Review of single-cell RNA-seq data clustering for cell-type identification and characterization

SHIXIONG ZHANG,[1,2] XIANGTAO LI,[3] JIECONG LIN,[2] QIUZHEN LIN,[4] and KA-CHUN WONG[2]

[1]School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[2]Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
[3]School of Artificial Intelligence, Jilin University, Jilin 130012, China
[4]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

ABSTRACT

In recent years, the advances in single-cell RNA-seq techniques have enabled us to perform large-scale transcriptomic profiling at single-cell resolution in a high-throughput manner. Unsupervised learning such as data clustering has become the central component to identify and characterize novel cell types and gene expression patterns. In this study, we review the existing single-cell RNA-seq data clustering methods with critical insights into the related advantages and limitations. In addition, we also review the upstream single-cell RNA-seq data processing techniques such as quality control, normalization, and dimension reduction. We conduct performance comparison experiments to evaluate several popular single-cell RNA-seq clustering approaches on simulated and multiple single-cell transcriptomic data sets.

Keywords: single-cell RNA-seq; clustering; cell types

BACKGROUND

With the unabated progress in high-throughput sequencing technologies, single-cell RNA-seq has become a powerful approach to simultaneously measure cell-to-cell expression variability of thousands or even hundreds of thousands of genes (Shapiro et al. 2013; Grun et al. 2015) at single-cell resolution. Such high-throughput transcriptomic profiling can capture the gene transcriptional activities to reveal cell identities and functions (Patel et al. 2014; Kiselev et al. 2019) and discover cell types (Shalek et al. 2014; Xu and Su 2015; Zeisel et al. 2015) or even rare cell types (Grun et al. 2015; Jiang et al. 2016a; Van Unen et al. 2017). Hence, one of the most common goals of those single-cell studies is to identify cell subpopulations under different contexts (Yang et al. 2017). The gene expression patterns of those subpopulations help us distinguish various cell types and functions, identifying different cell types.

Diverse computational approaches based on data clustering have emerged to interpret and understand single-cell RNA-seq data (Jiang et al. 2016a; Lin et al. 2017b; Yang et al. 2017; Wolf et al. 2018; Zheng et al. 2019). The advances in single-cell clustering have also initiated the development of multiple atlas projects such as the Mouse Cell Atlas (Han et al. 2018), Aging Drosophila Brain Atlas (Davie et al. 2018), and Human Cell Atlas (Rozenblatt-Rosen et al. 2017). However, several technical challenges are still involved in single-cell RNA-seq clustering. Low-quality cells (have very few genes or an aberrantly high gene count), amplification biases, and other confounding factors can affect the downstream clustering performance. In addition, given the whole transcriptome range of RNA-seq (the large number of genes assayed in scRNA-seq, that is, high dimensionality), the curse of dimensionality (distances between data points [that is, cells] become similar and not reliable for identifying cell groups) should be expected (Andrews and Hemberg 2018). Thus, data preprocessing steps including quality control, normalization, and dimensional reduction have become necessary before downstream interpretation. In addition, tissue heterogeneity can also affect the ability of single-cell RNA-seq clustering performance to detect rare cell types (Grun et al. 2015; Jiang et al. 2016a; Van Unen et al. 2017).

In this study, we review the recently developed computational clustering approaches for understanding and interpreting single-cell RNA-seq data. We also review the upstream single-cell RNA-seq data preprocessing steps such as quality control, row/column normalization, and

Corresponding author: sxzhang7-c@my.cityu.edu.hk
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.078965.121. Freely available online through the *RNA* Open Access option.

dimension reduction before clustering is performed. Four roughly classified categories of single-cell RNA-seq clustering methods and its applications are discussed in terms of strengths and limitations, including *k*-means clustering, hierarchical clustering, community-detection-based clustering, and density-based clustering. Figure 1 depicts the workflow of single-cell RNA-seq data clustering by data processing (quality control, normalization, and dimension reduction) and clustering methods. The strengths and limitations are discussed in the following sections to guide selection of different tools. In addition, we conduct several experiments on simulated and real single-cell RNA-seq data sets to evaluate and compare those clustering methods.

DATA PREPROCESSING

Given the technical variations and noises, data preprocessing is essential for unsupervised cluster analysis on single-cell RNA-seq data. Quality control is performed to remove the low-quality transcriptomic profile due to capture inefficiency; the single-cell RNA-seq reads should be normalized to remove any amplification bias, sample variation, and other technical confounding factors; and dimensional reduction is conducted to project the high-dimensional single-cell RNA-seq data into low-dimensional space. Those upstream steps could have substantial impacts on downstream tasks. Therefore, a myriad of tools has been developed to address the above issues.

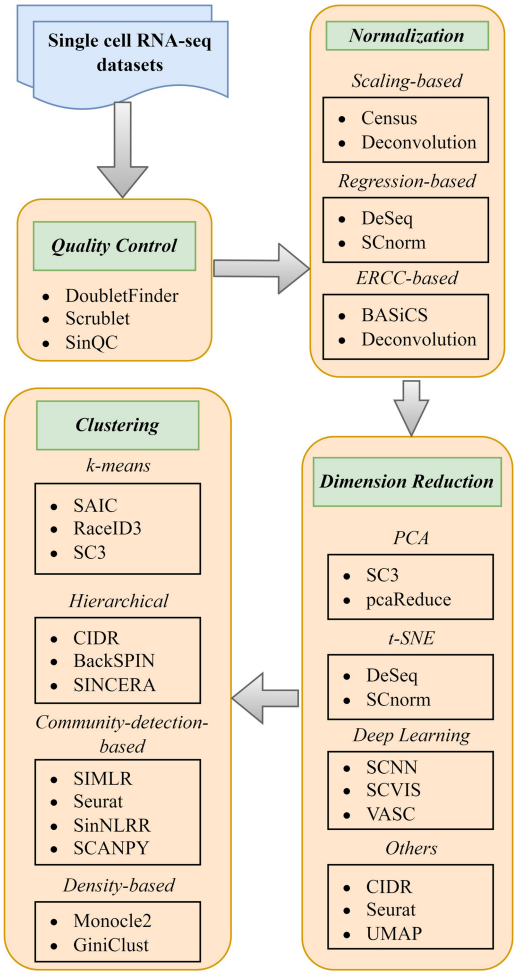

FIGURE 1. Workflow of single-cell RNA-seq data preprocessing and clustering.

Quality control

Low-quality cells or empty droplets will often have very few genes, and cell doublets may exhibit an aberrantly high gene count. Usually, the community filters cells that have gene counts over 2500 or less than 200 (Satija et al. 2015; Jiang et al. 2016a). In addition, low-quality cells often exhibit extensive mitochondrial contamination, and then filter cells that have >5% mitochondrial counts will be filtered (Jiang et al. 2016b).

Scrublet (Wolock et al. 2019) provides guidance on parameter threshold selection while there is no rule of thumb included in DoubletFinder. DoubletFinder outperformed Scrublet in terms of detection accuracy that impacts on downstream analyses such as clustering, while it has no advantage in computational efficiency and stability.

SinQC integrates both gene expression patterns and sample sequencing library qualities to detect low-quality cells (Jiang et al. 2016b). Gene expression patterns are used to classify cells as either gene expression outliers assumed as poor quality or cells of the main population considered as good quality (Jiang et al. 2016b).

Normalization

Technical artifacts or experimental noise (e.g., batch effect, insufficient counts, and zero inflation) of high-throughput transcriptomic sequencing may result in differences in expression measurements between samples (e.g., cells) (Cole et al. 2019). Several studies have revealed that those obvious differences can have a large impact on clustering (Finak et al. 2015; Butler et al. 2018; Haghverdi et al. 2018). Therefore, normalization is essential for adjusting the differences in expression levels across different samples, replicates, or even batches. The state-of-the-art normalization methods have been developed for addressing those issues.

We review three kinds of normalization methods as follows: (i) Scaling methods. Lun et al. (2016) proposed a strategy to normalize single-cell RNA-seq data with zero counts. Census (Qiu et al. 2017b) converts conventional per-cell measures of relative expression values to transcript counts without the need for any spike-in standard or unique molecular identifiers, eliminating much of the apparent technical variability in single-cell experiments; (ii) regression-based methods. DESeq proposed by Anders and Huber (2010) adopts local regression to link the

variance and mean of negative binomial distribution over the observed counts, resulting in balanced differentially expressed genes. SCnorm (Bacher et al. 2017) uses quantile regression to estimate the dependence of transcript expression on sequencing depth and scale factors to provide normalized expression estimates; (iii) methods based on spike-in External RNA Controls Consortium (ERCC). Ding et al. (2015) presented a normalization tool to remove technical noise and compute the true gene expression levels based on spike-in ERCC. BASiCS (Vallejos et al. 2015) can identify and remove the high and low levels of technical noise (counts). In addition to the above methods, the very simple and commonly used method is to transform read counts using a logarithm with a pseudocount such as one (Xu and Su 2015; Lin et al. 2017b; Butler et al. 2018).

However, those normalization methods also suffer from limitations caused by diverse assumptions and experimental protocols. The scaling methods cannot account for individual batch effects; the regression-based methods are sensitive to batch effects; and ERCC-based methods are not suitable for endogenous and spiked-in transcripts (Risso et al. 2014; Vallejos et al. 2017; Cole et al. 2019). Sctransform proposed to utilize the Pearson residuals from regularized negative binomial regression as a covariate in a generalized linear model to remove the influence of technical characteristics while preserving biological heterogeneity (Hafemeister and Satija 2019). Notably, it does not include the pseudocount addition or log-transformation and has been integrated in Seurat.

Dimension reduction

Recent advances in single-cell RNA-seq have contributed to measuring large-scale expression data sets with hundreds of thousands of transcripts while it also brings both opportunities and challenges in data analysis. Such high-dimensional gene expression data is unprecedentedly rich and should be well-explored. However, the past clustering methods may be unable to process and interpret such large-scale data. Hence, it is necessary to project the high-dimensional data to a lower-dimensional space using dimension reduction that can improve and refine the clustering results. In this section, we review several commonly used dimension reduction methods including principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) algorithm, Uniform Manifold Approximation and Projection (UMAP), deep learning models, and others. In a recent study, Sun et al. (2019) compared 18 different dimensionality reduction methods based on PCA, t-SNE, and UMAP, and evaluated the computational scalability.

PCA

PCA is a typical linear projection method that projects a set of possibly correlated variables into a set of linearly orthogonal variables (principal components). Due to its conceptual simplicity and efficiency, PCA has been widely used in single-cell RNA-seq processing (Shalek et al. 2014; Buettner et al. 2015; Usoskin et al. 2015; Jiang et al. 2016a; Zurauskiene and Yau 2016; Kiselev et al. 2017). Notably, SC3 (Kiselev et al. 2017) applied PCA to transform the distance matrices as the input of consensus clustering; Shalek et al. (2014) used PCA for single-cell RNA-seq data spanning several experimental conditions. In addition, some extended and improved PCA-based methods have been developed including pcaReduce (Zurauskiene and Yau 2016), which applied PCA iteratively to provide low-dimensional principal component representations; Usoskin et al. (2015) proposed an unbiased iterative PCA-based process to identify distinct large-scale expression data patterns. However, PCA cannot capture the nonlinear relationships between cells because of the high levels of dropout and noise (Kiselev et al. 2019).

t-SNE

t-SNE is the most commonly used nonlinear dimension reduction method which can uncover the relationships between cells. t-SNE converts data point similarity into probability and minimizes Kullback–Leibler divergence by gradient descent until convergence. In single-cell RNA-seq data analysis, t-SNE has become a cornerstone of dimension reduction and visualization for high-dimensional single-cell RNA-seq data (Zeisel et al. 2015; Ntranos et al. 2016; Prabhakaran et al. 2016; Li et al. 2017; Lin et al. 2017b; Butler et al. 2018; Haghverdi et al. 2018; Zhang et al. 2018; Linderman et al. 2019). Especially, Linderman et al. (2019) developed a fast interpolation-based t-SNE that dramatically accelerates the processing and visualization of rare cell populations for large data sets. Nonetheless, t-SNE also involves several limitations. First, t-SNE tends to get stuck in local optima since its loss function is nonconvex. Second, t-SNE requires hyperparameter tuning that may lead to unwanted results. Third, t-SNE is nondeterministic, namely, different runs with the same hyperparameters may produce different results.

UMAP

UMAP (McInnes et al. 2018) is a widely used technique for dimension reduction. UMAP provides increased speed and better preservation of data global structure for high dimensional data sets. It has been verified that it outperforms t-SNE (Becht et al. 2019), since t-SNE suffers from limitations such as loss of large-scale information (the inter-cluster relationships), slow computation time and inability to meaningfully represent very large data sets. UMAP preserves as much of the local and more of the global data structure, with a shorter running time.

Deep learning models

In recent years, deep learning models (neural networks and variational auto-encoders) have shown superior performance in interpenetrating complex high-dimensional data. SCNN (Lin et al. 2017a) tested various neural network architectures and incorporated prior biological knowledge to obtain the reduced dimension representation of single-cell expression data. SCVIS (Ding et al. 2018) and VASC (Wang and Gu 2018) are both based on variational auto-encoders, which can capture nonlinear relationships between cells and visualize the low-dimensional embedding in single-cell gene expression data. Up to now, those methods demonstrated superior ability of interpretation and compatibility on high-dimensional single-cell RNA-seq data.

Other methods

In addition, there are also other dimensional reduction methods such as CIDR (Lin et al. 2017b) applied principal coordinate analysis, which preserves the distance information in low-dimension space from its high-dimension space. Seurat (Butler et al. 2018) is a toolkit for analysis of single-cell RNA sequencing data and provides many dimension reduction methods such as PCA and t-SNE.

CLUSTERING METHODS FOR SINGLE-CELL RNA-SEQ

Diverse types of clustering methods have been developed for detecting cell types from single-cell RNA-seq data. Those methods can be roughly classified into four categories, including *k*-means clustering, hierarchical clustering, community-detection-based clustering, and density-based clustering. We review several computational applications of those clustering methods with their strengths and limitations. Tables 1 and 2 illustrate the overview of the state-of-the-art clustering methods on single-cell RNA-seq data. Table 1 summarizes the state-of-the-art clustering methods with its types (*k*-means, hierarchical, spectral, louvain, and density-based), strengths, and limitations. Table 2 summarizes the state-of-the-art clustering methods with complexity, automatic determination of clusters, scalability, and detection of rare clusters.

k-means clustering

k-means clustering is the most popular clustering approach, which iteratively finds a predefined number of *k* cluster centers (centroids) by minimizing the sum of the squared Euclidean distance between each cell and its closest centroid. In addition, it is suitable for large data sets since it can scale linearly with the number of data points (Lloyd 1982).

Several clustering tools based on *k*-means have been developed for interpreting single-cell RNA-seq data. SAIC (Yang et al. 2017) utilizes an iterative *k*-means clustering to identify the optimal subset of signature genes that separate single cells into distinct clusters. pcaReduce (Zurauskiene and Yau 2016) is a hierarchical clustering method while it relies on *k*-means results as the initial clusters. RaceID (Grun et al. 2015) applies *k*-means to unravel the heterogeneity of rare intestinal cell types (Tibshirani et al. 2001).

However, *k*-means clustering is a greedy algorithm that may fail to find its global optimum; the predefined number of clusters k can affect the clustering results; and another disadvantage is its sensitivity to outliers since it tends to identify globular clusters, resulting in failures in detecting of rare cell types.

To overcome the above drawbacks, SC3 (Kiselev et al. 2017) integrated individual *k*-means clustering results

TABLE 1. Overview of the state-of-the-art clustering methods on single-cell RNA-seq data, Part 1

Method	Type	Strengths	Limitations
SAIC	*k*-means	Low complexity; scalable to large data	Sensitive to outliers; no estimation of number of clusters
RaceID	*k*-means	Sensitive to rare cell types; estimation of number of clusters	Not suitable for rare cell types
pcaReduce	*k*-means/hierarchical	Hierarchy solutions	Not stable
SC3	*k*-means/hierarchical	High accuracy; estimation of number of clusters	High complexity; not scalable to large data
CIDR	Hierarchical	Sensitive to dropout	High complexity
BackSPIN	Hierarchical	Simultaneously cluster genes and cells	High complexity
SIMLR	Spectral	Suitable for data with heterogeneity and noise	Not scalable to large data
SinNLRR	Spectral	Suitable for noise data	No estimation of number of clusters
SCANPY	Louvain	Low complexity; scalable to large data	May not find small community
Seurat	Louvain	Low complexity; scalable to large data	May not find small community
GiniClust	Density-based	Available for detection of rare cell types	Not sensitive to large clusters

TABLE 2. Overview of the state-of-the-art clustering methods on single-cell RNA-seq data, Part 2

Method	Complexity	Automatic determination of clusters	Scalability	Detection of rare clusters
SAIC	O(N∗ K∗ T)	×	✓	×
RaceID	O(N∗ K∗ T)	✓	✓	✓
pcaReduce	O(N∗ K∗ T)/O(N^3)	✓	×	×
SC3	O(N∗ K∗ T)/O(N^3)	✓	×	×
CIDR	$O(N^3)$	✓	✓	✓
BackSPIN	$O(N^3)$	×	×	×
SIMLR	$O(N^3)$	✓	×	×
SinNLRR	$O(N^3)$	✓	✓	×
SCANPY	$O[N\log(N)]$	✓	✓	×
Seurat	$O[N\log(N)]$	✓	✓	×
GiniClust	$O[N\log(N)]$	✓	×	✓

with different initial conditions as the consensus clusters. RaceID2 (Grun et al. 2016) replaced the *k*-means clustering with *k*-medoids clustering that use 1-Pearson's correlation instead of Euclidean distance as the clustering distance metric. RaceID3 (Herman and Grun 2018), as the advanced version of RaceID2, added feature selection and introduced Random Forest to reclassify *k*-means clustering results. RaceID3 introduced a function to compute the average within-cluster dispersion up to a number of clusters helping to determine the suitable cluster number, which is different from other *k*-means-based methods.

Hierarchical clustering

Hierarchical clustering is another widely used clustering algorithm on single-cell RNA-seq data. There are two types of hierarchical strategies, including: (i) agglomerative clustering, in which the individual cells are progressively merged into clusters according to distance measures; (ii) divisive clustering, in which each cluster is split into small groups recursively. These two strategies build a hierarchical structure among the cells/genes and enable the improvement in finding rare cell types as small clusters. Hierarchical clustering does not require predetermining the number of clusters and making assumptions for the distributions of single-cell RNA-seq data. Hence, many single-cell RNA-seq clustering methods have adopted it as part of the computational component.

CIDR (Lin et al. 2017b) integrates both dimension reduction and clustering based on hierarchical clustering into single-cell RNA-seq analysis and uses implicit imputation process for dropout effects; it provides a stable estimation of pairwise cell distances. BackSPIN (Zeisel et al. 2015) developed a biclustering method based on divisive hierarchical clustering and sorting points into neighborhoods (SPIN) (Tsafrir et al. 2005) to simultaneously cluster genes and cells. The number of splits needs to be set manually in BackSPIN. Although intensive splits can improve the clustering resolution, it is prone to over-partition. pcaReduce (Zurauskiene and Yau 2016) is an agglomerative hierarchical clustering approach with PCA which provides clustering results in a hierarchical representation. SINCERA (Guo et al. 2015) is a simple pipeline adopted hierarchical clustering with centered Pearson's correlation and average linkage method to identify cell types.

The agglomerative hierarchical clustering has a time complexity (the amount of time taken by an algorithm to run) of $O(N^3)$, whereas divisive clustering is $O(2^N)$. Although hierarchical clustering can reveal the hierarchical relations among cells/genes and does not require setting the number of clusters, it has high time complexity.

Community-detection-based clustering

Given the limitations of *k*-means and hierarchical clustering methods in large-scale data sets, community-detection-based clustering has been increasingly popular recently. Community detection is crucial in sociology, biology, and other systems that can be represented as graphs with nodes and edges. For single-cell RNA-seq data, nodes refer to cells and edge weights are represented by cell–cell pairwise distances. The idea of graph-based clustering is to delete the branch with maximum weights (cell–cell pairwise distances) in a dense graph (cell relationship network). There are three commonly used approaches for community-detection-based (graph-based) clustering including clique algorithm, spectral clustering, and Louvain algorithm (Blondel et al. 2008).

A clique is a set of points fully connected to each other in a graph and represents a cluster (community). Although finding cliques in a graph is NP-complete (NP-complete denotes a nondeterministic Turing machine accepted in polynomial time complete; a problem is NP-complete if answers can be verified quickly), some studies have been conducted to address it, such as heuristic optimization. SNN-Clip (Xu and Su 2015) was proposed to leverage

the concept of shared nearest neighbor to calculate cell similarity (Zhang et al. 2009) for finding all quasi-cliques since the shared nearest neighbor graph is sparse. SNN-Clip does not require specifying the number of clusters manually but it is nonscalable and the resultant clusters are not stable.

Spectral clustering is a widely used clustering method recently. It is designed to be adaptive to data distribution by relying on the eigenvalues of the cell similarity matrix. Nonetheless, the spectral clustering's time complexity is $O(N^3)$. SIMLR (Wang et al. 2017) is an analytic framework for dimension reduction, clustering, and visualization of single-cell RNA-seq data. It is a method specifically designed for single-cell RNA-seq. SIMLR combines spectral clustering with multiple kernel similarity measures for clustering expression data generated from cross-platform and cross-condition experiments. In addition, SIMLR has an advantage in processing large-scale data sets with heavy noise (large meaningless data often causes the algorithm to miss patterns in data). SinNLRR (Zheng et al. 2019) was proposed to impose a nonnegative and low rank structure on the cell similarity matrix and then apply spectral clustering to detect cell types. However, the spectral clustering requires users to set the number of clusters in the data.

Louvain (Blondel et al. 2008) is the most popular community detection algorithm and widely used for single-cell RNA-seq data. It recursively merges communities into a single node and executes the modularity clustering on the condensed graphs. The time complexity of Louvain is $O[Nlog(N)]$, which is lower than other community-detection-based algorithms. SCANPY (Wolf et al. 2018) is a scalable toolkit for single-cell RNA-seq analysis and its clustering section is based on the Louvain algorithm. SCANPY has advantages in scaling its computation with the number of cells (over one million). Seurat (Satija et al. 2015) also applied the Louvain algorithm to cluster the cell types for the mapping of cellular localization.

Leiden (Traag et al. 2019) is an improvement of the Louvain algorithm. Louvain has three limitations: (i) there are limits to the precision of community division; (ii) the density of node distribution within a grouping will affect the identification of subpopulations; and (iii) badly connected communities. Leiden addresses the above important shortcomings, is faster to find better partitions and provides explicit guarantees and bounds. Leiden has been implemented in Seurat, Scanpy, and Monocle since it was proposed in 2020.

Density-based clustering

Density-based clustering methods separate data space into highly dense clusters. It can learn clusters with arbitrary shapes and identify noise (outliers). The most popular density-based clustering algorithm is DBSCAN (Ester et al. 1996). DBSCAN does not need to predetermine the number of clusters and its time complexity is $O[Nlog(N)]$. However, DBSCAN requires the user to set two parameters including ε (eps) and the minimum number of points required to form a dense region (minPts) (Ester et al. 1996) that will affect its clustering results. Jiang et al. (2016a) developed GiniClust, detecting rare cell types from single-cell gene expression data, and its clustering method is based on DBSCAN. GiniClust is effective in finding rare cell types since it can be adaptively adjusted to set a lower ε. However, such a design may lead to unreasonable large cell clusters. Monocle2 (Qiu et al. 2017a) also applied density peak clustering to identify the differentially expressed genes between cells.

Deep learning–based clustering

Recently, a few deep learning–based methods have been developed to deal with the clustering problem. DESC (Li et al. 2020) projected the high-dimensional scRNA-seq data into a low-dimensional level by autoencoder (an unsupervised learning technique in which we leverage neural networks for the task of representation learning) and conducted an iterative soft-clustering procedure to reduce the influence of batch effect. However, the DESC is designed for the batch correction since the objective of a soft-clustering procedure is to move each cell from two batch data sets to its nearest cluster. scziDesk (Chen et al. 2020) proposed to use a denoising autoencoder to characterize scRNA-seq data and then build a soft self-training *k*-means algorithm to cluster the cell population. scVAE (Gronbech et al. 2020) is based on variational autoencoders and could achieve better clustering and representation of the cells from both scRNA-seq and bulk data.

EXPERIMENTAL EVALUATIONS FOR CLUSTERING METHODS

In this section, we conduct independent experiments to evaluate several widely used single-cell RNA-seq clustering methods. These clustering methods contain RaceID3 (Herman and Grun 2018), Monocle2 (Qiu et al. 2017a), SIMLR (Wang et al. 2017), Seurat (Satija et al. 2015), SC3 (Kiselev et al. 2017), and CIDR (Lin et al. 2017b). We applied six single-cell RNA-seq clustering methods on four simulated single-cell RNA-seq data sets with different numbers of cells, 2k, 8k, 12k, and 100k, and five public single-cell RNA-seq data sets (GSE6552, GSE74672, SCP345, SCP916, and GSE162086 with cell type annotations). For evaluation and comparison, we introduce three commonly used metrics including adjusted Rand index, running time, and homogeneity score to measure the clustering performance and efficiency, respectively. Source code implemented in R programming language and raw data for all figures can be found at https://github.com/Alexzsx/Single-cell-RNA-seq-Clustering.

The parameter settings of the cluster methods on both data sets are tabulated in Supplemental Table S1. We fed all genes into each clustering method (package). Some of the processing steps have been implicitly incorporated into the clustering methods. The number of PC features were chosen based on the default setting given by individual methods. In particular, we would like to note that most parameters were chosen based on the default setting given by individual methods. Some of the processing steps have been implicitly incorporated into the clustering methods. RaceID3 filters the low-quality cells/genes for the purpose of quality control. Monocle2 provides a scaling-based normalization method and many dimension reduction methods (e.g., t-SNE and PCA). SIMLR provides impute function, normalization, and dimension reduction (t-SNE). Seurat filters the low-quality cells/genes, normalizes the expression values by "LogNormalize," removes cells with high proportion of mitochondrial gene expression, as well as some extremum cells, and runs PCA for dimension reduction. SC3 filters the underexpressed and overexpressed genes and conducts PCA on cell–cell distances. CIDR integrates PCA for dimension reduction.

Evaluation metrics for clustering

Since the single-cell RNA-seq clustering is an unsupervised learning task in most studies, three common metrics, adjusted Rand index, running time, and homogeneity score, are introduced for the evaluation.

The adjusted Rand index (ARI) proposed by Hubert and Arabie (1985) can be used to measure the similarity between the clustering results of interest and the true clustering. However, ARI is widely applied as the metric of single-cell RNA-seq clustering only when the cell labels are available (Xu and Su 2015; Ntranos et al. 2016; Aibar et al. 2017; Kiselev et al. 2017; Lin et al. 2017b). Given a set of n cells and two clusters ($X = \{X_1, X_2, \ldots, X_s\}$ partitioned by the clustering method, and $Y = \{Y_1, Y_2, \ldots, Y_r\}$ partitioned by annotated cell types) of these cells, the overlap between the two clusters can be summarized in a contingency table with s rows and r columns. The ARI is defined as below.

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

where $n_{ij} = |X_i \cap {}^{Y}{}_{j}|$ denotes the values from the contingency table; $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$ represent the ith row sums and jth column sums of the contingency table, respectively. $ARI = 1$ indicates a perfect overlap between clusters X and Y, while $ARI = 0$ indicates random clustering.

The homogeneity score (Rosenberg and Hirschberg 2007) evaluates the performance of clustering results with regard to the ground truth. It is defined as:

$$\text{Homogeneity} = \frac{I(X,\ Y)}{H(Y)} \quad (2)$$

where $H(Y)$ is the entropy of Y and $I(X, Y)$ is the mutual information of X and Y. It is bounded between 0 and 1. *Homogeneity* = 1 indicates all of its clusters contain only data points from a single class, while low values indicate that clusters contain mixed known groups.

In addition, running time is usually measured to evaluate the algorithm efficiency. High efficiency is an important feature since the single-cell RNA-seq data usually come up with thousands of cells and genes.

Performance on simulated single-cell RNA-seq data sets

In simulated single-cell RNA-seq data sets, there are four data sets with different numbers of cells: 2k, 8k, 12k, and 100k. We set the number of genes as 10k for all data sets. In addition, there are four cell types in each data set. Figure 2 shows the ARI, homogeneity scores, and running time of RaceID3, Monocle2, SIMLR, Seurat, SC3, and CIDR on simulated data sets for performance comparison. The results show that CIDR, RaceID3, and SC3 exhibit the best ARI (=1) and homogeneity (close to 1) performance among the six methods on small data sets (below 10k), while SC3 took more time on 2k cells, and all these three methods took over 18 min on 8k cells. When the cell number increased to 12k, CIDR and SC3 still maintained their advantages on three metrics and RaceID3 came to a time-consuming method with 9.17 h. However, when the cell number exceeds 100k, only SC3 and Seurat can still run successfully, and SC3 still performs best with ARI (=1), homogeneity (=0.999), and running time (21.78 min). Hence, the results show that RaceID3, CIDR, Monocle2, and SIMLR are not suitable for large data. SC3, SIMLR, and Monocle2 cannot provide an accurate estimation of the cluster count and has to be determined manually. Seurat, Monocle2, and RaceID3 require the user to adjust multiple parameters to achieve the best clustering performance, which limits user friendliness.

Performance on real single-cell RNA-seq data sets

In this section, we performed comparison experiments to evaluate the performances of those single-cell RNA-seq clustering methods on five publicly available data sets. Those data sets are from Mouse embryo stem (GSE65525; 2717 cells) (Klein et al. 2015), Mouse brain (GSE74672; 2881 cells) (Romanov et al. 2016), Human blood (SCP345; 13316 cells) (Tran et al. 2021), Mouse tissues (SCP916; 12648 cells) (Sanderson et al. 2020), and Human T Cell (GSE162086; 104417 cells) (Bacher et al. 2020). Supplemental Table S2 summarizes the statistics of the above five single-cell RNA-seq data sets with varying

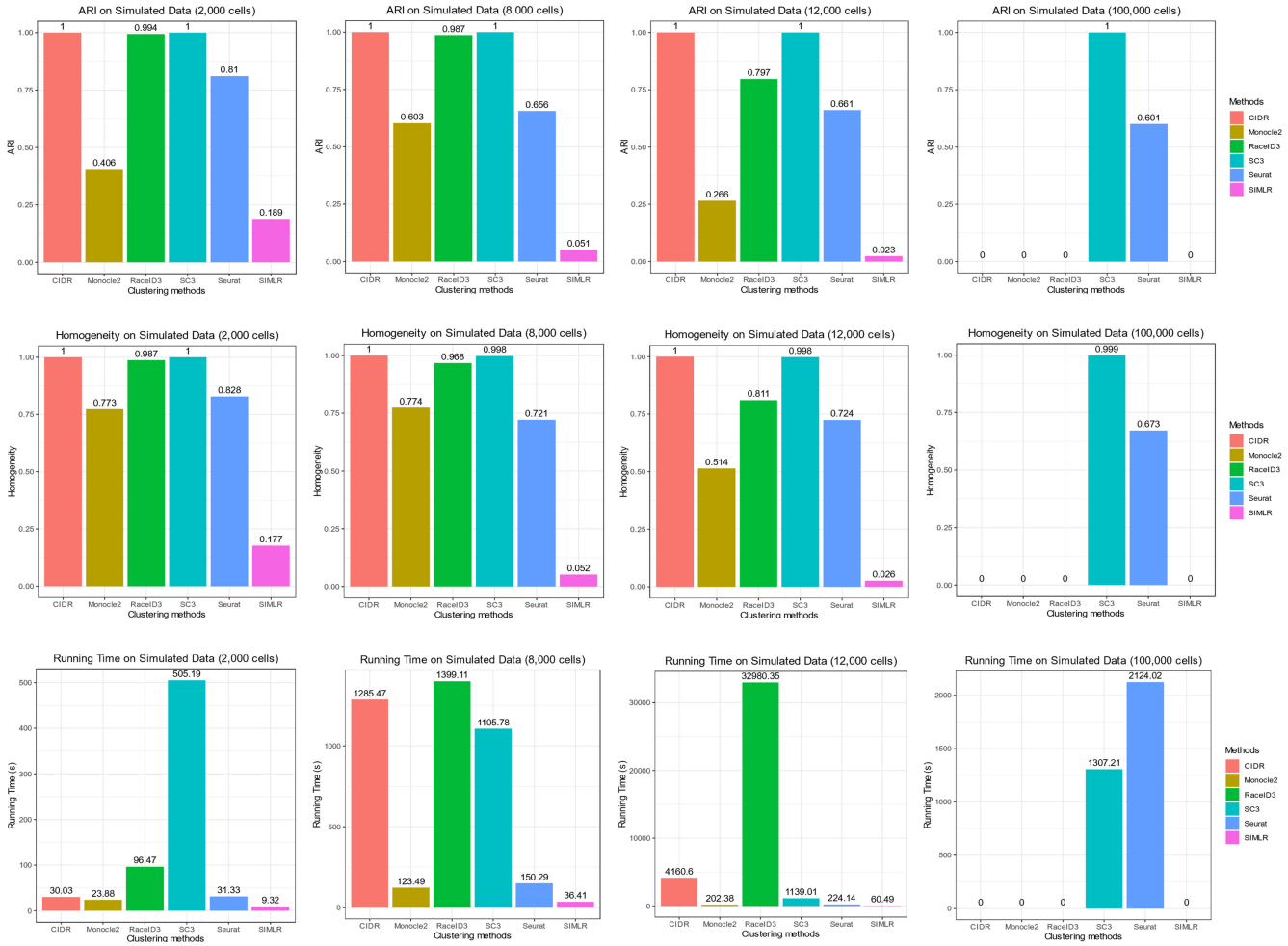

FIGURE 2. Comparison of clustering performance on simulated single-cell RNA-seq data sets. The *x*-axis represents the clustering methods. The *y*-axis denotes the ARI, homogeneity scores, or running time of clustering results by RaceID3 (Herman and Grun 2018), Monocle2 (Qiu et al. 2017a), SIMLR (Wang et al. 2017), Seurat (Satija et al. 2015), SC3 (Kiselev et al. 2017), and CIDR (Lin et al. 2017b).

degrees of number of genes, number of cells (varying from 2k to 100k), and number of cell types.

Figure 3 deploys the ARI, homogeneity scores, and running time of RaceID3, Monocle2, SIMLR, Seurat, SC3, and CIDR on GSE65525, GSE74672, SCP345, SCP916, and GSE162086 for performance comparison. In the Mouse embryo stem data set (GSE65525), SC3 obtained an ARI of 1 and homogeneity score of 1 and can correctly identify all cells for each cell type. Monocle2 was the fastest with a running time of 42.52 sec. In the Mouse brain data set (GSE74672), RaceID3 and SC3 outperformed the other clustering methods across the ARI, homogeneity, and running time. SIMLR still took the longest running time. When the number of cells is increased over ten thousand, RaceID3 achieved the best clustering performance on ARI (=0.757) and homogeneity scores (0.759) in Human blood (SCP345; 13316 cells). Seurat was the best efficient clustering method, while its clustering performance was far behind RaceID3 and SC3.

The fourth column from Figure 3 shows the comparison results of clustering methods on Mouse tissues (SCP916; 12648 cells). Although Seurat outperformed other methods in all metrics (ARI, homogeneity scores, and running time), the clustering performance is not good (ARI = 0.122, Homogeneity = 0.304) except for the running time. When the number of cells exceeds 100k, only Seurat could run successfully, but clustering performance is very bad. The experimental results from these five real data sets show that RaceID3, SC3, and Seurat usually have better clustering accuracy in small data sets while Seurat is scalable to larger single-cell RNA-seq data sets. Seurat was the fastest across all single-cell RNA-seq data sets except GSE65525. The visualization results are directly obtained from t-SNE. Figure 4 shows that SIMLR and Seurat perform well in GSE65525. Results from Supplemental Figure S1 show that all methods result in different degrees of undesirable overlap between clusters in GSE74672. When the cell numbers and cell types increase,

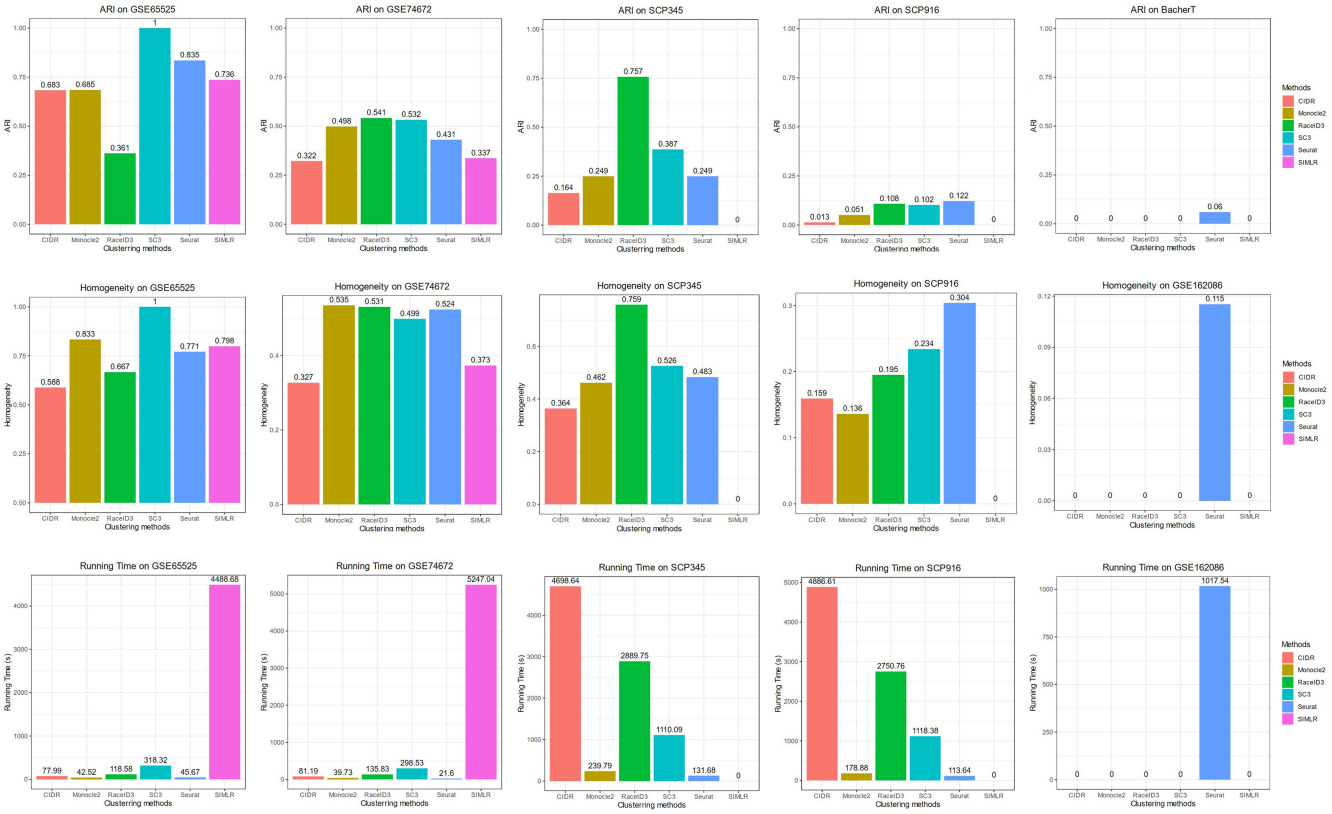

FIGURE 3. Comparison of clustering performance on real single-cell RNA-seq data sets. The *x*-axis represents the clustering methods. The *y*-axis denotes the ARI, homogeneity scores, or running time of clustering results by RaceID3 (Herman and Grun 2018), Monocle2 (Qiu et al. 2017a), SIMLR (Wang et al. 2017), Seurat (Satija et al. 2015), SC3 (Kiselev et al. 2017), and CIDR (Lin et al. 2017b).

most clustering methods did not cluster the correct cell types as shown in Supplemental Figure S2.

We conducted an evaluation of rare cell population identification across all five clustering methods on the mouse brain data set (GSE74672). The mouse brain data set contains six cell types including two rare cell types (of 48 microglia cells and 71 vsm cells, respectively) (Fig. 5). Seurat could identify almost both rare cell clusters distinguished from other cell clusters, and only a very few cells were clustered into other clusters (Seurat in Fig. 5). However, the rest of the evaluated clustering methods cannot resolve the rare cell clusters well, such as both RaceID3 and SIMLR, which merged most two rare cells into one group (i.e., Cluster 4 in RaceID3 and Cluster 8 in SIMLR, respectively). Note that RaceID3 is specifically developed for rare cell identification but still does not perform well. CIDR and SC3 split both common and rare cell clusters into smaller subclusters (Fig. 5). It is suggested that Seurat is qualified for rare cell-specific detection in the case of large single-cell RNA-seq data sets.

Data normalization performance on two single-cell RNA-seq data sets

The focus of this section is to compare three representative single-cell specific data normalization methods reviewed in the section "Normalization." Of these, SCnorm (Bacher et al. 2017), scran (Lun et al. 2016), and Seurat (Satija et al. 2015) are defined and applied according to their respective packages. We evaluate the data normalization methods on two publicly available single-cell RNA-seq data sets, namely mouse embryo stem (GSE65525; 2717 cells), containing four groups of cells and Mouse embryonic data (GSE29087; 92 cells), containing two groups of cells. In addition, GSE29087 has eight spike-in genes. All methods were performed using the default settings of their respective R packages.

The normalization results are displayed in Supplemental Figure S3 by the t-SNE visualization. For Mouse embryonic data (GSE29087; 92 cells), Seurat presents the clearest division between the two cell types. SCnorm misclassified several points, while scran seemingly divided the data points into three groups. All normalization methods showed noticeable improvement over the raw data. For Mouse embryo stem (GSE65525; 2717 cells), Seurat divided the 2717 cells into three groups with a mixed cell type of d2 and d3. However, it could be clustered into four real cell types when applied to the clustering step as shown in Figure 4. scran is unable to separate those four cell types and there is no difference between raw data and scran results. SCnorm could not be executed on GSE65525 since it exceeded thousands of cells.

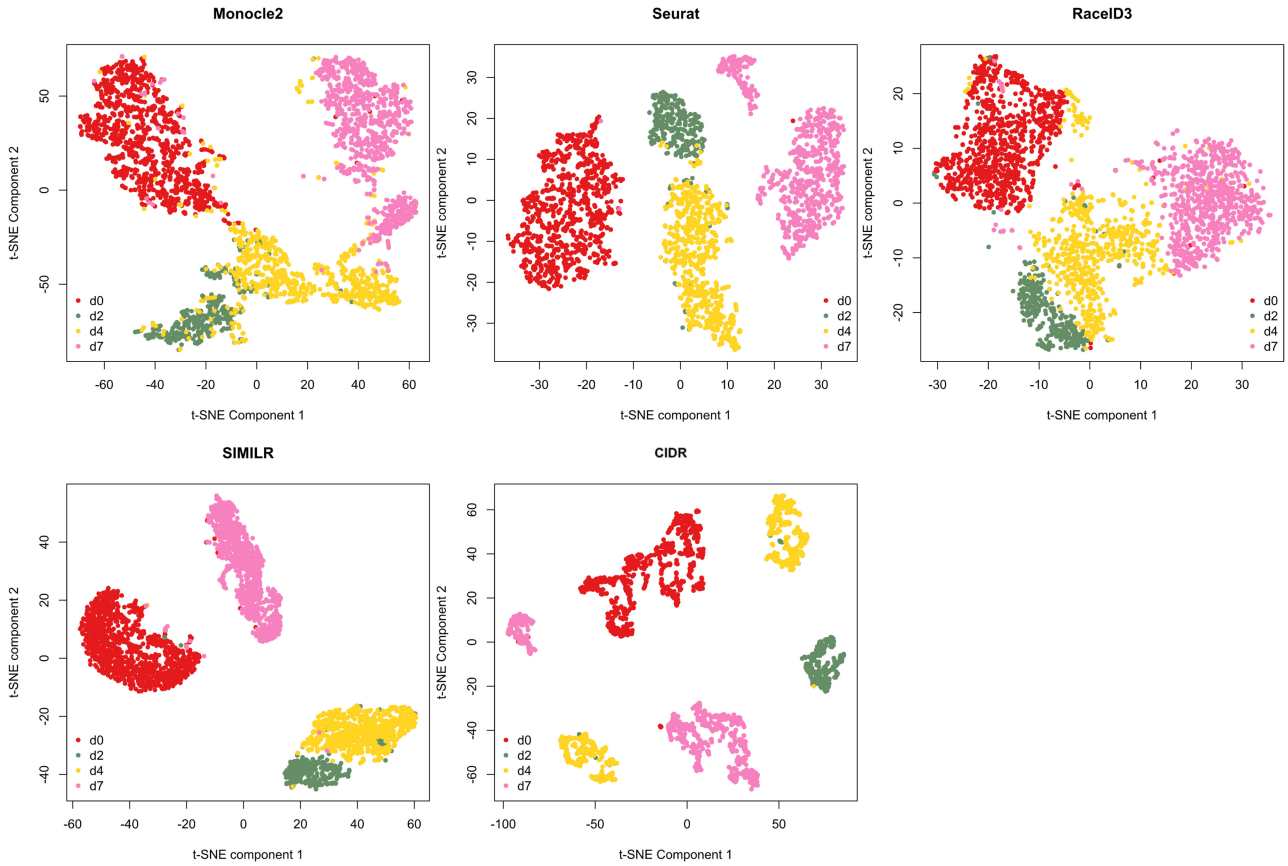

FIGURE 4. Visualization of clustering performance on mouse embryonic stem single-cell RNA-seq large-scale data set (GSE65525) by SIMLR (Wang et al. 2017), RaceID3 (Herman and Grun 2018), Seurat (Satija et al. 2015), CIDR (Lin et al. 2017b), and Monocle2 (Qiu et al. 2017a).

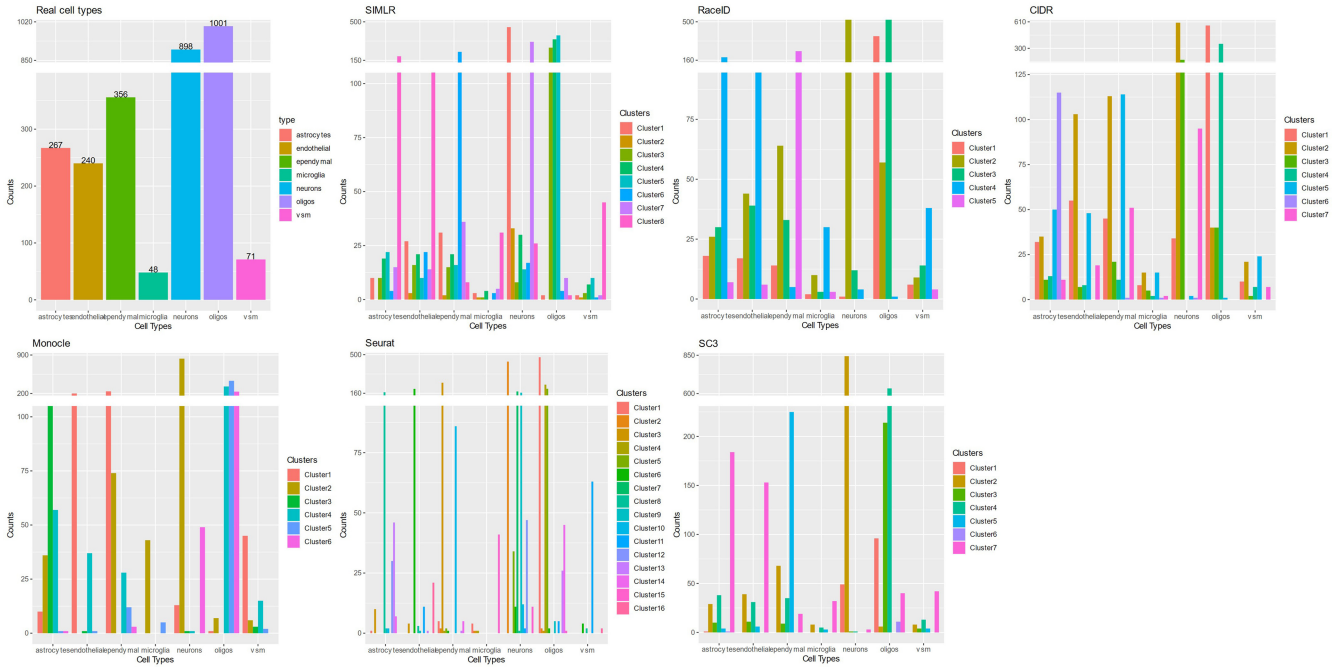

FIGURE 5. Visualization of identification performance of rare cell populations on mouse brain data set (GSE74672) by RaceID3 (Herman and Grun 2018), Monocle2 (Qiu et al. 2017a), SIMLR (Wang et al. 2017), Seurat (Satija et al. 2015), SC3 (Kiselev et al. 2017), and CIDR (Lin et al. 2017b).

According to the evaluation results, we recommend Seurat as the first choice for single-cell RNA-seq data normalization, since it is scalable to varying data sizes and effective. SCnorm and scran are also sound options on smaller single-cell RNA-seq data sets.

CONCLUSIONS

Single-cell RNA-seq data analysis is a crucial component in whole-transcriptome studies. In particular, data clustering is the central component of single-cell RNA-seq analysis. Clustering results can affect the performance of downstream analysis including identifying rare or new cell types, gene expression patterns that are predictive of cellular states, and functional implications of stochastic transcription. There are several related studies for the performance evaluation of clustering methods on single-cell RNA-seq data (Duo et al. 2018; Freytag et al. 2018). Those studies focused on assessing the methods for clustering single-cell RNA-seq data, while the data preprocessing steps may not be included in the respective discussion section, although it could have significantly influenced the downstream clustering performance. Therefore, in this study, we reviewed several clustering methods. In addition, the upstream RNA-seq data preprocessing steps have also been reviewed since those steps can significantly affect the downstream clustering performance. Lastly, our performance comparison experiments have also been conducted, revealing independent insights into the state-of-the-art methods without any conflict of interest.

Quality control ensures that the data quality is sufficient for downstream analysis. Most clustering packages evaluated in this study have included quality control steps such as Seurat, CIDR, and RaceID3. Those basic quality control steps are sufficient to find outlier peaks in the number of genes, the count depth, and the fraction of mitochondrial reads. We recommend leveraging Seurat, CIDR, and RaceID3 directly. According to our evaluation results, we recommend Seurat as the first choice since its simple processing performs very well on both small and large single-cell RNA-seq data sets. According to the experimental evaluation for the clustering methods, we recommend the SC3 and RaceID3 methods as the first choice for single-cell RNA-seq clustering. SC3 and RaceID3 usually have better clustering accuracy and are scalable to small single-cell RNA-seq data sets. When the cell numbers increase, we recommend Seurat as the first choice for single-cell RNA-seq clustering.

However, each clustering method has its drawbacks. For instance, *k*-means clustering requires users to determine the number of clusters and is sensitive to outliers; hierarchical clustering has high complexity and may be unsuitable for large-scale single-cell RNA-seq data; community-detection-based clustering cannot provide the estimation of the number of clusters and is unsuitable for small communities; and density-based clustering has advantages in detecting rare cell types with a sacrifice on large cluster performance.

In addition to those limitations, there are still some technical challenges in single-cell RNA-seq clustering. With the advanced development of single-cell RNA-seq techniques, the single-cell data sets are becoming extremely high-dimensional and sparse. Although some methods can deal with those data in a time span of hours, such as SIMLR, visualization of those data is still a challenge. Moreover, the low dimensionality of expression profiles implies intensive gene coexpression signatures that may inspire us to develop new clustering methods for low-dimensional data to interpret cell types (Crow and Gillis 2018). Advanced data integration and analysis approaches are needed for both basic research and clinical studies in the coming years.

Finally, with the advanced development of multiomics integration, extra single-cell omics (such as epigenomic, spatial, and proteomics) information has enriched the single-cell data and benefited the accuracy of clustering. However, the current clustering methods may not be scalable to the integrated data sets. New clustering methods for single-cell integration data are expected in the future.

DATA DEPOSITION

Source code implemented in R programming language and data sets can be found at https://github.com/Alexzsx/Single-cell-RNA-seq-Clustering.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This research was substantially sponsored by research projects (grant no. 62102294, grant no. 32170654, and grant no. 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. This project was substantially funded by the Strategic Interdisciplinary Research Grant of City University of Hong Kong (project no. 2021SIRG036). The work described in this paper was substantially supported by a grant from the Health and Medical Research Fund, the Food and Health Bureau, and The Government of the Hong Kong Special Administrative Region (07181426). The work described in this paper was partially supported by grants from the City University of Hong Kong (CityU 11203520, CityU 11203221).

Author contributions: S.Z. and K.W. designed the research; S.Z., X.L., and J.L. performed the research; S.Z. and Q.L. analyzed the data; and S.Z. and K.W. wrote the manuscript.

REFERENCES

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14:** 1083. doi:10.1038/nmeth.4463

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106. doi:10.1186/gb-2010-11-10-r106

Andrews TS, Hemberg M. 2018. Identifying cell populations with scRNASeq. *Mol Aspects Med* **59:** 114. doi:10.1016/j.mam.2017.07.002

Bacher R, Chu L-F, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. 2017. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14:** 584. doi:10.1038/nmeth.4263

Bacher P, Rosati E, Esser D, Martini GR, Saggau C, Schiminsky E, Dargvainiene J, Schröder I, Wieters I, Khodamoradi Y, et al. 2020. Low-avidity CD4+ T cell responses to SARS-CoV-2 in unexposed individuals and humans with severe COVID-19. *Immunity* **53:** 1258. doi:10.1016/j.immuni.2020.11.016

Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37:** 38. doi:10.1038/nbt.4314

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* **2008:** P10008. doi:10.1088/1742-5468/2008/10/P10008

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33:** 155. doi:10.1038/nbt.3102

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36:** 411. doi:10.1038/nbt.4096

Chen L, Wang W, Zhai Y, Deng M. 2020. Deep soft K-means clustering with self-training for single-cell RNA sequence data. *NAR Genom Bioinform* **2:** lqaa039. doi:10.1093/nargab/lqaa039

Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, Purdom E, Dudoit S, Yosef N. 2019. Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell Syst* **8:** 315. doi:10.1016/j.cels.2019.03.010

Crow M, Gillis J. 2018. Co-expression in single-cell analysis: saving grace or original sin? *Trends Genet* **34:** 823. doi:10.1016/j.tig.2018.07.007

Davie K, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft Ł, Aibar S, Makhzami S, Christiaens V, Bravo González-Blas C, et al. 2018. A single-cell transcriptome atlas of the aging Drosophila brain. *Cell* **174:** 982. doi:10.1016/j.cell.2018.05.057

Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A, Wang W. 2015. Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* **31:** 2225. doi:10.1093/bioinformatics/btv122

Ding J, Condon A, Shah SP. 2018. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun* **9:** 2002. doi:10.1038/s41467-018-04368-5

Duo A, Robinson MD, Soneson C. 2018. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* **7:** 1141. doi:10.12688/f1000research.15666.2

Ester M, Kriegel HP, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd international conference on knowledge discovery and data mining*, pp. 226–231.

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16:** 278. doi:10.1186/s13059-015-0844-5

Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. 2018. Comparison of clustering tools in R for medium-sized 10x genomics single-cell RNA-sequencing data. *F1000Res* **7:** 1297. doi:10.12688/f1000research.15809.1

Gronbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. 2020. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36:** 4415. doi:10.1093/bioinformatics/btaa293

Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525:** 251. doi:10.1038/nature14966

Grun D, Muraro MJ, Boisset J-C, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, et al. 2016. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19:** 266. doi:10.1016/j.stem.2016.05.010

Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. 2015. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol* **11:** e1004575. doi:10.1371/journal.pcbi.1004575

Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20:** 1. doi:10.1186/s13059-019-1874-1

Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36:** 421. doi:10.1038/nbt.4091

Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, et al. 2018. Mapping the mouse cell atlas by microwell-seq. *Cell* **172:** 1091. doi:10.1016/j.cell.2018.02.001

Herman JS, Grun D. 2018. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat Methods* **15:** 379. doi:10.1038/nmeth.4662

Hubert L, Arabie P. 1985. Comparing partitions. *J Classif* **2:** 193. doi:10.1007/BF01908075

Jiang L, Chen H, Pinello L, Yuan G-C. 2016a. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* **17:** 144. doi:10.1186/s13059-016-1010-4

Jiang P, Thomson JA, Stewart R. 2016b. Quality control of single-cell RNA-seq by SinQC. *Bioinformatics* **32:** 2514. doi:10.1093/bioinformatics/btw176

Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20:** 273. doi:10.1038/s41576-018-0088-9

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14:** 483. doi:10.1038/nmeth.4236

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161:** 1187. doi:10.1016/j.cell.2015.04.044

Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, et al. 2017. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49:** 708. doi:10.1038/ng.3818

Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly MP, Hu G, Li M. 2020. Deep learning enables accurate

clustering with batch effect removal in single-cell RNA-seq analysis. *Nat Commun* **11:** 2338. doi:10.1038/s41467-019-13993-7

Lin C, Jain S, Kim H, Bar-Joseph Z. 2017a. Using neural networks for reducing the dimensions of single-cell RNA-seq data. *Nucleic Acids Res* **45:** e156. doi:10.1093/nar/gkx681

Lin P, Troup M, Ho JWK. 2017b. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* **18:** 59. doi:10.1186/s13059-017-1188-0

Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. 2019. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* **16:** 243. doi:10.1038/s41592-018-0308-4

Lloyd S. 1982. Least squares quantization in PCM. *IEEE Trans Inform Theory* **28:** 129. doi:10.1109/TIT.1982.1056489

Lun AT, Bach K, Marioni JC. 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* **17:** 75. doi:10.1186/s13059-016-0947-7

McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: uniform manifold approximation and projection. *J Open Source Softw* **3:** 861. doi:10.21105/joss.00861

Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. 2016. Fast and accurate single-cell RNA-seq analysis by clustering of transcript compatibility counts. *Genome Biol* **17:** 112. doi:10.1186/s13059-016-0970-8

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, et al. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344:** 1396. doi:10.1126/science.1254257

Prabhakaran S, Azizi E, Carr A, Pe'er D. 2016. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *JMLR Workshop Conf Proc* **48:** 1070–1079.

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. 2017a. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14:** 979. doi:10.1038/nmeth.4402

Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. 2017b. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14:** 309. doi:10.1038/nmeth.4150

Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* **32:** 896. doi:10.1038/nbt.2931

Romanov RA, Zeisel A, Bakker J, Girach F, Hellysaz A, Tomer R, Alpár A, Mulder J, Clotman F, Keimpema E, et al. 2016. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* **20:** 176. doi:10.1038/nn.4462

Rosenberg A, Hirschberg J. 2007. V-Measure: a conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, Prague, Czech Republic. Association for Computational Linguistics. https://aclanthology.org/D07-1043

Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. 2017. The human cell atlas: from vision to reality. *Nature* **550:** 451. doi:10.1038/550451a

Sanderson SM, Xiao Z, Wisdom AJ, Bose S, Liberti MV, Reid MA, Hocke E, Gregory SG, Kirsch DG, Locasale JW. 2020. The Na^+/K^+ ATPase regulates glycolysis and modifies immune metabolism in tumors. *bioRxiv* doi:10.1101/2020.03.31.018739

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33:** 495. doi:10.1038/nbt.3192

Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, et al. 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510:** 363. doi:10.1038/nature13437

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14:** 618. doi:10.1038/nrg3542

Sun S, Zhu J, Ma Y, Zhou X. 2019. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* **20:** 269. doi:10.1186/s13059-019-1898-6

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc B (Stat Methodol)* **63:** 411. doi:10.1111/1467-9868.00293

Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9:** 5233. doi:10.1038/s41598-019-41695-z

Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T. 2021. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat Commun* **12:** 1029. doi:10.1038/s41467-021-21312-2

Tsafrir D, Tsafrir I, Ein-Dor L, Zuk O, Notterman DA, Domany E. 2005. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics* **21:** 2301. doi:10.1093/bioinformatics/bti329

Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, Hjerling-Leffler J, Haeggström J, Kharchenko O, Kharchenko PV, et al. 2015. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat Neurosci* **18:** 145. doi:10.1038/nn.3881

Vallejos CA, Marioni JC, Richardson S. 2015. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* **11:** e1004333. doi:10.1371/journal.pcbi.1004333

Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* **14:** 565. doi:10.1038/nmeth.4292

Van Unen V, Höllt T, Pezzotti N, Li N, Reinders MJT, Eisemann E, Koning F, Vilanova A, Lelieveldt BPF. 2017. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat Commun* **8:** 1740. doi:10.1038/s41467-017-01689-9

Wang D, Gu J. 2018. VASC: dimension reduction and visualization of dingle-cell RNA-seq data by deep variational autoencoder. *Genomics Proteomics Bioinformatics* **16:** 320. doi:10.1016/j.gpb.2018.08.003

Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. 2017. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **14:** 414. doi:10.1038/nmeth.4207

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19:** 15. doi:10.1186/s13059-017-1382-0

Wolock SL, Lopez R, Klein AM. 2019. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* **8:** 281. doi:10.1016/j.cels.2018.11.005

Xu C, Su Z. 2015. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31:** 1974. doi:10.1093/bioinformatics/btv088

Yang L, Liu J, Lu Q, Riggs AD, Wu X. 2017. SAIC: an iterative clustering approach for analysis of single cell RNA-seq data. *BMC Genomics* **18:** 689. doi:10.1186/s12864-017-4019-5

Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347:** 1138. doi:10.1126/science.aaa1934

Zhang S, Xu M, Li S, Su Z. 2009. Genome-wide de novo prediction of *cis*-regulatory binding sites in prokaryotes. *Nucleic Acids Res* **37:** e72. doi:10.1093/nar/gkp248

Zhang JM, Fan J, Fan HC, Rosenfeld D, Tse DN. 2018. An interpretable framework for clustering single-cell RNA-seq datasets. *BMC Bioinformatics* **19:** 93. doi:10.1186/s12859-018-2092-7

Zheng R, Li M, Liang Z, Wu F-X, Pan Y, Wang J. 2019. SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics* **35:** 3642–3650. doi:10.1093/bioinformatics/btz139

Zurauskiene J, Yau C. 2016. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17:** 140. doi:10.1186/s12859-016-0984-y