00
Goolam	0.69	0.98	0.98	0.06	1.00
Camp1	0.68	0.91	0.03	0.06	0.13
Li	0.79	0.95	0.01	0.40	0.62
Patel	0.00	1.00	0.21	0.13	1.00
Zeisel	0.81	0.50	0.01	0.15	0.01

FIGURE 3. Clustering results on eight datasets. (a) Benchmarking of our method against three other methods, with the optimization parameter *s* set to 10. (b) The cluster stability and zero ratios of dataset by the LAK, Tsne+hclust, pcaReduce, and SC3 methods. (c) The visualization of the clustering results of the dataset Biase. (d) Heatmap on dataset Biase. Differentially expressed genes were detected by the t-test and we selected the top 10 highly differential expressed genes of each cluster. The gaps separated clusters and genes.

Treutlein, Zeisel, and Li datasets are 90%, 81%, and 79% respectively, which means that dropouts are highly possible in these three datasets. The cluster stability is calculated by the frequency of the most frequently appearing solutions in the 100 results of each method run. Also, the results show that the stability of the LAK method is obviously higher than pcaReduce and t-SNE+hierarchical clustering methods from all datasets. Our method performs better in datasets Camp1, Li, and Zeisel than SC3, which got higher stability. In particular, the zero ratios of datasets Zeisel, Camp1 are 81% and 68%, the number of cells is 3005 and 777, and the stability of LAK and SC3 methods are 50%, 91% and 1%, 13% respectively on the datasets. Therefore, we infer that our LAK method performs better than the SC3 method in terms of stability on the larger and potentially more heavily dropouts of datasets.

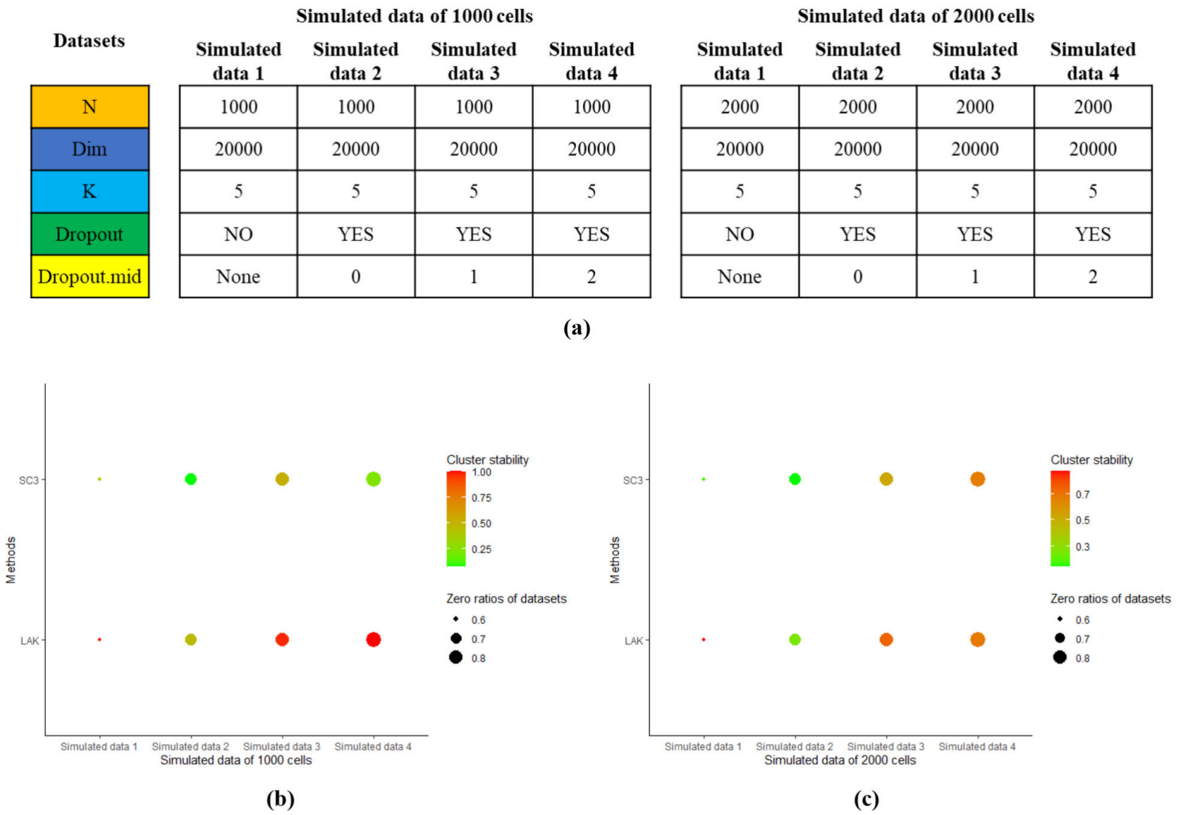
Fig. 3c shows the visualization of the clustering results obtained by our clustering method for the dataset Biase. It is concluded from the figures that our clustering results are highly precise, the distance between the type and the type is large, and the distance between cells and cells in each class is small. Our method plays best in the dataset Biase. Also, the visualizations of the datasets are shown in Fig. S1 (Supplemental File 1).

Fig. 3d is based on the seven datasets to display our differentially expressed genes through the heatmap, which intuitively shows the high discrepancy of the expressed genes selected by our method based on t-test. In particular, the differences of genes expressed were far more obvious in Biase, Goolam, and Yan datasets. In particular, the figures of datasets are shown in Fig. S2 (Supplemental File 1). Also, the differentially expressed gene names and control conditions shown in Supplemental Table S1.

C. SIMULATED SINGLE-CELL DATA

To better verify the stability of our clustering method, we implemented our clustering method and SC3 to perform clustering analysis on 8 sparse high-dimensional simulated single-cell datasets that suffered from different degrees of dropout events. We studied four simulated single-cell datasets of 1000 cells and four simulated single-cell datasets of 2000 cells, all of which were divided into five clusters and set to influence by different degrees dropout events. The details are shown in Fig. 4a. We applied our method to eight simulated single-cell datasets and the results were assessed by ARI.

As shown in Fig. 4b, there are the cluster stability and zero ratios of simulated single-cell datasets of 1000 cells by



Datasets	Simulated data of 1000 cells				Simulated data of 2000 cells			
	Simulated data 1	Simulated data 2	Simulated data 3	Simulated data 4	Simulated data 1	Simulated data 2	Simulated data 3	Simulated data 4
N	1000	1000	1000	1000	2000	2000	2000	2000
Dim	20000	20000	20000	20000	20000	20000	20000	20000
K	5	5	5	5	5	5	5	5
Dropout	NO	YES	YES	YES	NO	YES	YES	YES
Dropout.mid	None	0	1	2	None	0	1	2

(a)

(b)

(c)

FIGURE 4. **Clustering results on simulated single-cell data. (a) Detailed information on simulated single-cell data. N, the number of cell samples. Dim, the genetic dimension. K, the number of clusters specified to the simulate single-cell data. Dropout, whether the simulated data contains dropouts. NO, means there is no dropouts for datasets. YES, means there is dropouts for datasets. Dropout.mid, parameter control the point at which the probability is equal to 0.5. (b), (c) The cluster stability and zero ratios of simulated single-cell datasets compare the LAK method with the SC3 method. The color of points means the cluster stability. The size of points means the zero ratios of datasets.**

LAK and SC3 methods. The results show that the stability of our method is significantly better than the SC3 with different amounts of dropouts. Especially, the stability of the SC3 method is lower in simulated data 4 with a large number of dropouts, while the stability of SC3 method is higher. The stability of SC3 method is lower in simulated data 1 with no dropouts, while the stability of our method is higher. Fig. 4c shows that the stability of the LAK and SC3 methods are high on the four datasets, among which the stability of the LAK method is significantly higher than the stability of the SC3 method. To sum up, we conclude that our method is more stable than the SC3 method on most large datasets, including the large datasets with a large number of dropouts.

D. MATCHING WITH MARKER GENES AND CELL TYPEES IDENTIFICATION

To validate the accuracy of our method from a biological point of view, we also ran the LAK on the Zeisel dataset. The cluster number is 9, which is provided by the author. Also, the rest of our pipeline was used in the following step. LAK selected 5819 genes among all 18879 genes, and we only used LAK selected genes in the differentially expressed genes analysis.

Based on the above clustering results on the Zeisel dataset, we identified differentially expressed genes for each cluster using the one-tailed Welch's t-test. This test is commonly utilized when the algorithm in the two groups can be assumed by two independent normal distributions. Here, we used it to test the hypothesis that a specific gene has the same mean expression in the cells of two different groups. For the Cluster c, we calculated the p-values of each gene based on its expression between the cells in the Cluster c and not in. We selected the top 100 genes with the minimum p-value, as differentially expressed genes for every cluster.

Comparing our differentially expressed genes with the marker genes provided by the author (*Thy1*, *Gad1*, *Tbr1*, *Spink8*, *Mbp*, *Aldoc*, *Aif1*, *Cldn5*, *Acta2*), we find that most of our clusters can be matched up with the unique marker provided by the author. The Cluster 6 shares the same marker *Mbp* with the Cluster 2, and the marker *Acta2* has no matching cluster. We suppose that we may have a more sensitive result, the Cluster 2 and 6 could be classified as one. Except for these, each of our clusters has a one-to-one match with a marker, showing a high precision and the potential for investigating the underlying biological meaning of our pipeline.

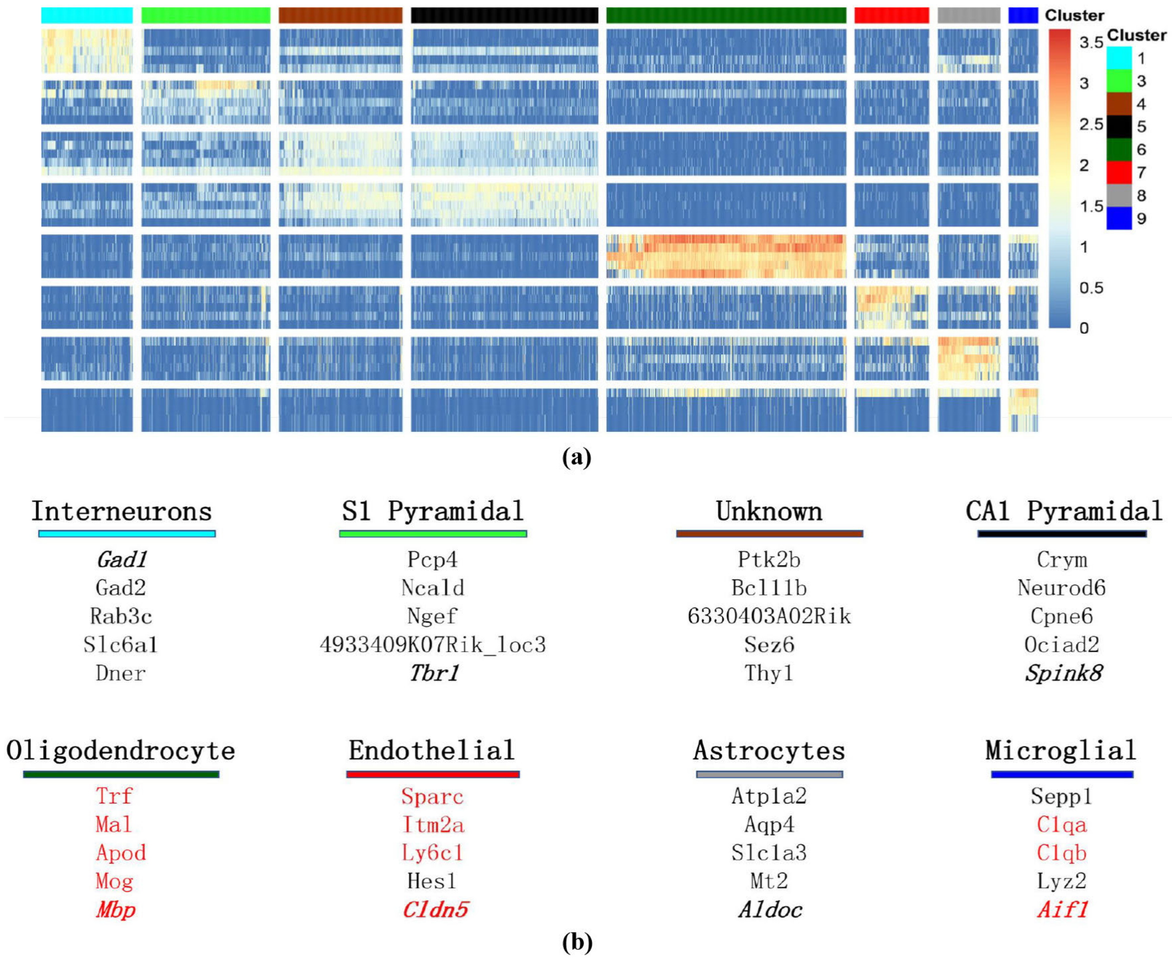


FIGURE 5. (a) Heatmap of clusters and marker genes found by the LAK. The Cluster 2 has no marker gene in our filtering condition, so its results are not shown. Only the top 5 highly expressed marker genes are shown for each cluster (except the Cluster 2). The fifth gene of each cluster is replaced by one of the marker genes (*Thy1*, *Gad1*, *Tbr1*, *Spink8*, *Mbp*, *Aldoc*, *Aif1*, *Cldn5*, *Acta2*) provided by Zeisel. Top 5 marker genes list of the Cluster 1 already includes *Gad1*, so its fifth gene *Dner* is not replaced. *Acta2* is not in the top 100 DE gene lists of any cluster. (b) Cell types identified by the LAK corresponding to each cluster. Lines under the cell types indicates cluster with unique color. Marker genes provided by Zeisel are indicated with bold italic types, and genes in red are the known markers we collected from literature. The Cluster 4 has no marker gene in literature, and the marker *Thy1* provided by the Zeisel was the marker of interneurons, the S1 pyramidal and CA1 pyramidal at the same time, so we assigned the Cluster 4 as Unknown due to the lack of marker information. Based on the heat map above, the Cluster 4 shares similar marker region with the Cluster 5, so we think its cell type should be close to the CA1 pyramid.

In each DE gene list, we chose genes with the most significant difference as marker genes. The median and mean expression of marker genes in the Cluster c is 10 times and 3 times more than that not in the Cluster c, respectively. Under this condition, the Cluster 2 did not find a corresponding marker gene, so its results won't be included. Fig. 5a shows the top 5 highly expressed marker genes found by LAK for each cluster. The *Apoe* appeared in two marker gene lists of different clusters (8 and 9), so we removed it. The fifth gene of each cluster was replaced by a marker gene provided by the author, except the Cluster 1, which already included the *Gad1*.

In addition to that, we consulted some literatures [39]–[42] and found that there were many coincidences between markers found by the LAK and known, so we assigned the clusters to the known cell types accordingly. Specifically, we assigned the Cluster 1 as interneurons by the *Gad1* [41]; the Cluster 3 as the S1 pyramidal by the *Tbr1* [33]; the Cluster 5 as the CA1 pyramidal by the *Spink8* [33]; the Cluster 6 as the oligodendrocyte by the *Trf*, *Mal*, *Mbp*, *Mog* and *Apod* [33], [39], [41]; the Cluster 7 as the endothelial cells by the *Sparc*, *Itm2a*, *Ly6c1* and *Cldn5* [33], [39], [40]; the Cluster 8 as the astrocytes by the *Aldoc* [33]; the Cluster 9 as the microglial cell by the *C1qa*, *C1qb*, and *Aif1* [33], [40], [42]. In the Cluster 4, we did not find any known cell types marker supported by literature among the first five filtered differentially expressed genes, so we assigned the cluster 5 as unknown. Based on the heat map (Fig. 5a), the Cluster 4 shares a similar marker region with the Cluster 5, so we think its cell type should be close to the CA1 pyramid. The integrated results are shown in Fig. 5b.

E. MATCHING WITH KNOWN TISSUES

To verify the accuracy of our clustering method from the perspective of actual data analysis, we ran the LAK algorithm on a new dataset that composed by 5 scRNA-seq datasets of different human tissues, and these data are sequenced by independent experiments. We respectively obtained neurons cells from the Darmanis [43], embryo cells developed from the Yan, liver cells from the Camp, pancreas cells from the Xin [44], and blood cells from the Pollen [45]. We took 100 samples (if the original data samples less than 100, take the original sample size; if the original data samples more than 100, take 100 samples) for each dataset, and merged them into a new dataset. We only keep genes that are shared in different datasets. Then the LAK algorithm with the cluster number set to 5 was implemented on this new dataset.

Based on the clustering results of the new dataset, we tested the accuracy of LAK by calculating the ARI value. Suppose s_0 is the parameter value automatically selected by the LAK algorithm when run for the first time, and we then ran the LAK algorithm 100 times fixing s to s_0, and calculated the ARI values according to the real cell types. These 101 ARI values had two different results of 0.995 (84 times) and 0.871 (17 times) (Fig. S3 shows in Supplemental File 1). Therefore, we can conclude that the LAK method has high accuracy.

IV. DISCUSSION AND CONCLUSIONS

In this study, we presented a novel algorithm for scRNA-seq analysis that determines the appropriate number of clusters and separate single cells into distinct groups. To certify the accuracy of our method, we applied our pipeline to eight publicly available datasets, and we estimated the consistency of the original author's results and ours by calculating the ARI values. In addition to that, we also implemented our pipeline to the Zeisel dataset to validate our method by comparing the differentially expressed genes in our clusters and marker genes provided by the author. The high ARI values when the LAK was implemented in a new dataset sampled from different human tissues also suggest that our method is high in accuracy.

Effective analysis of high-dimensional and sparse scRNA-seq data requires an efficient information extraction algorithm, which is also the key to determine the accuracy of clustering. The dimension reduction methods, such as the PCA and T-SNE, mix features, and the data after dimension reduction cannot correspond with the original gene, which makes the results difficult to explain. Our method solves this problem to some extent and also provides some methodological insights. The LAK takes an adaptive feature subset selection algorithm and produces interpretable results, that is, making it clear which genes have a decisive impact on the inter-cluster differences.

One of the current difficulties of single-cell clustering research is how to determine the cluster number k. It is impossible to carry out the further work with uncertain cluster number for most of the currently popular methods. Besides, the inaccurate estimation of the k will also make the clustering result ambiguous and extremely unstable. The LAK adopted the Gap statistics as a method to determine the cluster number, which has some randomness. The single-cell RNA-seq data needs not only a more stable method for determining the number of clusters of the high-dimensional data, but also a clustering algorithm that reduces the sensitivity to different initial cluster numbers. Even the wrong number of clusters should not throw too bad results. For example, when the clustering algorithm takes a larger cluster number, the same type of cells should be subdivided into clusters, and different types of cells should not be grouped.

As single-cell datasets have become larger and larger over time, in theory, more cells could provide us with more information, which should be an opportunity to design better clustering algorithms. However, this rapid growth of data volume will seriously challenge numerous previous popular methods, causing problems such as the slow convergence speed, insufficient memory, and a decline in accuracy. Our method may have similar problems. For example, when the number of cells exceeds 10,000, the calculation of the gap statistics will become quite slow if run on a single personal computer. Distributed computing frameworks, such as the Hadoop and Spark [46], may be helpful in solving this problem. After the number of sequenced cells in one dataset increases to 100,000, the distributed computation may be the only solution.

APPENDIX

Figure S1-S3 is available at Supplemental File 1. Supplemental Table S1 is available at Supplemental Table 1. In addition, our algorithm is implemented with R. All scripts to the figures in this paper are available in the GitHub repository (https://github.com/HIT-biostatistical/LAK).

ACKNOWLEDGMENT

(Jiao Hua and Hongkun Liu contributed equally to this work.)

REFERENCES

[1] H. Cai, "Sex difference and smoking predisposition in patients with COVID-19," *Lancet Respiratory Med.*, vol. 8, no. 4, p. e20, Apr. 2020.

[2] W. Sungnak, N. Huang, C. Bécavin, M. Berg, R. Queen, M. Litvinukova, C. Talavera-López, H. Maatz, D. Reichart, F. Sampaziotis, K. B. Worlock, M. Yoshida, and J. L. Barnes, "SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes," *Nature Med.*, vol. 26, no. 5, pp. 681–687, Apr. 2020.

[3] M. Liao, Y. Liu, J. Yuan, Y. Wen, G. Xu, J. Zhao, L. Cheng, J. Li, X. Wang, F. Wang, and L. Liu, "Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19," *Nature Med.*, vol. 26, pp. 842–844, May 2020.

[4] C. G. K. Ziegler, S. J. Allon, S. K. Nyquist, I. M. Mbano, V. N. Miao, C. N. Tzouanas, Y. Cao, A. S. Yousif, J. Bals, B. M. Hauser, and J. Feldman, "SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues," *Cell*, vol. 181, no. 5, pp. 1016.e19–1035.e19, May 2020.

[5] K. Q. Lao, F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, B. Tuch, J. Bodeau, and A. Siddiqui, ''mRNA-sequencing whole transcriptome analysis of a single cell on the SOLiD system,'' *J. Biomol. Techn.*, vol. 20, no. 5, pp. 266–271, Dec. 2009.

[6] E. Shapiro, T. Biezuner, and S. Linnarsson, ''Single-cell sequencing-based technologies will revolutionize whole-organism science,'' *Nature Rev. Genet.*, vol. 14, no. 9, pp. 618–630, Jul. 2013.

[7] T. Kalisky, S. Oriel, T. H. Bar-Lev, N. Ben-Haim, A. Trink, Y. Wineberg, I. Kanter, S. Gilad, and S. Pyne, ''A brief review of single-cell transcriptomic technologies,'' *Briefings Funct. Genomics*, vol. 17, no. 1, pp. 64–76, Jan. 2018.

[8] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, ''Challenges in unsupervised clustering of single-cell RNA-seq data,'' *Nature Rev. Genet.*, vol. 20, no. 5, pp. 273–282, Jan. 2019.

[9] D. Sinha, A. Kumar, H. Kumar, S. Bandyopadhyay, and D. Sengupta, ''DropClust: Efficient clustering of ultra-large scRNA-seq data,'' *Nucleic Acids Res.*, vol. 46, no. 6, p. e36, Jan. 2018.

[10] M. Barron, S. Zhang, and J. Li, ''A sparse differential clustering algorithm for tracing cell type changes via single-cell RNA-sequencing data,'' *Nucleic Acids Res.*, vol. 46, no. 3, p. e14, Feb. 2018.

[11] H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan, M. Wong, P. J. Choi, L. J. K. Wee, A. M. Hillmer, I. B. Tan, P. Robson, and S. Prabhakar, ''Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors,'' *Nature Genet.*, vol. 49, no. 5, pp. 708–718, Mar. 2017.

[12] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, ''Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning,'' *Nature Methods*, vol. 14, no. 4, pp. 414–416, Mar. 2017.

[13] L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan, ''GiniClust: Detecting rare cell types from single-cell gene expression data with gini index,'' *Genome Biol.*, vol. 17, no. 1, p. 144, Jul. 2016.

[14] O. Stegle, S. A. Teichmann, and J. C. Marioni, ''Computational and analytical challenges in single-cell transcriptomics,'' *Nature Rev. Genet.*, vol. 16, no. 3, pp. 133–145, Jan. 2015.

[15] T. Tian, J. Wan, Q. Song, and Z. Wei, ''Clustering single-cell RNA-seq data with a model-based deep learning approach,'' *Nature Mach. Intell.*, vol. 1, no. 4, pp. 191–198, Apr. 2019.

[16] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, and A. R. Green, ''SC3: Consensus clustering of single-cell RNA-seq data,'' *Nature Methods*, vol. 14, no. 5, pp. 483–486, 2017.

[17] Z. Sun, L. Chen, H. Xin, Y. Jiang, Q. Huang, A. R. Cillo, T. Tabib, J. K. Kolls, T. C. Bruno, R. Lafyatis, D. A. A. Vignali, K. Chen, Y. Ding, M. Hu, and W. Chen, ''A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies,'' *Nature Commun.*, vol. 10, no. 1, pp. 1–10, Apr. 2019.

[18] L. Zhu, J. Lei, L. Klei, B. Devlin, and K. Roeder, ''Semisoft clustering of single-cell data,'' *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 2, pp. 466–471, Jan. 2019.

[19] X. Ren, L. Zheng, and Z. Zhang, ''SSCC: A novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data,'' *Genomics, Proteomics Bioinf.*, vol. 17, no. 2, pp. 201–210, Apr. 2019.

[20] J. Žurauskienė and C. Yau, ''PcaReduce: Hierarchical clustering of single cell transcriptional profiles,'' *BMC Bioinf.*, vol. 17, no. 1, p. 140, Mar. 2016.

[21] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, ''SINCERA: A pipeline for single-cell RNA-seq profiling analysis,'' *PLOS Comput. Biol.*, vol. 11, no. 11, Nov. 2015, Art. no. e1004575.

[22] R. Tibshirani, ''Regression shrinkage and selection via the lasso,'' *J. Roy. Stat. Soc., B Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[23] H. R. Frost and C. I. Amos, ''Gene set selection via LASSO penalized regression (SLPR),'' *Nucleic Acids Res.*, vol. 45, no. 12, p. e114, May 2017.

[24] B. Khalfaoui and J.-P. Vert, ''DropLasso: A robust variant of lasso for single cell RNA-seq data,'' 2018, *arXiv:1802.09381*. [Online]. Available: http://arxiv.org/abs/1802.09381

[25] R. Tibshirani, G. Walther, and T. Hastie, ''Estimating the number of clusters in a data set via the gap statistic,'' *J. Roy. Stat. Soc., B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, May 2001.

[26] V. Y. Kiselev, A. Yiu, and M. Hemberg, ''Scmap: Projection of single-cell RNA-seq data across data sets,'' *Nature Methods*, vol. 15, no. 5, pp. 359–362, Apr. 2018.

[27] F. H. Biase, X. Cao, and S. Zhong, ''Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing,'' *Genome Research.*, vol. 24, no. 11, pp. 1787–1796, Aug. 2014.

[28] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, ''Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq,'' *Nature*, vol. 509, no. 7500, pp. 371–375, Apr. 2014.

[29] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang, ''Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells,'' *Nature Struct. Mol. Biol.*, vol. 20, no. 9, pp. 1131–1139, Aug. 2013.

[30] M. Goolam, A. Scialdone, S. J. L. Graham, I. C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J. C. Marioni, and M. Zernicka-Goetz, ''Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos,'' *Cell*, vol. 165, no. 1, pp. 61–74, Mar. 2016.

[31] J. G. Camp, K. Sekine, T. Gerber, H. Loeffler-Wirth, H. Binder, M. Gac, S. Kanton, J. Kageyama, G. Damm, D. Seehofer, L. Belicova, M. Bickle, R. Barsacchi, R. Okuda, E. Yoshizawa, M. Kimura, H. Ayabe, H. Taniguchi, T. Takebe, and B. Treutlein, ''Multilineage communication regulates human liver bud development from pluripotency,'' *Nature*, vol. 546, no. 7659, pp. 533–538, Jun. 2017.

[32] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suva, A. Regev, and B. E. Bernstein, ''Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma,'' *Science*, vol. 344, no. 6190, pp. 1396–1401, Jun. 2014.

[33] A. Zeisel, A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, ''Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq,'' *Science*, vol. 347, no. 6226, pp. 1138–1142, Mar. 2015.

[34] L. van der Maaten and G. Hinton, ''Visualizing data using t-SNE,'' *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[35] S. H. Yip, P. Wang, J.-P.-A. Kocher, P. C. Sham, and J. Wang, ''Linnorm: Improved statistical analysis for single cell RNA-seq expression data,'' *Nucleic Acids Res.*, vol. 45, no. 22, p. e179, Sep. 2017.

[36] D. M. Witten and R. Tibshirani, ''A framework for feature selection in clustering,'' *J. Amer. Statistical Assoc.*, vol. 105, no. 490, pp. 713–726, 2010.

[37] L. Hubert and P. Arabie, ''Comparing partitions,'' *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[38] B. L. Welch, ''The generalization of 'student's' problem when several different population variances are involved,'' *Biometrika.*, vol. 34, nos. 1–2, pp. 28–35, Jan. 1947.

[39] A. B. Rosenberg, C. M. Roco, R. A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. T. Graybuck, D. J. Peeler, S. Mukherjee, W. Chen, S. H. Pun, D. L. Sellers, B. Tasic, and G. Seelig, ''Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding,'' *Science*, vol. 360, no. 6385, pp. 176–182, Apr. 2018.

[40] P. T. Shah, J. A. Stratton, M. G. Stykel, S. Abbasi, S. Sharma, K. A. Mayr, K. Koblinger, P. J. Whelan, and J. Biernaskie, ''Single-cell transcriptomics and fate mapping of ependymal cells reveals an absence of neural stem cell function,'' *Cell*, vol. 173, no. 4, pp. 1045–1057, May 2018.

[41] U. Wilhelmsson, D. Andersson, Y. de Pablo, R. Pekny, A. Ståhlberg, J. Mulder, N. Mitsios, T. Hortobágyi, M. Pekny, and M. Pekna, ''Injury leads to the appearance of cells with characteristics of both microglia and astrocytes in mouse and human brain,'' *Cerebral Cortex*, vol. 27, no. 6, pp. 3360–3377, Apr. 2017.

[42] E. W. Ye, P. Lin, Y. Zuo, X. Li, and W. Hong, ''Detecting activated cell populations using single-cell RNA-seq,'' *Neuron*, vol. 96, no. 2, pp. 313–329, Oct. 2017.

[43] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. H. Gephart, B. A. Barres, and S. R. Quake, ''A survey of human brain transcriptome diversity at the single cell level,'' *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 23, pp. 7285–7290, Jun. 2015.

[44] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada, ''RNA sequencing of single human islet cells reveals type 2 diabetes genes,'' *Cell Metabolism*, vol. 24, no. 4, pp. 608–615, Oct. 2016.

[45] A. A. Pollen *et al.*, "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature Biotechnol.*, vol. 32, no. 10, pp. 1053–1058, Aug. 2014.

[46] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," *HotCloud*, vol. 10, no. 10, p. 95, Jun. 2010.

JIAO HUA was born in Henan, China, in 1996. She is currently pursuing the master's degree in applied statistics with the Harbin Institute of Technology, China. Her research interes includes the field of bioinformatics.

HONGKUN LIU was born in Liaoning, China, in 1981. He received the master's degree in computer science and technology from Jilin University, in 2006. He is currently engaged in data analysis related work in the Network and Information Center, Ocean University of China.

BOYANG ZHANG was born in Beijing, China. He received the M.Sc. degree in applied statistics from the Harbin Institute of Technology, in 2019.

SHUILIN JIN was born in Hebei, China, in 1980. He received the B.Sc. degree in probability and statistics from Jilin University, Jilin, in 2003, and the D.Sc. degree in mathematics from Jilin University, in 2009. His research interests include bioinformatics, computational biology, and other related fields.

In 2009, he became a Postdoctoral Fellow at the Harbin Institute of Technology, where he is currently a Professor. In 2013 and 2019, he visited Harvard University. He has presided over the National Natural Science Foundation of China, Youth Fund projects, and so on. He has published more than 30 high-level academic papers in Proceedings of the National Academy of Sciences of the United States of America such as PNAS, nucleic research, molecular neurobiology, and BMC bioinformatics, among which one was selected as an ESI highly-cited paper with an H factor of 11.

Dr. Jin is currently a Reviewer of the Math Review, a Reviewer of the Zentralblatt Math, a communication reviewer of various journals, and a Communication Reviewer of the Natural Science Foundation.