TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 15/15 pp772–789
DOI: 10.26599/TST.2020.9010028
Volume 26, Number 5, October 2021

# A Data-Driven Clustering Recommendation Method for Single-Cell RNA-Sequencing Data

Yu Tian, Ruiqing Zheng, Zhenlan Liang, Suning Li, Fang-Xiang Wu, and Min Li*

**Abstract:** Recently, the emergence of single-cell RNA-sequencing (scRNA-seq) technology makes it possible to solve biological problems at the single-cell resolution. One of the critical steps in cellular heterogeneity analysis is the cell type identification. Diverse scRNA-seq clustering methods have been proposed to partition cells into clusters. Among all the methods, hierarchical clustering and spectral clustering are the most popular approaches in the downstream clustering analysis with different preprocessing strategies such as similarity learning, dropout imputation, and dimensionality reduction. In this study, we carry out a comprehensive analysis by combining different strategies with these two categories of clustering methods on scRNA-seq datasets under different biological conditions. The analysis results show that the methods with spectral clustering tend to perform better on datasets with continuous shapes in two-dimension, while those with hierarchical clustering achieve better results on datasets with obvious boundaries between clusters in two-dimension. Motivated by this finding, a new strategy, called QRS, is developed to quantitatively evaluate the latent representative shape of a dataset to distinguish whether it has clear boundaries or not. Finally, a data-driven clustering recommendation method, called DDCR, is proposed to recommend hierarchical clustering or spectral clustering for scRNA-seq data. We perform DDCR on two typical single cell clustering methods, SC3 and RAFSIL, and the results show that DDCR recommends a more suitable downstream clustering method for different scRNA-seq datasets and obtains more robust and accurate results.

**Key words:** single-cell RNA-sequencing (scRNA-seq); cellular heterogeneity; cell type identification; data latent shape; clustering

## 1 Introduction

Cells can be considered as the fundamental units of living organisms[1]. The construction of a comprehensive cell atlas would help researchers list all cell types in human bodies, identify where the cells are located and distinguish different states and developmental stages of cells. The atlas would further help to identify biological markers and signatures for diseases and provide a better understanding on the basis of system biology. To achieve this goal, the Human Cell Atlas Project[2, 3] is proposed, which focuses on constructing a reference map of all human cell types. Furthermore, cellular heterogeneity is a prerequisite for maintaining the development of a biological system, regulating homeostasis and responding to external perturbations[4]. In the process of analyzing cellular heterogeneity[5], the single-cell RNA-sequencing (scRNA-seq) has become one of the most powerful techniques[6–8].

• Yu Tian, Ruiqing Zheng, Zhenlan Liang, Suning Li, and Min Li are with the Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China. E-mail: tian_yu@csu.edu.cn; rqzheng@csu.edu.cn; liangzhenlan@csu.edu.cn; suninglsn@163.com; limin@mail.csu.edu.cn.

• Fang-Xiang Wu is with the Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada. E-mail: faw341@mail.usask.ca.

† Yu Tian and Ruiqing Zheng contribute equally to this paper.

∗ To whom correspondence should be addressed.

Manuscript received: 2021-03-03; accepted: 2021-03-23

Recently, a lot of scRNA-seq clustering methods have been proposed to identify cell types. These methods apply different preprocessing strategies such as similarity learning[9], dropout imputation[10], and dimensionality reduction[11]. Specifically, dropout events caused by failures in amplification in the RNA-seq experiment have become a prominent problem in scRNA-seq data analysis. To deal with this phenomenon, Lin et al.[10] incorporated the dropout imputation with a robust weighted distance calculation strategy. Instead of treating all the zero counts as dropout entries, scImpute[12] only imputed the one with a high dropout probability by referring to the expression levels in other similar cells. Additionally, dimensionality reduction[13] plays an important role as one of the characteristics of scRNA-seq data is of high dimensionality. Becht et al.[14] proposed UMAP, which is based on the manifold theory and topological data analysis. It can preserve the global structure in a superior runtime. ZIFA[11] built a latent variable model by incorporating the presence of zero-inflation with a factor analysis framework. These two dimensionality reduction methods can be applied with a classic clustering algorithm to partition cells, and the performances have been proved to be good on scRNA-seq data. Gene selection is also significant, and Guo et al.[15] designed SINCERA to solve this problem. The method selected genes in the expression matrix based on the abundancy and selectivity of gene expression. Moreover, a critical step in clustering is to learn an accurate cell-to-cell similarity matrix. Jiang et al.[16] obtained a dissimilarity matrix by using the gene differential pattern among all cell pairs to construct the differentiability correlation between cells. TCC-based clustering[17] adopted the Jensen–Shannon distance to build an affinity matrix according to the transcript-compatibility count quantification. RAFSIL[18] defined the similarity by counting the frequency of two cells falling into the same leaf in a random forest classifier. In addition, subspace clustering[19, 20] has also been successfully applied in cell type identification. SinNLRR[21] and AdaptiveSSC[22] both used subspaces to learn the similarity between cells. Butler et al.[23] identified the highly variable features and constructed a KNN graph based on the Euclidean distance in latent spaces, and the edge weights between any two cells were defined based on the Jaccard similarity. Furthermore, in order to improve the robustness and generalization ability of clustering, a series of ensembled methods have been proposed. A multi-kernel based similarity learning strategy named SIMLR[9] was proved to have good performance on cell partitions. Based on SIMLR, MPSSC[24] learned a new similarity matrix by imposing a sparse structure on the doubly stochastic affinity matrix. SC3[25] assembled multiple clustering results obtained based on different (dis)similarity measurements and dimensionality reductions, and the results were used to calculate a consensus matrix. SAME-clustering[26] combined a maximally diverse subset of four clustering solutions obtained from five individual clustering methods, then the subset was combined with the expectation-maximization (EM) algorithm to build an ensemble clustering solution. Among all these methods, we find that hierarchical clustering[10, 15, 16, 18, 25, 27–29] and graph-based clustering[30–34] such as spectral clustering and Louvain community detection algorithm are the most popular approaches in the downstream clustering analysis[9, 12, 21–24, 35]. Additionally, density-based clustering is also widely used in scRNA-seq data analysis for the identification of outlier cells[36, 37]. Here, we choose several classic clustering methods that are popularly applied in scRNA-seq clustering for a correlation analysis. Based on the result, hierarchical clustering and spectral clustering are selected for the follow-up experiments.

In this study, we carry out a comprehensive analysis and combine visualization to compare hierarchical clustering and spectral clustering on scRNA-seq datasets under different biological conditions. Results show that the preprocessing strategies with spectral clustering tend to perform better on datasets with continuous shapes in two-dimension (we would use continuous shapes as the simplified representation), while those with hierarchical clustering achieve better results on datasets with obvious boundaries between clusters in two-dimension (we would use classification structures as the simplified representation). Based on this finding, a new strategy is developed to quantitatively evaluate the latent representative shape (we use QRS as the simplified representation) of an scRNA-seq dataset to distinguish whether it has clear boundaries or not. A data-driven clustering recommendation method, called DDCR, is proposed to recommend hierarchical clustering or spectral clustering for scRNA-seq data. We perform DDCR on two typical single cell clustering methods, SC3 and RAFSIL. The results show that DDCR recommends a more suitable downstream clustering method for different scRNA-seq datasets, and the recommendation improves the overall results of cell type identification.

# 2 Method

## 2.1 Datasets

In this study, we collect 12 well-annotated scRNA-seq datasets from AarryExpress[38] and GEO database[39] and carry out a comparative analysis on the collected datasets. The 12 scRNA-seq datasets range from hundreds to thousands in size and are classified into two categories according to their biological backgrounds. Specifically, datasets like T cells and B cells in lymphocytes[40] would be identified as cells containing specific functional subsets, while datasets with a dynamic development process from individual stem cells to multiple lineages[41] would be identified as cells undergoing differentiation. Here, we collect six datasets with cells undergoing differentiation and six datasets with cells containing specific functional subsets. Furthermore, different units are used to compute the gene expression values, such as fragments per kilobase of transcript per million mapped reads (FPKM) and transcripts per kilobase of transcript per million mapped reads (TPM). The cell type labels of each dataset obtained and validated from the prior biological studies are used as pre-annotations to evaluate the performances of the comparative analysis. Here, we consider the cell labels as gold standards if the cells are from different stages or lines, while the labels assigned by other techniques such as the computational methods are considered as silver standards[25]. Details of the datasets are described in Table 1.

## 2.2 Comparative analysis

### 2.2.1 Correlation analysis of classic clustering methods

Clustering is a key step in scRNA-seq downstream analysis. Many classic clustering techniques have been applied in scRNA-seq data clustering, such as the hierarchical clustering, graph-based clustering, and density-based clustering. Among them, density-based clustering[53] is mainly used for the identification of rare cells, and several parameters need to be tuned in this algorithm to obtain a specified clustering result. Therefore, we choose three other popular clustering methods, hierarchical clustering[54], spectral clustering[30], and Louvain algorithm[31], for the correlation analysis experiment. The experiment compares the performances of these methods by combining them with the similarity matrix calculated based on the correlation distance. Results show that hierarchical clustering and graph-based clustering (i.e., spectral clustering and Louvain algorithm) perform obviously different on scRNA-seq datasets under different biological conditions, which means these two methods are the least relevant. For spectral clustering and Louvain algorithm which are both graph-based and have similar performances, since the results of spectral clustering are relatively better and parameters need to be tuned in Louvain to obtain a specified clustering result, we finally choose spectral clustering in these two methods.

### 2.2.2 Hierarchical clustering and spectral clustering

Based on the results of the correlation analysis, we select hierarchical clustering and spectral clustering, which both are the most popular approaches in single cell clustering analysis, for the comparative experiments. These two methods cluster data points based on different strategies and theories. In hierarchical clustering, each point starts as a cluster, then these clusters are merged recursively. We would get a dendrogram and

**Table 1 Details of 12 published datasets analyzed.**

| Datasets | Cells | Genes | Number of groups | Label standard | Units | Species |
|---|---|---|---|---|---|---|
| Ting | 114 | 14405 | 5 | Sliver | RPM | Mus musculus[42] |
| Buettner | 182 | 8989 | 3 | Gold | FPKM | Mus musculus[43] |
| Pollen | 249 | 14805 | 11 | Gold | TPM | Mus musculus[44] |
| Ginhoux | 251 | 11834 | 3 | Sliver | RPKM | Mus musculus[45] |
| LaManno | 337 | 14703 | 13 | Sliver | UMI | Homo sapiens[46] |
| Darmanis | 420 | 22085 | 8 | Sliver | CPM | Homo sapiens[47] |
| Leng | 460 | 19084 | 4 | Gold | TPM | Homo sapiens[48] |
| Camp | 465 | 18999 | 6 | Sliver | FPKM | Homo sapiens[49] |
| Gokce | 1208 | 16379 | 10 | Sliver | TPM | Mus musculus[50] |
| Nestorowa | 1645 | 3991 | 3 | Sliver | UMI | Mus musculus[51] |
| Close | 1733 | 23045 | 4 | Sliver | TPM | Homo sapiens[52] |
| Zeisel | 3005 | 4412 | 9 | Sliver | UMI | Mus musculus[27] |

branches of the hierarchical tree represent the clustering result. Spectral clustering utilizes the spectrum of the constructed similarity matrix to partition data points. Based on different partition rules, these two methods have their own characteristics. For example, hierarchical clustering can get different clustering results by manually cutting the dendrogram, however, there is no perfect definition of a cluster boundary in this algorithm, which would result in the failure of complex tasks. In contrast, spectral clustering can handle complex distributions of data points, but its performance heavily depends on the reliability of the similarity matrix. In this study, we would compare these two clustering methods in the cell type identification task based on scRNA-seq data.

### 2.2.3 Preprocessing strategies in single cell heterogeneity analysis

The comparative analysis experiments apply hierarchical clustering and spectral clustering to substitute the downstream clustering analysis of original methods. Here, we select 10 typical different preprocessing strategies in scRNA-seq heterogeneity analysis including similarity learning, dropout imputation, and dimensionality reduction. Specifically, CIDR and SINCERA are selected as the dropout imputation and the gene selection strategy. Also, two dimensionality reduction strategies, UMAP and ZIFA, and six similarity learning strategies including spearman correlation coefficient, SC3, RAFSIL, SIMLR, MPSSC, and SEURAT are selected. Then we incorporate the preprocessing strategies of these methods with hierarchical clustering and spectral clustering. The parameter settings of each method are based on their initial default values. Additionally, spectral clustering needs to be applied on the basis of a similarity matrix, which the strategies of gene selection and dimensionality reduction cannot get. To solve this problem and ensure that the comparison is completed under same conditions, similarity matrices are calculated by applying the correlation distance on the processed matrices obtained from SINCERA, UMAP, and ZIFA.

### 2.2.4 Evaluation metrics

We use two common metrics in clustering methods evaluation on scRNA-seq data, normalized mutual information (NMI)[55] and adjusted rand index (ARI)[56], to evaluate the performances of hierarchical clustering and spectral clustering with different preprocessing strategies. NMI and ARI are defined as follows:

$$\mathrm{NMI}\left(F_1, F_2\right)=\frac{I\left(F_1, F_2\right)}{\left[H\left(F_1\right)+H\left(F_2\right)\right]/2} \tag{1}$$

$$\mathrm{ARI}=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}} \tag{2}$$

where $F_1$ and $F_2$ represent the true labels of cells and the predicted labels calculated by clustering methods, respectively. $I\left(F_1, F_2\right)$ denotes the mutual information between $F_1$ and $F_2$, and $H\left(F\right)$ is the entropy of those two elements. $n_{ij}$ represents the number of cells that belong to both $F_{1i}$ and $F_{2j}$, $a_i$ is the number of cells in $F_{1i}$, and $b_j$ is the number of cells in $F_{2j}$. Based on the theories of these two metrics, a larger value of NMI or ARI indicates a better clustering performance.

In addition to NMI and ARI, we define a new metric called fake neighbor rate (FNR) to evaluate the accuracy of the similarity matrix computed by each preprocessing strategy. Given a similarity matrix $S$, the $k$ nearest neighbors of each cell are obtained by sorting each row of $S$ in descending order. For each cell $i$ and its $k$ nearest neighbors, they are labeled the same class for we assume that both of them should belong to a same cluster. Then FNR is defined by calculating the proportion of cells that belong to the same class as their $k$ nearest neighbors in the assigned labels but do not belong to the same class in the true labels. Here, we set the nearest neighbor $k=\{1,5,10,15,20\}$ to evaluate the accuracy and robustness of each similarity matrix. Based on the theory of FNR, a smaller value indicates a better similarity learning performance.

### 2.2.5 Comparison results of two clustering methods

We apply the 10 preprocessing strategies with both hierarchical clustering (the linkage criteria is unified as the ward linkage) and spectral clustering on the 12 collected datasets under different biological conditions, and Figs. 1 and 2 show the corresponding results of the comparative analysis upon NMI. We also use FNR to evaluate the accuracy of the similarity matrix computed by each strategy. The results of ARI and FNR are given in the supplementary materials. From the comparison results, we find that, overall, the methods on datasets with cells containing specific functional subsets generally obtain better results than those on datasets with cells undergoing differentiation. We speculate that the mature cells with specific functions may have significant biological signals to differentiate the subgroups, which makes it easier to obtain more accurate partitions. As shown in Fig. 1, for cells undergoing differentiation that belong to a continuous process and have fuzzy

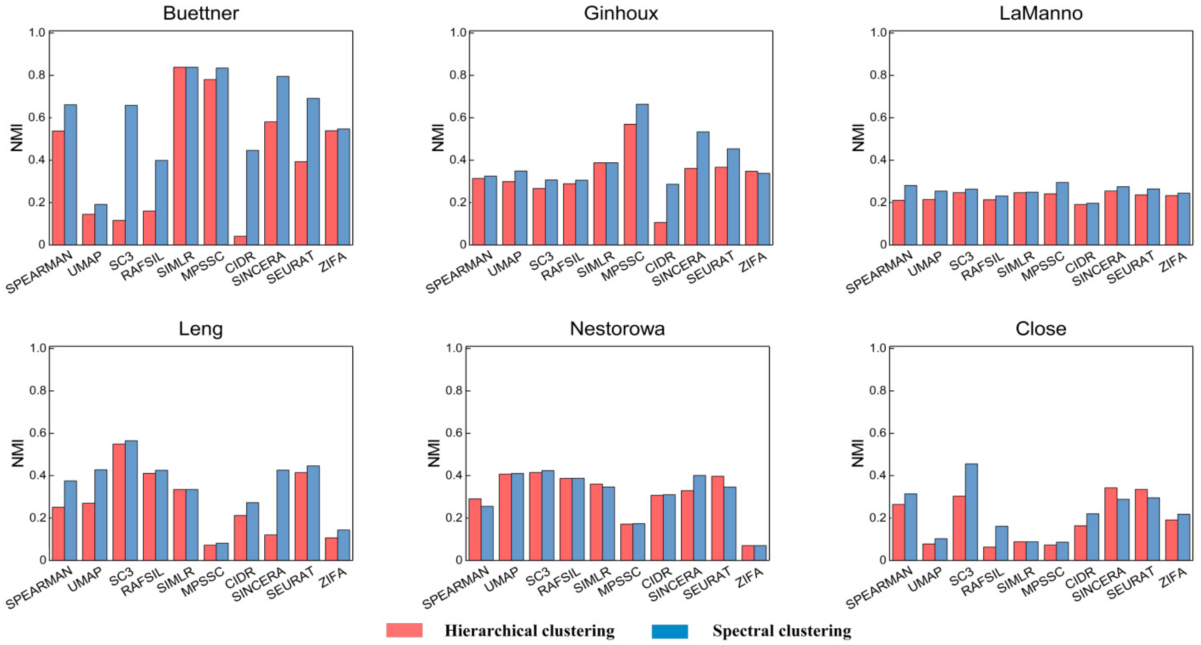


**Fig. 1 Comparison results of hierarchical clustering and spectral clustering combined with 10 preprocessing strategies on cells undergoing differentiation.**

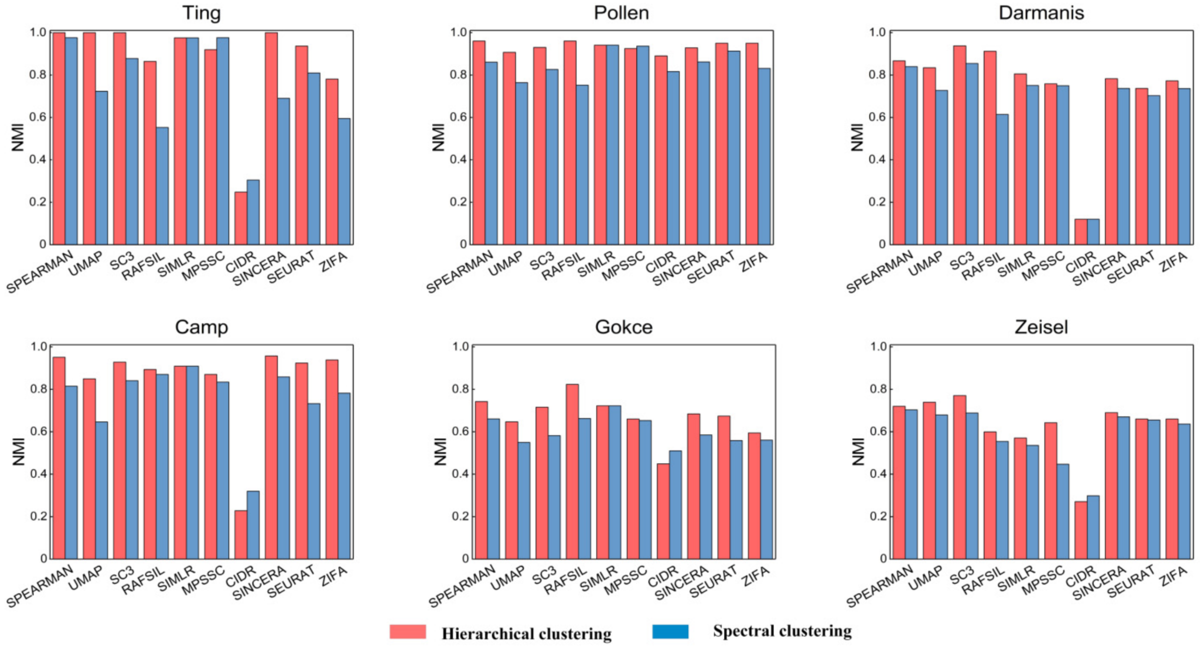


**Fig. 2 Comparison results of hierarchical clustering and spectral clustering combined with 10 preprocessing strategies on cells containing specific functional subsets.**

boundaries, we find that the methods with spectral clustering generally performs better than hierarchical clustering among most of the 10 preprocessing strategies. For cells containing specific functional subsets, we find in Fig. 2 that the methods with hierarchical clustering perform better among seven or nine strategies on all six datasets. Based on the general differences between these two kinds of data, we intuitively suppose that the latent shapes of cells undergoing differentiation and cells containing specific functional subsets may affect the performances of hierarchical clustering and spectral clustering. In order to verify our hypotheses, we conduct further experiments to visualize the latent shapes of scRNA-seq data in two-dimension.

Visualization is a significant tool to reflect the distributions of cells in low dimension[57]. In this study, we visualize each dataset to reveal its latent shape and analyze the relationships between the data shape and the performances of different clustering methods. We choose two commonly used methods, UMAP and t-SNE[58], to visualize datasets in two-dimension with pre-annotations. Figures 3 and 4 show the visualization results on two types of cells (i.e., cells undergoing differentiation and cells containing specific functional subsets). As expected, datasets with cells undergoing differentiation shown in Fig. 3 tend to display a continuous shape, which indicates the dynamic process of differentiation and development. On the contrary, Fig. 4 shows that datasets with cells containing specific functional subsets have clear classification structures and obvious boundaries between clusters. Based on the comparison and visualization results, we can draw two conclusions: (1) Generally, datasets with obvious classification structures in two-dimension can be clustered better than other datasets. (2) The methods with spectral clustering tend to perform better on datasets with continuous shapes, while those with hierarchical clustering achieve better results on datasets with obvious classification structures.

## 2.3 QRS

According to the conclusions, a quantitative measurement called QRS is developed to determine whether the latent representative shape of the scRNA-seq dataset has clear boundaries or not. Firstly, given a scRNA-seq data denoted as matrix $X = [x_1, x_2, \dots, x_n]$ with $m$ genes and $n$ cells, where $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]^{\mathrm{T}}$ represents the expressions of $m$ genes in cell $i$, QRS reduces the dimension of the expression matrix $X$ into two-dimension by UMAP to construct a matrix $Y$. Compared to t-SNE, UMAP is faster and more suitable for scRNA-seq data[12, 59]. For UMAP, we use the implementation provided by the *uwot* R package with default values for all datasets. In order to identify the data latent shape on a unified scale, min-max normalization is applied to map the data into the range from 0 to 1. The equation is defined as follows:

$$Y' = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{3}$$

where $Y_{\min}$ and $Y_{\max}$ are the minimum and maximum values of the expression respectively. Then, we use the minimum spanning tree[60] algorithm to construct the overall skeleton of the data distribution in two-dimension, and distinguish the latent shape of the data by cutting the tree. To build the tree, QRS defines the distance matrix by calculating the Euclidean distance between cells. After getting the distance matrix, QRS builds the Euclidean minimum spanning tree by applying the fast EMST Dual-Tree Boruvka algorithm[61, 62], and the *emstreeR* R package is used to implement this algorithm. The constructed minimum spanning tree can connect all cells in each dataset together, without any cycles and with the minimum possible total edge value. Based on this principle, we suppose that edges in the tree with larger values than a certain threshold are most likely to be the inter-cluster edges. Here, QRS defines the threshold to qualitatively distinguish whether a dataset has clear boundaries between clusters or not. We assume that each cell is evenly dropped on a 1 × 1 plane, and per cell resolution (pCR) is defined by the side length of each cell square. The equation of pCR is defined as follows:

$$\mathrm{pCR} = 1/\sqrt{N} \tag{4}$$

where $N$ is the numbers of cells. Then according to pCR, the threshold is defined as follows:

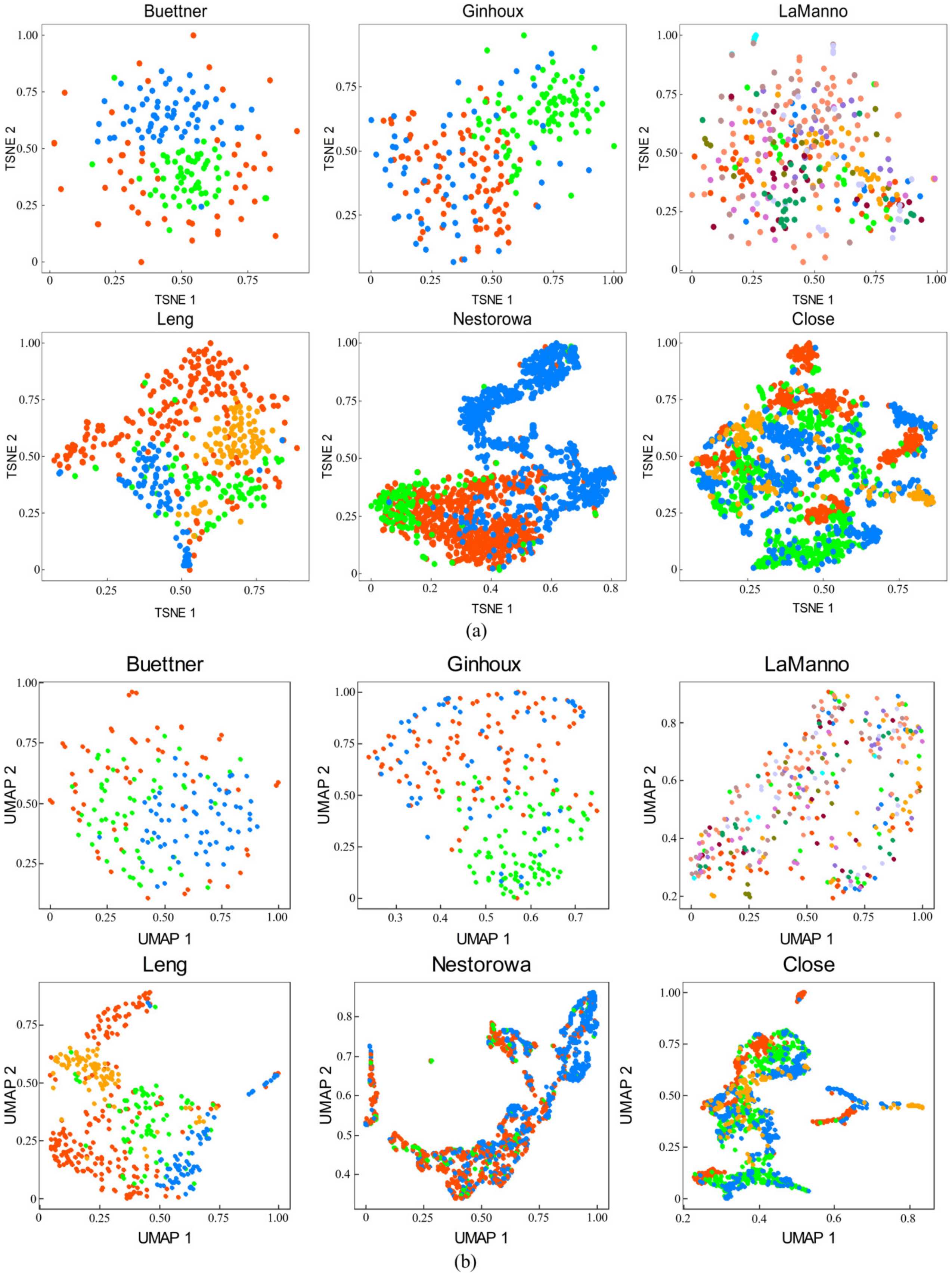
$$\mathrm{Threshold} = \lambda \times \mathrm{pCR} \tag{5}$$

In QRS, we set $\lambda = 5$ as the default in the following experiments. If all the edge values in the tree are smaller than the threshold, we consider the boundaries between clusters are not clear and the datasets have continuous shapes. Otherwise, QRS would consider whether the dataset really has an obvious classification structure by cutting the edges with values that are larger than the threshold and rejudging the balance of the cluster sizes after cutting. We use the ratio of 2 to 8 as the standard for measuring cluster balance. For the clusters formed after cutting, if the proportion of the largest cluster to the second largest cluster exceeds the standard, we consider the dataset has continuous shapes, otherwise the dataset would be classified as having an obvious classification structure.

## 2.4 DDCR

Based on the latent shape identified by QRS, a data-driven clustering recommendation method called DDCR is proposed to select suitable downstream clustering methods from hierarchical clustering and spectral clustering. A brief workflow of DDCR is shown in Fig. 5.

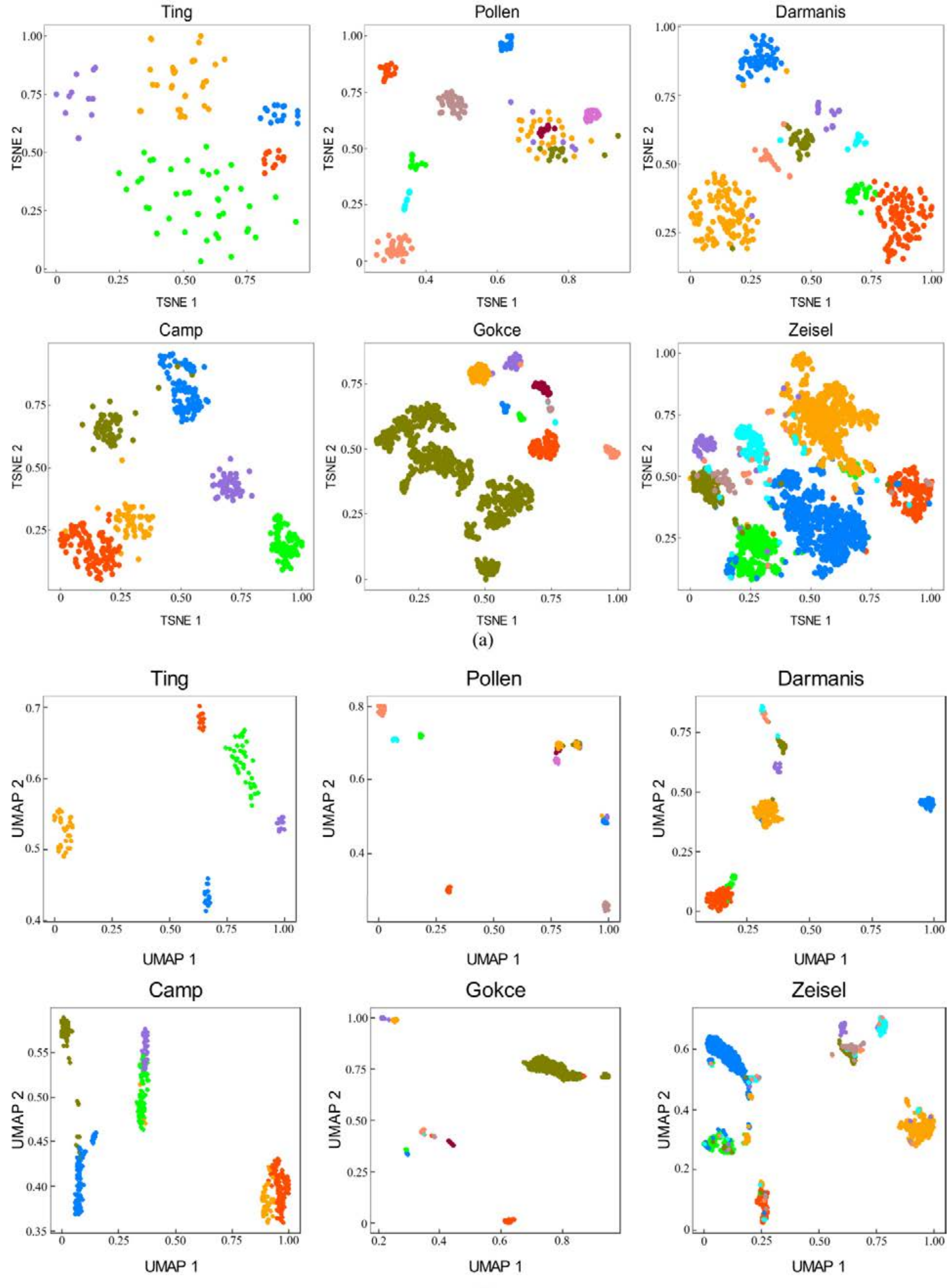
Taking the expression matrix as input, DDCR firstly performs gene filtering. Specifically, if the expression values of a gene in all cells are zero, it will be removed.



(a)

(b)

**Fig. 3 Visualizations of cells undergoing differentiation based on t-SNE (a) and UMAP (b).**



(a)

(b)

**Fig. 4 Visualizations of cells containing specific functional subsets based on t-SNE (a) and UMAP (b).**

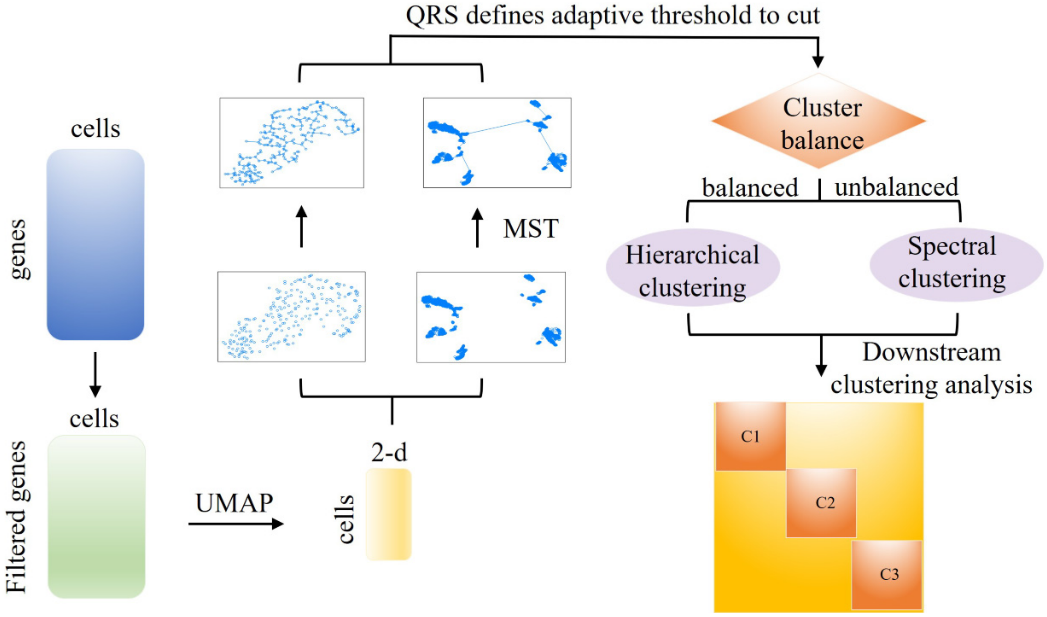


**Fig. 5 Framework of DDCR. DDCR takes the expression matrix as input and applies gene filtering and normalization. Next, QRS is applied by performing dimensionality reduction and the minimum spanning tree algorithm to distinguish whether the dataset has clear boundaries or not. Based on the identified data latent shape, DDCR recommends hierarchical clustering or spectral clustering as the downstream clustering method for the dataset.**

For further analysis, normalization is performed to prevent the highly expressed genes from affecting the study. Next, QRS is applied by performing dimensionality reduction and the minimum spanning tree algorithm on the filtered matrix. The algorithm builds a tree to connect cells in each dataset with the minimum possible total edge values, and the edges in the tree with larger values than the threshold are cut. By comparing the numbers of edges to cut and the cluster balance after cutting, QRS would distinguish whether the dataset has clear boundaries or not. Finally, based on the identified data latent shape, for datasets with continuous shapes, DDCR recommends spectral clustering as the downstream clustering method, while hierarchical clustering is recommended for datasets with obvious classification structures.

## 3 Result

Combined with the biological backgrounds and visualization results, we find that, in general, the methods with spectral clustering tend to perform better on datasets with continuous shapes, while those with hierarchical clustering achieve better results on datasets with obvious classification structures. Though cells undergoing differentiation tend to display a continuous shape, there are still some datasets displaying obvious classification structures. In order to prove that QRS can accurately identify the data latent shapes, and furthermore, to validate the effectiveness of DDCR comprehensively, in addition to the 12 datasets collected for the comparative analysis, we select another eight scRNA-seq datasets as validation sets for the further performance evaluation. The eight datasets including four datasets with cells undergoing differentiation and four datasets with cells containing specific functional subsets. All these datasets are downloaded from the same source as before, and the details are described in Table 2. For the four counts datasets with cells undergoing differentiation, the expression values are computed by using the preprocessing pipeline of the scran[71] and scater[72] Bioconductor packages[73].

### 3.1 Data latent shape identification by QRS

We integrate the total 20 scRNA-seq datasets under different biological backgrouds to differentiate data latent shapes by QRS. As the results of identifications described in Table 3, most latent shapes identified by QRS are consistent with the types of datasets except for Nakamura, Horns, Petropoulos, and Park. These four datasets with cells undergoing differentiation are classified as having obvious classification structures, and the correctness of the identifications would be further verified by the results of DDCR.

**Table 2 Details of eight added validation datasets analyzed.**

| Datasets | Cells | Genes | Number of groups | Label standard | Units | Species |
|---|---|---|---|---|---|---|
| Nakamura | 182 | 4320 | 7 | Sliver | Counts | Mus musculus[63] |
| Lin | 402 | 9438 | 16 | Gold | TPM | Mus musculus[64] |
| Horns | 454 | 2621 | 14 | Sliver | Counts | Drosophila[65] |
| Usoskin | 622 | 17 772 | 4 | Sliver | RPM | Mus musculus[66] |
| Chu | 1018 | 19 072 | 7 | Gold | TPM | Homo sapiens[67] |
| Petropoulos | 1289 | 8772 | 5 | Gold | Counts | Homo sapiens[68] |
| Baron | 1866 | 14 878 | 13 | Sliver | UMI | Mus musculus[69] |
| Park | 2701 | 2441 | 3 | Sliver | Counts | Mus musculus[70] |

**Table 3 Results of the data latent shapes identified by QRS**

| Datasets | Cells | Threshold | Number of edges to cut | Balance of clusters | Data latent shape | Cell type |
|---|---|---|---|---|---|---|
| Ting | 114 | 0.4683 | 1 | Balanced | Classification | Subtypes |
| Buennter | 182 | 0.3706 | 0 | – | Continuous | Differentiation |
| Nakamura | 182 | 0.3706 | 2 | Balanced | Classification | Differentiation |
| Pollen | 249 | 0.3169 | 2 | Balanced | Classification | Subtypes |
| Ginhoux | 251 | 0.3156 | 0 | – | Continuous | Differentiation |
| LaManno | 337 | 0.2724 | 0 | – | Continuous | Differentiation |
| Lin | 402 | 0.2494 | 5 | Balanced | Classification | Subtypes |
| Darmanis | 420 | 0.2440 | 2 | Balanced | Classification | Subtypes |
| Horns | 454 | 0.2347 | 1 | Balanced | Classification | Differentiation |
| Leng | 460 | 0.2331 | 0 | – | Continuous | Differentiation |
| Camp | 465 | 0.2319 | 2 | Balanced | Classification | Subtypes |
| Usoskin | 622 | 0.2005 | 1 | Balanced | Classification | Subtypes |
| Chu | 1018 | 0.1567 | 2 | Balanced | Classification | Subtypes |
| Gokce | 1208 | 0.1439 | 4 | Balanced | Classification | Subtypes |
| Petropoulos | 1289 | 0.1393 | 2 | Balanced | Classification | Differentiation |
| Nestorowa | 1645 | 0.1233 | 1 | Unbalanced | Continuous | Differentiation |
| Close | 1733 | 0.1201 | 1 | Unbalanced | Continuous | Differentiation |
| Baron | 1866 | 0.1151 | 2 | Balanced | Classification | Subtypes |
| Park | 2701 | 0.0962 | 1 | Balanced | Classification | Differentiation |
| Zeisel | 3005 | 0.0912 | 3 | Balanced | Classification | Subtypes |

### 3.2 Recommendation of DDCR

In current single cell clustering methods, computing an accurate cell-to-cell similarity matrix is one of the most critical steps and many approaches have been proposed to solve this problem. According to the FNR results in the comparative analysis, we obtain the performances of the similarity matrices computed by 10 different preprocessing strategies. As the results of FNR given in the supplementary materials shown, with the increasing numbers of nearest neighbors $k$, the similarity matrices learned by RAFSIL generally obtain more accurate and robust performances. Additionally, although the similarity matrices learned by SC3 perform not well when $k = 1$, the values of FNR do not increase sharply like other methods with the change of $k$. Basically, the similarity matrices learned by SC3 can achieve the same superior performances as the matrices learned by RAFSIL, overall. Therefore, in this section, we firstly apply SC3 and RAFSIL with both hierarchical clustering and spectral clustering to assess the correctness of QRS. Furthermore, to validate the effectiveness of DDCR, we apply DDCR to recommend hierarchical clustering or spectral clustering as the downstream clustering method for SC3 and RAFSIL, and then compare the corresponding results of the modified methods with the original ones. In the original clustering methods, both of them use different ensemble strategies to construct a robust (dis)similarity matrix, and apply the hierarchical clustering to partition cells into clusters. All the comparison results including the evaluation of QRS and the NMI and ARI of these four comparison methods (i.e., SC3, SC3-DDCR, RAFSIL, and RAFSIL-DDCR) are given in the Appendix.

In the experimental results of the evaluation on QRS, both SC3 and RAFSIL with spectral clustering perform better on datasets identified as having continuous shapes by QRS, while those with hierarchical clustering achieve better results on datasets identified as having obvious classification structures. It should be noted that the four datasets, Nakamura, Guo, Petropoulos, and Park, whose types are cells undergoing differentiation but identified as classification by QRS, achieve better performances with hierarchical clustering. These results further verify the correctness of QRS and the rationality of our recommendation. Furthermore, as original SC3 and RAFSIL use hierarchical clustering as the downstream clustering method, we find that these two methods with DDCR achieve better performances than the original ones on datasets with continuous shapes. Moreover, in order to show more intuitively that DDCR improves the clustering performances of datasets with continuous shapes, we draw the comparison results of these six datasets in the form of histograms in Figs. 6 and 7. Based on these results, we validate that DDCR can recommend a more suitable downstream clustering method for different scRNA-seq datasets and obtain more robust and accurate results.

## 4 Conclusion

Hierarchical clustering and spectral clustering are the most popular downstream clustering approaches in the scRNA-seq clustering analysis. However, due to the complex backgrounds of scRNA-seq data, like the cells undergoing differentiation and cells containing specific functional subsets, it is not trivial to select the best clustering method for different kinds of data.

In this study, we carry out a comprehensive analysis to evaluate the performances of hierarchical clustering and spectral clustering on scRNA-seq datasets under different biological conditions by using 10 different preprocessing strategies. The experimental results show that the methods with spectral clustering tend to perform better on datasets with continuous shapes in two-dimension, while those with hierarchical clustering achieve better results on datasets with obvious boundaries between clusters in two-dimension. Based on this finding, a new strategy, called QRS, is developed to quantitatively evaluate the latent representative shape of a dataset and to distinguish whether it has clear boundaries or not. Finally, a data-driven clustering

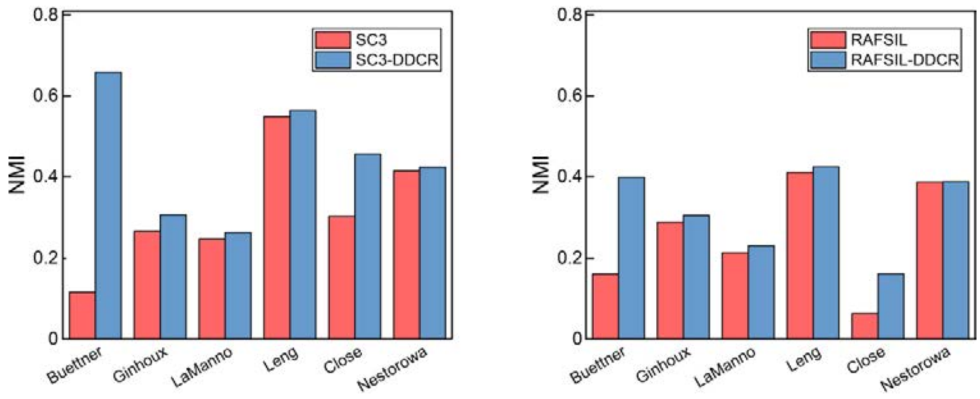


**Fig. 6 NMI of SC3, SC3-DDCR, RAFSIL, and RAFSIL-DDCR on the datasets with continuous shapes in two-dimension.**

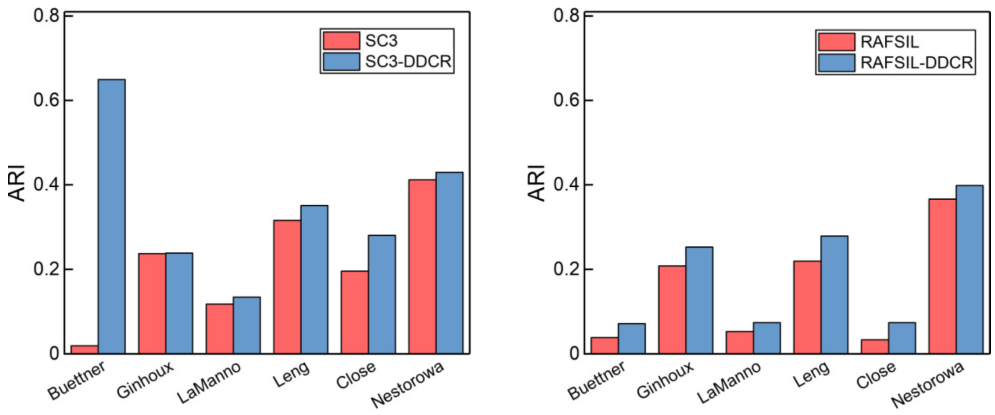


**Fig. 7 ARI of SC3, SC3-DDCR, RAFSIL, and RAFSIL-DDCR on the datasets with continuous shapes in two-dimension.**

recommendation method, called DDCR, is proposed to recommend hierarchical clustering or spectral clustering as the downstream clustering method for scRNA-seq data. We perform DDCR on two typical single cell clustering methods, SC3 and RAFSIL, to evaluate its performance, results show the accuracy of QRS on identifying data latent shapes, and further verify that DDCR can recommend a more suitable downstream clustering method for different scRNA-seq datasets which improves the overall results of clustering analysis. However, noise in gene expressions may affect the accuracy of the data latent shapes identification. In the future, we can introduce some prior biological information such as marker genes and gene regulatory relationship[74, 75] to assist in a more accurate extraction of informative features from scRNA-seq data under different biological backgrounds. Furthermore, the increasing scale of scRNA-seq data brings a challenge to the efficiency of current methods, and approaches such as data partitioning or sampling[76, 77] may provide a possible way to solve this problem.

## Appendix

We apply the 10 preprocessing strategies with both hierarchical clustering (the linkage criteria is unified as the ward linkage) and spectral clustering on the 12 collected datasets under different biological conditions, and use FNR to evaluate the accuracy of the similarity matrix computed by each strategy. The results of FNR are given in Fig. A1. Next, we collect another eight datasets for validation and integrate the total 20 scRNA-seq datasets to prove that QRS can accurately identify the data latent shapes, and furthermore, validate the correctness of DDCR comprehensively. The results are shown in Tables A1–A4.

### Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. U19A2064), the Hunan Provincial Science and Technology Program (No. 2019CB1007), the Fundamental Research Funds for the Central Universities, CSU (No. 2282019SYLB004), and the Fundamental Research Funds for the Central Universities of Central South University (No. 2020zzts593).

### References

[1] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, et al., Science forum: The human cell atlas, *eLife*, vol. 6, p. e27041, 2017.

[2] A. Regev, S. Teichmann, O. Rozenblatt-Rosen, M. Stubbington, K. Ardlie, I. Amit, P. Arlotta, G. Bader, C. Benoist, M. Biton, et al., The human cell atlas white paper, arXiv preprint arXiv: 1810.05192, 2018.

[3] O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, and S. A. Teichmann, The human cell atlas: from vision to reality, *Nature*, vol. 550, no. 7677, pp. 451–453, 2017.

[4] Y. H. Choi and J. K. Kim, Dissecting cellular heterogeneity using single-cell RNA sequencing, *Molecules and Cells*, vol. 42, no. 3, p. 189, 2019.

[5] E. A. A. Alaoui, S. C. K. Tekouabou, S. Hartini, Z. Rustam, H. Silkan, and S Agoujil, Improvement in automated diagnosis of soft tissues tumors using machine learning, *Big Data Mining and Analytics*, vol. 4, no. 1, pp. 33–46, 2021.

[6] A. E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, Single-cell RNA-seq: Advances and future challenges, *Nucleic Acids Research*, vol. 42, no. 14, pp. 8845–8860, 2014.

[7] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data, *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282, 2019.

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

[9] B. Wang, J. 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, 2017.

[10] P. J. Lin, M. Troup, and J. W. K. Ho, CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data, *Genome Biology*, vol. 18, pp. 59, 2017.

[11] E. Pierson and C. Yau, ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis, *Genome Biology*, vol. 16, pp. 241, 2015.

[12] W. V. Li and J. J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data, *Nature Communications*, vol. 9, pp. 997, 2018.

[13] Z. G. Wang, X. Xiao, and S. Rajasekaran, Novel and efficient randomized algorithms for feature selection, *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 208–224, 2020.

[14] E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, *Nature Biotechnology*, vol. 37, pp. 38–44, 2019.

[15] M. Z. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, SINCERA: A pipeline for single-cell RNA-Seq profiling analysis, *PLoS Computational Biology*, vol. 11, no. 11, p. e1004575, 2015.

[16] H. Jiang, L. L. Sohn, H. Y. Huang, and L. N. Chen, Single cell clustering based on cell-pair differentiability correlation and variance analysis, *Bioinformatics*, vol. 34, no. 21, pp. 3684–3694, 2018.

[17] V. Ntranos, G. M. Kamath, J. M. Zhang, L. Pachter, and D. N. Tse, Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts, *Genome*

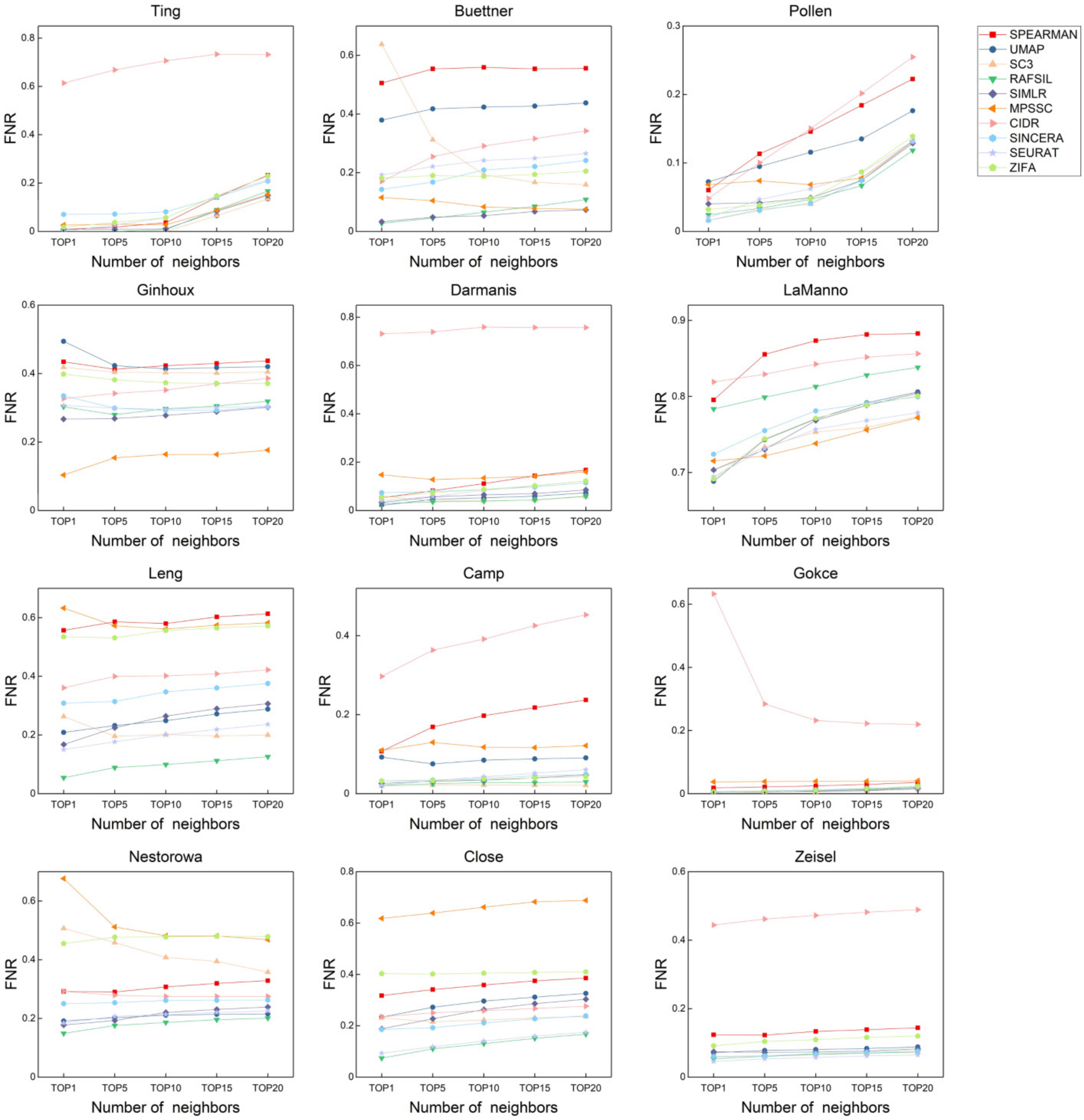


**Fig. A1 FNR of the similarity matrices computed by 10 preprocessing strategies on 12 scRNA-seq datasets.**

Biology, vol. 17, pp. 112, 2016.

[18] M. B. Pouyan and D. Kostka, Random forest based similarity learning for single cell RNA sequencing data, *Bioinformatics*, vol. 34, no. 13, pp. i79–i88, 2018.

[19] G. C. Liu, Z. C. Lin, and Y. Yu, Robust subspace segmentation by low-rank representation, in *Proc. 27th Int. Conf. Machine Learning*, Madison, WI, USA, 2010, pp. 663–670.

[20] R. Vidal and P. Favaro, Low rank subspace clustering (LRSC), *Pattern Recognition Letters*, vol. 43, pp. 47–61, 2014.

[21] R. Q. Zheng, M. Li, Z. L. Liang, F. X. Wu, Y. Pan, and J. X. Wang, SinNLRR: A robust subspace clustering method for cell type detection by non-negative and low-rank representation, *Bioinformatics*, vol. 35, no. 19, pp. 3642–3650, 2019.

[22] R. Q. Zheng, Z. L. Liang, X. Chen, Y. Tian, C. Cao, and M. Li, An adaptive sparse subspace clustering for cell type identification, *Frontiers in Genetics*, vol. 11, pp. 407, 2020.

[23] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nature Biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[24] S. Park and H. Y. Zhao, Spectral clustering based on learning similarity matrix, *Bioinformatics*, vol. 34, no. 12, pp. 2069–2076, 2018.

[25] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al., SC3: Consensus clustering of single-cell RNA-seq data, *Nature Methods*, vol. 14, no. 5, pp. 483–486, 2017.

**Table A1 NMI of SC3 and RAFSIL with hierarchical clustering and spectral clustering.**

| Dataset | QRS | SC3-HC | SC3-SC | RAFSIL-HC | RAFSIL-SC |
|---|---|---|---|---|---|
| Ting | Classification | 1 | 1 | **0.8641** | 0.7724 |
| Buettner | Continuous | 0.1157 | **0.6552** | 0.1605 | **0.5349** |
| Nakamura | Classification | **0.9026** | 0.7991 | **0.8992** | 0.8379 |
| Pollen | Classification | **0.9300** | 0.7654 | **0.9604** | 0.8617 |
| Ginhoux | Continuous | 0.2666 | **0.3124** | 0.2890 | **0.3145** |
| LaManno | Continuous | 0.2592 | **0.2711** | 0.2134 | **0.2309** |
| Lin | Classification | **0.8039** | 0.7646 | **0.8161** | 0.7036 |
| Darmanis | Classification | **0.9309** | 0.8220 | **0.9119** | 0.8266 |
| Horns | Classification | **0.8630** | 0.7837 | **0.8824** | 0.8745 |
| Leng | Continuous | 0.5354 | **0.5589** | 0.4108 | **0.5937** |
| Camp | Classification | **0.9393** | 0.6709 | **0.8938** | 0.6355 |
| Usoskin | Classification | **0.8156** | 0.6818 | **0.9168** | 0.6133 |
| Chu | Classification | **0.9084** | 0.8356 | **0.9168** | 0.8947 |
| Gokce | Classification | **0.7508** | 0.7067 | **0.8666** | 0.6108 |
| Petropoulos | Classification | **0.6998** | 0.6111 | **0.5281** | 0.5278 |
| Nestorowa | Continuous | 0.4149 | **0.4242** | 0.3873 | **0.3878** |
| Close | Continuous | 0.4271 | **0.4555** | 0.0650 | **0.3587** |
| Baron | Classification | **0.5408** | 0.5151 | **0.6776** | 0.6334 |
| Park | Classification | **0.9072** | 0.6548 | **0.5154** | 0.3480 |
| Zeisel | Classification | **0.7734** | 0.6495 | **0.5993** | 0.5300 |

**Table A2 ARI of SC3 and RAFSIL with hierarchical clustering and spectral clustering.**

| Dataset | QRS | SC3-HC | SC3-SC | RAFSIL-HC | RAFSIL-SC |
|---|---|---|---|---|---|
| Ting | Classification | 1 | 1 | **0.7405** | 0.6224 |
| Buettner | Continuous | 0.0186 | **0.6489** | 0.0384 | **0.3974** |
| Nakamura | Classification | **0.8957** | 0.7012 | **0.9052** | 0.7888 |
| Pollen | Classification | **0.9045** | 0.3804 | **0.9413** | 0.6906 |
| Ginhoux | Continuous | 0.2374 | **0.2474** | 0.2081 | **0.2295** |
| LaManno | Continuous | 0.1248 | **0.1331** | 0.0529 | **0.0737** |
| Lin | Classification | **0.5895** | 0.5441 | **0.5586** | 0.4882 |
| Darmanis | Classification | **0.9550** | 0.6769 | **0.9423** | 0.7417 |
| Horns | Classification | **0.7978** | 0.5343 | **0.8022** | 0.7541 |
| Leng | Continuous | 0.3142 | **0.3486** | 0.2193 | **0.4692** |
| Camp | Classification | **0.9451** | 0.3809 | **0.8184** | 0.8180 |
| Usoskin | Classification | **0.8453** | 0.5594 | **0.9358** | 0.6463 |
| Chu | Classification | **0.7671** | 0.6161 | **0.7734** | 0.7446 |
| Gokce | Classification | **0.4670** | 0.4131 | **0.9034** | 0.2406 |
| Petropoulos | Classification | **0.5346** | 0.4295 | **0.4324** | 0.3622 |
| Nestorowa | Continuous | 0.4149 | **0.4297** | 0.3662 | **0.3986** |
| Close | Continuous | 0.2730 | **0.2802** | 0.0323 | **0.2440** |
| Baron | Classification | **0.3263** | 0.2856 | **0.4020** | 0.3283 |
| Park | Classification | **0.8688** | 0.4830 | **0.6405** | 0.1777 |
| Zeisel | Classification | **0.8309** | 0.5609 | **0.4922** | 0.4712 |

[26] R. Huh, Y. C. Yang, Y. C. Jiang, Y. Shen, and Y. Li, SAME-clustering: Single-cell aggregated clustering via mixture model ensemble, *Nucleic Acids Research*, vol. 48, no. 1, pp. 86–95, 2020.

[27] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. Q. He, C. Betsholtz, et al., Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.

[28] J. Žurauskienė and C. Yau, pcaReduce: Hierarchical clustering of single cell transcriptional profiles, *BMC Bioinformatics*, vol. 17, p. 140, 2016.

[29] J. M. Zhang, J. Fan, H. C. Fan, D. Rosenfeld, and D. N.

**Table A3 Comparison results NMI and ARI of between SC3 and SC3-DDCR.**

| Dataset | NMI of SC3 | NMI of SC3-DDCR | ARI of SC3 | ARI of SC3-DDCR |
|---|---|---|---|---|
| Ting | 1 | 1 | 1 | 1 |
| Buettner | 0.1157 | **0.6552** | 0.0186 | **0.6489** |
| Nakamura | 0.9026 | 0.9026 | 0.8957 | 0.8957 |
| Pollen | 0.9300 | 0.9300 | 0.9045 | 0.9045 |
| Ginhoux | 0.2666 | **0.3124** | 0.2374 | **0.2474** |
| LaManno | 0.2592 | **0.2711** | 0.1248 | **0.1331** |
| Lin | 0.8039 | 0.8039 | 0.5895 | 0.5895 |
| Darmanis | 0.9309 | 0.9309 | 0.9550 | 0.9550 |
| Horns | 0.8630 | 0.8630 | 0.7978 | 0.7978 |
| Leng | 0.5354 | **0.5589** | 0.3142 | **0.3486** |
| Camp | 0.9393 | 0.9393 | 0.9451 | 0.9451 |
| Usoskin | 0.8156 | 0.8156 | 0.8453 | 0.8453 |
| Chu | 0.9084 | 0.9084 | 0.7671 | 0.7671 |
| Gokce | 0.7508 | 0.7508 | 0.4670 | 0.4670 |
| Petropoulos | 0.6998 | 0.6998 | 0.5346 | 0.5346 |
| Nestorowa | 0.4149 | **0.4242** | 0.4149 | **0.4297** |
| Close | 0.4271 | **0.4555** | 0.2730 | **0.2802** |
| Baron | 0.5408 | 0.5408 | 0.3263 | 0.3263 |
| Park | 0.9072 | 0.9072 | 0.8688 | 0.8688 |
| Zeisel | 0.7734 | 0.7734 | 0.8309 | 0.8309 |

**Table A4 Comparison results NMI and ARI of between RAFSIL and RAFSIL-DDCR.**

| Dataset | NMI of RAFSIL | NMI of RAFSIL-DDCR | ARI of RAFSIL | ARI of RAFSIL-DDCR |
|---|---|---|---|---|
| Ting | 0.8641 | 0.8641 | 0.7405 | 0.7405 |
| Buettner | 0.1605 | **0.6552** | 0.0384 | **0.3974** |
| Nakamura | 0.8992 | 0.8992 | 0.9052 | 0.9052 |
| Pollen | 0.9604 | 0.9604 | 0.9413 | 0.9413 |
| Ginhoux | 0.2890 | **0.3145** | 0.2081 | **0.2295** |
| LaManno | 0.2134 | **0.2309** | 0.0529 | **0.0737** |
| Lin | 0.8161 | 0.8161 | 0.5586 | 0.5586 |
| Darmanis | 0.9119 | 0.9119 | 0.9423 | 0.9423 |
| Horns | 0.8824 | 0.8824 | 0.8022 | 0.8022 |
| Leng | 0.4108 | **0.5937** | 0.2193 | **0.4692** |
| Camp | 0.8938 | 0.8938 | 0.8184 | 0.8184 |
| Usoskin | 0.9168 | 0.9168 | 0.9358 | 0.9358 |
| Chu | 0.9168 | 0.9168 | 0.7734 | 0.7734 |
| Gokce | 0.8666 | 0.8666 | 0.9034 | 0.9034 |
| Petropoulos | 0.5281 | 0.5281 | 0.4324 | 0.4324 |
| Nestorowa | 0.3873 | **0.3878** | 0.3662 | **0.3986** |
| Close | 0.0650 | **0.3587** | 0.0323 | **0.2440** |
| Baron | 0.6776 | 0.6776 | 0.4020 | 0.4020 |
| Park | 0.5154 | 0.5154 | 0.6405 | 0.6405 |
| Zeisel | 0.5993 | 0.5993 | 0.4922 | 0.4922 |

Tse, An interpretable framework for clustering single-cell RNA-Seq datasets, *BMC Bioinformatics*, vol. 19, pp. 93, 2018.

[30] L. U. Von, A tutorial on spectral clustering, *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[31] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[32] Y. Zhang, B. Wu, Y. Liu and J. Lv, Local community detection based on network motifs, *Tsinghua Science and Technology,* vol. 24, no. 6, pp. 716–727, 2019.

[33] B. Zhao, J. Wang, M. Li, F. Wu, and Y. Pan, Detecting protein complexes based on uncertain graph model,

*IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 11, no. 3, pp. 486–497, 2014.

[34] X. Meng, J. Xiang, R. Zheng, F. Wu, and M. Li, DPCMNE: Detecting protein complexes from protein-protein interaction networks via multi-level network embedding, *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* doi:10.1109/TCBB.2021.3050102.

[35] Z. Liang, M. Li, R. Zheng, Y. Tian, X. Yan, J. Chen, F. X. Wu, and J. Wang, Cell type detection based on sparse subspace representation and similarity enhancement, *Genomics, Proteomics & Bioinformatics*, https://doi.org/10.1016/j.gpb.2020.09.004.

[36] L Jiang, H Chen, L Pinello and GC Yuan, GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index, *Genome Biology,* vol. 17, no. 1, pp. 1–13, 2016.

[37] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, *Nature Biotechnology,* vol. 32, no. 4, p. 381, 2014.

[38] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, et al., ArrayExpress—A public repository for microarray gene expression data at the EBI, *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2013.

[39] R. Edgar, M. Domrachev, and A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Research,* vol. 30, no. 1, pp. 207–210, 2002.

[40] J. Zhao, S. Zhang, Y. Liu, X. He, M. Qu, G. Xu, H. Wang, M. Huang, J. Pan, Z. Liu, Z. Li, L. Liu, and Z. Zhang, Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human, *Cell Discovery,* vol. 6, no. 1, pp. 1–19, 2020.

[41] S. L. Goldman, M. MacKay, E. Afshinnekoo, A. M. Melnick, S. Wu, and C. E. Mason, The impact of heterogeneity on single-cell sequencing, *Frontiers in Genetics,* vol. 10, p. 8, 2019.

[42] D. T. Ting, B. S. Wittner, M. Ligorio, N. V. Jordan, A. M. Shah, D. T. Miyamoto, N. Aceto, F. Bersani, B. W. Brannigan, K. Xega, et al., Single-cell rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells, *Cell Reports,* vol. 8, no. 6, pp. 1905–1918, 2014.

[43] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpoi, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells, *Nature Biotechnology,* vol. 33, no. 2, pp. 155–160, 2015.

[44] A. A. Pollen and T. J. Nowakowski, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex, *Nature Biotechnology,* vol. 32, no. 10, p. 1053, 2014.

[45] A. Schlitzer, V. Sivakamasundari, J. Chen, H. R. B. Sumatoh, J. Schreuder, J. Lum, B. Malleret, S. Zhang, A. Larbi, F. Zolezzi, et al., Identification of cdc1- and cdc2-committed dc progenitors reveals early lineage priming at the common dc progenitor stage in the bone marrow, *Nature Immunology,* vol. 16, no. 7, pp. 718–728, 2015.

[46] G. La Manno, D. Gyllborg, S. Codeluppi, K.Nishimura, C.Salto, A. Zeisel, L. E. Borrm, S. R. W. Stott, E. M. Toledo, et al, Molecular diversity of midbrain development in mouse, Human, and Stem Cells, *Cell,* vol. 167, no. 2, pp. 566–580, 2016.

[47] 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, *PNAS,* vol. 112, no. 23, pp. 7285–7290, 2015.

[48] N. Leng, L. F. Chu, C. Barry, Y. Li, J. Choi, P. Jiang, R. M. Stewart, J. Thomson, and C Kendziorski, Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments, *Nature Methods,* vol. 12, no. 10, p. 947, 2015.

[49] J. G. Camp, K. Sekine, T. Gerber, H. Loeffler-Wirth, H. Binder, M. Gac, S. Kanton, J. Kageyama, G. Damm, D. Seehofer, L. Belicova, et al., Multilineage communication regulates human liver bud development from pluripotency, *Nature,* vol. 546, no. 7659, pp. 533–538, 2017.

[50] D. Gokie, G. M. Stanley, B. Treutlein, N. F. Neff, J. G. Camp, R. C. Malenka, P. E. Rothwell, M. V. Fuccillo, T. C. Südhof, and S. R. Quake, Cellular Taxonomy of the Mouse Striatum as Revealed by Single Cell RNA Sequencing, *Biophysical Journal,* vol. 16, no. 4, pp. 1126–1137, 2016.

[51] S. Nestorowa, F. K. Hamey, S. B. Pijuan, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, and B. Gottgens, A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation, *Blood,* vol. 128, no. 8, pp. e20-e31, 2016.

[52] J. L. Close, Z. Z. Yao, B. P. Levi, J. A. Miller, T. E. Bakken, V. Menon, J. T. Ting, A. Wall, A. R. Krostag, E. R. Thomsen, et al., Single-cell profiling of an in vitro model of human interneuron development reveals temporal dynamics of cell type production and maturation, *Neuron*, vol. 93, no. 5, pp. 1035–1048, 2017.

[53] M. Ester, H. P. Kriegel, J. Sander, and X. W. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proc. $2^{nd}$ Int. Conf. Knowledge Discovery and Data Mining*, Portland, OR, USA, 1996, pp. 226–231.

[54] A. Smoliński, B. Walczak, and J. W. Einax, Hierarchical clustering extended with visual complements of environmental data set, *Chemometrics and Intelligent Laboratory Systems*, vol. 64, no. 1, pp. 45–54, 2002.

[55] A. Strehl and J. Ghosh, Cluster ensembles—A knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[56] S. Wagner and D. Wagner, *Comparing Clusterings—AnOverview*. Karlsruhe, Germany: University at Karlsruhe,

2017.

[57] H. Cho, B. Berger, and J. Peng, Generalizable and scalable visualization of single-cell data using neural networks, *Cell Systems*, vol. 7, no. 2, pp. 185–191, 2018.

[58] L. V. der Maaten and G Hinton, Visualizing data using tSNE, *Journal of machine learning research,* vol. 9, no. 11, pp. 2579-2605, 2008.

[59] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv: 1802.03426, 2018.

[60] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. Cambridge, MA, USA: MIT Press, 2001, pp. 561–579.

[61] W. B. March, P. Ram, and A. G. Gray, Fast euclidean minimum spanning tree: Algorithm, analysis, and applications, in *Proc. 16th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Washington, DC, USA, 2010, pp. 603–612.

[62] R. R. Curtin, J. R. Cline, N. P. Slagle, W. B. March, P. Ram, N. A. Mehta, and A. G. Gray, MLPACK: A scalable C++ machine learning library, *Journal of Machine Learning Research*, vol. 14, pp. 801–805, 2013.

[63] T. Nakamura, I. Okamoto, K. Sasaki, Y. Yabuta, C. Iwatani, H. Tsuchiya, Y. Seita, S. Nakamura, T. Yamamoto, and M. Saitou, A developmental coordinate of pluripotency among mice, monkeys and humans, *Nature*, vol. 537, no. 7618, pp. 57–62, 2016.

[64] C. Lin, S. Jain, H. Kim, and Z. Bar-Joseph, Using neural networks for reducing the dimensions of single-cell RNA-seq data, *Nucleic Acids Research*, vol. 45, no. 17, p. e156, 2017.

[65] H. J. Li, F. Horns, B. Wu, Q. J. Xie, J. F. Li, T. C. Li, D. J. Luginbuhl, S. R. Quake, and L. Q. Luo, Classifying *Drosophila* olfactory projection neuron subtypes by single-cell RNA sequencing, *Cell*, vol. 171, no. 5, pp. 1206–1220, 2017.

[66] D. Usoskin, A. Furlan, S. Islam, H. Abdo, P. Lönnerberg, D. H. Lou, J. Hjerling-Leffler, J. Haeggström, O. Kharchenko, P. V. Kharchenko, et al., Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing, *Nature Neuroscience*, vol. 18, no. 1, pp. 145–153, 2015.

[67] L. F. Chu, N. Leng, J. Zhang, Z. Hou, D. Manott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson, Singlecell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm, *Genome Biology,* vol. 17, no. 1, pp. 1–20, 2016.

[68] S. Petropoulos, D. Edsgärd, B. Reinius, Q. L. Deng, S. P. Panula, S. Codeluppi, A. P. Reyes, S. Linnarsson, R. Sandberg, and F. Lanner, Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos, *Cell*, vol. 165, no. 4, pp. 1012–1026, 2016.

[69] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, and A. M. Klein, A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure, *Cell Systems*, vol. 3, no. 4, pp. 346–360, 2016.

[70] J. Park, R. Shrestha, C. X. Qiu, A. Kondo, S. Z. Huang, M. Werth, M. Y. Li, J. Barasch, and K. Suszták, Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease, *Science*, vol. 360, no. 6390, pp. 758–763, 2018.

[71] A. T. L. Lun, D. J. McCarthy, and J. C. Marioni, A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor, *F1000 Research*, vol. 5, pp. 2122, 2016.

[72] D. J. McCarthy, K. R. Campbell, A. T. L. Lun, and Q. F. Wills, Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R, *Bioinformatics*, vol. 33, no. 8, pp. 1179–1186, 2017.

[73] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, A comparison of single-cell trajectory inference methods, *Nature Biotechnology*, vol. 37, no. 5, pp. 547–554, 2019.

[74] R. Q. Zheng, M. Li, X. Chen, S. Y. Zhao, F. X. Wu, Y. Pan, and J. X. Wang, An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 347–354, 2021.

[75] X. Chen, M. Li, R. Q. Zheng, S. Y. Zhao, F. X. Wu, Y. H. Li, and J. X. Wang, A novel method of gene regulatory network structure inference from gene knock-out expression data, *Tsinghua Science and Technology*, vol. 24, no. 4, pp. 446–455, 2019.

[76] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, A survey of data partitioning and sampling methods to support big data analysis, *Big Data Mining and Analytics,* vol. 3, no. 2, pp. 85–101, 2020.

[77] L. Wang and W. Fan, A multilevel splitting algorithm for quick sampling, *Tsinghua Science and Technology,* vol. 26, no. 4, pp. 417–425, 2021.

**Yu Tian** received the BS degree in network engineering from Changsha University of Science and Technology, Changsha, China, in 2017. She is currently working toward the MS degree in computer science at the School of Computer Science and Engineering, Central South University, Changsha, China. Her current research interests include machine learning and bioinformatics.

**Ruiqing Zheng** received the BS and the MS degrees in computer science from Central South University, Changsha, China, in 2013 and 2016, respectively. He is currently working toward the PhD degree at the School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include bioinformatics and system biology.

**Zhenlan Liang** received the BS degree in information and computing science from Hunan Normal University, Changsha, China, in 2017. She is currently working toward the PhD degree at the School of Computer Science and Engineering, Central South University, Changsha, China. Her research interests include bioinformatics and machine learning.

**Suning Li** is a research assistant in Hunan Provincial Key Lab of Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China. Her research interest is bioinformatics.

**Fang-Xiang Wu** received the BS and MS degrees in applied mathematic from Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first PhD degree in control theory and its applications from Northwestern Polytechnical University, Xi'an, China, in 1998, and the second PhD degree in biomedical engineering from University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004–2005, he worked as a postdoctoral fellow in the Laval University Medical Research Center (CHUL), Quebec City, Canada. He is currently a professor at the Division of Biomedical Engineering and the Department of Mechanical Engineering, U of S. His current research interests include computational and systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, and applications of control theory to biological systems. He is serving as the editorial board member of five international journals, the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals.

**Min Li** received the BS degree in communication engineering and the MS and PhD degrees in computer science from Central South University, Changsha, China, in 2001, 2004, and 2008, respectively. She is currently a professor and vice dean at the School of Computer Science and Engineering, Central South University, Changsha, China. Her main research interests include bioinformatics and system biology.