mathematics

MDPI

Article

SCM Enables Improved Single-Cell Clustering by Scoring Consensus Matrices

Yilin Yu and Juntao Liu *

School of Mathematics and Statistics, Shandong University, Weihai 264209, China; sofiayuyu@163.com
* Correspondence: juntaosdu@126.com or juntao@sdu.edu.cn

Abstract: Single-cell clustering facilitates the identification of different cell types, especially the identification of rare cells. Preprocessing and dimensionality reduction are the two most commonly used data-processing methods and are very important for single-cell clustering. However, we found that different preprocessing and dimensionality reduction methods have very different effects on single-cell clustering. In addition, there seems to be no specific combination of preprocessing and dimensionality reduction methods that is applicable to all datasets. In this study, we developed a new algorithm for improving single-cell clustering results, called SCM. It first automatically searched for an optimal combination that corresponds to the best cell type clustering of a given dataset. It then defined a flexible cell-to-cell distance measure with data specificity for cell-type clustering. Experiments on ten benchmark datasets showed that SCM performed better than almost all the other seven popular clustering algorithms. For example, the average ARI improvement of SCM over the second best method SC3 even reached 29.31% on the ten datasets, which demonstrated its great potential in revealing cellular heterogeneity, identifying cell types, depicting cell functional states, inferring cellular dynamics, and other related research areas.

Keywords: single-cell clustering; data processing; dimensionality reduction; distance measure

MSC: 62P10; 92-08

Citation: Yu, Y.; Liu, J. SCM Enables Improved Single-Cell Clustering by Scoring Consensus Matrices. *Mathematics* **2023**, *11*, 3785. https://doi.org/10.3390/math11173785

Academic Editors: Junseok Kim and Jianjun Paul Tian

Received: 2 August 2023
Revised: 29 August 2023
Accepted: 1 September 2023
Published: 3 September 2023

1. Introduction

Single-cell sequencing technology refers to the sequencing of single-cell genomes or transcriptomes to obtain genomic, transcriptomic, or other multiomics information to reveal cell population differences and cellular evolutionary relationships [1]. By mapping the cells of different organisms at the resolution of a single cell, this technology has made significant contributions to better exploring cellular heterogeneity across various organisms and has played an important role in research fields such as oncology, developmental biology, microbiology, and neuroscience [2]. Traditional bulk RNA sequencing technology detects the average expression levels of cell populations, ignoring the gene expression of individual cells, which leads to the inability to study cell heterogeneity [3]. In contrast, single-cell RNA sequencing (scRNA-seq) technology acquires transcriptome information at the resolution of single cells, enables the observation of cells with higher precision and better identifies rare cell types, which is helpful to stimulate new ideas for disease treatments [4].

A common strategy used to study the identification of cell types is unsupervised clustering, and various single-cell clustering methods have been developed. For instance, Xu et al. [5] introduced the SNN-cliq algorithm, which used a quasiclique-based clustering algorithm to cluster cells on a shared nearest neighbour (SNN) network constructed from single-cell scRNA-seq datasets. Seurat [6] combined the shared nearest neighbour graph with the Louvain community detection algorithm to identify cell types. Based on the consensus clustering strategy, the popular SC3 [7] algorithm integrated the k-means clustering results with different distance and different dimensionality reduction methods to carry out

the "consensus" identification of cell types. Yang et al. [8] designed the SAFE algorithm for ensemble clustering based on the hypergraph model. Semisoft clustering was adopted in SOUP [9] to identify both pure and intermediate cell types. CIDR [10] improved the accuracy of clustering by imputing the distance matrix. PcaReduce [11] combined PCA and hierarchical clustering to classify cell types. SIMLR [12] performs clustering by generating a similarity matrix based on multikernel learning.

Although great efforts have been made to develop clustering methods, biological and technical noise, redundant information and the high dimensionality of scRNA-seq data greatly discourage the accurate identification of cell types [13,14]. Therefore, effective data preprocessing and dimensionality reduction before cell type clustering are two basic and essential steps for downstream analysis [15,16]. For example, SC3 and CIDR apply the *log* transformation for data preprocessing before cell type clustering. Both SCANPY [17] and Seurat preprocess the expression matrix via the *z-score* transformation. SINCERE [18] normalized gene expressions using the *z-score* method. For data dimensionality reduction, various methods have been designed based on specific issues. The most popular principal component analysis (PCA) [19] approach reduces the dimensionality by extracting principal components while reducing the loss of data information as much as possible. The Laplacian eigenmap [20] approach is a manifold dimensionality reduction algorithm that preserves the structure of local sample points in the low-dimensional space. The combination of these two dimensionality reduction methods was employed in SC3. Becht et al. [21] developed the UMAP method, which was designed based on Riemannian geometry and algebraic topology to perform dimension reduction and data visualization. In contrast to linear dimensionality reduction methods, such as PCA and LDA [22], UMAP can map structural information from high-dimensional space directly into low-dimensional space, and it pays more attention to preserving the topology structure with neighbouring samples. SOUP utilizes nonnegative matrix factorization (NMF) [23] to reduce dimensionality, and multidimensional scaling (MDS) [24] is applied in ascending [25] for dimensionality reduction.

In practice, existing single-cell clustering methods usually apply fixed preprocessing and dimensionality reduction methods across datasets rather than applying a flexible approach. Nevertheless, different datasets actually have different requirements for preprocessing and dimensionality reduction. In summary, determining the appropriate preprocessing and dimensionality reduction methods for different datasets is an urgent problem to be solved, and the solution will greatly improve the clustering accuracy.

In this study, we developed a novel SCM algorithm by first searching for the optimal preprocessing and dimensionality reduction methods based on a given dataset by using a newly designed *f-value* approach. Based on the selected optimal combination of preprocessing and dimensionality reduction methods, the optimal consensus matrix can be generated accordingly. Since the cell-to-cell distance metric is also a crucial step in cell type clustering, we developed a flexible distance measure with data specificity in this study. Based on the optimal consensus matrix and the flexible distance, a hierarchical clustering was employed to complete the final cell type clustering step (see Figure 1). The clustering accuracy of SCM was evaluated on ten benchmark datasets, and the results showed that SCM performed much better than all the compared clustering methods.

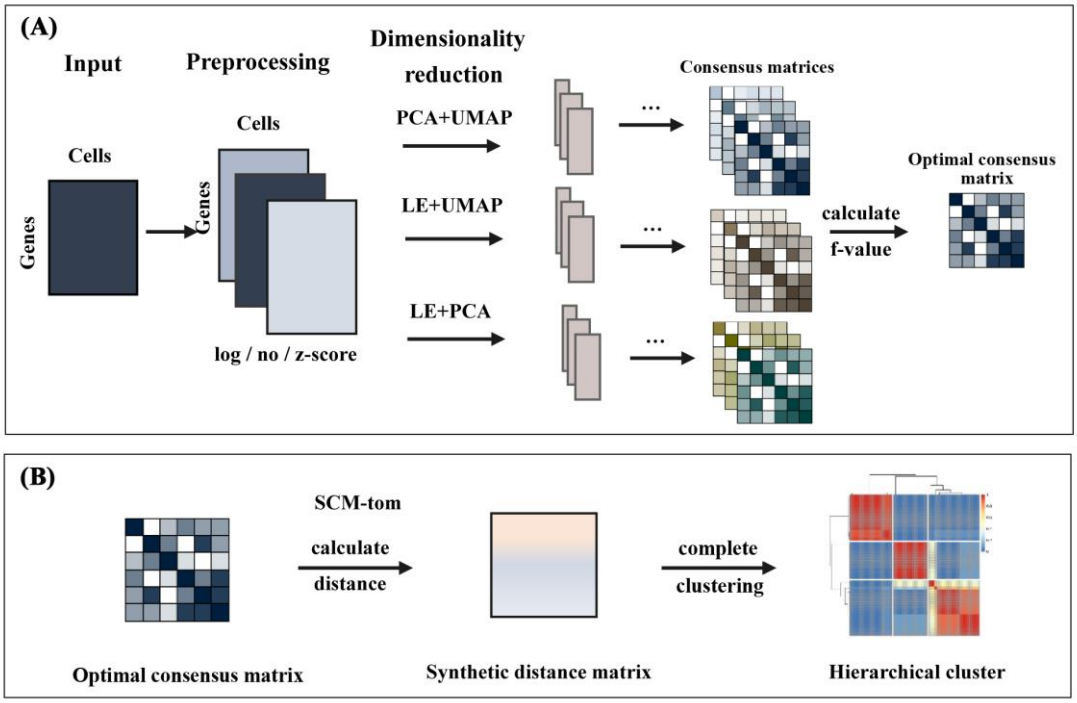

Figure 1. Overview of SCM. SCM is composed of two main parts: (**A**) Given a single-cell gene expression matrix, nine consensus matrices are then constructed and the optimal one is determined with the largest *f-value*. (**B**) A flexible distance measure termed SCM-tom is defined with data specificity. Based on the distance matrix, hierarchical clustering is applied to complete the final cell type clustering step.

2. Results

2.1. Effects of Different Combinations of Preprocessing and Dimensionality Reduction Methods on Cell Type Clustering

In this study, the commonly used *log* transformation, no transformation and *z-score* transformation were employed as the candidates for the preprocessing methods. Three popular techniques, PCA, Laplacian Eigenmaps (LE), and UMAP, were applied as dimensionality reduction methods. The combinations of preprocessing methods and dimensionality reduction methods are shown in Table 1. Then, based on the framework of SC3, nine consensus matrices and the corresponding clustering results were generated with the nine combinations in Table 1.

Table 1. Summary of the nine combinations of preprocessing and dimensionality reduction methods.

Method Labels	Preprocessing Method	Dimensionality Reduction Method
x_1	log transformation	PCA + UMAP
x_2	log transformation	LE + UMAP
x_3	log transformation	LE + PCA
x_4	no transformation	PCA + UMAP
x_5	no transformation	LE + UMAP
x_6	no transformation	LE + PCA
x_7	z-score transformation	PCA + UMAP
x_8	z-score transformation	LE + UMAP
x_9	z-score transformation	LE + PCA

Based on the SC3 framework, the nine combinations of preprocessing and dimensionality reduction methods (Table 1) were run on all ten datasets (see the Methods section). Considering the instability of SC3 clustering, each combination was run 50 times, and the average of the 50 ARI values was used to evaluate the performance of the clustering results.

From the running results, different preprocessing and dimensionality reduction methods were found to have quite different effects on cell type clustering (Figure 2). Taking the Buettner dataset as an example, when the preprocessing method was set to the *log* transformation, the ARI values corresponding to x_1, x_2, and x_3 in Table 1 are 0.77, 0.35, and 0.05, respectively. For the Usoskin dataset, if the dimensionality reduction method was set to PCA + UMAP, the ARI values corresponding to x_1, x_4, x_7 in Table 1 were 0.85, 0.06, and 0.82, respectively. The best combination of preprocessing and dimensionality reduction for the Buettner dataset was x_7; however, on the Zeisel dataset, the x_7 combination had the worst performance. Therefore, it is apparent that different datasets exhibited large differences in sensitivity to different combinations of preprocessing and dimensionality reduction methods.

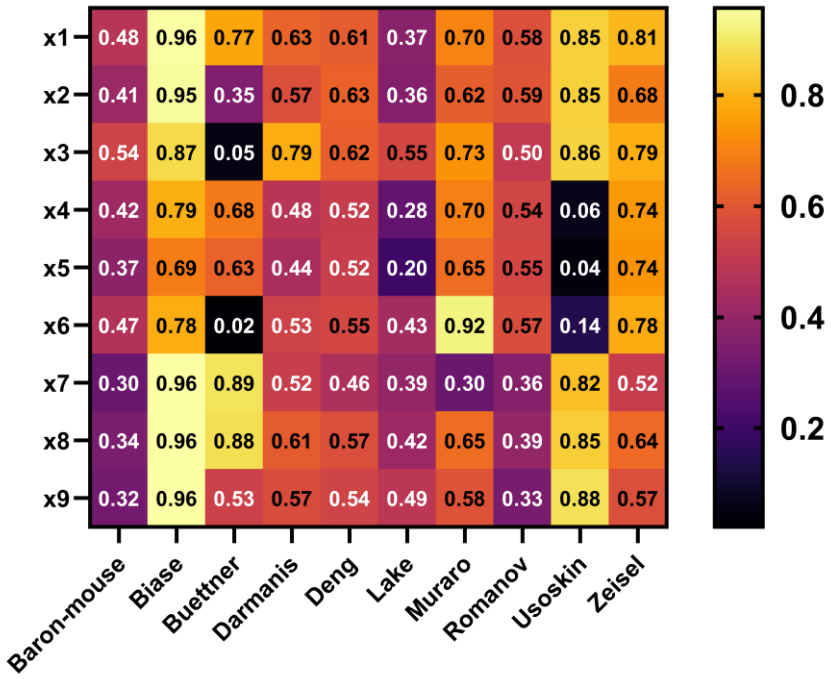

Figure 2. ARI values of the clustering results with different combinations of preprocessing and dimensionality reduction methods (Table 1). The rows represent different combinations and the columns indicate different datasets.

2.2. Performance of the f-Value for the Selection of the Optimal Combination

In this study, we developed an *f-value* approach (see the Methods section for details) to automatically determine the optimal combination of preprocessing methods and dimensionality reduction methods with data specificity. In practice, the ARI value of the optimal combination selected by the *f-value* may be slightly lower than that of the real best combination. In addition, although the selected optimal combination was not the best, its ARI value was ranked highly. To fully evaluate the effectiveness of the *f-value*, we first defined and calculated a so-called optimal ARI interval with an error range of e, termed $OAI(e)$, for a given dataset as follows.

For each dataset, the largest ARI value under the nine combinations was recorded as M_{ARI}, and then $OAI(e)$ was defined as $[M_{ARI} - e, M_{ARI}]$. Based on the definition, if the ARI value under a combination belongs to the interval $OAI(e)$ and ranks at the top p value among the nine ARI values, it was considered that the combination is the optimal one with an error range of e and a rank of p. Due to the instability of SC3, the nine combinations

were run 50 times on each dataset. Then, the accuracy of the selected optimal combination is defined by the following formula.

$$ACC = \frac{\sum_{i=1}^{50} count(i)}{50}$$

where $count(i) = 1$ if the selected optimal combination belongs to the interval $OAI(e)$ and ranks at the top p value, and $count(i) = 0$ otherwise.

We evaluated the performance of the *f-value* in selecting the optimal combinations on the ten datasets under different error ranges e (0.05, 0.03, and 0.01) and top ranks p (3 and 2). The results showed that the *f-value* demonstrated high accuracy within a small error range on most datasets (Figure 3). For example, when the error range e was set to 0.01 and p was set to 2, the accuracies of the *f-value* on the Darmanis, Buettner, and Baron-mouse datasets were 100%, and they were very close to 100% on the Usoskin, Biase, and Lake datasets. When p was set to 2, as error range e increased from 0.01 to 0.03, the accuracy of the *f-value* on some datasets showed a significant upwards trend. For example, the accuracy on the Zeisel dataset increased from 50% to 86%. Moreover, the accuracies on the Biase and Lake datasets reached 100%. When the error range e was set to 0.03 and p was set to 3, the accuracies on the Zeisel, Usoskin, and Muraro datasets increased to 100%. It can be seen that a slight increase in error range e may significantly increase the accuracy of the *f-value* on some datasets, which, to some extent, indicates that the difference between the combination with the highest *f-value* and the combination with the highest ARI may be very small and imperceptible. Based on the above results, the combination with the highest *f-value* obtained excellent performance on most datasets and can help to accurately select the optimal combination of preprocessing and dimensionality reduction methods within a small error range.

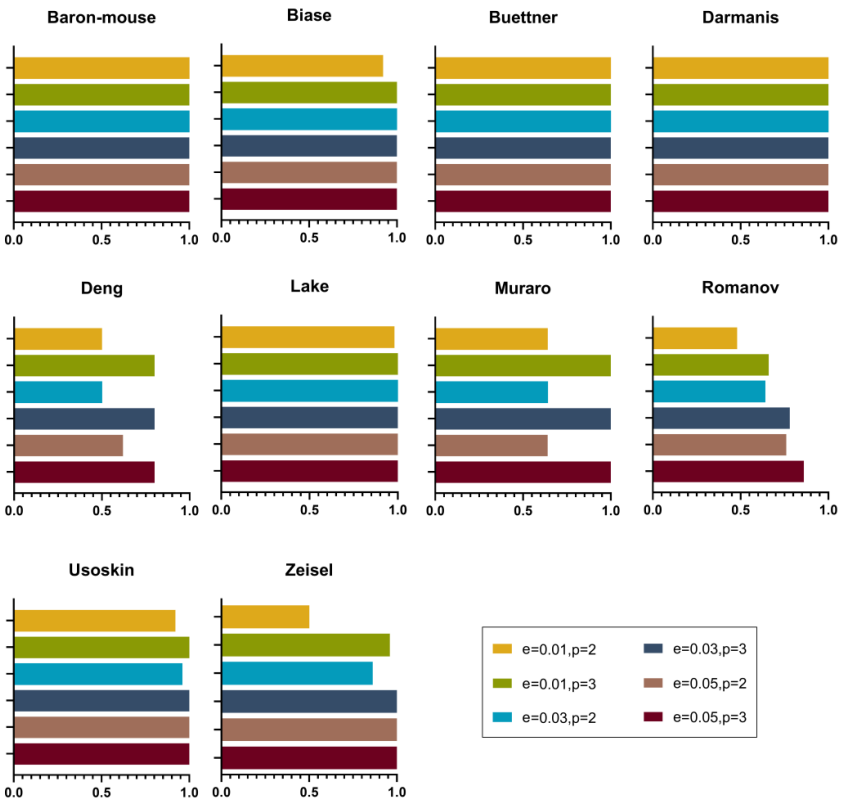

Figure 3. The accuracy of the *f-value* in selecting the optimal combinations. Each bar represents the accuracy of the *f-value* with different error ranges e and top ranks p.

2.3. Accuracy Comparison between the Selected Optimal Combination and the Nine Other Combinations

In this section, a comparison of the ARI values of the selected optimal combination of preprocessing and dimensionality reduction methods with the nine other combinations is performed. The model was run on each dataset 50 times separately, and the error bars of the ARI values were drawn accordingly (Figure 4). For a fair comparison, the Euclidean distance measure was applied to all the combinations.

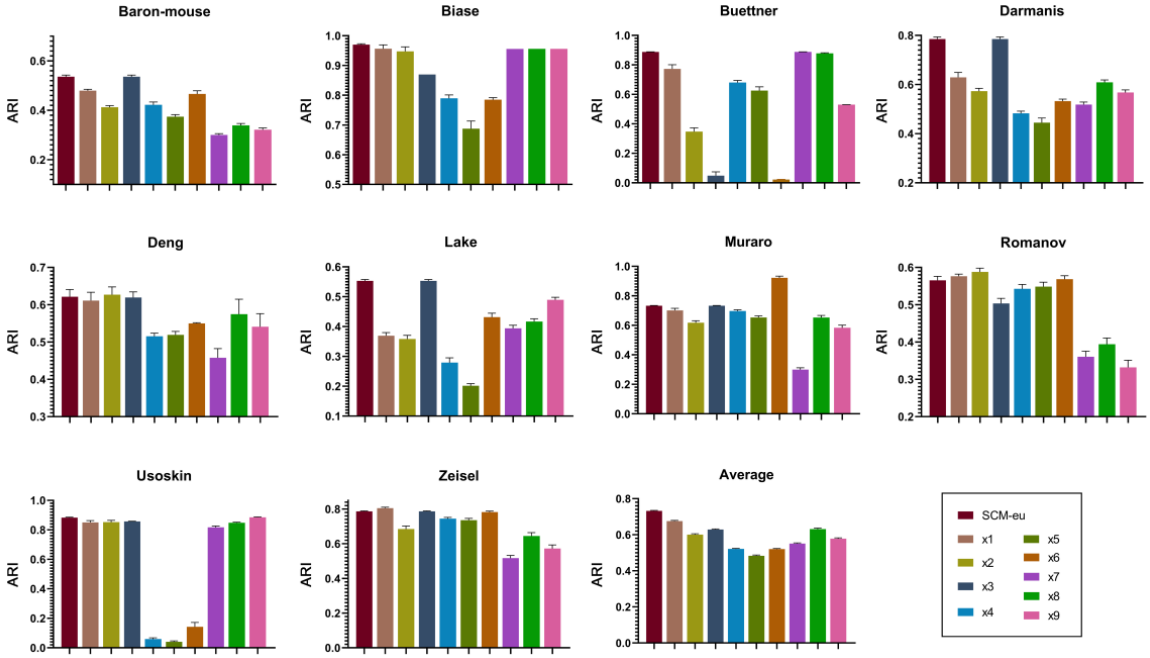

Figure 4. The performance of SCM among the nine combinations is shown in Table 1. The first bar represents the ARI value corresponding to the optimal consensus matrix with SC3's default distance calculation method. The bars x_1–x_9 symbolize the nine combinations (Table 1). Specifically, the blue bar (x_3) depicts the results of SC3 with default parameters. Each panel represents a dataset. The last panel shows the average ARI value over the ten datasets.

As shown by the comparison, the *f-value* can always identify the best combination on most datasets (Figure 4). In comparison with the default SC3 framework corresponding to x_3, the selected optimal combination was more conducive to improving the clustering efficiency. For example, on the Baron-mouse, Biase, Buettner, Darminis, Lake, and Usoskin datasets, the average ARI values of the selected combinations were or approached the highest among all nine combinations. On the Muraro dataset, although the selected combination was not the best one, its rank was high, and its ARI value was higher than those of all the others except for the best one. On the other datasets, such as the Deng, Romanov, and Zeisel datasets, the ARI values of the selected combination were only slightly lower than those of the best one, and its rank was high. In summary, the *f-value* effectively improved the accuracy of cell type clustering, which could facilitate novel discoveries in scRNA-seq analysis. In addition, we showed the heatmaps of selected optimal consensus matrices on the ten datasets (see Supplementary Figures S10–S19).

2.4. Performance Evaluation of the Reconstructed Flexible Distance

After determining the optimal combination of preprocessing and dimensionality reduction methods, we redefined a new cell-to-cell distance (termed SCM-tom) for the optimal consensus matrix to replace the default Euclidean distance of SC3. To evaluate the performance of SCM-tom, we made a comparison between the two distance measures.

Each distance measure was run 50 times, and the average ARI value was calculated for the accuracy comparison.

The comparison showed that the new SCM-tom distance measure performed much better than the Euclidean distance, named SCM-eu, and the average accuracy improvement of SCM-tom over SCM-eu was 4.77% on the ten datasets (Figure 5). On the Muraro, Baron-mouse, Deng, and Romanov datasets, the improvement even reached 11.79–27.89%. Based on the ten datasets, we found that the Euclidean distance consistently performed worse, possibly because it was unable to effectively capture differential patterns between cells. Overall, the new distance metric demonstrated great advantages over the traditional Euclidean distance in improving cell type clustering, possibly because it overcame the information loss problem experienced by the Euclidean distance metric, thereby further enhancing the clustering performance.

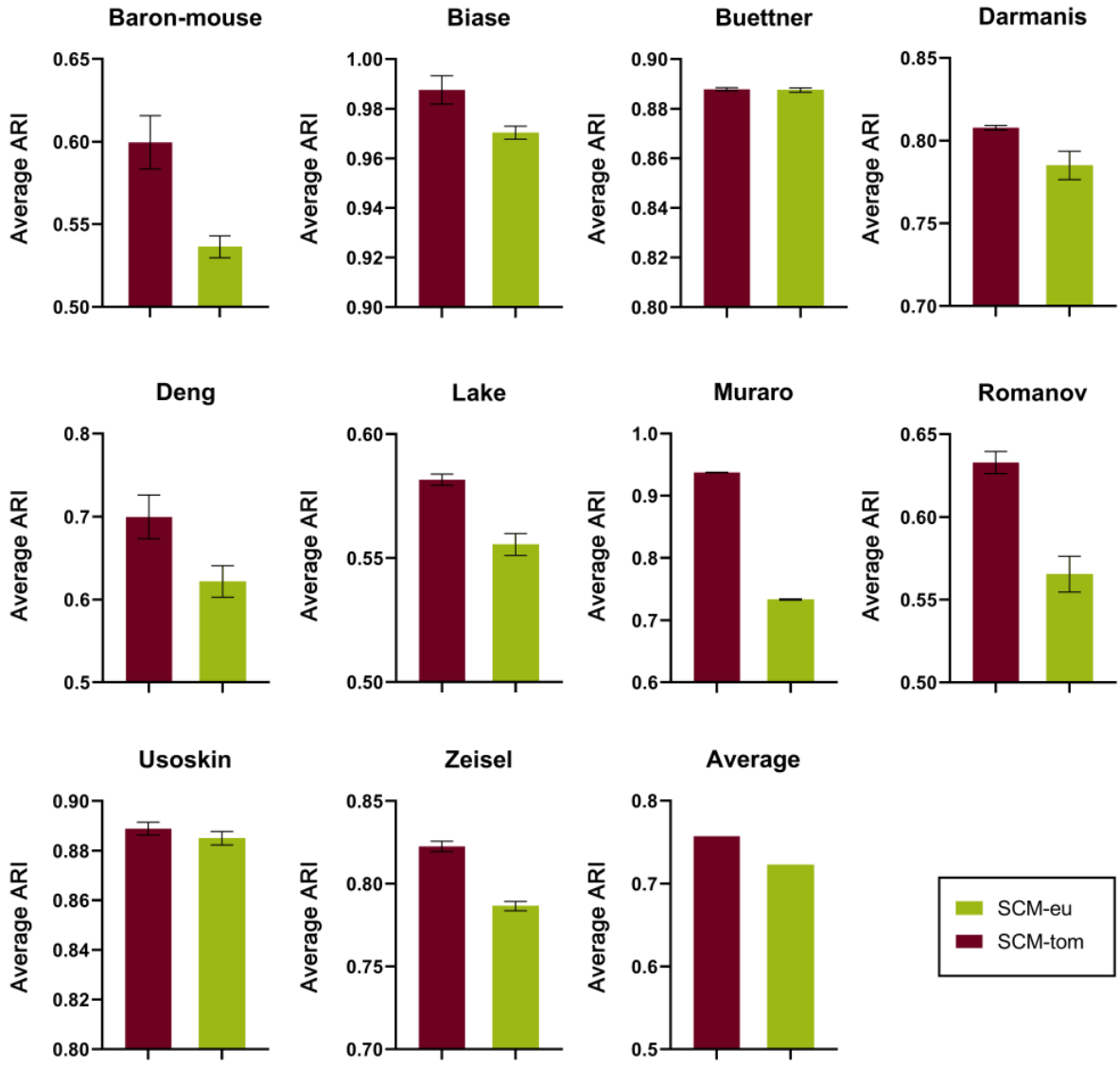

Figure 5. The performance of SCM-tom and SCM-eu. Each panel represents a dataset. The last panel shows the average ARI value over the ten datasets.

2.5. Performance Comparison between SCM and Other Popular Clustering Algorithms

To evaluate the performance of SCM and other state-of-the-art cell type clustering methods, we compared SCM with seven popular single-cell clustering algorithms on the ten datasets, including SC3, Seurat, Soup, CIDR, pcaReduce, SIMLR, and tSNE + kmeans (see Supplementary Notes for parameter setup and versions of the clustering methods).

Due to the instabilities of SCM, SC3, and pcaReduce, each of them was run 50 times, and the average ARI value of the 50 clustering results was utilized for the performance comparison.

After comparison, the results showed that SCM had the best performance on all datasets, and the average accuracy improvement of SCM over the second best method reached 29.31% (Figure 6). In particular, the accuracy improvements of SCM over the other compared methods on the Baron-mouse, Darmanis, Muraro, Deng, and Romanov datasets were 21.95–145.40%, 10.17–171.05%, 27.79–325.79%, 8.02–129.16%, and 37.63–99.35%, respectively. Specifically, SCM largely improved SC3 on all ten datasets by optimizing the framework. For example, on the Buettner dataset, the ARI value of SC3 was no more than 0.01, while the ARI value of SCM reached 0.89. On the Muraro dataset, the ARI value of SCM was 0.93, whereas that of SC3 was only 0.73, and most of the other clustering algorithms were concentrated at approximately 0.4. As a whole, the SCM algorithm effectively selected the optimal combination of preprocessing and dimensionality reduction methods and then utilized a reliable cell-to-cell distance measure, which greatly improved the clustering performance.

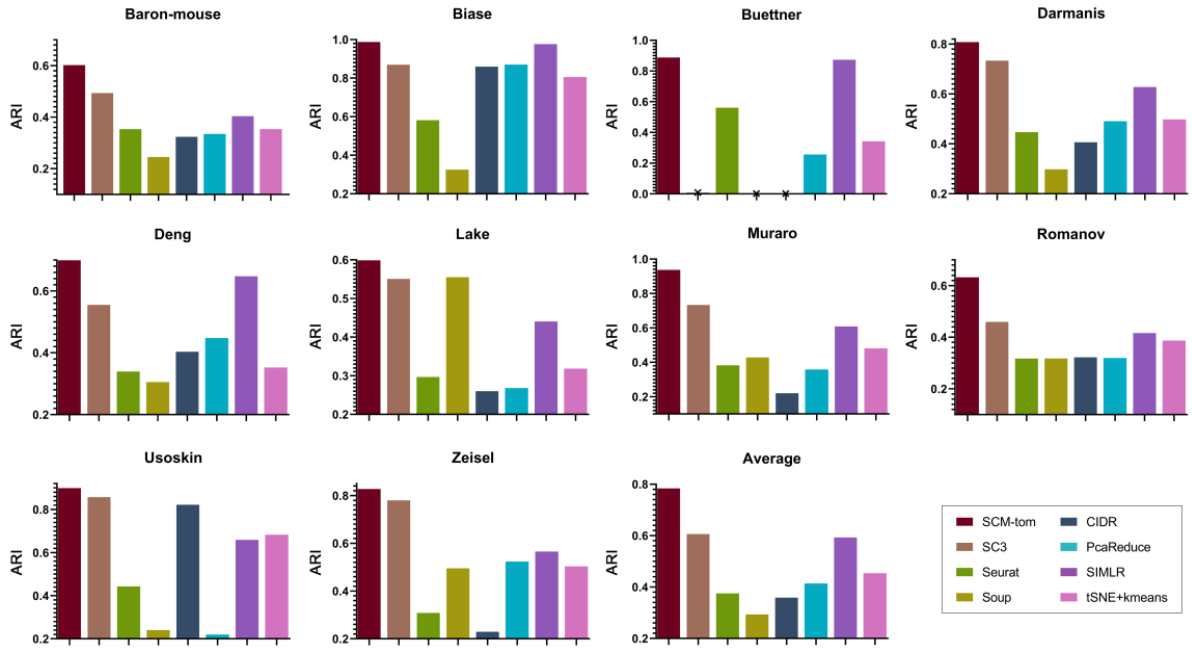

Figure 6. Performance of different algorithms on the ten datasets. Each panel represents a dataset. Some bars are not shown on the Buettner dataset because the corresponding ARI values were no more than 0.1. The last panel shows the average ARI value for each algorithm on the ten datasets.

2.6. Visualization of the Clustering Results

For the purpose of illustrating the clustering results more intuitively, we used tSNE to visualize the clustering results of each clustering algorithm. The clustering results of the Zeisel dataset are displayed in Figure 7. Visualizations of the remaining datasets are provided in Supplementary Figures S1–S10. Each point in the figure represents a cell, and the cells were labelled by a clustering method and marked by different colours. From the visualizations, it was clear that the clustering results of SCM highly overlapped with the real labels. Focusing specifically on the circled results in the figure, SCM provided clearer boundaries than the author's true labels. However, cells in the circle were erroneously clustered in one cell type by SC3, SIMLR, and tSNE + kmeans. Compared with the clustering results of other algorithms, SCM exhibited high accuracy in the partitioning of cell populations. In conclusion, SCM can accurately distinguish cell types and greatly improve the clustering efficiency.

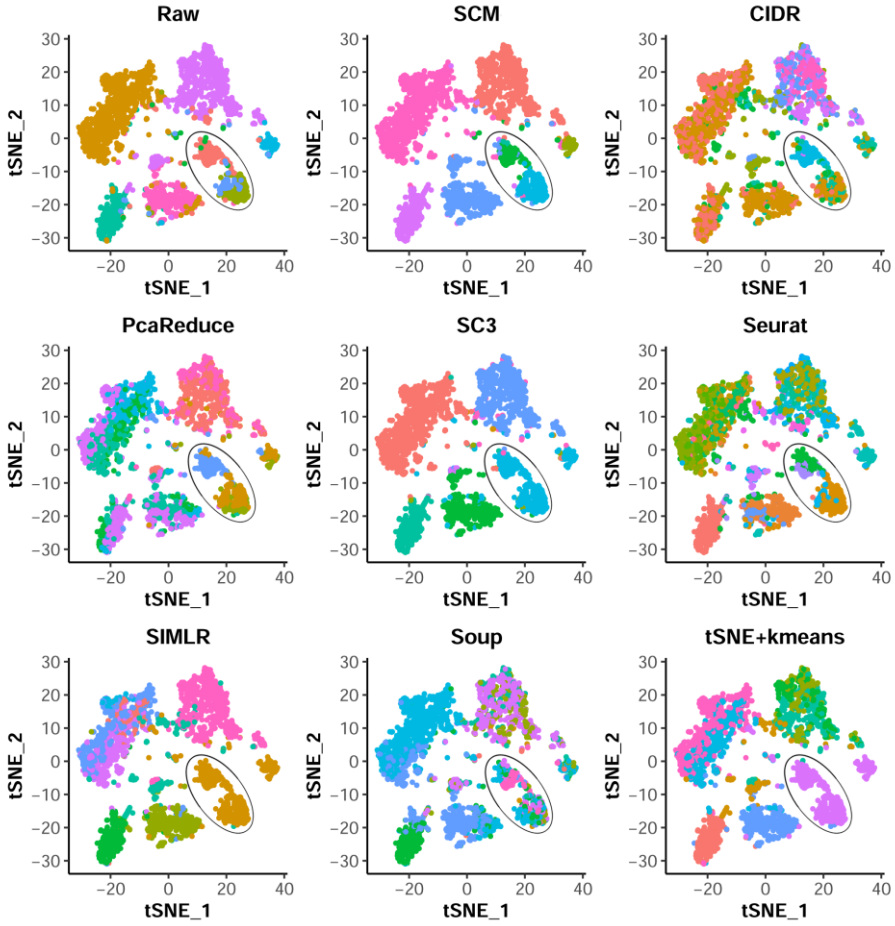

Figure 7. Visualizations of the clustering results of different clustering methods on the Zeisel dataset. Each panel represents a clustering algorithm. The first panel represents the true labels annotated by the author of the dataset. Cells are colored according to their true cellular labels.

3. Discussion

The clustering of cells, as a common analytical approach, plays an increasingly powerful role in single-cell studies. Therefore, achieving accurate clustering of single-cell data is very important for research in the field of bioinformatics. In this study, we proposed a novel single-cell clustering algorithm by scoring consensus matrices and redefining a more reliable cell-to-cell distance measure. Experiments on ten benchmark datasets showed that SCM performed much better than all the compared clustering algorithms, which demonstrates its great potential in revealing cellular heterogeneity and identifying new cell types. The work and innovations in this paper were mainly based on the following two points.

First, we designed the *f-value* scoring mechanism to achieve the flexible selection of the optimal combinations of preprocessing and dimensionality reduction methods based on data specificity. The experimental results confirmed the strong effectiveness of the *f-value* in identifying the optimal combination. Second, a flexible distance measure named SCM-tom was reconstructed with data specificity by calculating both the direct and indirect distances of cells and fully capturing the topological structure information among cells. The experimental results again confirmed the great advantages of SCM-tom over the most used Euclidean distance.

Although the great advantages of SCM in single-cell clustering have been demonstrated, further improvements to SCM can still be made in the future. For example, the current version of SCM only considered three preprocessing methods and three pairs of dimensionality reduction methods, which may limit the optimal selection of SCM. In addition, the current version only utilized gene expression data, which may also limit its

performance in some cases. In the future, more preprocessing methods and three pairs of dimensionality reduction methods and multiomics data will be added to achieve more accurate cell type clustering.

The software was developed to be user-friendly and is expected to play a crucial role in new discoveries of single-cell clustering using scRNA-seq, especially in complex human diseases, such as cancers, and in the discovery of new cell types.

4. Methods

4.1. Datasets

A total of ten public scRNA-seq datasets were used to evaluate the performance of the clustering methods, and the true cell types in each dataset were validated in previous studies. The ten datasets were Biase [26], Deng [27], Darmanis [28], Muraro [29], Usoskin [30], Romanov [31], Zeisel [32], Lake [33], Buettner [34], and Baron-mouse [35]. The numbers of cells (ranging from 56 to 3042) and genes (ranging from 8989 to 25,734) in these datasets are presented in Supplementary Table S1.

4.2. Evaluation Metrics

The clustering accuracy in this study was measured by the adjusted rank index (ARI) [36]. Here, $U = \{u_1, u_2,..., u_p\}$ was the true partition of the cells into P clusters, and $V = \{v_1, v_2,..., v_k\}$ was the partition calculated by a clustering algorithm. The ARI of these two partitions is defined as follows.

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{n}{2}}$$

$$n = \sum_{i=1}^{P} n_i = \sum_{j=1}^{K} n_j$$

where n_{ij} is the number of objects that are in both class u_i and cluster v_j; n_i and n_j are the numbers of objects in class u_i and cluster v_j, respectively. The value of ARI belongs to $[-1, 1]$, and a larger ARI value represents a better clustering result.

4.3. Preprocessing of the Gene Expression Matrix

Three commonly used preprocessing methods, including a *log* transformation, no transformation, and a *z-score* transformation, were employed to analyse the effects of different preprocessing methods on cell type clustering. Given a gene expression matrix $X_{G\times N} = \{x_{ij}\}$ with G genes and N cells, data preprocessing is performed as follows.

$$log \text{ transformation}: \; x'_{ij} = \log_2(x_{ij} + 1)$$

$$\text{no transformation}: \; x'_{ij} = x_{ij}$$

$$z - score \text{ transformation}: \; x'_{ij} = \frac{x_{ij} - u_i}{\sigma_i}$$

where u_i and σ_i represent the mean and standard deviation of the i-th row of $X_{G\times N}$, respectively.

As illustrated in Figure 8, three processed gene expression matrices were obtained after the above three preprocessing methods. Afterwards, the Euclidean distance, Pearson distance, and Spearman distance between two cells were calculated for each processed expression matrix. Consequently, three distance matrices can be obtained for each processed expression matrix.

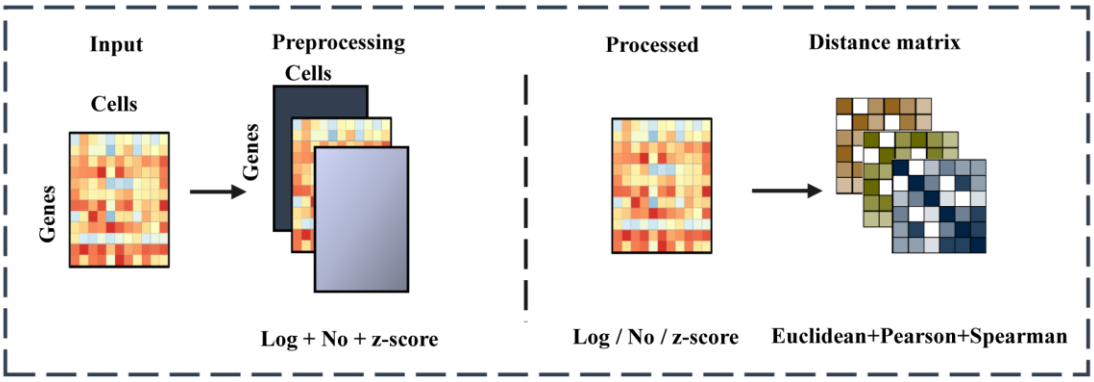

Figure 8. Overview of data preprocessing.

4.4. Data Dimensionality Reduction

After preprocessing of the gene expression matrix, each pair of the three dimensionality reduction methods (PCA + UMAP, LP + UMAP, and LP + PCA) was applied to each distance matrix. Based on the pipeline of SC3, each combination of a preprocessing method and a pair of dimensionality reduction methods can yield a consensus matrix (Figure 9).

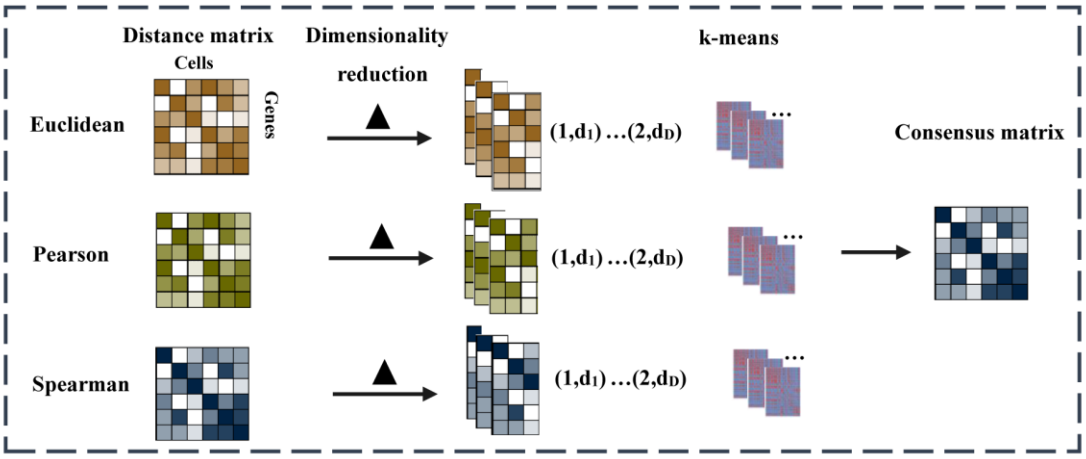

Figure 9. The process of calculating a consensus matrix from a pre-processed gene expression matrix.

As mentioned above, each processed gene expression matrix can generate three distance matrices. Dimensionality reduction was performed on the three distance matrices via each of the three pairs of dimensionality reduction methods. After each dimensionality reduction process for each distance matrix, $d_1, d_2, \ldots, d_D$ dimensions were retained (4–7% of the original dimensions). Then, each distance matrix generated $2 \times d_D$ dimensionality reduction results under the pair of dimensionality reduction methods. Therefore, three distance matrices under a pair of dimensionality reduction methods can generate a total of $3 \times 2 \times d_D$ dimensional reduction results. Then, SCM utilized k-means clustering for each dimensional reduction result and fused them into a consensus matrix using consensus clustering (see Supplementary Methods). Thus, nine consensus matrices were finally generated under the nine combinations of preprocessing methods and dimensionality reduction methods in Table 1.

4.5. Selection of the Optimal Preprocessing and Dimensionality Reduction Methods

After preprocessing and dimensionality reduction, a total of nine consensus matrices were constructed; hence, nine corresponding clustering results were obtained. In this study, the F-statistic [37] was utilized to quantitatively measure the quality of a consensus matrix. In detail, given a consensus matrix $Y_{N\times N} = \{Y_{ij}\}$, $Y_{ij} \in [0, 1]$ and its corresponding clustering result R, the *f-value* $F(Y)$ of consensus matrix Y was calculated as follows.

Step 1. Calculating the *f-value* of each row of the consensus matrix.

Suppose that each column of a consensus matrix represents a cell and that the rows denote the different states of the cells. Then, the quality of a consensus matrix should be determined by all the states, and therefore, we first calculate the *f-value* of each row of a consensus matrix to measure its cell type discrimination intensity. Then, the *f-values* of all the rows are effectively fused into the *f-value* of the whole consensus matrix.

First, the between-cluster variance of the i-th row of consensus matrix Y is calculated by the following formula:

$$var\_b(i) = \sum_{j=1}^{k} n_j (\overline{Y}_i - \overline{Y}_{ij})^2,$$

where n_j is the number of cells (columns) in the j-th cluster of R; k is the number of clusters in R; $\overline{Y}_i$ is the mean of the i-th row of Y, and $\overline{Y}_{ij}$ is the mean of the i-th row of Y in the j-th cluster of R.

Then, the in-cluster variance of the i-th row of consensus matrix Y is computed as follows:

$$var\_a(i) = \sum_{j=1}^{N} (\overline{Y}_i - Y_{ij})^2,$$

$$var\_i(i) = var\_a(i) - var\_b(i),$$

where N is the number of cells, Y_{ij} is the value of the i-th row of Y and $var\_a(i)$ denotes the entire variance from both the between-cluster and in-cluster distances.

Finally, the *f-value* of the i-th row of consensus matrix Y is generated by the following formula:

$$F(i) = \frac{var\_b(i)/df_1}{var\_i(i)/df_2}$$

$$df_1 = k - 1, \quad df_2 = N - k$$

where df_1 and df_2 are the degrees of freedom of $var\_b(i)$ and $var\_i(i)$, respectively.

Step 2. Generating the *f-value* of a consensus matrix.

After the calculation of *f-value* $F(i)$ for each row of consensus matrix Y, the mean and standard deviation of the *f-values* $F(1)$, $F(2)$,..., $F(N)$ are, respectively, calculated and denoted by $mean(F_Y)$ and $dev(F_Y)$. In this study, we proposed a hypothesis that a higher quality consensus matrix has larger row *f-values* with smaller variance. Therefore, the final *f-value* of consensus matrix Y is calculated by the following formula.

$$F(Y) = \sqrt[\alpha]{mean(F_Y)} - \lambda \sqrt[\alpha]{dev(F_Y)}$$

where λ and α are set to 0.5 and 5, respectively, by default.

Based on the definition of the *f-value* for each row of a consensus matrix, the larger the $F(i)$ value is, the smaller the in-cluster cell-to-cell variance and the larger the between-cluster variance, which leads to higher quality values in the i-th row. Accordingly, the larger the $F(Y)$ value is, the higher quality the whole consensus matrix is. After computing the *f-value* of the nine consensus matrices, the one with the largest *f-value* is determined as the optimal consensus matrix.

4.6. Definition of a Flexible Cell-to-Cell Distance SCM-Tom

The SC3 method applies the fixed Euclidean distance on the consensus matrix. In this paper, a flexible distance measure termed SCM-tom was developed to reconstruct more reliable distances between cells with data specificity. The new distance measure was improved from the weighted correlation network analysis (WGCNA) approach [38] as follows.

(1) Since the value Y_{ij} in consensus matrix Y represents the probability of the i-th and j-th cells belonging to the same cell type, the original consensus matrix is defined as the initial correlation coefficient matrix S:

$$S_{ij} = \left|cor(i,j)\right| = Y_{ij}$$

(2) The β index is introduced to increase the variability between the correlation coefficients by constructing a new adjacency matrix $A = \{\alpha_{ij}\}$.

$$\alpha_{ij} = S_{ij}^{\beta}$$

where α_{ij} redefines the association strength of the two cells i and j, and the default value of β in this study is set to 8.

(3) The connectivity strength k_j of each cell j with the other cells is obtained by the following formula.

$$k_j = \sum_{i=1, i\neq j}^{n} \alpha_{ij}$$

In this study, the relationship between two cells i and j was determined not only by their direct association strength α_{ij} but also by their indirect connections via their common neighbours. For example, if both cells i and j have strong association strengths α_{iu} and α_{uj} with cell u, it is assumed that cells i and j have a close relationship. Based on the above assumption, the topological overlap matrix $TOM\ \Omega = \{\omega_{ij}\}$ is generated as follows.

$$\omega_{ij} = \frac{\sum_{u} \alpha_{iu}\alpha_{uj} + \alpha_{ij}}{\min\{k_i, k_j\} + 1 - \alpha_{ij}}$$

The topological overlap matrix focuses on primary and secondary associations and describes the connection relationship between two cells. Based on the above definition, the value ω_{ij} indicates the similarity between the i-th and j-th cells at the level of both direct and indirect associations. The larger the value ω_{ij} is, the closer the two cells are to each other.

Finally, the new flexible distance matrix $D = \{d_{ij}\}$ is obtained by the following formula.

$$d_{ij} = 1 - \omega_{ij}$$

Therefore, d_{ij} represents the distance between the two cells i and j and is in the range of 0 and 1, and a larger value indicates a greater difference between two cells. Based on the SCM-tom distance measure, hierarchical clustering is performed to complete the final cell clustering step.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math11173785/s1, Supplementary Materials.pdf. This file contains Supplemental Methods: Consensus clustering; Supplementary Table: Table S1 (11 Datasets used for validations); Supplementary Figures: Figures S1–S9 (Visualizations of the clustering results on nine datasets); Figures S10–S19 (Heatmaps of optimal consensus matrices on the ten datasets).

Author Contributions: Conceived and designed the experiments: J.L. Performed the experiments: Y.Y. and J.L. Analysed the data: Y.Y. and J.L. Contributed reagents/materials/analysis tools: J.L. Wrote the paper: Y.Y. and J.L. Designed the software used in analysis: Y.Y. Oversaw the project: J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, grant number 2020YFA0712400, and the National Natural Science Foundation of China, grant number 62272268. And the APC was funded by the National Key R&D Program of China. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement: The source code of SCM and the ten datasets used in this study are available at https://sourceforge.net/projects/transcriptomeassembly/files/SCM/ (accessed on 31 May 2023).

Conflicts of Interest: The authors declare that they have no competing interest.

References

1. Potter, S.S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* **2018**, *14*, 479–492. [CrossRef] [PubMed]
2. Tang, X.; Huang, Y.; Lei, J.; Luo, H.; Zhu, X. The single-cell sequencing: New developments and medical applications. *Cell Biosci.* **2019**, *9*, 53. [CrossRef]
3. Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **1999**, *6*, 281–297. [CrossRef] [PubMed]
4. Hedlund, E.; Deng, Q. Single-cell RNA sequencing: Technical advancements and biological applications. *Mol. Asp. Med.* **2018**, *59*, 36–46. [CrossRef]
5. Xu, C.; Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **2015**, *31*, 1974–1980. [CrossRef] [PubMed]
6. Stuart, T.; Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **2019**, *20*, 257–272. [CrossRef] [PubMed]
7. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R.; et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **2017**, *14*, 483–486. [CrossRef]
8. Yang, Y.; Huh, R.; Culpepper, H.W.; Lin, Y.; Love, M.I.; Li, Y. SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics* **2018**, *35*, 1269–1277. [CrossRef]
9. Zhu, L.; Lei, J.; Klei, L.; Devlin, B.; Roeder, K. Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. USA* **2018**, *116*, 466–471. [CrossRef] [PubMed]
10. Lin, P.; Troup, M.; Ho, J.W.K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **2017**, *18*, 59. [CrossRef]
11. Žurauskienė, J.; Yau, C. pcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* **2016**, *17*, 1–11. [CrossRef] [PubMed]
12. Wang, B.; Ramazzotti, D.; De Sano, L.; Zhu, J.; Pierson, E.; Batzoglou, S. SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning. *Proteomics* **2017**, *18*, 1700232. [CrossRef] [PubMed]
13. Petegrosso, R.; Li, Z.; Kuang, R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings Bioinform.* **2020**, *21*, 1209–1223. [CrossRef] [PubMed]
14. Bacher, R.; Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **2016**, *17*, 1–14. [CrossRef] [PubMed]
15. Ding, J.; Condon, A.; Shah, S.P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **2018**, *9*, 2002. [CrossRef] [PubMed]
16. Sun, S.; Zhu, J.; Ma, Y.; Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.* **2019**, *20*, 269. [CrossRef] [PubMed]
17. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*, 1–5. [CrossRef] [PubMed]
18. Guo, M.; Wang, H.; Potter, S.S.; Whitsett, J.A.; Xu, Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput. Biol.* **2015**, *11*, e1004575. [CrossRef]
19. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
20. Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **2003**, *15*, 1373–1396. [CrossRef]
21. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I.W.H.; Ng, L.G.; Ginhoux, F.; Newell, E.W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38–44. [CrossRef]
22. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
23. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef] [PubMed]
24. Torgerson, W.S. Multidimensional scaling: I. Theory and method. *Psychometrika* **1952**, *17*, 401–419. [CrossRef]
25. Senabouth, A.; Lukowski, S.W.; Hernandez, J.A.; Andersen, S.B.; Mei, X.; Nguyen, Q.H.; E Powell, J. ascend: R package for analysis of single-cell RNA-seq data. *GigaScience* **2019**, *8*, giz087. [CrossRef]
26. Biase, F.H.; Cao, X.; Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **2014**, *24*, 1787–1796. [CrossRef]
27. Deng, Q.; Ramsköld, D.; Reinius, B.; Sandberg, R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **2014**, *343*, 193–196. [CrossRef] [PubMed]
28. Darmanis, S.; Sloan, S.A.; Zhang, Y.; Enge, M.; Caneda, C.; Shuer, L.M.; Gephart, M.G.H.; Barres, B.A.; Quake, S.R. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 7285–7290. [CrossRef] [PubMed]

29. Muraro, M.J.; Dharmadhikari, G.; Grün, D.; Groen, N.; Dielen, T.; Jansen, E.; van Gurp, L.; Engelse, M.A.; Carlotti, F.; de Koning, E.J.; et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **2016**, *3*, 385–394.e3. [CrossRef]
30. Usoskin, D.; Furlan, A.; Islam, S.; Abdo, H.; Lönnerberg, P.; Lou, D.; Hjerling-Leffler, J.; Haeggstrom, J.Z.; Kharchenko, O.; Kharchenko, P.V.; et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **2015**, *18*, 145–153. [CrossRef] [PubMed]
31. Romanov, R.A.; Zeisel, A.; Bakker, J.; Girach, F.; Hellysaz, A.; Tomer, R.; Alpár, A.; Mulder, J.; Clotman, F.; Keimpema, E.; et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **2017**, *20*, 176–188. [CrossRef] [PubMed]
32. Zeisel, A.; Munoz-Manchado, A.B.; Codeluppi, S.; Lonnerberg, P.; La Manno, G.; Jureus, A.; Marques, S.; Munguba, H.; He, L.; Betsholtz, C.; et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **2015**, *347*, 1138–1142. [CrossRef] [PubMed]
33. Lake, B.B.; Ai, R.; Kaeser, G.E.; Salathia, N.S.; Yung, Y.C.; Liu, R.; Wildberg, A.; Gao, D.; Fung, H.-L.; Chen, S.; et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **2016**, *352*, 1586–1590. [CrossRef] [PubMed]
34. Buettner, F.; Natarajan, K.N.; Casale, F.P.; Proserpio, V.; Scialdone, A.; Theis, F.J.; Teichmann, S.A.; Marioni, J.C. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cell. *Nat. Biotechnol.* **2015**, *33*, 155–160. [CrossRef]
35. Baron, M.; Veres, A.; Wolock, S.L.; Faust, A.L.; Gaujoux, R.; Vetere, A.; Ryu, J.H.; Wagner, B.K.; Shen-Orr, S.S.; Klein, A.M.; et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **2016**, *3*, 346–360.e4. [CrossRef]
36. Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
37. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics: Methodology and Distribution*; Kotz, S., Johnson, N.L., Eds.; Springer: New York, NY, USA, 1992; pp. 66–70.
38. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef]