Genomics Proteomics Bioinformatics 19 (2021) 282–291

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

SSRE: Cell Type Detection Based on Sparse Subspace Representation and Similarity Enhancement

Zhenlan Liang [1], Min Li [1,*], Ruiqing Zheng [1], Yu Tian [1], Xuhua Yan [1], Jin Chen [2] Fang-Xiang Wu [3], Jianxin Wang [1]

[1] *School of Computer Science and Engineering, Central South University, Changsha 410083, China*
[2] *College of Medicine, University of Kentucky, Lexington, KY 40536, USA*
[3] *Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada*

Received 30 October 2019; revised 13 August 2020; accepted 29 October 2020
Available online 27 February 2021

Handled by Luonan Chen

KEYWORDS

Single-cell RNA sequencing;
Clustering;
Cell type;
Similarity learning;
Enhancement

Abstract Accurate identification of **cell types** from **single-cell RNA sequencing** (scRNA-seq) data plays a critical role in a variety of scRNA-seq analysis studies. This task corresponds to solving an unsupervised **clustering** problem, in which the similarity measurement between cells affects the result significantly. Although many approaches for cell type identification have been proposed, the accuracy still needs to be improved. In this study, we proposed a novel single-cell clustering framework based on **similarity learning**, called SSRE. SSRE models the relationships between cells based on subspace assumption, and generates a sparse representation of the cell-to-cell similarity. The sparse representation retains the most similar neighbors for each cell. Besides, three classical pairwise similarities are incorporated with a gene selection and **enhancement** strategy to further improve the effectiveness of SSRE. Tested on ten real scRNA-seq datasets and five simulated datasets, SSRE achieved the superior performance in most cases compared to several state-of-the-art single-cell clustering methods. In addition, SSRE can be extended to visualization of scRNA-seq data and identification of differentially expressed genes. The matlab and python implementations of SSRE are available at https://github.com/CSUBioGroup/SSRE.

Introduction

With the recent emergence of **single-cell RNA sequencing** (scRNA-seq) technology, numerous scRNA-seq datasets have been generated, which brings unique challenges for advanced omics data analysis [1,2]. Unlike bulk sequencing averaging

* Corresponding author.
E-mail: limin@mail.csu.edu.cn (Li M).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2020.09.004

the expression of mass cells, scRNA-seq technique quantifies gene expression at the single-cell resolution. Single-cell techniques promote a wide variety of biological topics such as cell heterogeneity, cell fate decision, and disease pathogenesis [3–5]. Among all the applications, cell type identification plays a fundamental role and its performance has a substantial impact on downstream studies [6]. However, identifying cell types from scRNA-seq data is still a challenging problem. The traditional clustering methods cannot work well on scRNA-seq data because of the high noise rate and high dropouts [7]. Therefore, new efficient and reliable clustering methods for cell type identification are urgent and meaningful.

In recent studies, several novel clustering approaches for detecting cell types from scRNA-seq data have been proposed. Among these methods, cell types are mainly decided on the basis of learned cell-to-cell similarity. For example, single-cell interpretation via multikernel learning (SIMLR) [8] visualizes and clusters cells using multi-kernel similarity learning [9], which performs well on grouping cells. Shared nearest neighbor (SNN)-Cliq [10] firstly constructs a distance matrix based on the Euclidean distance, and then introduces the shared k-nearest neighbors (KNN) model to redefine the similarity. SNN-Cliq provides both the estimation of cluster number and the clustering results by searching for quasi-cliques. Moreover, Corr [11] defines the cell-pair differentiability correlation instead of computing primary (dis)similarity like Pearson correlation and Euclidean distance. RAFSIL [12] divides genes into multiple clusters, and makes dimension reduction on each gene cluster. Then, RAFSIL concatenates the informative features obtained from each gene cluster. Finally, RAFSIL applies the random forest to calculate the similarities for each cell recursively. Besides, nonnegative matrix factorization (NMF) determines the cell types in the latent space [13], while SinNLRR [14] and AdaptiveSSC [15] learn the similarity matrix with nonnegative low rank and sparse constraints. Instead of learning a specific similarity, some researchers have turned to use ensemble learning that focuses on the consensus of multiple clustering methods [16,17].

Even though many approaches have been applied to cell type identification, most of them are sensitive to noise, especially for the high-dimensional data. They generally compute the similarity between two cells merely considering the gene expression of these two cells [18]. In this study, we developed SSRE, a novel method for cell type identification. It focuses on similarity learning, in which the cell-to-cell similarity is measured by considering more similar neighbors. SSRE computes the linear representation between cells based on sparse subspace theory, and thus generates a sparse representation of cell-to-cell similarity [19]. Moreover, motivated by the observations that each similarity measurement can represent data from a different aspect [16,20], SSRE incorporates three classical pairwise similarities into similarity learning. In order to reduce the effect of irrelevant features and improve the overall accuracy, SSRE designs a two-step procedure, *i.e.*, 1) adaptive gene selection and 2) similarity enhancement. The experimental results show that when combined with spectral clustering, the learned similarities by SSRE can reveal the block structure of scRNA-seq data reliably. Also, the experimental results on ten real scRNA-seq datasets and five simulated scRNA-seq datasets show that SSRE achieves higher accuracy of cell type detection in most cases than the compared popular approaches. Moreover, SSRE can be easily extended to other scRNA-seq tasks such as differential expression analysis and data visualization.

Method

Framework of SSRE

We introduce the overview of SSRE briefly. A schematic diagram of SSRE is shown in **Figure 1**, and detailed steps of SSRE are introduced later in this section. Given a scRNA-seq expression matrix, SSRE first removes genes whose expression levels are zero in all the cells. Then, the informative genes are selected based on the sparse subspace representation (SSR), Pearson

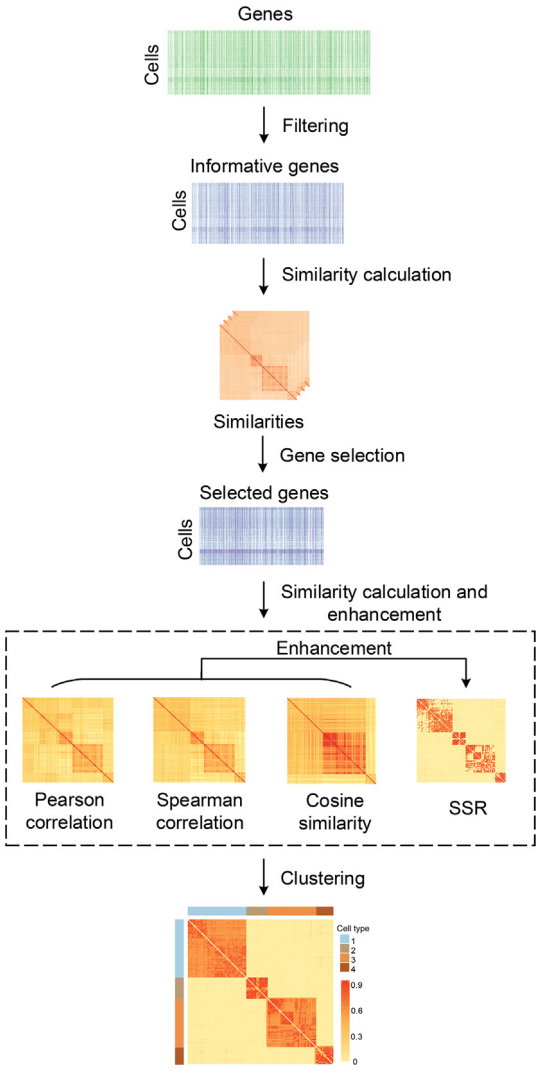

Figure 1 The schematic diagram of SSRE
SSRE consists of five main parts, including gene filtering, similarity calculation, gene selection, similarity enhancement, and clustering. The original input is a gene expression matrix. After filtering, four similarity measurements (Pearson correlation, Spearman correlation, cosine similarity, and SSR) are applied to select informative genes. The selected gene expression matrix is then used as input to the subsequent process for single-cell clustering. SSRE, single-cell clustering framework based on similarity learning; SSR, sparse subspace representation.

correlation, Spearman correlation, and cosine similarity. With the preprocessed gene expression matrix, SSRE learns SSR for each cell simultaneously. Then, SSRE derives an enhanced similarity matrix from the learned SSR similarity and the other three pairwise similarities. Finally, SSRE uses the enhanced similarity to identify cell types and visualize data.

Sparse subspace representation

The estimation of the similarity (or distance) matrix is a crucial step in clustering [8]. If the similarity matrix is well generated, it could be relatively easier to distinguish the cluster. In this study, we adopted sparse subspace theory [19] to compute the linear representation between cells and generate a sparse representation of the cell-to-cell similarity. Some subspace-based clustering methods have been successfully applied to computer vision field, and have been proved to be highly robust in corrupted data [21,22]. For scRNA-seq data, the sparse representation of cell-to-cell similarity is measured by considering the linear combination of similar neighbors. This tends to catch global structure information and generate more reliable similarity than traditional similarity measurement. The specific calculation processes are described as follows.

Mathematically, a scRNA-seq dataset with p genes and n cells can be denoted as $X = [x_1, x_2, \cdots, x_n] \in R^{p \times n}$, where $x_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]^T$ indicates the expression profiles of the p genes in cell i. Its linear representation coefficient matrix $C = [c_1, c_2, \cdots, c_n] \in R^{n \times n}$ satisfies the equation $X = XC$. According to the assumption that the expression of a cell can be represented by other cells in the same type, only the similarity of cells in the same cluster is non-zero. It also means that the coefficient matrix C is usually sparse. With the relaxed sparse constraint, the coefficient matrix C can be computed by solving an optimization problem as follows:

$$\begin{aligned} &min \frac{1}{2\lambda} \|X - XC\|_F^2 + \|C\|_1 \\ &s.t., diag(C) = 0 \end{aligned} \tag{1}$$

where $||\cdot||_F$ denotes the Frobenius norm which calculates the square root of sum of all squared elements, and constraint $diag(C) = 0$ prevents the cells from being represented by themselves, while λ is a penalty factor. An efficient approach to solve Equation (1) is the alternating direction method of multipliers (ADMM) [23]. We rewrite Equation (1) as follows:

$$\begin{aligned} &min \frac{1}{2\lambda} \|X - XZ\|_F^2 + \|C\|_1 \\ &s.t., Z - C = 0, diag(C) = 0 \end{aligned} \tag{2}$$

where Z is an auxiliary matrix. According to the model of ADMM, the augmented Lagrangian with auxiliary matrix Z and penalty parameter $(\gamma) > 0$ for the optimization Equation (2) is

$$\begin{aligned} \mathscr{L}_{\frac{1}{\gamma}}(Z, C, Y) = &\frac{1}{2\lambda} \|X - XZ\|_F^2 + \|C\|_1 + tr(Y^T(Z - C)) \\ &+ \frac{1}{2\gamma} \|C - Z\|^2 \end{aligned} \tag{3}$$

where Y is the dual variable. The derivation of its update can be found in section 1 of File S1. Matrix C is the target sparse representation matrix. To keep the symmetry and nonnegative nature of similarity matrix, the element of SSR is calculated as $sim_{sparse}(i,j) = |c_{ij}| + |c_{ji}|$.

Data preprocessing and gene selection

Before used to calculate SSR, the original data needs to be preprocessed. Various data preprocessing methods have been used in the previous studies, such as gene filtering [12,16], feature selection [24,25], and imputation [26,27]. In this study, we first removed genes with zero expression in all of cells and applied L_2-norm to each cell to eliminate the expression scale difference between different cells. Then, we computed the preliminary SSR with the normalized gene expression matrix, and adopted the Laplacian score [28] on SSR to assess the contribution that genes make to cell-to-cell similarity learning. According to the Laplacian scores, we selected significant genes for the following study. Genes with higher Laplacian scores are considered as more informative in distinguishing cell types [8]. Besides the SSR, we also considered three additional pairwise similarities, *i.e.*, Pearson correlation, Spearman correlation, and cosine similarity, to evaluate the importance of genes (denoted as $sim_{pearson}$, $sim_{spearman}$, and sim_{cosine}, respectively). For each similarity, we ranked genes in descending order by the Laplacian score and selected the top t genes as an important gene set that is denoted by G_1. The determination of the threshold t can be formulated as

$$\begin{aligned} &min\ var\ (LS_{G1}) + var\ (LS_{G2}) \\ &s.t.\ 0.1 \times p < |G_1| < 0.5 \times p \end{aligned} \tag{4}$$

where $G_1 = [g_1, g_2, \cdots, g_{t-1}]$ and $G_2 = [g_t, g_{t+1}, \cdots, g_p]$ denote two gene sets divided by t. The LS_{G1} and LS_{G2} are the Laplacian scores of genes in sets G_1 and G_2, respectively, and $|*|$ is the cardinality of a set. The $var(*)$ indicates variance of a set while p is the number of genes. Finally, we recomputed sim_{sparse}, $sim_{pearson}$, $sim_{spearman}$, and sim_{cosine} based on the intersection of four selected important gene sets. In the next section, we introduce an enhancement strategy to further improve the learned SSR sim_{sparse}.

Similarity enhancement

The SSR sim_{sparse} may suffer from the high-level technical noise in the data resulting in underestimation. Inspired by the consensus clustering and resource allocation, we further enhanced sim_{sparse} by integrating multiple pairwise similarities including $sim_{pearson}$, $sim_{spearman}$, and sim_{cosine}. These pairwise similarities partially reveal the local information between cells.

We imputed the missing values in sim_{sparse} according to their nearest neighbors' information. We firstly defined a target similarity matrix P as follows:

$$P(x_i, x_j) = \begin{cases} 1, x_j \in KNN(x_i) \\ 0, else \end{cases} \tag{5}$$

where $KNN(x_i)$ indicates the KNN of cell x_i. Then we marked the similarity $sim_{sparse}(x_i, x_j)$ between cells x_i and x_j as a missing value when it is zero in the sim_{sparse} but $P(x_i, x_j) = 1$ in at least one pairwise similarity matrix. Let $Isim_{sparse} = O^{n \times n}$ denotes the initial matrix to be imputed where n indicates the number of cells. For a marked missing value, the similarity

$Isim_{sparse}(x_i, x_j)$ was computed by the modified Weighted Adamic/Adar [29,30]. It was formulated as follows:

$$Isim_{sparse}(x_i, x_j) = \sum_{x_z \in CN(x_i, x_j)} \frac{sim_{sparse}(x_i, x_z) + sim_{sparse}(x_j, x_z)}{|\Gamma(x_z)|} \tag{6}$$

where $|\Gamma(x_z)|$ indicates the number of neighbors of cell x_z, and $CN(x_i, x_j)$ denotes the set of common neighbors of cell x_i and x_j. Note that the imputed similarity $Isim_{sparse}(x_i, x_j)$ is zero when $CN(x_i, x_j) = \varnothing$. At the end, an enhanced and more comprehensive SSR matrix $Esim_{sparse}$ was computed as $Esim_{sparse} = Isim_{sparse} + Isim_{sparse}{}^T + sim_{sparse}$.

Spectral clustering

Spectral clustering is a typical clustering technique that divides multiple objects into disjoint clusters depending on the spectrum of the similarity matrix [31]. Compared with the traditional clustering algorithms, spectral clustering is advantageous in model simplicity and robustness. In this study, we performed spectral clustering on the final enhanced SSR $Esim_{sparse}$. The inputs of spectral clustering are the cell-to-cell similarity matrix and the cluster number. The detailed introduction and analysis of spectral clustering could be found in previous studies [31,32].

Datasets

Datasets used in this study consist of two parts, real scRNA-seq datasets and simulated scRNA-seq datasets. We collected ten real scRNA-seq datasets that vary in terms of species, tissues, and biological processes, from public databases or published studies. The scale of these ten datasets varies from dozens to thousands, and the gene expression levels of them were computed by different units. The details of these real datasets are described in **Table 1**. Four datasets (*i.e.*, Treutlein [33], Deng [34], Ting [35], and Macosko [36] datasets) of these ten datasets were downloaded from the data subdirectory of MPSSC tool (https://github.com/ishspsy/project/tree/master/MPSSC). The Yan [37] and Goolam [38] datasets were collected from the popular single-cell consensus clustering (SC3) software package (https://github.com/hemberg-lab/SC3). The Song [39], Engel [40], and Haber [41] datasets were obtained via Gene Expression Omnibus [42] database (GEO: GSE85908, GSE74597, and GSE92332, respectively; https://www.ncbi.nlm.nih.gov/geo/), and the Vento [43] dataset was downloaded from ArrayExpress [44] (ArrayExpress: E-MTAB-6678; https://www.ebi.ac.uk/arrayexpress/). In addition, we used Splatter [45] to simulate five scRNA-seq datasets for more comprehensive analysis. They either have different size or different sparsity. We set *group.prob* to (0.65, 0.25, 0.1) for all simulated datasets, and changed the scale and sparsity by adjusting *nCells* and *dropout.mid*, respectively. The other parameters were set to default. The sample sizes of the five simulated datasets are 1000, 1000, 1000, 500, and 1500, and the corresponding sparsity is 0.61, 0.8, 0.94, 0.94, and 0.94, respectively.

scRNA-seq clustering methods

For performance comparison, we took the original SSR, native spectral clustering (SC), and eight state-of-the-art clustering methods (*i.e.*, SIMLR [8], MPSSC [20], Corr [11], SNN-Cliq [10], NMF [13], SC3 [16], dropClust [46], and Seurat [47]) as comparison. Among these methods, SIMLR, MPSSC, Corr, and SNN-Clip focus on similarity learning. Both SIMLR and MPSSC learn a representative similarity matrix from multi-Gaussian-kernels with different resolutions. Corr introduces a cell-pair differentiability correlation to relieve the effect of dropouts. SNN-Cliq applies the SNN to redefine the pairwise similarity. NMF detects the type of cells by projecting the high dimensional data into a latent space, in which each dimension of the latent space denotes a specific type. SC3 is a typical and powerful consensus clustering method. It obtains clusters by applying different upstream processes, and desires the final clusters to fit better. DropClust is a clustering algorithm designed for large-scale single-cell data, and it exploits an approximate nearest neighbor search technique to reduce the time complexity of analyzing large-scale data. Seurat, a popular R package for single-cell data analysis, obtains cell groups based on KNN-graph and Louvain clustering. Moreover, SC [32] with the Pearson correlation is considered as a baseline.

Metric of performance evaluation

We evaluated the proposed approach using two common metrics, *i.e.*, normalized mutual information (NMI) [48] and adjusted rand index (ARI) [49]. They have been widely used to assess clustering performance. Both NMI and ARI evaluate the consistency between the obtained clustering and pre-annotated labels, and have slightly different emphasis [50]. Given the real labels $L1$ and the clustering labels $L2$, NMI is calculate as

$$\mathrm{NMI}(L1, L2) = \frac{I(L1, L2)}{[H(L1) + H(L2)]/2} \tag{7}$$

$I(L1, L2)$ is the mutual information between $L1$ and $L2$, and H denotes entropy. For ARI, given $L1$ and $L2$, it is computed as

$$\mathrm{ARI}(L1, L2) = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_{ij}\binom{n_{ij}}{2}\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2} + \sum_j\binom{b_j}{2}\right] - [\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}} \tag{8}$$

where n_{ij} is the number of cells in both group $L1_i$ and group $L2_j$. The a_i and b_j denote the number of cells in group $L1_i$ and group $L2_j$, respectively.

Results and discussion

SSRE can greatly improve the clustering accuracy

In order to evaluate the performance of SSRE comprehensively, we first applied it on ten pre-annotated real scRNA-seq datasets and compared its performance with the original SSR, SC, and eight state-of-the-art clustering methods. See details in the Method section. Then, we tested all these methods on five simulated datasets for further comparison. In our experiments, for a fair comparison, we set the number of clusters to the

Table 1 The details of real scRNA-seq datasets used in this study

Dataset	No. of cells	No. of genes	No. of groups	Unit	Ref.
Treutlein	80	959	5	FPKM	[33]
Yan	90	20,214	7	RPKM	[37]
Deng	135	12,548	7	RPKM	[34]
Goolam	124	40,315	5	CPM	[38]
Ting	114	14,405	5	RPM	[35]
Song	214	27,473	4	TPM	[39]
Engel	203	23,337	4	TPM	[40]
Haber	1522	20,108	9	TPM	[41]
Vento	5418	33,693	38	HTSeq-count	[43]
Macosko	6418	12,822	39	UMI	[36]

Note: FPKM, fragments per kilobase of exon model per million mapped fragments; RPKM, reads per kilobase of exon model per million mapped reads; CPM, counts of exon model per million mapped reads; RPM, reads of exon model per million mapped reads; TPM, transcripts per kilobase of exon model per million mapped reads; UMI, unique molecular identifier.

Table 2 NMI values of all analyzed methods across ten real datasets

Method	Treutlein	Yan	Deng	Goolam	Ting	Song	Engel	Haber	Vento	Macosko
SC	0.71	0.69	0.63	0.72	0.89	0.51	0.71	0.40	0.70	0.80
SNN-Cliq	0.64	0.76	0.78	0.62	0.73	0.54	0.31	0.24	0.51	0.55
SIMLR	0.69	0.79	**0.84**	0.56	0.98	0.67	0.74	0.40	0.64	0.72
SC3	0.73	0.81	0.72	0.72	**1.00**	**0.73**	**0.81**	0.05	0.66	0.83
NMF	0.67	0.64	0.68	0.55	0.60	0.52	0.70	0.05	0.68	0.72
MPSSC	0.80	0.76	0.76	0.56	0.98	0.60	0.55	0.17	0.40	0.71
Corr	0.64	0.81	0.72	0.56	0.71	0.60	0.29	–	–	–
dropClust	**0.82**	0.76	0.73	0.81	0.91	0.61	0.29	0.43	0.67	0.71
Seurat	0.53	0.72	0.68	0.62	0.80	0.71	0.72	**0.62**	0.69	0.62
SSR	0.73	0.86	0.79	0.69	**1.00**	0.69	0.76	0.52	0.70	0.84
SSRE	**0.82**	**0.92**	0.81	**0.83**	**1.00**	**0.73**	0.77	0.53	**0.72**	**0.87**

Note: SC, native spectral clustering; SNN, shared nearest neighbor; SIMLR, single-cell interpretation via multikernel learning; SC3, single-cell consensus clustering; NMF, nonnegative matrix factorization; SSR, sparse subspace representation; SSRE, single-cell clustering framework based on similarity learning. "–" indicates unreachable. The bold value is the highest value in each column.

number of pre-annotated types for all methods except SNN-Cliq and Seurat because SNN-Cliq and Seurat do not need the number of clusters as input. The other parameters in all the methods were set to the default as described in the original papers. **Table 2** and **Table 3** summarize the NMI and ARI values of all methods on ten real scRNA-seq datasets, respectively. The results of Corr in large datasets are unreachable because of the high computational complexity. As shown in Table 2 and Table 3, the proposed method SSRE outperformed all other methods in most cases. SSRE achieved the best or tied first on seven datasets upon NMI and ARI. Meanwhile, SSRE ranked the second on three datasets based on NMI and two datasets based on ARI. It demonstrates that SSRE obtains more reliable results independent to the scale and the biological conditions of scRNA-seq data. Moreover, SSRE performed better than SSR on nine of the ten datasets in terms of NMI and ARI, which illustrates the effectiveness of the enhancement strategy in SSRE. Results of simulation experiment are shown in Tables S1 and S2. SSRE achieved the better performance overall, which shows the good stability of SSRE. SSRE is slightly time-consuming compared with some methods such as SC and dropClust, but its running time is still in a reasonable range. More detailed descriptions can be found in section 2 of File S1.

Estimating number of clusters is another key step in most clustering methods, which affects the accuracy of clustering method. SSRE performed eigengap [32] on the learned similarity matrix to estimate the number of clusters. Eigengap is a typical cluster number estimation method. It determines the number of clusters by calculating max gap between eigenvalues

Table 3 ARI values of all analyzed methods across ten real datasets

Method	Treutlein	Yan	Deng	Goolam	Ting	Song	Engel	Haber	Vento	Macosko
SC	0.59	0.44	0.33	0.54	0.89	0.49	0.67	0.19	0.37	0.52
SNN-Cliq	0.26	0.49	0.54	0.20	0.55	0.27	0.13	0.00	0.03	0.07
SIMLR	0.51	0.60	**0.67**	0.30	0.98	0.55	0.67	0.21	0.38	0.52
SC3	0.65	0.71	0.47	0.54	**1.00**	0.70	0.71	0.09	0.40	0.77
NMF	0.47	0.42	0.44	0.30	0.29	0.31	0.62	0.06	0.45	0.51
MPSSC	0.61	0.60	0.48	0.40	0.98	0.50	0.48	0.10	0.16	0.43
Corr	0.56	0.71	0.53	0.32	0.50	0.41	0.13	—	—	—
dropClust	**0.88**	0.62	0.46	0.59	0.89	0.58	0.24	0.24	0.45	0.45
Seurat	0.57	0.64	0.53	0.53	0.73	0.66	0.69	**0.43**	0.46	0.33
SSR	0.51	0.79	0.56	0.49	**1.00**	0.63	0.74	0.31	0.45	0.73
SSRE	0.62	**0.91**	0.65	**0.67**	**1.00**	**0.75**	**0.75**	0.32	**0.47**	**0.86**

Note: The bold value is the highest value in each column.

of a Laplacian matrix. To assess reliability of the estimation in different methods, we compared the estimated numbers with pre-annotated numbers. The results are summarized in Table S3. Besides SSRE and SSR, another four methods which also focus on similarity learning were selected for comparison. More experimental details can be seen in section 3 of File S1.

Analysis of parameter setting

In SSRE, four parameters are required to be set by users, *i.e.*, penalty coefficients λ and γ in solving SSR sim_{sparse}, gene selection threshold t, and the number of nearest neighbors k in similarity enhancement procedure. In this study, the selection of the threshold t was determined adaptively by solving Equation (4). The number of nearest neighbors k was set to $0.1 \times n$ (n is the number of cells) for small datasets with less than 5000 cells and set to 100 for other larger datasets. The other two parameters λ and γ in augmented Lagrangian (we used $1/\lambda$ and $1/\gamma$ in the coding implementation) were proportionally set as:

$$1/\gamma = \rho/\lambda, \rho = min_j\{max_i\{m_{ij}\}\} \tag{9}$$

where m_{ij} is the element of matrix $M = X^T X$. The m_{ij} is equivalent to the cosine similarity between cells x_i and x_j. This is same as previous work [19]. In our experiments, ρ/λ was set to a constant. For a given dataset, the larger value of ρ leads to the larger value of λ, which will result in the sparser matrix C. It means that the value of ρ can control the sparsity of matrix C adaptively in different datasets. Moreover, to validate the effect of penalty coefficient λ in clustering results, we tested SSRE with ρ/λ from 2 to 30 with the increment of 2 on all real datasets. We found that SSRE's performance was basically stable when ρ/λ is in the interval of 6 and 20. The results are shown in **Figure 2** and Figure S1. In our study, we set ρ/λ to 10 and $1/\lambda = \rho/\lambda$ as default for all datasets.

Application of SSRE in visualization

One of the most valuable aims in single-cell analysis is to identify new cell types or subtypes [6]. Visualization is an effective

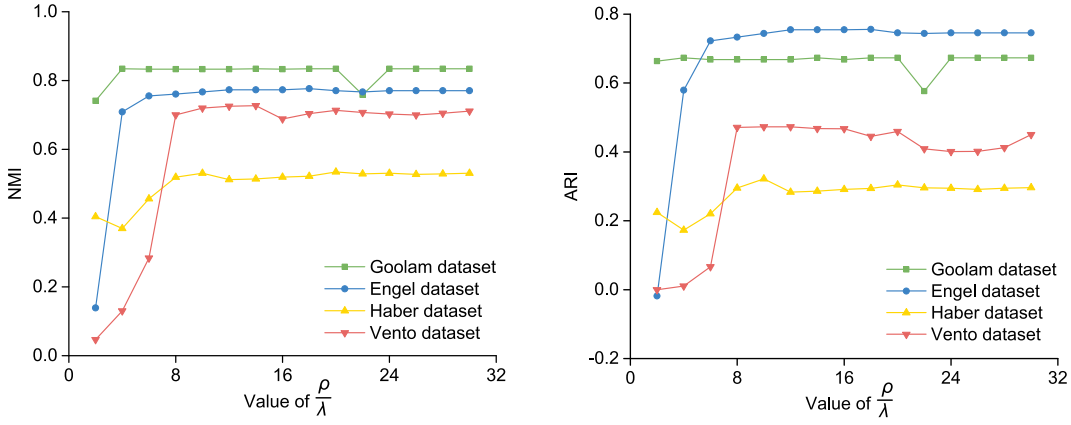

Figure 2 Clustering performance of SSRE with different parameter settings
The change of clustering performance of SSRE *versus* the value of parameter ρ/λ on four datasets (*i.e.*, Goolam dataset [38], Engel dataset [40], Haber dataset [41], and Vento dataset [43]) is shown here. The change of NMI values (**A**) and ARI values (**B**). NMI, normalized mutual information; ARI, adjusted rand index.

tool to intuitively display subgroups of all cells. The t-distributed stochastic neighbor embedding (t-SNE) [51] is one of the most popular visualization methods, and it has been proved to be powerful in scRNA-seq data. In our study, we performed a modified t-SNE on the similarities learned by different methods for visualization. We focused on two datasets, Goolam and Yan, and selected the native t-SNE, Corr, SIMLR, MPSSC, SSR, and SSRE for comparison. In Goolam dataset [38], cells were derived from mouse embryos in five differentiation stages: 2-cell, 4-cell, 8-cell, 16-cell, and 32-cell. The visualization results of Goolam dataset are shown in **Figure 3**A. As shown in Figure 3A, SSRE placed cells with the same type together and distinguished cells with different types clearly. And, although SIMLR can clearly distinguish groups from each other, some cells with the same type were separated. The second dataset Yan [37] was obtained from human

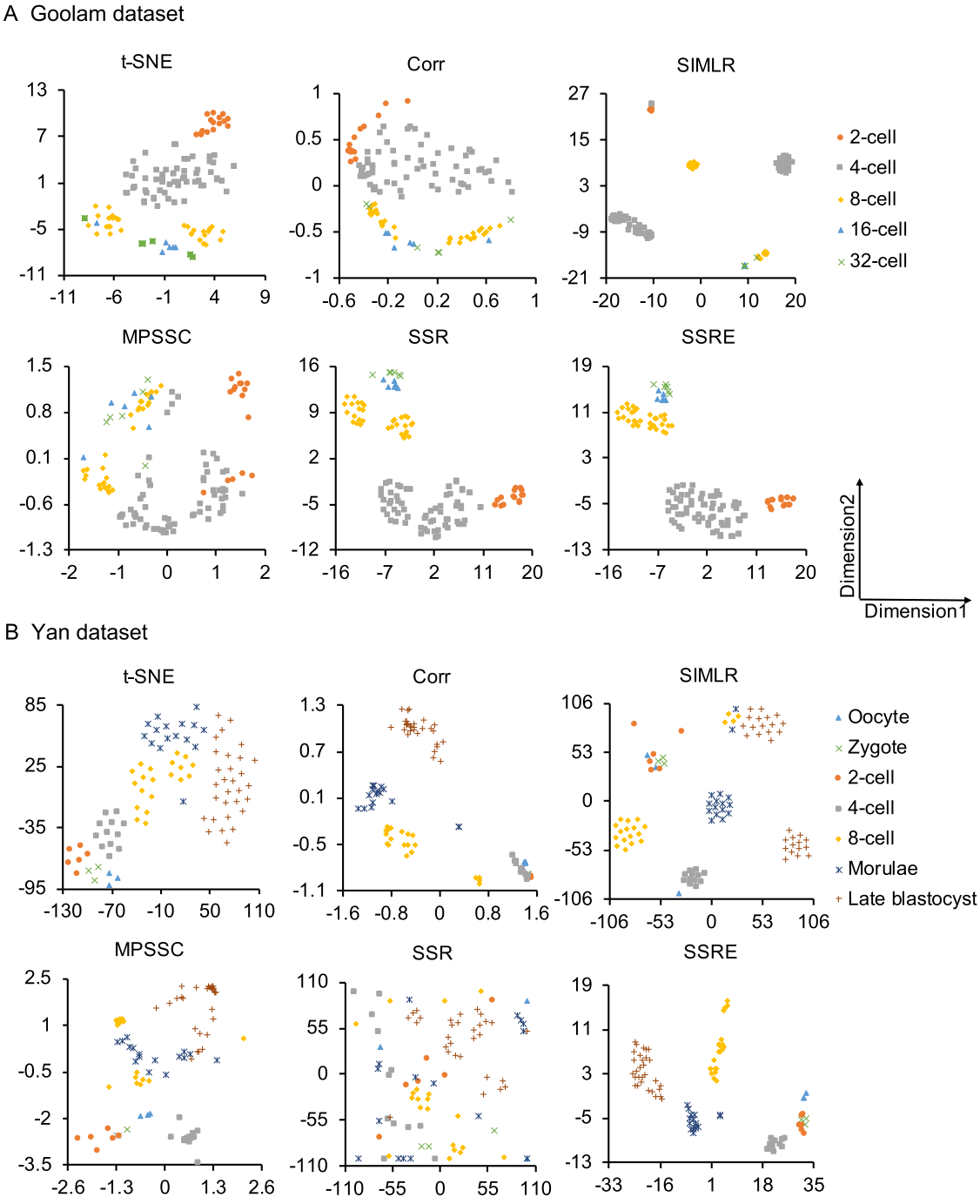

Figure 3 Visualization of Goolam and Yan datasets using different methods

Two datasets are visualized by t-SNE, Corr, SIMLR, MPSSC, SSR, and SSRE, respectively. **A.** The clustering results from Goolam dataset [38]. **B.** The clustering results from Yan dataset [37]. Each point in the figure represents a cell. Different colors and shapes indicate different cell types. t-SNE, t-distributed stochastic neighbor embedding; SIMLR, single-cell interpretation via multikernel learning.

pre-implantation embryos. It involves seven primary stages of preimplantation development: metaphase II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula, and late blastocyst. As shown in Figure 3B, Corr, SIMLR, and SSRE had a better overall performance than other methods. However, the four cell types, *i.e.*, oocyte, zygote, 2-cell, and 4-cell, were mixed totally in Corr, and mixed partially in SIMLR. Moreover, SIMLR also divided the cells with same type into different groups that were generally far away from each other. SSRE clusters cells more accurately, according to oocyte, 2-cell, and other cell types, than the competing methods.

Application of SSRE in identifying differentially expressed genes

The predicted clusters may potentially enable enhanced downstream scRNA-seq data analysis in biological sights. As a demonstration, we aimed to detect significantly differentially expressed genes (DEGs) based on the clustering results. Specifically, we applied the Kruskal-Wallis test [52] to the gene expression profiles with the inferred labels. The Kruskal-Wallis test, a non-parametric method, is often used for testing that if two or more groups are from the same distribution. We used the R function kruskal.test to perform the Kruskal-Wallis test. Then we detected DEGs according to the P value. The significant P value ($P < 0.01$) of a gene indicates that the gene's expression in at least one group stochastically dominates one other group. We took the Yan [37] dataset as an example to analyze the DEGs. The details of Yan have been introduced above. Figure S2 shows the heat map of gene expression of the top 50 most significantly DEGs identified. Notice that genes *NLRP11*, *NLRP4*, *CLEC10A*, *H1FOO*, *GDF9*, *OTX2*, *ACCSL*, *TUBB8*, and *TUBB4Q* have been reported in previous studies [37,53], which were also identified by SSRE. Genes *CLEC10A*, *H1FOO*, and *ACCSL* were reported as the markers of 1-cell stage cells (zygote) of human early embryos, while *NLRP11* and *TUBB4Q* are the markers of 4-cell stage cells [54]. Genes *GDF9* and *OTX2* are the markers of germ cell and primitive endoderm cell, respectively [55,56]. Genes *H1FOO* and *GDF9* were marked as the potential stage-specific genes in the oocyte and the blastomere of 4-cell stage embryos [57]. Certain *PRAMEF* family genes were reported as ones with transiently enhanced transcription activity in 8-cell stage. *MBD3L* family genes were identified as 8-cell stage-specific genes during the human embryo development in the previous studies [58,59]. All these are part of the top 50 significantly DEGs detected by SSRE.

Conclusion

Identifying cell types from single-cell transcriptome data is a meaningful but challengeable task because of the high-level noise and high dimension. The ideal identification of cell types enables more reliable characterizations of a biological process or phenomenon. Otherwise, it will introduce additional biases. Many approaches from different perspectives have been proposed recently, but the accuracy of cell type identification is still far from expectation. In this study, we presented SSRE, a similarity learning-based computational framework for cell type identification. Besides three classical pairwise similarities, SSRE computed the SSR of cells based on the subspace theory. Moreover, a gene selection process and an enhancement strategy were designed based on the characteristics of different similarities to learn more reliable similarities. SSRE greatly improved the clustering performance by appropriately combining multiple similarity measurements and adopting the embedding of sparse structure. The systematic performance evaluations on multiple scRNA-seq datasets showed that SSRE achieves superior performance among all competing methods. Furthermore, with the further downstream analyses, it is demonstrated that the learned similarity and inferred clusters can potentially be applied to more exploratory analyses, such as identifying gene markers and detecting new cell subtypes. In addition, for more flexible use, users can choose one or two of the three pairwise similarities mentioned in this study to perform gene selection and similarity enhancement procedures, and all three are used by default. Nonetheless, the proposed computational framework still can be improved in future study. One limitation of SSRE is relatively time-consuming in large-scale datasets; therefore, parallel computing is a possible strategy to accelerate the framework [60]. And more informative genes can be extracted or other biological information, such as gene functions [61] and gene regulatory relationships [62,63], can be incorporated to distinguish cell types. In addition, with the emergence of single-cell multi-omics data, it will be a possible trend to design corresponding multi-view clustering models to integrate the multi-omics data for cell type identification [64,65].

Code availability

The matlab and python implementations of SSRE are available at https://github.com/CSUBioGroup/SSRE.

CRediT author statement

Zhenlan Liang: Conceptualization, Methodology, Validation, Writing - original draft. **Min Li:** Supervision, Methodology, Writing - review & editing. **Ruiqing Zheng:** Methodology, Writing - original draft. **Yu Tian:** Data curation, Validation. **Xuhua Yan:** Software. **Jin Chen:** Writing - review & editing. **Fang-Xiang Wu:** Writing - review & editing. **Jianxin Wang:** Writing - review & editing. All authors read and approved the final manuscript.

Competing interests

The authors have declared that they have no competing interests.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (NSFC)-Zhejiang Joint Fund for the Integration of Industrialization and Information (Grant No. U1909208); the 111 Project, China (Grant No. B18059); the Hunan Provincial Science and Technology Program, China (Grant No. 2019CB1007); the Fundamental Research Funds for the Central Universities-Freedom Explore Program of Central South University, China (Grant No. 2019zzts592);

and the Natural Science Foundation, USA (Grant No. 1716340).

Supplementary materials

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2020.09.004.

ORCID

0000-0001-6556-071X (Zhenlan Liang)
0000-0002-0188-1394 (Min Li)
0000-0001-6372-6798 (Ruiqing Zheng)
0000-0002-3907-4651 (Yu Tian)
0000-0002-3183-3342 (Xuhua Yan)
0000-0001-6721-3199 (Jin Chen)
0000-0002-4593-9332 (Fang-Xiang Wu)
0000-0003-1516-0480 (Jianxin Wang)

References

[1] Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 2014;42:8845–60.
[2] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16:133–45.
[3] Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 2015;33:155–60.
[4] Guo G, Huss M, Tong GQ, Wang C, Sun LL, Clarke ND, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell 2010;18:675–85.
[5] Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 2018;18:35–45.
[6] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273–82.
[7] Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science 2002;297:1183–6.
[8] Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 2017;14:414–6.
[9] Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics 2004;20:2626–35.
[10] Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 2015;31:1974–80.
[11] Jiang H, Sohn L, Huang H, Chen L. Single cell clustering based on cell-pair differentiability correlation and variance analysis. Bioinformatics 2018;34:3684–94.
[12] Pouyan MB, Kostka D. Random forest based similarity learning for single cell RNA sequencing data. Bioinformatics 2018;34:i79–88.
[13] Shao C, Höfer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. Bioinformatics 2017;33:235–42.
[14] Zheng R, Li M, Liang Z, Wu FX, Pan Y, Wang J. SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. Bioinformatics 2019;35:3642–50.
[15] Zheng R, Liang Z, Chen X, Tian Y, Cao C, Li M. An adaptive sparse subspace clustering for cell type identification. Front Genet 2020;11:407.
[16] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14:483–6.
[17] Yang Y, Huh R, Culpepper HW, Lin Y, Love MI, Li Y. SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data. Bioinformatics 2019;35:1269–77.
[18] Lin P, Troup M, Ho JW. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol 2017;18:59.
[19] Elhamifar E, Vidal R. Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 2013;35:2765–81.
[20] Park S, Zhao H. Spectral clustering based on learning similarity matrix. Bioinformatics 2018;34:2069–76.
[21] Elhamifar E, Vidal R. Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 2013;35:2765–81.
[22] Vidal R, Favaro P. Low rank subspace clustering (LRSC). Pattern Recognit Lett 2014;43:47–61.
[23] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc; 2011.
[24] Feng Z, Wang Y. Elf: extract landmark features by optimizing topology maintenance, redundancy, and specificity. IEEE-ACM Trans Comput Biol Bioinform 2020;17:411–21.
[25] Feng Z, Ren X, Fang Y, Yin Y, Huang C, Zhao Y, et al. scTIM: Seeking Cell-Type-Indicative Marker from single cell RNA-seq data by consensus optimization. Bioinformatics 2020;36:2474–85.
[26] Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods 2018;15:539–42.
[27] Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr A, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;174:716–29.
[28] He X, Cai D, Niyogi P. Laplacian score for feature selection. Adv Neural Inf Process Syst 2005;18:507–14.
[29] Murata T, Moriyasu S. Link prediction of social networks based on weighted proximity measures. IEEE WIC ACM Int Conf Web Intell 2007:85–8.
[30] Pech R, Hao D, Cheng H, Zhou T. Enhancing subspace clustering based on dynamic prediction. Front Comput Sci 2019;13:802–12.
[31] Bach FR, Jordan MI. Learning spectral clustering. Adv Neural Inf Process Syst 2004;16:305–12.
[32] von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007;17:395–416.
[33] Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–5.
[34] Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 2014;343:193–6.
[35] Ting DT, Wittner BS, Ligorio M, Jordan NV, Shah AM, Miyamoto DT, et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep 2014;8:1905–18.
[36] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161:1202–14.
[37] Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 2013;20:1131–9.
[38] Goolam M, Scialdone A, Graham SJ, Macaulay IC, Jedrusik A, Hupalowska A, et al. Heterogeneity in *Oct4* and *Sox2* targets biases cell fate in 4-cell mouse embryos. Cell 2016;165:61–74.

[39] Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, et al. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Mol Cell 2017;67:148–61.
[40] Engel I, Seumois G, Chavez L, Samaniego-Castruita D, White B, Chawla A, et al. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. Nat Immunol 2016;17:728–39.
[41] Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, et al. A single-cell survey of the small intestinal epithelium. Nature 2017;551:333–9.
[42] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
[43] Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature 2018;563:347–53.
[44] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31:68–71.
[45] Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 2017;18:1–15.
[46] Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. dropClust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Res 2018;46:e36.
[47] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.
[48] Strehl A, Ghosh J. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 2002;3:583–617.
[49] Wagner S, Wagner D. Comparing clusterings: an overview. Karlsruhe: Universität Karlsruhe, Fakultät für Informatik; 2007, p. 1–19.
[50] Romano S, Vinh NX, Bailey J, Verspoor K. Adjusting for chance clustering comparison measures. J Mach Learn Res 2016;17:4635–66.
[51] Lvd M, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[52] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.
[53] Madissoon E, Töhönen V, Vesterlund L, Katayama S, Unneberg P, Inzunza J, et al. Differences in gene expression between mouse and human for dynamically regulated genes in early embryo. PLoS One 2014;9:e102949.
[54] Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 2013;500:593–7.
[55] Pennetier S, Uzbekova S, Perreau C, Papillier P, Mermillod P, Dalbiès-Tran R. Spatio-temporal expression of the germ cell marker genes *MATER*, *ZAR1*, *GDF9*, *BMP15*, and *VASA* in adult bovine tissues, oocytes, and preimplantation embryos. Biol Reprod 2004;71:1359–66.
[56] Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell 2016;165:1012–26.
[57] Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat Protoc 2010;5:516–35.
[58] Wang Y, Zhao C, Hou Z, Yang Y, Bi Y, Wang H, et al. Unique molecular events during reprogramming of human somatic cells to induced pluripotent stem cells (iPSCs) at naïve state. Elife 2018;7:e29518.
[59] Töhönen V, Katayama S, Vesterlund L, Sheikhi M, Antonsson L, Filippini-Cattaneo G, et al. Transcription activation of early human development suggests *DUX4* as an embryonic regulator. bioRxiv 2017:123208.
[60] Kumar S, Singh M. A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. Big Data Min Anal 2019;2:240–7.
[61] Li H-D, Xu Y, Zhu X, Liu Q, Omenn GS, Wang J. Clustermine: a knowledge-integrated clustering approach based on expression profiles of gene sets. J Bioinform Comput Biol 2020;18:2040009.
[62] Zheng R, Li M, Chen X, Zhao S, Wu F, Pan Y, et al. An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines. IEEE-ACM Trans Comput Biol Bioinform 2021;18:347–54.
[63] Aibar S, González-Blas CB, Moerman T, Imrichova H, Hulselmans G, Rambow F, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017;14:1083–6.
[64] Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci U S A 2018;115:7723–8.
[65] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 2019;177:1873–87.