IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 22, 2020, accepted September 4, 2020, date of publication September 8, 2020,
date of current version September 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022718

GRACE: A Graph-Based Cluster Ensemble Approach for Single-Cell RNA-Seq Data Clustering

JIHONG GUAN, RUI-YI LI, AND JIASHENG WANG
Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Corresponding author: Jiasheng Wang (13wjs@tongji.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61772367, and in part by the Joint Funds of the National Natural Science Foundation of China (NSFC) under Grant U1936205.

ABSTRACT Rapid development of single cell RNA sequencing (scRNA-seq) technology has accelerated the exploration in biomedical researches. One of the focal interests in scRNA-seq data analysis is to classify cells into different types, which significantly assists in studying inter-cellular heterogeneity, such as cell types, cell states, and cell lineages, at the resolution of single cells. Although a number of tailored approaches have been developed for scRNA-seq data, their performance varies with different datasets and their clustering accuracy need to be improved. In this paper, we propose a novel ensemble clustering framework for scRNA-seq data called **GRACE** (**GRA**ph-based **C**luster **E**nsemble approach). First, we construct a highly reliable graph network for single cells by combining the clustering outcomes from five leading scRNA-seq data clustering methods. Then, we remeasure the relationships between cells by exploring the topology structure of network using random walk distance. Finally, we build a hierarchical cell-tree and obtain the clustering labels by cutting the tree structure into an appropriate number of sub-trees. Experimental results on twelve benchmark datasets show that GRACE has the higher clustering accuracy and is more robust among a variety of datasets than the state-of-the-art individual approaches. In addition, the graph structure of the network which is built upon the ensemble clusters is more reliable than the networks which are constructed according to the conventional similarity metrics.

INDEX TERMS Single cell RNA-seq data, ensemble cluster, random walk distance, graph theory, hierarchical clustering.

I. INTRODUCTION

As the basic structural and functional unit of organisms, single cells store the important genetic information [1]. In the process of cell proliferation and differentiation, a number of factors, such as cell state [2], micro-environment of cells [3], and the regulation of internal procedures of cells, lead to the heterogeneity of cells [4]. Previously, the technology of large population sequencing often analyzes tens of thousands cells altogether, where the expression value of gene is the average score of all the cells. It thus usually highlights the cell types with large populations and belies the rare cell types such as stem cells and cancer cells [5], [6]. Fortunately, the single cell RNA sequencing (scRNA-seq) technology can overcome this issue and promote the study of cellular heterogeneity [6], [7]. Clustering analysis, which can group cells according to gene expression patterns, is essential in order to mining the underlying information of scRNA-seq data. Related studies on clustering analysis have been applied to many focal research interests, such as discovering cell types [8], [9], reconstructing cell development tracks, fate decisions [10], [11], and establishing spatial models of complex tissues [12].

Clustering analysis has always been a focal research interest in data mining and machine learning [13]. Up to now, the traditional classic clustering algorithms, such as k-means [14], DBSCAN (Density-Based Spatial Clustering of Application with Noise) [15], CLIQUE (Clustering In QUEst) algorithm [16], spectral clustering [17] and hierarchical clustering [18] are still wildly used. At the same time, clustering algorithms are advancing to pursue higher clustering accuracy and efficiency. Following the idea that clusters are the high density regions in the feature space separated by low density regions, a density-based method has been proposed, which bases on fast searching and finding density peaks [19]. To handle high-dimensional realistic data, some advanced subspace clustering algorithms are proposed

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

from novel perspectives, such as eliminating the effect of the errors from the linear projection space [20], combing with deep neural networks architectures [21], [22]. In addition, many ensemble clustering approaches have been developed to achieve better clustering results and greater robustness. For example, Huang *et al.* have developed a serials of ensemble clustering algorithms by factor graph [23], probability trajectories [24], locally weighting base clusterings [25], exploring cluster-wise similarities via random walks [26] and integrating ultra-scalable spectral clustering [27]. In the area of scRNA-seq clustering analysis, many tailored methods have been developed to overcome the challenges posed by the inherent nature of scRNA-seq data, such as zero inflation (dropouts) [28], over-dispersion [29] and amplification bias [30], and we will briefly review some of the major approaches.

Kiselev *et al.* proposed a consensus clustering method named single-cell consensus clustering (SC3), which adopts three measurements to calculate the similarity between cells and two ways for feature reduction. By applying k-means clustering algorithm on each branching data, they construct a consistent matrix of cells and then use hierarchical clustering to obtain the final clustering results [31]. Hierarchical clustering is also applied by SINCERA [32], Clustering through Imputation and Dimensionality Reduction (CIDR) [33], and cellTree [34]. The main difference of these approach is the method to measure the similarity between samples, where SINCERA uses Pearson correlation coefficient, cellTree employs Chi-Square distance, and CIDR applies the square of the Euclidean distance. In addition, CIDR improves the clustering efficiency by an implicit imputation approach to alleviate the impact of dropouts in scRNA-seq data. Sun *et al.* proposed a probability model based method DIMM-SC, which assumes that the data is generated by k polynomial distribution whose parameters follow the Dirichlet prior distribution, and solves the parameters with maximum likelihood estimation [35]. Some researches apply bi-clustering methods to scRNA-seq data. BackSPIN splits the similarity matrix of samples and assigns genes to each sub-matrix iteratively in order to cluster cells and genes simultaneously [36]. There are some clustering algorithms specially developed to detect rare cell types, such as giniClust and RaceID, which can group the datasets with uneven population distributions [37], [38]. Recently, some deep learning methods have been proposed with the emergence of big data. Lopez *et al.* proposed a method named scVI for processing scRNA-seq data which applies the deep generation model like Variational Auto-Encoder to implement the low-dimensional representation of data and then clusters low-dimensional data using k-means [39].

In addition to the methods mentioned above, another import type of clustering methods is graph theory-based approaches. The purpose is to segment the graph network, trying to make the edge weights (similarities) within the sub-graphs as high as possible while the edge weights connecting different sub-graphs as low as possible. For example, clusters can be formed by splitting the smallest spanning tree found on the network or using minimum-cut algorithm to finding pre-defined sub-graphs [40], [41]. Chen *et al.* proposed an approach called SNNCliq, which identifies clusters by a quasi-clique-based clustering algorithm on a graph constructed based on shared nearest neighbor (SNN) [42]. Seurat is a graph modularity optimization-based clustering method. It constructs a graph network of cells with SNN similarity and then optimizes the modularity function to determine clusters [12]. Wang *et al.* proposed the Single-cell Interpretation via Multi-kernel LeaRning (SIMLR), which constructs the graph of cells based on the similarity learned from multiple kernels and uses spectral clustering algorithm on the graph for clustering [43].

Previous methods mostly use conventional similarity metrics (such as Euclidean distance, Pearson correlation coefficient) or the second order similarity (such as SNN that considers 1-hop neighbors) to measure the similarity of single cells. The graph division approaches which rely on these similarities to find dense sub-graphs loss the graph topology properties since they don't consider the higher-order neighboring information, such as k-hop ($k \geqslant 2$) neighbors. Besides, the methods are often optimized for the specific dataset. Therefore, the outputs of those methods are unstable among different scRNA-seq datasets and it is hard for users to select an appropriate methods to apply. At the same time, the clustering accuracy also needs to be improved.

To address above problems, we propose a novel cluster ensemble approach called GRACE, which is a graph theory based clustering method. First, we construct a graph network using the predicting results of multiple basic clustering methods. Then, we build a tree of cells based on random walk distance on graph which can be considered as a higher order similarity that takes the high-order topological information into consideration. Finally, we obtain the clusters by cutting the tree structure according to the average distance of intra-clusters. Experimental results on twelve benchmark datasets show that GRACE outperforms the state-of-the-art methods.

The rest of this paper is organized as follows: Section II and III introduce the twelve real scRNA-seq datasets and the clustering performance metrics. Section IV presents our method in detail. Section V talks about the experimental results. Section VI concludes this paper.

II. BENCHMARK DATASETS

Twelve scRNA-seq datasets are collected from publicly available platforms, such as ArrayExpress [44], Gene Expression Omnibus (GEO) [45], and Sequence Read Archive [46]. The brief information about these data are listed in Table 1, in which the header of ''#Cells'', ''#Genes'' and ''#Cluster'' indicate the number of cells (instances), genes (features), and clusters, respectively. Datasets are named by the accession numbers provided in the original publications. In addition, these datasets are collected from some representative sequencing platforms, including SMART-seq2 [47], [48], sci-RNA-seq [49], 10X Genomics [50], and

TABLE 1. Overview of twelve scRNA-seq datasets.

Dataset	#Cells	#Genes	#Clusters	Sequencing protocols
E-MTAB-2600 [52]	704	30768	3	SMART-seq2
GSE65525 [51]	2717	24175	4	inDrop
GSE108097 [48]	2746	20670	16	Microwell-seq
GSE98561 [49]	4186	13488	10	sci-RNA-seq
SRP073767 [8]	4271	16653	8	10X
GSE60361 [36]	3005	19972	9	Quantitative scRNA-seq
GSM2230757 [53]	1937	20125	14	inDrop
GSM2230758 [53]	1724	20125	14	inDrop
GSM2230759 [53]	3605	20125	14	inDrop
GSM2230760 [53]	1303	20125	14	inDrop
GSM2230761 [53]	822	14878	13	inDrop
GSM2230762 [53]	1064	14878	13	inDrop

Droplet-based protocols [28], [51]. All of these data have 'gold-standard' (deemed as true) cluster labels assigned to each single cell, and the different clusters indicate the diverse cell groups.

To investigate the influence of culture condition in cellular self-renewal and pluripotency state, researchers sequence mouse Embryonic stem cells (mESCs) accross three different conditions: serum, 2i, and the alternative ground state a2i. E-MTAB-2600 is a dataset of mESCs, in which the three clusters corresponds to the three different conditions and there are 704 cells in total and 30768 genes are sequenced in each cell. This data is available in ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2600/).

GSE65525 is also a single cell data of mESCs which contains 2717 cells. GSE65525 reveals the population structure and the heterogeneous onset of cell differentiation after Leukemia Inhibitory Factor (LIF) withdrawal in mESCs. We downloaded the read count matrices of mESCs sample 1, mouse ES cells LIF - 2 days, mouse ES cells LIF - 4 days and mouse ES cells LIF - 7 days, and put all cells together. Distinct clusters are sets of cells with different days after LIF withdrawal. This data is downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525).

The SRP073767 dataset is provided by the 10X scRNA-seq platform, which profiles the transcriptome of the peripheral blood mononuclear cells (PBMCs) from a healthy donor. The total number of cells is 4217 classified in 8 types. This data is downloaded from the website of 10X genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k).

Since the brain function is relies on a diverse set of differentiated cell types, including neurons, glia, and vasculature. The authors of the GSE60361 data used large-scale scRNA-seq to classify cells from mouse somatosensory cortex and hippocampal CA1 region. The 9 clusters indicates distinct cell types in mouse cortex. There are 3005 single cells in total and 19972 genes are sequenced. This data can be downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE60361).

The mouse bladder cells data is in the Mouse Cell Atlas project GSE108097, which obtains sequenced data by applying Microwell-seq, a high-throughput and low-cost scRNA-seq platform. The authors identify 16 cell types in mouse bladder tissue, which are considered to be different clusters in our experiment. In this data, there are 4186 instances and 13488 features. This data is provided by the authors (https://figshare.com/s/865e694ad06d5857db4b).

GSE98561 is a dataset of the worm neuron cells dataset which is profiled by sci-RNA-seq. The authors profiled about 4000 neural cells from the nematode Caenorhabditis elegans at the L2 larval stage and identified the cell types. After removing the cells with the "Unclassified" labels, we thus obtained 10 cell types (that is, the 10 clusters). This data is available in GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi) with the accession number GSE98561.

The last six data sets in Table 1 are from a super-dataset GSE84133, in which four of them (i.e. GSM2230757, GSM2230758, GSM2230759, GSM2230760) are single cell data of human pancreatic islets and two (i.e. GSM2230761, GSM2230762) are mouse pancreatic islets. Clusters in these datasets indicates different endocrine cell types, including rare ghrelin-expressing epsilon-cells, exocrine cell types, vascular cells, Schwann cells, quiescent and activated pancreatic stellate cells, and four types of immune cells. These datasets could be downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133).

III. PERFORMANCE EVALUATION METRICS

In our experimental results, Adjusted Rand Index (ARI) is used to evaluate the clustering performance, which is widely used when the sample labels of ground truth are given [54]. ARI calculates the agreement between the ground truth and the predict clustering labels and the calculation can be defined as

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

where n_{ij} is the value at the i^{th}-row and the j^{th}-column in the contingency table, a_i is the sum of the i-th row of the contingency table, b_j is the sum of the j-th column of the contingency table, and $\binom{n}{2}$ denotes a binomial coefficient.

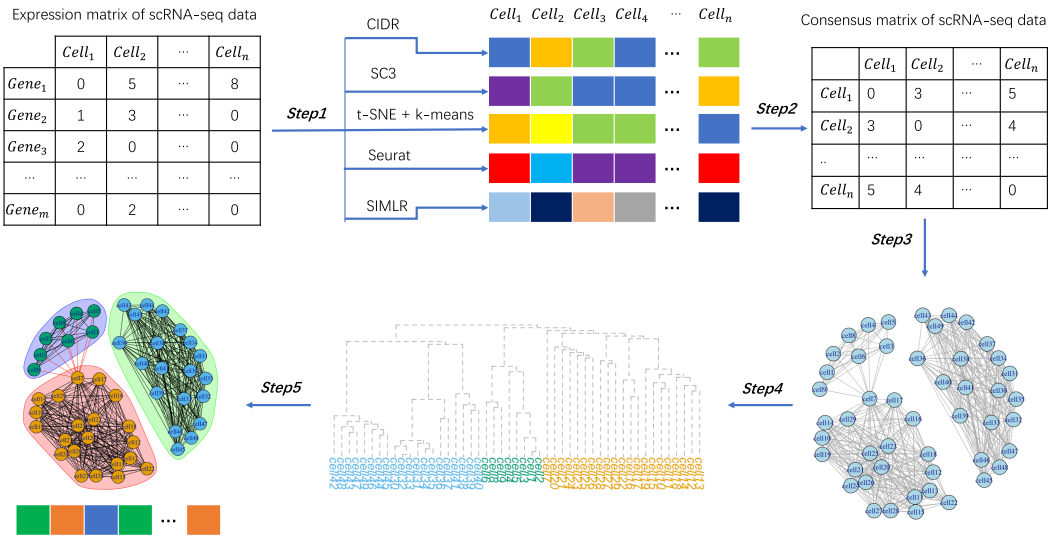

FIGURE 1. **Overview of GRACE. The first step is grouping the scRNA-seq data using CIDR, SC3, t-SNE+k-means, Seurat, and SIMLR. From the five clustering outcomes, we compute a consensus matrix representing the cell-to-cell relationships. Then a graph is constructed where nodes represents the single cells and weights of edges indicating each pair cells' similarity. Based on the random walk distance, we create a hierarchical tree of single cells. Finally, we implement clustering by cutting the tree-structure into sub-trees.**

We also evaluate the clustering performance according to other intuitive metrics, *Homogeneity score*, *Completeness score* and *V-measure*, that apply the conditional entropy analysis [55]. Given the ground truth class labels, Homogeneity score calculates the levels of whether each predicted cluster contains unique members of a single class of cells. The Completeness score indicates whether all the members of a given cell group are assigned to the same predicted cluster. V-measure is a balance metric of the Homogeneity and Completeness scores with computing the harmonic mean of them. Formally, the scores are defined as

$$\begin{cases} h = 1 - \dfrac{H(C|K)}{H(C)} \\ c = 1 - \dfrac{H(K|C)}{H(K)} \\ v = 2\dfrac{h \cdot c}{h + c} \end{cases} \tag{2}$$

where C and K is the set of ground-truth and the predicted cell labels, respectively. $H(C|K)$ indicates the conditional entropy of the classes given the cluster assignments and $H(C)$ means the entropy of the classes. Specifically,

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n} \cdot log\frac{n_{c,k}}{n_k} \tag{3}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n} \cdot log\frac{n_c}{n} \tag{4}$$

where n is the total number of cells. n_c and n_k are the number of cells which belong to the class c and the cluster k. $n_{c,k}$ is the number of cells from class c assigned to cluster k.

IV. METHODS

In this section, we show the whole picture of GRACE. First, we give a general description of GRACE and the five scRNA-seq clustering methods. Next, we describe the main steps in cluster ensemble process in detail, including the computation of the consensus matrix and random walk distance, the construction of the graph network and the hierarchical tree structure, and the estimation of the number of clusters. Finally, we summarize GRACE to the pseudo-code for a formal description.

A. OVERVIEW OF GRACE

Figure 1 shows the overview of our GRACE method. Generally, GRACE is composed by five parts. First, we take a scRNA-seq data gene expression matrix as input and use five clustering methods, which are CIDR [32], SC3 [31], t-SNE+k-means [37], Seurat [56], and SIMLR [43], to obtain five sets of clustering solutions. Second, the five individual solutions are combined into a $n \times n$ consensus matrix that representing the relationship between cells, where n represents the number of single cells. Third, basing on consensus matrix, a graph network is created where the nodes represent the cells and the weights of edges are set as the similarity value between pair of cells. Fourth, we measure the relationships between cells based on the random walk on the graph. Each node is assumed to be a cluster and the nodes are gradually grouped into a hierarchical tree under a specific merging criteria. Finally, we cut the hierarchical tree into the appropriate number of sub-trees and calculate the optimal clustering outcomes.

B. FIVE STATE-OF-THE-ART scRNA-SEQ DATA CLUSTERING METHODS

1) CIDR

The dropout event in scRNA-seq data is a big problem in the computational analysis. Lin *et al.* developed a clustering algorithm called CIDR which can be regarded as a fast principal coordinate analysis (PCoA)-like algorithm considering dropout events [32]. Since previous researches [30], [57] have shown that the possibility of a gene expression value being loss is inversely correlated with the true expression levels, CIDR reduces the dropout-induced zero inflation by imputing the zero values of dropout candidate genes which are collected from the zero peaks in the distribution of the log-transformed expression profile. Then, CIDR performs the dimension reduction approach, PCoA, on the imputed dissimilarity matrix of cells. Finally, CIDR applies the hierarchical clustering on the first few principal coordinates.

Besides, CIDR estimates the number of clusters k based on the Calinski-Harabasz index (CHI), also known as the variance ratio criterion [58]. By calculating the ratio of the sum of between-clusters dispersion and inter-cluster dispersion under different k, CIDR selects the most optimal k with the highest CHI score which indicates that the clusters are dense and well separated.

2) SC3

To achieve high accuracy and robust clustering solutions for scRNA-seq data, Kiselev *et al.* proposed a consensus clustering approach SC3 by combining multiple clustering results [31]. In the pre-processing step, SC3 filters out less-informative genes in scRNA-seq data expression profile. SC3 adopts three metrics, the Euclidean distance, the Pearson correlation coefficient, and the Spearman correlation coefficient, to calculate the similarity between each pair of cells. After that, SC3 applies two dimensionality reduction approaches, PCA and the Laplacian transform, on the three similarity matrix with two methods. Then, SC3 uses k-means on the data matrices to get different clustering results. Finally, SC3 constructs a consensus matrix of cells which combines the clustering outcomes and applies the hierarchical clustering on it for the final clustering labels. A hybrid SC3 approach is designed for the large datasets which groups 30% of cells using SC3, trains the support vector machine (SVM) with the clustering labels, and finally assigns the labels to the remaining cells.

To estimate the number of clusters, SC3 implements a random matrix theory (RMT) based approach, where the number of clusters is determined by the number of eigenvalues that are significantly different from the Tracy–Widom distribution [59], [60].

3) SIMLR

In scRNA-seq data clustering analysis, a key issue is selecting the appropriate similarity metrics for the cell-to-cell relationships. Wang *et al.* proposed a framework called SIMLR which learns a distance metric by combining multiple kernels to fit the structure of a specific scRNA-seq data [43]. SIMLR uses the Gaussian kernels with various hyper-parameters for the kernel construction. Assuming that the learned similarity matrix should have an approximate block-diagonal structure, where the cells have larger similarities to other cells within the same blocks, SIMLR applies an alternating convex optimization method to solve the objective optimization function which enforces the low rank constraint on the similarity matrix. Other computational analysis tasks such as visualization, dimension reduction, gene prioritization and clustering are all conducted on the learned similarity matrix. In the clustering task, SIMLR adopts the spectral clustering algorithm on the similarity matrix [17].

The number of clusters in SIMLR is determined by a heuristic approach based on the gap statistic [61].

4) SEURAT

Seurat is developed to identify and interpret the heterogeneity of single cells and integrate the diverse types of single-cell data [56], [62]. The approach identifies sub-populations of cells through unsupervised graph-based clustering. It calculates the k-nearest neighbors for each cell and then construct a shared nearest neighbor (SNN) graph in which the nodes represent the cells and the weights of the edges are the similarities between the cells. After that, it applies the smart local moving (SLM) algorithm to detect community on the SNN graph [63]. The SLM algorithm starts with a network in which each node is assigned to its own singleton community. It improves community structure by community merging and individual node movements to construct the final solution.

5) T-SNE+k-MEANS

It has been shown that dimension reduction before clustering is helpful for the improvement of scRNA-seq data clustering accuracy [31], [32], [57]. The method of "t-SNE+k-means" has successfully been applied in the rare cell types identification [37]. It reduces the high dimensional scRNA-seq data into a lower dimensional subspace by t-SNE algorithm and clusters the lower-dimensional data with k-means. In addition, the number of clusters are estimated by ADPclust which calculates the local density of samples and search for the cluster centers from estimated density peaks [64].

C. GRAPH-BASED CLUSTER ENSEMBLE METHOD

1) CONSENSUS MATRIX COMPUTATION

The consensus matrix is computed with the inferred cell labels from the five individual scRNA-seq clustering methods assuming that the more approaches divided two cells into the same cluster, the more similar the two cells are. Formally, we define the consensus matrix C as a $n \times n$ matrix, where n indicates the number of cells and the element c_{ij} in C is equal to the number of the scRNA-seq method that classifies cell i and j into the same cluster. In this case, the value of the elements in the consensus matrix ranges from 0 to 5.

2) GRAPH NETWORK CONSTRUCTION

Building a reasonable network for the single cells is a fundamental task since there is no actual network among them. The connectivity between two nodes usually depends whether the nodes are similar enough. Consequently, we construct the network of the single cells according to the consensus matrix C where the value reflects a high-level integrated similarity between the cells since it is derived from multiple high-performing clustering outcomes. The graph can be defined as $G = \{V, E\}$, where V and E is the set of nodes and edges, respectively. Here, G is an undirected weighted graph and w_{ij} is the weight of edge between sample i and j. To build a highly reliable network, we impose a constraint that $w_{ij} = c_{ij}$ only if $c_{ij} >= 3$, otherwise $w_{ij} = 0$.

3) RANDOM WALK DISTANCE ON GRAPH

In general, if two nodes are directly connected or have many common neighbors, the probability that they belong to the same cluster is high [42], [56]. While from the perspective of the random walk on graph, two nodes are more similar if they have similar walking paths on the network. Therefore, we measure the relationships between cells based on their walking paths using random walk algorithm [65].

A typical random walk model on a regular graph is that, at each step, the walker at current location jumps to another site according to some probability distribution. In a simple random walk approach, the walker can only jump to adjacent positions of the graph to form a walking path [66]. From the graph constructed above, we compute a transition matrix M, in which the element $m_{ij} = \frac{w_{ij}}{Deg(i)}$, $Deg(i) = \sum_{j=1}^{n_i} w_{ij}$ and n_i is the number of neighborhoods connecting with node i.

If a walker goes from the i^{th} cell in scRNA-seq data graph, the initial probability $P_{i.}^0$ will be an $n \times 1$ vector where only the i^{th} value is 1 and the others are 0. For each step of the walker walking on the graph, the probability vector is updated following $P^{t+1} = M^T P^t$, where P_{ij}^t is the probability of which the walker goes from cell i to cell j in t steps. Previous studies have shown that if the length of steps t tends towards infinity in the random walk process, the probability of being on a vertex j only depends on the degree of vertex j and is irrelevant to the starting vertex i. Therefore, it is important to choose an appropriate length of steps. If t is too small, the data is insufficient to depict the topology of graph. On the other hand, if t is too large, the system will result in a stationary distribution. In our experiments, we set $t = 4$ that is empirically advised in the previous study [65].

The random walk distance between cell i and cell j can be calculated as

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^t - P_{jk}^t)^2}{Deg(k)}}, \tag{5}$$

where n is the number of cells and P_{ik}^t is the probability of a walker going from cell i to cell k in t steps.

4) HIERARCHICAL TREE-STRUCTURE OF SINGLE CELLS

Next, we construct a tree-structure of cells. As shown in equation (6), the distance between a single cell k and a cluster C is defined as the average random walk distance from each node of C to node k, and $|C|$ is the number of nodes in cluster C. Equation (7) calculates the distance between clusters C_i and C_j, which means the difference of the random walk distance about nodes in these two clusters.

$$d_{kC} = \frac{1}{|C|} \sum_{i \epsilon C} P_{ik}^t \tag{6}$$

$$d_{C_i C_j} = \sqrt{\sum_{k=1}^{n} \frac{(d_{kC_i} - d_{kC_j})^2}{Deg(k)}} \tag{7}$$

Initially, we divide the cells into separate groups where each group only has one cell and each cell is put in one group. To form the tree-structure, the criteria is needed for selecting two groups to be combined each time. Here, we adopt the strategies from Ward's method [67]. We define the growth of the average intra-cluster distance before and after the union of each two adjacency cluster C_i and C_j as equation (8), where $C_u = C_i \cup C_j$ and there is at least one edge between these two groups.

$$\Delta\sigma(C_i, C_j) = \frac{1}{n} \left(\sum_{k \in C_u} d_{kC_u}^2 - \sum_{k \in C_i} d_{kC_i}^2 - \sum_{k \in C_j} d_{kC_j}^2 \right) \tag{8}$$

Then two clusters with the smallest value of $\Delta\sigma$ is selected to be merged.

5) ESTIMATION THE NUMBER OF CLUSTERS

After the tree is constructed, we need to determine the number of divisions, a.k.a., the number of clusters [15]. Here we define the average intra-cluster distance of K groups as

$$\sigma_K = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} d_{iC_k}^2, \tag{9}$$

where C_k is the k^{th} cluster. Then the growth rate η_K can be calculated as

$$\eta_K = \frac{\sigma_{K+1} - \sigma_K}{\sigma_K - \sigma_{K-1}}, \tag{10}$$

and the optimal number of clusters K is that satisfies the maximum value of η_K.

D. PSEUDO-CODE OF GRACE

The pseudo-code of GRACE is shown in Algorithm (1). Line 1-3 is the process to construct a weighted graph of single cells. Line 4-11 is the process to build the tree-structure of clusters. Line 12-14 finish the structure of cutting and return the cluster labels.

V. RESULTS

A. DIFFERENCES AMONG THE CLUSTERING METHODS

Generally, clustering methods exhibit very different performance across different datasets. The reason lays in the

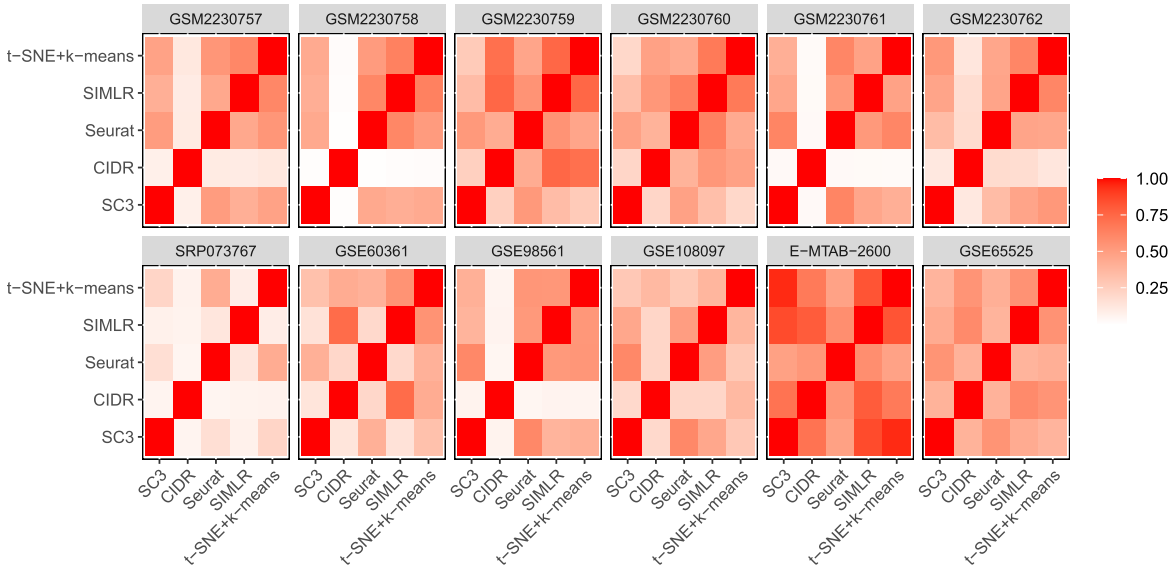

FIGURE 2. Peer-to-peer comparison on the similarity among the five scRNA-seq data clustering approaches on twelve scRNA-seq datasets. The similarity is derived from ARI of the predicting results.

Algorithm 1 Framework of GRACE

Input: The expression profile of scRNA-seq data, D; The number of samples(cells), N;

Output: The clustering labels of samples, L;

1: Cluster the scRNA-seq data D using the five clustering methods;
2: Calculate the consensus matrix from the clustering outcomes;
3: Construct the graph G of cells based on the consensus matrix;
4: Calculate the random walk distance of cells with equation (5);
5: Initialize the partition $P^0 = \{v_1, v_2, \ldots, v_N\}$;
6: **while** $k < N$ **do**
7: Calculate $d_{C_iC_j}, \forall\ C_i, C_j \in P^k$ with equation (6) and (7);
8: Select the two closest clusters to merge $C_3 = C_1 \cup C_2$ according to equation (8);
9: Update partition $P^{k+1} = \{\{P^k \setminus \{C_1, C_2\}\} \cup C_3\}$;
10: $k := k + 1$;
11: **end while**
12: Evaluate the optimal number of clusters K with equation (9) and (10);
13: Get the cluster label, setting L to P^{N-K};
14: **return** L.

fact that the method is usually optimized for some specific datasets [68]. To demonstrate this, We compare the similarity between the five state-of-the-art scRNA-seq data clustering approaches by computing the ARI of their predicting clustering results on twelve scRNA-seq datasets. As shown in Figure 2, the color from white to red indicates the ARI values ranges from 0 to 1. One can observe that the predicting outcomes of the five approaches are inconsistent (i.e. a lower ARI) on most datasets. The reason is that different methods capture different aspects of information from the complex and high-dimensional scRNA-Seq data.

B. IMPROVING CLUSTERING ACCURACY FROM INDIVIDUAL METHODS

In this section, we compare the ARI between GRACE and the five individual methods on twelve published datasets. The datasets are collected from different tissues by different sequencing technologies where the numbers of cells and numbers of cell types are totally different.

Table 2 shows the ARI of our method comparing with the five individual clustering methods on the twelve datasets. Among the twelve scRNA-seq datasets, GRACE produces the best results in ten datasets (GSM2230757, GSM2230758, GSM2230759, GSM2230760, GSM2230761, SRP073767, GSE60361, GSE98561, GSE108097 and E-MTAB-2600), and the second best in the other two datasets (GSM2230762 and GSE65525). We also calculate the average ARI of each method on all the datasets, which are listed in the last line of the table. The results show that GRACE outperforms all other methods.

Furthermore, we make a statistical rank of these methods across twelve datasets according to ARI. A higher ARI value corresponds a larger rank, and a larger rank represents a better clustering performance. As shown in Figure 3, our approach GRACE achieves the highest rank and performs significantly better than other five individual methods. One can observe that the individual clustering methods have an

TABLE 2. Similarity between predicted outcomes and gold-standard cluster labels is measured through ARI.

Dataset	CIDR	SC3	Seurat	SIMLR	t-SNE+k-means	GRACE
E-MTAB-2600	0.41	0.44	0.40	0.52	0.44	**0.53**
GSE65525	0.63	0.48	0.50	0.64	**0.86**	0.84
GSE108097	0.22	**0.57**	**0.57**	0.49	0.37	**0.57**
GSE98561	0.09	0.16	0.40	0.18	0.43	**0.48**
SRP073767	0.07	0.48	0.54	0.65	0.57	**0.76**
GSE60361	0.49	0.32	0.43	0.51	0.82	**0.86**
GSM2230757	0.08	0.34	0.49	0.57	0.52	**0.60**
GSM2230758	0.03	0.39	0.58	0.83	0.64	**0.86**
GSM2230759	0.78	0.34	0.57	0.86	0.83	**0.97**
GSM2230760	0.56	0.32	0.62	0.89	0.74	**0.93**
GSM2230761	0.03	0.42	0.66	0.45	0.59	**0.75**
GSM2230762	0.23	0.26	**0.54**	0.43	0.37	0.50
Average ARI	0.3	0.38	0.52	0.59	0.60	**0.72**

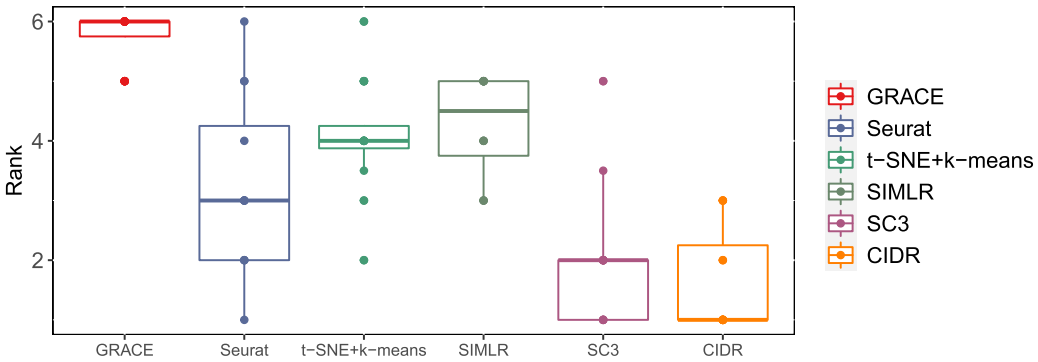

FIGURE 3. Performance rank of GRACE and the five individual clustering methods. Methods are ranked according to ARI on the twelve datasets. A higher rank indicates a better performance (6 is the best and 1 is the worst). The horizontal line inside each box represents the median.

unstable performance on different datasets, while our method is highly robust.

We also adopt other three metrics in clustering performance comparison. Figure 4 shows the Completeness scores, Homogeneity scores and V-measure of different approaches on twelve scRNA-seq datasets. Overall, GRACE performs the best in the Completeness score and V-measure. For the Homogeneity score, although SC3 is generally better than GRACE, the performance of SC3 is less stable. Compared with ''t-SNE+k-means'', other three methods (Seurat, SIMLR, and SC3) tend to obtain higher Homogeneity scores but lower Completeness scores. Among these methods, CIDR performs the worst for all the three metrics and is unstable dealing with different datasets.

In conclusion, individual clustering approaches exhibit the unstable performance in different datasets. GRACE improves the clustering accuracy and is more robust.

C. ADVANTAGES OF HIGH QUALITY GRAPH NETWORK

The reason why the ensemble cluster algorithm GRACE can improve the clustering accuracy is closely related to the way of the network construction. In order to verify that the network constructed by GRACE has higher reliability and is more conducive to cluster structure mining, we compared GRACE with other five conventional network construction approaches using similarly metrics. To be equitable, we only replace the network construction methods while keep other steps to cluster the scRNA-seq data in GRACE.

Here we list the five conventional methods to built network in our experiment.

1) Euclidean Distance (ED);
2) Manhattan Distance (MD);
3) Pearson Correlation Coefficient (PCC);
4) Spearman Correlation Coefficient (SCC);
5) Shared Nearest neighbors (SNN).

ED and MD are common distance metrics where a smaller distance between two cells indicates a greater similarity between them. The corresponding similarity can be set as $1-normalizeddistance$, where the normalized distance normalizes ED and MD into [0,1]. Literally, PCC and SCC demonstrate the similarity between cells. PCC is defined as the co-variance of two samples divided by the standard deviation of the two. The formula for calculating the SCC is similar to the calculation of the PCC, but the rank is substituted for the respective values.

Different from above four primary similarity indexes, SNN is called the ''second-order similarity'', which measures the similarity of the samples according to the number neighbors. Previous study show that SNN is more stable and robust for high-dimensional sparse data than the traditional distance metrics which generally results in a small value between samples [69]. In our experiments, we use the SNN similarity

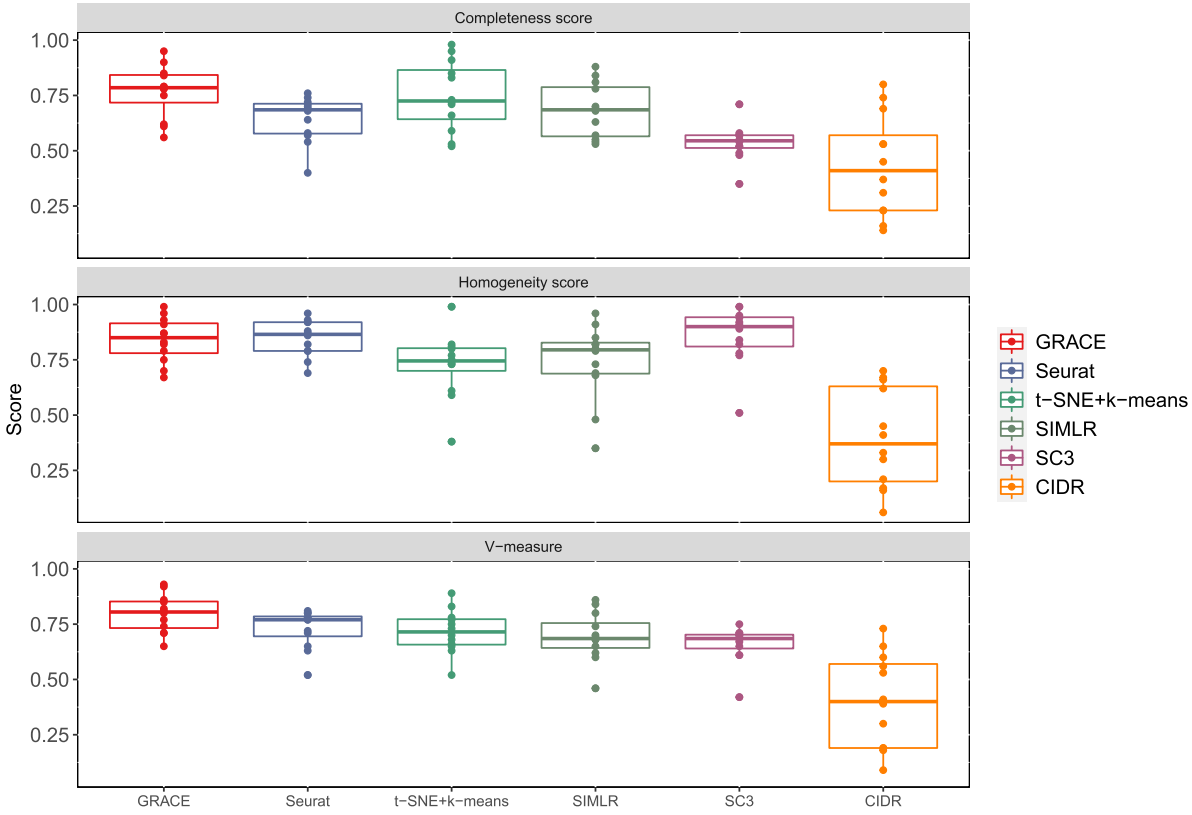

FIGURE 4. The Completeness score, Homogeneity score and V-measure of different approaches on twelve scRNA-seq datasets.

mentioned in [42] and corresponding similarity between cell i and j is defined as equation (11),

$$s_{ij}^{SNN} = k - \frac{min\left(Rank_i(SN) + Rank_j(SN)\right)}{2}, \qquad (11)$$

where k is the number of nearest neighbors of node, SN is the intersection of the k-nearest neighbors (kNN) of sample i and sample j, $Rank_i(SNN)$ is the rank of each node of SN in kNN for sample i, and $min()$ is a function computing the minimum value of a vector. For example, there are 3 neighbors are shared by sample i and j, and the rank of them in kNN is (1,2,4) and (1,3,4) respectively. Then w_{ij}^{SNN} gets its max value $k-1$ because the top ranking of SN is 1 for both sample i and j.

Figure 5 shows the clustering performance of graph-based methods with different network construction methods. One can observe that GRACE, which uses ensemble cluster results in building network, reaches the highest clustering accuracy on 10 scRNA-seq datasets. It performs much better than all the other methods on GSE60361, GSE98561, and GSE65525. For instance, our GRACE algorithm has a clustering accuracy of 0.86 on the data set GSE60361, while the highest ARI of several other methods is 0.35. Similarly, ARI of the GRACE algorithm reaches 0.84 on the dataset GSE65525, while ARI in several other methods range from 0.18 to 0.50. On average, GRACE obtains the highest ARI (0.72), followed by PCC(0.50) and ED (0.13) performs the worst.

In conclusion, the ensemble clustering based graph construction method is confirmed to have high quality outputs and provide more reliable graph information for us to detect communities.

D. RUNNING TIME OF GRACE

In this section, we test the the efficiency of GRACE. Since GRACE is an ensemble algorithm, its running time is always greater than that of each single algorithm it integrates. However, the overhead of GRACE is small. We list the running time of the five single clustering algorithms and the integration step in GRACE (overhead). To be fair, we use spaltter [70] to generate five simulated single cell datasets, where the number of features of each simulated data is set as 5000, while the sample size increased from 1000 to 5000 across the step size 1000. In simulating data, the parameters used by spaltter are estimated from one real dataset GSE60361.

The experiments are conducted on a desktop computer with 3.2GHz Intel Core i7 CPU, 16 GB 2400MHz DDR4 RAM, and Windows 10 operating system. Table 3 presents the time consumption of GRACE's ensemble step and other five individual methods. With the increase of sample size, the time consumption of SIMLR increases rapidly, followed by SC3, while the other 3 methods present shorter running time. Comparing to the clustering run-time, the overhead of

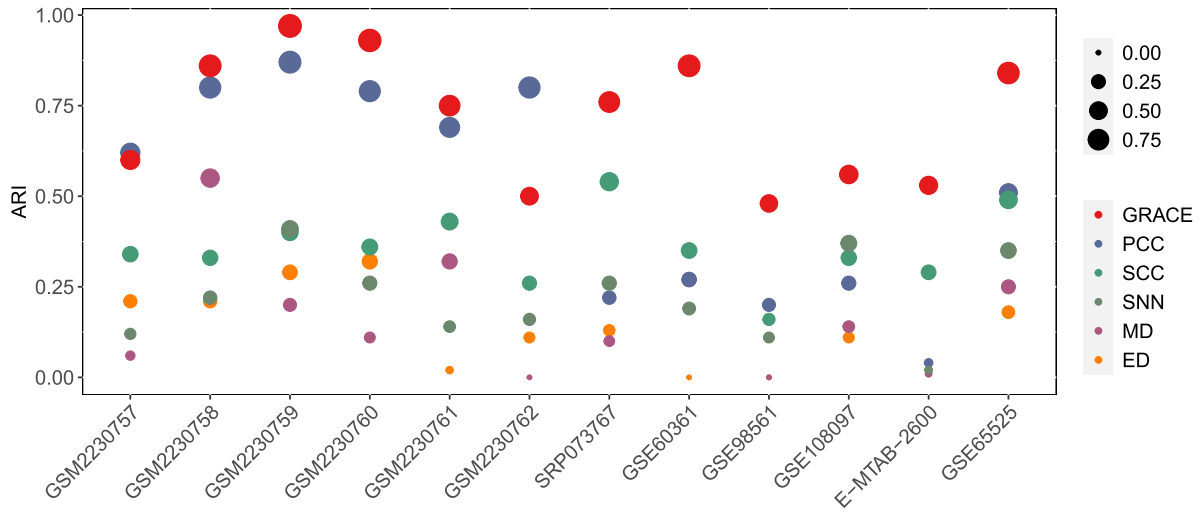

FIGURE 5. Performance comparison of GRACE with five network construction methods on twelve datasets.

TABLE 3. Time consumption (/seconds) of different methods with sample size increasing from 1000 to 5000.

#Cells	CIDR	SC3	Seurat	SIMLR	t-SNE+k-means	GRACE overhead
1000	4.17	65.89	6.02	46.7	16.69	0.83
2000	19.26	372.39	9.71	178.53	71.42	3.03
3000	59.61	346.87	14.95	494.4	156.21	7.69
4000	139.42	730.47	21.52	1191.58	291.81	32.6
5000	274.65	1294.04	34.42	2382.1	461.85	65.79

GRACE is small and acceptable. Even for the largest data with 5000 samples, the overhead is around 5%.

VI. CONCLUSION AND DISCUSSION

In the past decades, there has been increasing interests in scRNA-seq data analysis, where the growing clustering methods have helped to solve many research problems. However, experimental results show that the existing methods are not robust across multiple datasets and even perform poorly on some complex datasets [68]. Since the scRNA-seq data from different platforms or laboratories are always unlabeled and have limited additional information, it's difficult to determine which clustering approach is more appropriate. To address this problem, we propose a novel clustering framework GRACE, a graph-based cluster ensemble approach. By integrating the outcomes of five high-performing clustering methods for network construction, GRACE is able to build a more reliable graph network than using other conventional similarity metrics. What's more, the comparative analysis on twelve datasets shows that GRACE is highly robust and exhibits a competitive performance.

In our future work, we would like to apply GRACE to some specific disease related studies, such as disease-related cell types and biological pathway identification. Besides, we will also apply the graph constructed in GRACE to scRNA-seq data visualization. Moreover, the idea of ensemble clustering is not limited to clustering scRNA-seq data and may be useful to a wide range of clustering applications.

ACKNOWLEDGMENT

The authors would like to thank all DMB Lab members for providing suggestions and discussions on this work. They also want to show their appreciation to the studies which share the benchmark datasets and develop modern clustering approaches for scRNA-seq data.

REFERENCES

[1] M. Pavlovic, "Cell physiology: Liaison between structure and function," *Bioengineering*. 2015, pp. 23–35.
[2] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, 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 Biotechnol.*, vol. 33, no. 2, pp. 155–160, Feb. 2015.
[3] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suva, A. Regev, and B. E. Bernstein, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, Jun. 2014.
[4] P. Dalerba, T. Kalisky, D. Sahoo, P. S. Rajendran, M. E. Rothenberg, A. A. Leyrat, S. Sim, J. Okamoto, D. M. Johnston, and D. Qian, "Single-cell dissection of transcriptional heterogeneity in human colon tumors," *Nature Biotechnol.*, vol. 29, no. 12, pp. 1120–1127, Dec. 2011.
[5] T. Kalisky, P. Blainey, and S. R. Quake, "Genomic analysis at the single-cell level," *Annu. Rev. Genet.*, vol. 45, no. 1, pp. 431–445, Dec. 2011.
[6] E. Shapiro, T. Biezuner, and S. Linnarsson, "Single-cell sequencing-based technologies will revolutionize whole-organism science," *Nature Rev. Genet.*, vol. 14, no. 9, pp. 618–630, Sep. 2013.

[7] T. Kalisky and S. R. Quake, "Single-cell genomics," *Nature Methods*, vol. 8, no. 4, pp. 311–314, 2011.
[8] G. X. Y. Zheng *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Commun.*, vol. 8, Jan. 2017, Art. no. 14049.
[9] K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, S. A. McCarroll, C. L. Cepko, A. Regev, and J. R. Sanes, "Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics," *Cell*, vol. 166, no. 5, pp. 1308–1323, 2016.
[10] 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 Biotechnol.*, vol. 32, no. 4, pp. 381–386, Apr. 2014.
[11] J. D. Welch, A. J. Hartemink, and J. F. Prins, "SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data," *Genome Biol.*, vol. 17, no. 1, p. 106, Dec. 2016.
[12] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature Biotechnol.*, vol. 33, no. 5, pp. 495–502, May 2015.
[13] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
[14] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.
[15] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967.
[16] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, vol. 27, no. 2, pp. 94–105, 1998.
[17] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.
[18] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, vol. 96, no. 34, pp. 226–231.
[19] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[20] X. Peng, Z. Yu, Z. Yi, and H. Tang, "Constructing the L2-graph for robust subspace learning and subspace clustering," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1053–1066, Apr. 2017.
[21] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured AutoEncoders for subspace clustering," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, Oct. 2018.
[22] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, "Deep clustering with sample-assignment invariance prior," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, 2020.
[23] D. Huang, J. Lai, and C.-D. Wang, "Ensemble clustering using factor graph," *Pattern Recognit.*, vol. 50, pp. 131–142, Feb. 2016.
[24] D. Huang, J.-H. Lai, and C.-D. Wang, "Robust ensemble clustering using probability trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1312–1326, May 2016.
[25] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, May 2018.
[26] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Nov. 6, 2018, doi: 10.1109/TSMC.2018.2876202.
[27] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1212–1226, Jun. 2020.
[28] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nature Methods*, vol. 11, no. 2, pp. 163–166, Feb. 2014.
[29] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "A general and flexible method for signal extraction from single-cell RNA-seq data," *Nature Commun.*, vol. 9, no. 1, p. 284, Dec. 2018.
[30] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, pp. 740–742, Jul. 2014.
[31] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hemberg, "SC3: Consensus clustering of single-cell RNA-seq data," *Nature Methods*, vol. 14, no. 5, p. 483, 2017.
[32] P. Lin, M. Troup, and J. W. K. Ho, "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data," *Genome Biol.*, vol. 18, no. 1, p. 59, Dec. 2017.
[33] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, "SINCERA: A pipeline for single-cell RNA-seq profiling analysis," *PLOS Comput. Biol.*, vol. 11, no. 11, Nov. 2015, Art. no. e1004575.
[34] D. A. duVerle, S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda, "CellTree: An R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data," *BMC Bioinf.*, vol. 17, no. 1, p. 363, Dec. 2016.
[35] Z. Sun, T. Wang, K. Deng, X.-F. Wang, R. Lafyatis, Y. Ding, M. Hu, and W. Chen, "DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data," *Bioinformatics*, vol. 34, no. 1, pp. 139–146, Jan. 2018.
[36] A. Zeisel, A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, Mar. 2015.
[37] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden, "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature*, vol. 525, no. 7568, pp. 251–255, Sep. 2015.
[38] L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan, "GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index," *Genome Biol.*, vol. 17, no. 1, p. 144, Dec. 2016.
[39] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, Dec. 2018.
[40] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Trans. Comput.*, vol. C-20, no. 1, pp. 68–86, Jan. 1971.
[41] E. Hartuv and R. Shamir, "A clustering algorithm based on graph connectivity," *Inf. Process. Lett.*, vol. 76, nos. 4–6, pp. 175–181, Dec. 2000.
[42] C. Xu and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, Jun. 2015.
[43] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning," *Nature Methods*, vol. 14, no. 4, pp. 414–416, Apr. 2017.
[44] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, U. Sarkans, and A. Brazma, "ArrayExpress update–from bulk to single-cell expression data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D711–D715, Jan. 2019.
[45] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar, "NCBI GEO: Mining millions of expression profiles—Database and tools," *Nucleic Acids Res.*, vol. 33, pp. 562–566, Jan. 2005.
[46] R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," *Nucleic Acids Res.*, vol. 39, pp. 19–21, Nov. 2011.
[47] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg, "Full-length RNA-seq from single cells using smart-seq2," *Nature Protocols*, vol. 9, no. 1, pp. 171–181, Jan. 2014.
[48] X. Han, R. Wang, Y. Zhou, L. Fei, and G. Guo, "Mapping the mouse cell atlas by microwell-seq," *Cell*, vol. 173, no. 5, p. 1307, May 2018.
[49] J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, and J. Shendure, "Comprehensive single-cell transcriptional profiling of a multicellular organism," *Science*, vol. 357, no. 6352, pp. 661–667, Aug. 2017.
[50] X. Liu, Q. Xiang, F. Xu, J. Huang, N. Yu, Q. Zhang, X. Long, and Z. Zhou, "Single-cell RNA-seq of cultured human adipose-derived mesenchymal stem cells," *Sci. Data*, vol. 6, no. 1, Mar. 2019, Art. no. 190031.
[51] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015.

[52] J. K. Kim, A. A. Kolodziejczyk, T. Ilicic, S. A. Teichmann, and J. C. Marioni, "Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression," *Nature Commun.*, vol. 6, no. 1, p. 8687, Dec. 2015.
[53] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai, "A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure," *Cell Syst.*, vol. 3, no. 4, pp. 346–360, 2016.
[54] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.
[55] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Empirical Methods Natural Lang. Process.*, 2007, pp. 410–420.
[56] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature Biotechnol.*, vol. 36, no. 5, pp. 411–420, May 2018.
[57] E. Pierson and C. Yau, "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome Biol.*, vol. 16, no. 1, p. 241, Dec. 2015.
[58] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.-Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[59] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet.*, vol. 2, no. 12, p. e190, 2006.
[60] C. A. Tracy and H. Widom, "Level-spacing distributions and the airy kernel," *Commun. Math. Phys.*, vol. 159, nos. 1–2, pp. 151–174, Jan. 1994.
[61] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, and A. Sidow, "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival," *Nature Commun.*, vol. 9, no. 1, p. 4453, Dec. 2018.
[62] T. Stuart, A. Butler, P. J. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, p. 1888, 2019.
[63] L. Waltman and N. J. van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *Eur. Phys. J. B*, vol. 86, no. 11, p. 471, Nov. 2013.
[64] X.-F. Wang and Y. Xu, "Fast clustering using adaptive density peak detection," *Stat. Methods Med. Res.*, vol. 26, no. 6, pp. 2800–2811, Dec. 2017.
[65] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006.
[66] P. Revesz, *Random Walk Random Non-Random Environments*. Singapore: World Scientific, 2005.
[67] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?" *J. Classification*, vol. 31, no. 3, pp. 274–295, Oct. 2014.
[68] S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo, "Comparison of clustering tools in r for medium-sized 10x genomics single-cell RNA-sequencing data," *FResearch*, vol. 7, p. 1297, Aug. 2018.
[69] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in *Proc. Int. Conf. Sci. Stat. Database Manage.*, 2010, pp. 482–500.
[70] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: Simulation of single-cell RNA sequencing data," *Genome Biol.*, vol. 18, no. 1, p. 174, Dec. 2017.

JIHONG GUAN received the bachelor's degree from Huazhong Normal University, in 1991, the master's degree from the Wuhan Technical University of Surveying and Mapping (merged into Wuhan University in 2000), in 1998, and the Ph.D. degree from Wuhan University, in 2002. She is currently a Professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. Her research interests include databases, data mining, distributed computing, bioinformatics, and geographic information systems. She has extensively published more than 300 articles in domestic and international journals (including *Nature Communications*, IEEE TKDE/ TITS/TSC/TGRS/TCBB, NAR, and *Bioinformatics*) and conferences (including AAAI, ICDE, VLDB, SIGIR, RECOMB, and DASFAA).

RUI-YI LI is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, Shanghai, China. Her research interest is data analysis on scRNA-seq data.

JIASHENG WANG received the bachelor's degree in computer science and technology from Tongji University, Shanghai, China, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology. His research interests include data mining, sub-graph search on social networks, game theory, and human behavior.