Briefings in Bioinformatics, 00(0), 2021, 1–15

doi: 10.1093/bib/bbaa433
Articles

jSRC: a flexible and accurate joint learning algorithm for clustering of single-cell RNA-sequencing data

Wenming Wu, Zaiyi Liu, Xiaoke Ma

Corresponding author: Xiaoke Ma, School of Computer Science and Technology, Xidian University, Xi'an 710071, China.
Tel: +86 29 8820 2427; E-mail: xkma@xidian.edu.cn

Abstract

Single-cell RNA-sequencing (scRNA-seq) explores the transcriptome of genes at cell level, which sheds light on revealing the heterogeneity and dynamics of cell populations. Advances in biotechnologies make it possible to generate scRNA-seq profiles for large-scale cells, requiring effective and efficient clustering algorithms to identify cell types and informative genes. Although great efforts have been devoted to clustering of scRNA-seq, the accuracy, scalability and interpretability of available algorithms are not desirable. In this study, we solve these problems by developing a joint learning algorithm [a.k.a. joints sparse representation and clustering (jSRC)], where the dimension reduction (DR) and clustering are integrated. Specifically, DR is employed for the scalability and joint learning improves accuracy. To increase the interpretability of patterns, we assume that cells within the same type have similar expression patterns, where the sparse representation is imposed on features. We transform clustering of scRNA-seq into an optimization problem and then derive the update rules to optimize the objective of jSRC. Fifteen scRNA-seq datasets from various tissues and organisms are adopted to validate the performance of jSRC, where the number of single cells varies from 49 to 110 824. The experimental results demonstrate that jSRC significantly outperforms 12 state-of-the-art methods in terms of various measurements (on average 20.29% by improvement) with fewer running time. Furthermore, jSRC is efficient and robust across different scRNA-seq datasets from various tissues. Finally, jSRC also accurately identifies dynamic cell types associated with progression of COVID-19. The proposed model and methods provide an effective strategy to analyze scRNA-seq data (*the software is coded using MATLAB and is free for academic purposes; https://github.com/xkmaxidian/jSRC*).

Key words: single-cell RNA-seq; joint learning; sparse representation; dynamic cell clustering

Wenming Wu received the master degree in the School of Mathematics and Statistics, Henan University, China, in 2019. Currently, she is studying for a PhD degree at School of Computer Science and Technology in Xidian University. Her research interests include machine learning, data mining, bioinformatics and complex network analysis of bioinformatics.

Dr Liu received his MD degree from West China University of Medical Sciences in 1999, and received his PhD degree from Sichuan University in 2004, in Chengdu, Sichuan Province. Now, he is a Professor of Radiology and Supervisor of MD and chairman of the Department of Radiology of GDPH. He has been the PI of 10 scientific research grants, including the National Science Fund for Distinguished Young Scholar and the National Key Research and Development Program of China. Liu is the vice president of Youth Committee of Chinese Society of Radiology and vice president of Guangdong Society of Radiology.

Xiaoke Ma received his PhD degree in computer science from Xidian University in 2012. He was a postdoctor at the University of Iowa (USA) in 2012–2015. He is a full professor at the School of Computer Science and Technology, Xidian University (P.R. China). His research interests include machine learning, data mining and bioinformatics. He is an *ad hoc* reviewer for many international journals and publishes about 60 papers in the peer-reviewed international journals, such as IEEE Transactions Knowledge and Data Engineering, IEEE Transactions on Cybnetics, Pattern Recognition, Information Sciences, JSTAT, Physica A, Bioinformatics, PLoS Computational Biology, Nuclear Acids Research, IEEE Transactions Computational Biology and Bioinformatics and IEEE Transactions NanoBioScience.

Submitted: 29 September 2020; **Received (in revised form):** 21 December 2020

Introduction

Traditional RNA-sequencing technology exploits the gene expression by using mixed cells, masking the specificity of cells, whereas single-cell RNA-sequencing (scRNA-seq) overcomes the drawback by investigating the transcriptome of genes at cell level [1]. scRNA-seq transforms the paradigm of RNA-seq, which enables the possibility for the identification of existing, characterization of cells, classification of tumor sub-populations and investigation of cellular heterogeneity [2–4]. Clustering of scRNA-seq is a data-driven and unbiased strategy to classify cell types, shedding light on revealing the mechanisms of complex diseases and critical biological processes [5–7].

Clustering of scRNA-seq is highly nontrivial largely due to the noise, high-dimensionality and heterogeneity of data, and current algorithms aim at solving one or some of these issues [8–12]. On the high-dimensionality, scRNA-seq comprises the profiles of genome-wide genes, resulting in the so-called 'curse of dimensionality', where cell types are difficult to discriminate since similarity of cells is not reliable. Therefore, dimension reduction (DR) is the prerequisite for cell discovery [13]. Principle component analysis (PCA) is the most widely used approach for DR by projecting data into a low-dimensional space with an immediate purpose to discover the genes with the highest variance. There are also some variants of PCA. For example, SC3 [14] utilizes a subset of principle components, while PCAReduce [15] employs an iterative-based strategy to obtain the critical components by repeating PCA.

On the clustering of cells, various algorithms have been developed, where the difference lies in how to explore the features of cells and how to select clustering strategies. The most popular algorithm is *K*-means [3], which iteratively identifies the centers of clusters and assigns each cell to the nearest cluster. However, it saves the running time by sacrificing accuracy and robustness. To overcome the robustness problem, hierarchical clustering is also popular for the identification of cell types. For example, CIDR [16] adopts hierarchical clustering for scRNA-seq by imputing zeros during distance calculation, thereby improving the stability of estimation of cell distances in low-depth samples. To overcome the limitation of single algorithms, the ensemble learning, such as consensus clustering, is promising for clustering of scRNA-seq. Typical algorithms are SAME [1], SC3 [14] and SAFE [17], where the difference lies on the selection of algorithms and ensemble strategies. CellAssign [18] employs a probabilistic model to leverage prior knowledge of cell-type biomarker genes. Otherwise, many algorithms are developed, including DIMM-SC [19], SIMLR [20], SCANPY [10], SoptSC [21], CellBIC [22], BREM-SC [23], LCA [24] and CCSN [25].

These algorithms identify cell types by directly exploiting the expression profile of scRNA-seq, which are sensitive to the noise of data. Therefore, some algorithms aim to clustering cells by using the latent features in a latent space where cells are well separated. For example, SOUP [26] learns the clustering structure for pure and transitional cells via nonnegative matrix factorization (NMF), which leads to a better estimation of cell developmental trajectories. However, independence of DR and clustering fails to characterize patterns in data, resulting in unsatisfactory performance entirely. DRjCC [27] is the first joint learning algorithm for clustering of scRNA-seq, where DR and cell clustering are iteratively optimized. It improves the performance of cell clustering and accelerates the speed of convergence, which is one of the major motivations of this study.

However, DRjCC has several limitations. First, it fails to derive the implicit relation among cells and genes, particularly for the transitional cells, thereby resulting in the undesirable performance. Second, the interpretability of patterns obtained by DRjCC is lacking because it only exploits the sparsity of features. Third, it empirically selects the number of cell types. Therefore, there is a critical need for alternatives for DRjCC to further improve the performance of algorithms. Actually, the sparse representation (SR) learning addresses the interpretability of patterns by using few critical features [28, 29], which is successfully applied to image processing, data classification and other applications [29, 30]. Finally, current algorithms cannot automatically determine the number of cell types in scRNA-seq, requiring prior knowledge or empirically strategies.

In this study, we overcome these issues by proposing an accurate and flexible algorithm for clustering of scRNA-seq, which joints sparse representation and clustering (jSRC). To improve the scalability, the DR is employed as a component of algorithm. jSRC jointly learns the DR and clustering, where features are selected under the guidance of clustering of cells, thereby improving the accuracy of clustering. To improve the interpretability, the SR constraint is imposed on features based on the hypothesis that cells within the same type have the similar expression patterns on the biomarker genes. We formulate clustering of scRNA-seq into an optimization problem and design update rules to optimize the objective function, where the cell types and the number of cell types are automatically determined by the learned coefficient matrix. We apply jSRC to 15 scRNA-seq datasets with the number of single cells varies from 49 to 110 824 from various tissues and organisms. The experimental results demonstrate that jSRC significantly outperforms 12 state-of-the-art methods with 20.29% improvement in adjusted Rand index (ARI) on average. And jSRC automatically and precisely selects the number of cell types. Furthermore, the proposed algorithm also improves the interpretability and robustness. By using scRNA-seq of Coronavirus Disease-2019 (COVID-19), jSRC accurately identifies cell types and their dynamics associated with disease progression. jSRC reaches a good balance between the accuracy and running times since it takes less than 21 minutes for scRNA-seq with more than 11 000 cells. These result demonstrate that the proposed algorithm is a promising tool for scRNA-seq analysis.

Algorithm

In this section, we address the objective function, optimization, parameter selection and informative gene discovery of the proposed algorithm.

Objective function

The overview of jSRC is shown in Figure 1, which consists of three learning tasks, i.e.SR, DR and cell clustering. Thus, the objective function of jSRC is also composed of three corresponding costs.

Let capital, bold lower-case and lower-case letters be matrices, vectors and scalars, respectively. Given the expression levels of n cells measured on m genes, $X \in R^{m\times n}$ is the gene-by-cell normalized expression matrix. On SR learning, given a dictionary D, $\mathbf{x}$ is represented as a linear combination of columns of D with coefficient vector $\mathbf{a}$, which is formulated as

$$\min_{\mathbf{a}} \|\mathbf{x} - D\mathbf{a}\|^2 + \beta\|\mathbf{a}\|_1, \tag{1}$$

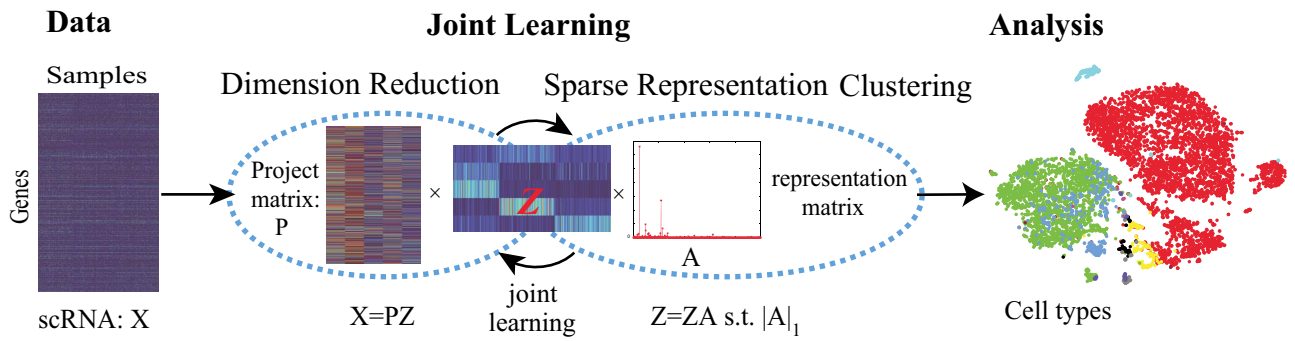

FIG. 1. Overview of the jSRC algorithm, which consists of SR learning, DR and cell clustering. DR identifies the critical features of cells under the guidance of clustering of cells, and SR learning is imposed on the features of cells to improve interpretability. And these three tasks are jointly learned to improve the accuracy. The cell types and informative genes are directly identified based on the learned features.

where $\|\mathbf{a}\|_1$ is l_1 norm to fulfill the sparse constraint, parameter β controls the relevant importance and $\|X\|$ is the Frobenius norm of X.

Moreover, we expect that cells belonging to the same cell type have similar expression patterns, where each cell can be represented by a few of other cells in the same type. SR is very suitable for recognizing the similarities and the relations of cells. The similarity, which is reflected by the value of SR coefficient, it can be used to cluster cells into different cell types. For each cell vector $\mathbf{x}_{.i}$, it represented by the dictionary of the all cells except $\mathbf{x}_{.i}$ itself.

$$\mathbf{x}_{.i} = a_1\mathbf{x}_{.1} + \cdots + a_{i-1}\mathbf{x}_{.i-1} + a_{i+1}\mathbf{x}_{.i+1} + \cdots + a_n\mathbf{x}_{.n}, \tag{2}$$

where $\mathbf{a} = [a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_n]$ is the representation coefficients of SR. Similar to [29] and [31], the cell $\mathbf{x}_{.i}$ will be represented as linearly as possible by the cells of the same type, different types of cells do not participate in the representation. The corresponding optimization problem can be written as

$$\min_{\mathbf{a}} \|\mathbf{x}_{.i} - X\mathbf{a}\|^2 + \beta\|\mathbf{a}\|_1, \text{s.t.} \quad \mathbf{a}_i = 0. \tag{3}$$

Notice that the second item in Equation 3 is similar to the LASSO regularization, which is also the L_1-norm constraint. However, LASSO regularization cannot guarantee the self-representation, resulting in trivial solution. To solve this problem, we impose constraint on $\mathbf{a}$, i.e. $\mathbf{a}_i$=0.

On the dimension reduction, the subspace learning is employed as

$$X \approx PZ, \tag{4}$$

where $P \in R^{m\times k}$ is the project matrix with k features and $Z \in R^{k\times n}$ is the feature matrix of cells in the projected space.

On cell clustering, we explore the intra-connection of cells, where a pair of cells with similar expression pattern is more likely to belong to the same type. Here, we simply consider the nearest neighbor of each cell $\mathbf{z}_{.i}$, i.e. $\mathbf{z}_{.j}$, where $j = arg\max_j |\mathbf{a}_{.i}|$. We assign them into the same cell type by minimizing the distance among them, i.e.

$$\begin{aligned} \mathcal{O}(Z) \quad &= \quad \sum_{i=1}^{n} \|\mathbf{z}_{.i} - \mathbf{z}_{.j}\|^2 \\ s.t. \quad & \quad j = arg\max_j |\mathbf{a}_{.i}|, \quad i \in \{1, \ldots, n\}. \end{aligned} \tag{5}$$

Therefore, combining Equations 3), 5 and 4, we reformulate the objective function of jSRC as

$$\begin{aligned} \min_{P,Z,A} \quad & \|X - PZ\|^2 + \|Z - ZA\|^2 \\ + \quad & \alpha\sum_{i=1}^{n} \|\mathbf{z}_{.i} - \mathbf{z}_{.j}\|^2 + \beta\|A\|_1 \\ s.t. \quad & j = arg\max_j |\mathbf{a}_{.i}|, diag(A) = 0, i \in \{1, \ldots, n\}. \end{aligned} \tag{6}$$

Optimization rules

Since Equation 6 is non-convex since it involves three variables P, Z and A, we optimize one variable by fixing the others until the termination criterion is reached. The Lagrange function L of Equation 6 is formulated as

$$\begin{aligned} L(P, Z, A) = & \|X - PZ\|^2 + \|Z - ZA\|^2 \\ & + \alpha\sum_{i=1}^{n} \|\mathbf{z}_{.i} - \mathbf{z}_{.j}\|^2 + \beta\|A\|_1. \end{aligned} \tag{7}$$

By fixing A and Z, Equation 7 is reformulated as minimization problem, i.e.

$$\min L(P) = \|X - PZ\|^2, \tag{8}$$

where the partial derivatives of $L(P)$ with respect to P is written as

$$\frac{\partial L}{\partial P} = -2(X - PZ)Z'. \tag{9}$$

According to KKT condition, by setting $\frac{\partial L(P)}{\partial P}$ =0, the update rule for P is obtained as

$$P \leftarrow P\frac{XZ'}{PZZ'}. \tag{10}$$

By fixing variable A and P, Equation 7 is reformulated as

$$\begin{aligned} L(Z) =& \|X - PZ\|^2 + \|Z - ZA\|^2 + \alpha \sum_{i=1}^{n} \|\mathbf{z}_{.i} - \mathbf{z}_{.j}\|^2 \\ =& \sum_{i=1}^{n} \|\mathbf{x}_{.i} - P\mathbf{z}_{.i}\|^2 + \sum_{i=1}^{n} \|\mathbf{z}_{.i} - Z\mathbf{a}_{.i}\|^2 \\ &+ \alpha \sum_{i=1}^{n} \|\mathbf{z}_{.i} - \mathbf{z}_{.j}\|^2. \end{aligned} \tag{11}$$

Notice that $L(Z)$ is differentiable with respect to $\mathbf{z}_{.i}$, i.e.

$$\frac{\partial L(Z)}{\partial \mathbf{z}_{.i}} = -P^{'}\mathbf{x}_{.i} + P^{'}P\mathbf{z}_{.i} + \mathbf{z}_{.i} - Z\mathbf{a}_{.i} + \alpha\mathbf{z}_{.i} - \alpha\mathbf{z}_{.j}. \tag{12}$$

Setting $\frac{\partial L(Z)}{\partial \mathbf{z}_{.i}}$=0, the update rule for $\mathbf{z}_{.i}$ is deduced as

$$\mathbf{z}_{.i} \leftarrow \mathbf{z}_{.i} \frac{P'\mathbf{x}_{.i} + Z\mathbf{a}_{.i} + \alpha\mathbf{z}_{.j}}{(P'P + I + \alpha I)\mathbf{z}_{.i}}, \tag{13}$$

where $Z = [\mathbf{z}_{.i}] = [\mathbf{z}_{.1}, \mathbf{z}_{.2}, \cdots, \mathbf{z}_{.i}, \cdots, \mathbf{z}_{.n}]$.

By fixing Z and P, let

$$\begin{aligned} J(\mathbf{a}_{.i}, \mathbf{b}_{.i}) =& \|\mathbf{z}_{.i} - Z\mathbf{a}_{.i}\|^2 + \beta\|\mathbf{b}_{.i}\|_1 \\ & s.t. \mathbf{a}_{.i} - \mathbf{b}_{.i} = 0. \end{aligned} \tag{14}$$

Equation 14 can be optimized by ADMM [32] by the augmented Lagrangian function as

$$\begin{aligned} L(\mathbf{a}_{.i}, \mathbf{b}_{.i}; \mathbf{t}_{.i}) =& \|\mathbf{z}_{.i} - Z\mathbf{a}_{.i}\|^2 + \beta\|\mathbf{b}_{.i}\|_1, \\ &+ \langle \mathbf{t}_{.i}, \mathbf{a}_{.i} - \mathbf{b}_{.i} \rangle + \sigma\|\mathbf{a}_{.i} - \mathbf{b}_{.i}\|^2, \end{aligned} \tag{15}$$

where $\sigma > 0$ be the penalty parameter and $\mathbf{t}_{.i} \in R^{n\times 1}$ is the Lagrange multiplier.

Equation 15 is solved using the following sub-problems:

$$\begin{cases} \mathbf{a}_{.i} \leftarrow \underset{\mathbf{a}_{.i}}{argmin} \|\mathbf{z}_{.i} - Z\mathbf{a}_{.i}\|^2 + \sigma\|\mathbf{a}_{.i} - \mathbf{b}_{.i} + \frac{\mathbf{t}_{.i}}{\sigma}\|^2, \\ \mathbf{b}_{.i} \leftarrow \underset{\mathbf{b}_{.i}}{argmin} \sigma\|\mathbf{a}_{.i} - \mathbf{b}_{.i} + \frac{\mathbf{t}_{.i}}{\sigma}\|^2 + \beta\|\mathbf{b}_{i}\|_1, \\ \mathbf{t}_{.i} \leftarrow \mathbf{t}_{.i} + \sigma(\mathbf{a}_{.i} - \mathbf{b}_{.i}). \end{cases}$$

According to ADMM, the update rules for sub-problem $\mathbf{a}_{.i}$ are formulated as

$$\mathbf{a}_{.i} \leftarrow \mathbf{a}_{.i} \frac{Z'\mathbf{z}_{.i} + \sigma\mathbf{b}_{.i} - T}{(Z'Z + \sigma II)\mathbf{a}_{.i}}, \tag{16}$$

and

$$P_{\Omega}(\mathbf{a}_{.i}) \in \{\mathbf{a}_{.i} \in R^n | \mathbf{a}_{ii} = 0, i = 1, 2, \cdots, n\}, \tag{17}$$

where $\Omega \in R^n$ is closed convex set to guarantee that $\mathbf{a}_{ii} = 0$.

Sub-problem $\mathbf{b}_{.i}$ has analytical solution as

$$\begin{aligned} \mathbf{b}_{.i} \leftarrow& \underset{\mathbf{b}_{.i}}{argmin} \sigma\|\mathbf{a}_{.i} - \mathbf{b}_{.i} + \frac{\mathbf{t}_{.i}}{\sigma}\|^2 + \beta\|\mathbf{b}_{i}\|_1 \\ =& S_{\beta/\sigma}\left(\mathbf{a}_{.i} + \frac{\mathbf{t}_{.i}}{\sigma}\right), \end{aligned} \tag{18}$$

where S is defined as

$$S_{\varepsilon}[x] = \begin{cases} x + \varepsilon, & if \quad x > \varepsilon, \\ x - \varepsilon, & if \quad x < -\varepsilon, \\ 0, & otherwise. \end{cases}$$

The optimization procedure is shown in Algorithm 1 and the convergence of jSRC is proven (Supplementary Materials).

Algorithm 1 The jSRC algorithm

Input: scRNA-seq data X, k.
Step 1. Initialize P, Z and A;
Step 2. Update the variable P according to Eq. (10);
Step 3. Update the variable Z according to Eq. (13);
Step 4. Update the variable $\mathbf{a}_{.i}$ according to Eqs. (16-17);
Step 5. Update the variable $\mathbf{b}_{.i}$ according to Eq. (18);
Step 6. Update $\mathbf{t}_{.i} \leftarrow \mathbf{t}_{.i} + \sigma(\mathbf{a}_{.i} - \mathbf{b}_{.i})$;
Step 7. Goto Step 2 until convergent.
Output: Clustering matrix R.

Cell type discovery

jSRC automatically obtains cell types from matrix A. Specifically, given representation $\mathbf{z}_{.i}$ of the i-th cell, the nearest cell $\mathbf{z}_{.j^*}$ is connected to it where $j^* = arg\max_j |\mathbf{a}_{.j}|$. In this case, the similarity network for cells is constructed. The connected components in the network correspond to the cell types and the number of connected components is the number of cell types.

Informative gene selection

Informative gene selection involves which genes have similar functions or co-expressed and how to extract them. Genes with similar expression patterns are clustered into modules to identify the functions of unknown genes or the unknown functions of genes. The genes in the same module tend to perform similar functions and participate in the same metabolic process or same cell pathway. In the scRNA-seq data, a network is constructed according to the SR-based gene clustering. Different modules from this network are identified, and the center genes of each module are selected for the informative genes.

Informative genes are defined as the representative ones within gene modules. jSRC first identifies gene modules using the projected matrix P, where the SR learning in Equation 3 is employed to obtain the coefficient matrix W for genes, i.e.

$$\min_{W} \|P - WP\|^2 + \beta\|W\|_1, \quad s.t. \quad diag(W) = 0, \tag{19}$$

where β is a parameter.

Table 1. Statistics of scRNA-seq datasets used for experiments, where #cells denotes the number of cells, #types corresponds the number of cell types and Organ is the tissues

Scale	Dataset	Species	#Cells	#Types	Organ	Platform	Ref.
Simulate	Splat1	Simulate	650	3	–	–	[35]
	Splat2	Simulate	2000	3	–	–	[35]
Moderate-scale	Biase	Mouse	49	3	Fetal brain	SMART-Seq SMARTer	[36]
	Ting	Mouse	187	7	Different tumous	scRNA-Seq, modified Tang2010 protocol	[37]
	Camp	Human	220	7	Fetal brain	SMART-Seq SMARTer	[40]
	Zheng	Human	500	3	Peripheral blood	10X Genomics	[8]
	Mouse1	Mouse	822	13	Pancreas	inDrop	[38]
	Mouse2	Mouse	1064	13	Pancreas	inDrop	[38]
	Human4	Human	1303	14	Pancreas	inDrop	[38]
	Zeisel	Mouse	3005	9	Mouse brain	Quantitative scRNA-seq with UMI	[4]
	Human3	Human	3605	14	Pancreas	inDrop	[38]
Large-scale	Birey	Human	11 838	14	Forebrain	Quantitative scRNA-seq with UMI	[41]
	COVID-19	Human	63 103	10	BALF	10x Genomics	[42]
	PBMC	Human	68 579	11	Peripheral blood	10X Genomics	[8]
	Tabula	Mouse	110 824	120	20 mouse organs	10X Genomics	[39]

Gene modules are identified by using hierarchical clustering based on matrix W. For each gene module, the eigenvalues of co-variance matrix of gene expression profiles are calculated, denoted by $\lambda_1 > \cdots > \lambda_m$. Informative genes in each module corresponds to the minimal value such that the contribution of top l eigenvalues is greater than a threshold δ, i.e. $l = \arg_t \sum_{i=1}^{t} \lambda_i / \sum_{i=1}^{m} \lambda_i \geq \delta$. We set δ=0.8 according to [15].

Parameter selection

jSRC involves three parameters k, α and β, where k is the number of features, α and β are regularization parameters.

Wu *et al.* [33] proposed the instability-based NMF model for the selection of k. Specifically, for each k, jSRC algorithm runs ι times and obtains ι basis matrices (denoted by $P_1, \cdots, P_\iota$). Given two matrices P_1 and P_2, matrix G is defined with the element g_{ij} as the cross correlation between the i-th column of matrix P_1 and the j-th column of matrix P_2. The dissimilarity between P_1 and P_2 is defined as

$$diss(P_1, P_2) = \frac{1}{2k}(2k - \sum_{j=1}^{k} max\mathbf{g}_j - \sum_{i=1}^{k} max\mathbf{g}_i), \tag{20}$$

where $\mathbf{g}_j$ denotes the j-th column of matrix G. The instability is the discrepancy of all the basis matrices for k, which is defined as

$$\Upsilon(k) = \frac{2}{\iota(\iota-1)} \sum_{1 \leq i \leq j \leq \iota} diss(P_1, P_2). \tag{21}$$

The k corresponding to the minimal $\Upsilon(k)$ is selected. We set α and β as the tuning parameters, which are selected empirically.

Materials

Performance evaluation

Given the predicted cluster labels c and the ground truth cluster labels c^*, ARI is defined as follows [34]:

$$ARI(c^*, c) = \frac{\sum_{e,t} \binom{n_{et}}{2} - \left[\sum_e \binom{n_e}{2} \sum_t \binom{n_t}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_e \binom{n_e}{2} + \sum_t \binom{n_t}{2}\right] - \left[\sum_e \binom{n_e}{2} \sum_t \binom{n_t}{2}\right] / \binom{n}{2}},$$

where n is the total number of single cells, n_e and n_t are the number of single cells in the predicted cluster e and in truth cluster t, respectively, and n_{et} is the number of single cells shared by e and t.

Accuracy (Acc) is also used to measure the performance of algorithms, which is defined as

$$Acc = \frac{1}{n} \sum_{i=1}^{n} \delta(c_i, c_i^*), \tag{22}$$

where c_i and c_i^* denote the predicted and truth label of i-th cell, respectively, and $\delta(x, y)$ is 1 if $x = y$, 0 otherwise.

Datasets

Fifteen scRNA-seq datasets are selected for experiments, including two simulated, nine moderate-scale and four large-scale ones. Specifically, the simulated datasets are Splat1 and Splat2 [35]. Moderate-scale datasets including five mouse scRNA-seq datasets, such as Biase [36], Ting [37], Zeisel [4], Mouse1 [38], Mouse2 [38] and four human datasets, including Zheng [8] for peripheral blood mononuclear cells, Camp [40] for fetal brain, Human3 and Human4 [38] for the pancreas. Large-scale datasets include Tabula [39] for mouse and Birey [41] for forebrain spheroids, PBMC [8] of peripheral blood mononuclear cells and COVID-19 [42] of bronchoalveolar lavage fluid (BALF) for human. The statistics of these benchmark datasets are summarized Table 1. Cells have different scales due to the sequencing depths and cell sizes, the transcript per million normalization is adopted as SOUP [26].

Results

jSRC automatically determines cell types and accelerate convergence

We first check convergence of jSRC by taking DR and SR learning in Equation 3 as baselines, where the relative error (max-min normalized error) is employed to measure convergence. How relative errors of algorithms as the number of iterations increases from 1 to 100 on the Zheng dataset is shown in Figure 2A, where jSRC is much faster than DR and SR. In detail, jSRC requires

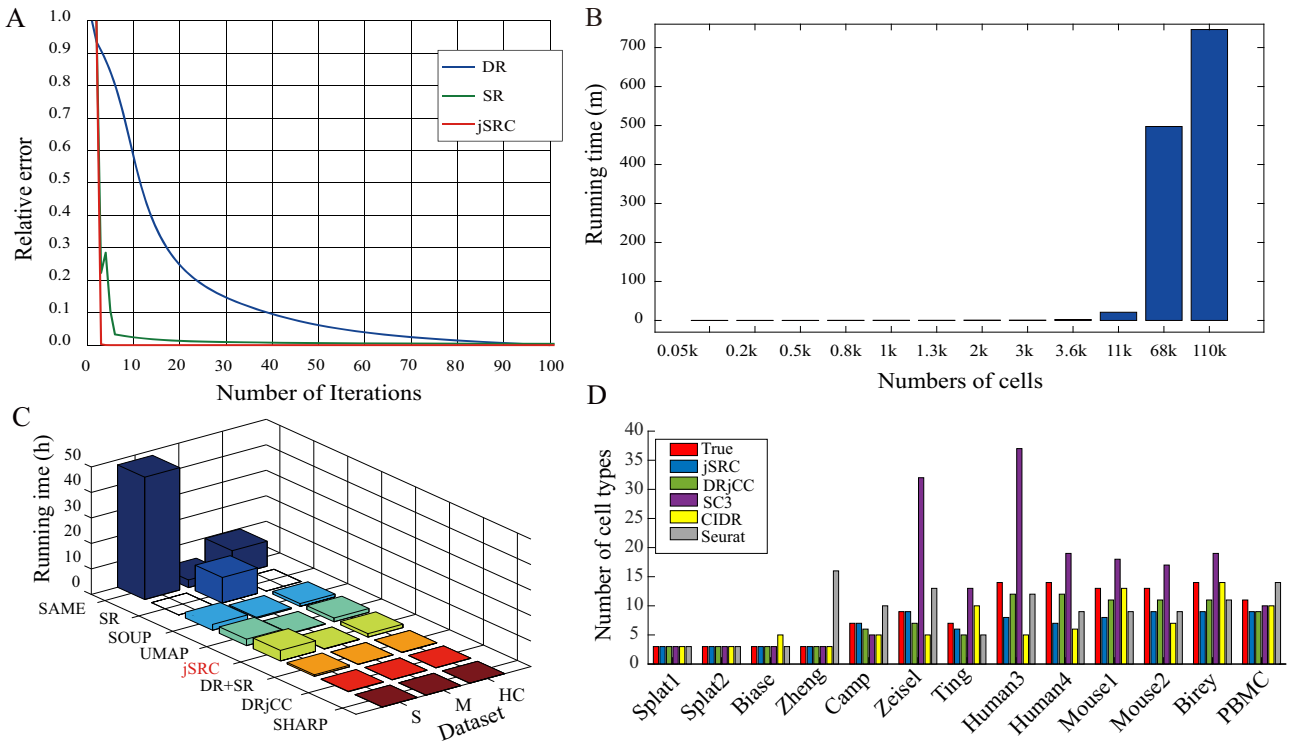

FIG. 2. Parameters selection: (**A**) the convergence of algorithms versus the number of iterations on the Zheng dataset, (**B**) scalability of jSRC on the scRNA-seq datasets with various number of cells, (**C**) running time of algorithms on the large-scale COVID-19 dataset and (**D**) performance of various algorithms on determining the number of cell types.

20 iterations to converge, whereas DR and SR take more than 80 iterations. Furthermore, the relative error of jSRC is much less than those DR and SR, implying that jSRC is more precise and faster than them. Then, we check the scalability of jSRC by investigating the running time on datasets with various number of cells, which is shown in Figure 2B. jSRC is efficient since it takes 21 minutes to clustering scRNA-seq with 11 000 cells. Then, we compare jSRC with state-of-the-art methods on the large-scale COVID-19 dataset as shown in Figure 2C, implying that jSRC is efficient for scRNA data. The convergence of jSRC is theoretically proved (Supplementary Materials).

How parameters α and β affect the accuracy of jSRC is investigated (Supplementary Figure S1A and B), and jSRC achieves the best performance on all these datasets at α=0.2 and β=0.6, where the SR learning and clustering reach a good trade-off. The number of features is determined by the instability of matrix factorization (Section Algorithm) as shown in Supplementary Figure S1C, where $k \geq 50$ is a good choice for larger-scale datasets, and 10 for small- and moderate-size datasets. Determining the number of clusters is challenging in clustering analysis. Vast majority of current algorithms empirically select it, whereas jSRC automatically learns the number of cell types from the coefficient matrix (Section Algorithm). The accuracy of various algorithms on the determination of cell types is also shown in Figure 2D and Supplementary Figure S2, where jSRC accurately predicts the number of cell types in six scRNA-seq datasets (two simulate scRNA-seq and four real biological scRNA-seq datasets).

jSRC improves the interpretability of patterns

jSRC jointly learns SR and clustering of cells, where features of cells are extracted under the guidance of clustering. It is natural to ask whether joint learning promotes the performance of feature selection in representation learning. To check the discriminative of features in the learned space, we compare the distributions of cells in the original and learned spaces by visualizing cells with t-SNE. The distributions of cells in Zheng dataset is shown in Figure 3A, where panel A1 is the scatter of cells in the original space and A2 in the learned space. Interestingly, cells in the learned space are more compact within the same types than those in the original space. Furthermore, cells belonging to various types are well separated in the learned space, whereas cells in the original space are mixed (rectangular box). To check whether jSRC is sensitive to dataset, we apply it to all datasets. The results are shown in Figure 4, demonstrating that jSRC is robust on the improvement of feature extraction.

jSRC expects that cells belonging to the same type are close to each in the projected space, which is the foundation for cell types (Algorithms). To quantify the closeness of cells, we calculate the similarity of cells in the original and learned space using the Euclidean distance on the Zheng dataset as shown in Figure 3B, where panel B1 is the similarity of cells in the original space and B2 in the learned space. The blue points in the diagonal blocks denote these cells that are correctly classified, whereas the red ones representing these are misclassified. The comparison of Figure 3B1 and B2 indicates that the number of misclassified cells in the learned is much less than that in the original space. The similar tendency occurs in other scRNA-seq dataset (Supplementary Figure S3).

Then, we ask why the mixed cells are well separated in the learned space (cells in the black box in Figure 3A1). The representation coefficient learned by jSRC indicates why cells are similar to each cell. For example, three representative mixed cells circled on Figure 3A1 are shown in Figure 3C, where each

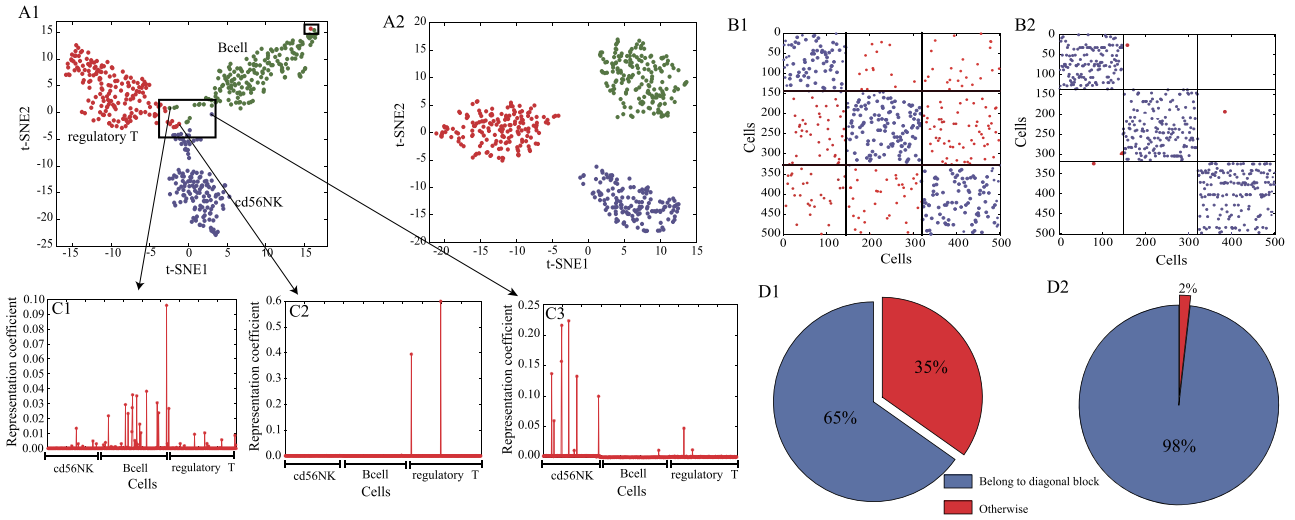

FIG. 3. jSRC improves interpretability of patterns: (**A**) visualization of cells using t-SNE representation in the original space (A1) and learned space by jSRC on the Zheng dataset (A2), where cell types are labeled with different colors; (**B**) cluster grid map on the Zheng dataset in the original space (B1) and learned space by jSRC (B2), respectively; (**C**) the schematic examples of mixed cells that are represented using few cells in the same type; and (**D**) the ratio is the percentage of cells that participate in the representation belong to the diagonal block on Zheng dataset, where (D1) is piechart in the original space and (D2) in the learned space.

of them can be represented by a few of cells within the same cell type (the height of bar denotes to the similarity between the mixed cell and the close cell in the same type). Then, we quantify the percentage of classified and misclassified cells. The ratios on the Zheng dataset is shown in Figure 3D, where panel D1 contains the ratios in the original space and D2 in the learned one. Specifically, 98% of the cells are correctly classified in the learned space, whereas only 65% of the cells are correctly classified in original space. The similar tendency repeats the other scRNA-seq datasets (Supplementary Figure S4).

Furthermore, jSRC also improves the interpretability of clustering of cells. The heatmap of coefficient matrix is shown in Supplementary Figure S5A, where each cell is represented by at most three cells within the same cell type. Supplementary Figure S5B–F shows the clustering result correlation heatmap of typical scRNA-seq cell clustering methods, including CIDR, *K*-means, SC3, Seurat and SOUP. The advantage of SR is that, for any cell, the strength and relation between a cell and the others are clearly quantified, whereas state-of-the-art methods fail to address these issue (Supplementary Figure S5). The interpretability of patterns obtained by jSRC provides clues for biologists for further studies.

jSRC significantly improves the accuracy of cell clustering

The above experiments prove that jSRC accelerates convergence, automatically determines the number of cell types and improves interpretability of patterns. Then, we check the performance of jSRC on the identification of cell types. Fifteen benchmark datasets are selected, including two artificial and thirteen biological datasets, which are summarized in Table 1. These datasets cover a wide spectrum of experimental technology, sequencing depth, tissues and heterogeneity of cells with the number of cells ranging from 49 to 110 824. Two common measurements are employed to fully characterize the performance of cell clustering, such as ARI and Acc (section performance evaluation).

Twelve state-of-the-art approaches are selected for a comparison to fully evaluate the performance of jSRC, including 2 DRs and 10 clustering methods. The former category includes UMAP [43] and DR, where UMAP is adopted because it outperforms t-SNE [44] and SCVIS [45]. In order to apply the DR algorithm to cell clustering, the *K*-means is selected as the post-processing technique. And the number of features for UMAP and DR are the same as jSRC. Clustering methods including *K*-means [3], SC3 [14], Seurat [9], CIDR [16], SOUP [26], SAME [1], SHARP [46], DRjCC [27], SR and dimension reduction+sparse representation (DR+SR). Since jSRC joints DR and SR, we independently execute two components (DR+SR) and SR alone. DRjCC is the first joint learning, and SC3, Seurat, *K*-means and CIDR are typical methods for scRNA-seq data. SC3, Seurat, CIDR and SAME automatically determine the number of cell types. For a fairer comparison, the others use the same number of cell types as jSRC.

We first benchmark jSRC using the two artificial scRNA-seq datasets [35], where Splat1 contains 650 cells with three cell types, and Splat2 has 2000 cells with three cell types. Performance of various algorithms on the simulated datasets in terms of ARI are shown in Figure 5A. These results indicate that jSRC cell clustering result outperforms other methods on simulated datasets. Specifically, ARIs of jSRC on Spalt1 and Splat2 are 0.677 and 0.924, respectively, whereas those of the best available methods are 0.545 and 0.894, respectively. jSRC is also superior to state-of-the-art in terms of accuracy (Figure 5B). These results prove that jSRC is more precise than the available methods on the simulated data.

Then, we ask whether jSRC is applicable to the moderate-scale real biological scRNA-seq datasets from various tissues of mouse and human (Table 1). Performance of various algorithms on the cells in terms of ARI is shown in Figure 5A, where jSRC has a similar performance with the best state-of-the-art methods on four datasets and significantly outperforms them on five datasets. Specifically, for the simple case, Zheng dataset [8] contains 500 human peripheral blood mononuclear cells (PBMCs), which consists of three cell types such as CD56+ NK cells, CD19+ B cells and CD4+/CD25+ regulatory T cells. SC3, SOUP, SAME, SHARP and jSRC outperform the others on the Zheng

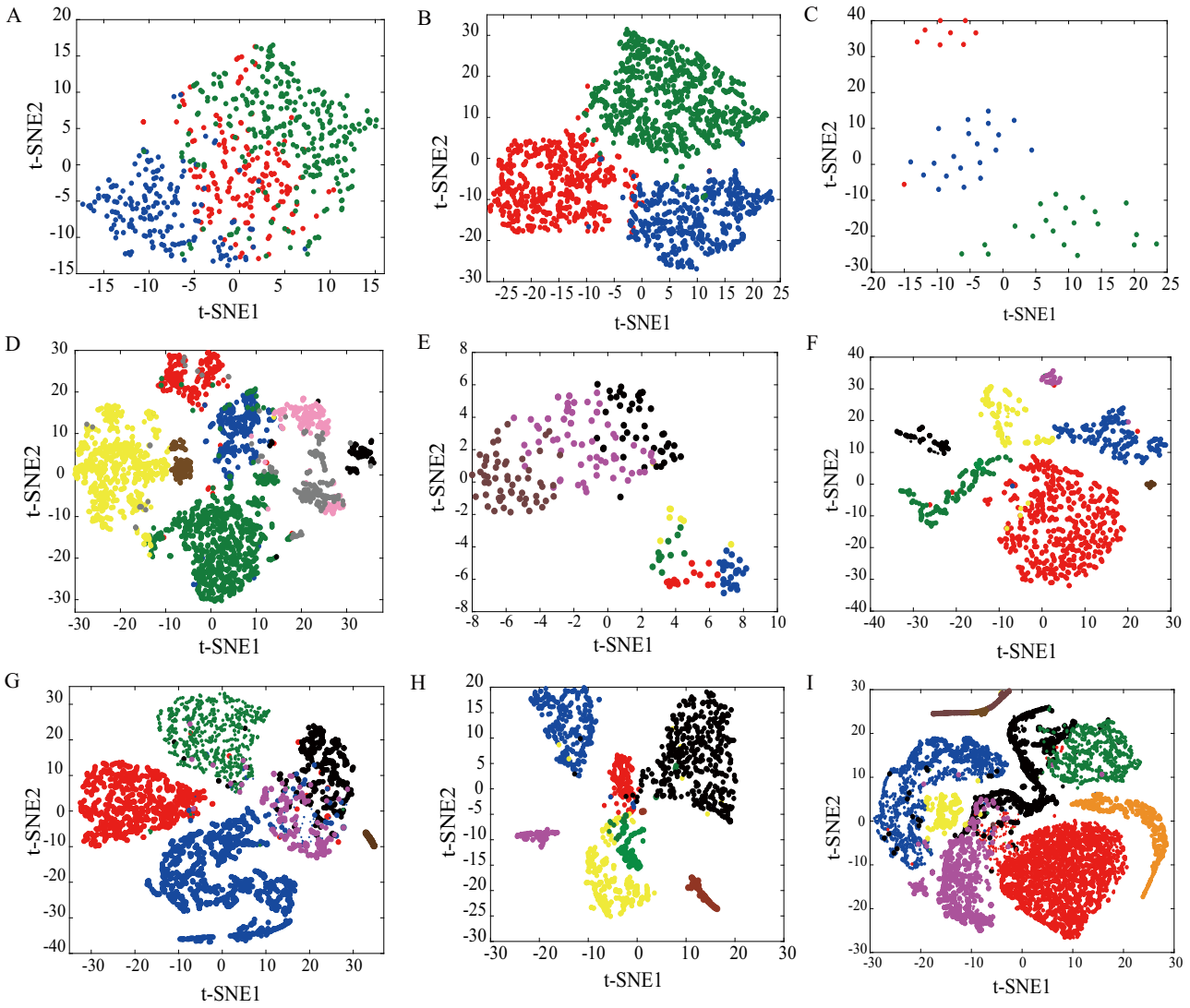

FIG. 4. Visualization of the cells using t-SNE in the learned space by jSRC on nine datasets: (**A**) Splat1, (**B**) Splat2, (**C**) Biase, (**D**) Zeisel, (**E**) Camp, (**F**) Mouse2, (**G**) Human3, (**H**) Human4 and (**I**) Birey.

dataset as shown in Figure 5. Moreover, they have a similar performance since the difference is subtle. The reason why these three algorithms achieve an excellent performance is that the mixture of various cell types in PBMCs is not heavy. For the challenging case, Camp dataset [40] comprises 220 fetal brain cells with 7 cell types such as apical progenitors (AP1, AP2), basal progenitors (BP1, BP2) and neurons (N1, N2, N3). Figure 5 indicates jSRC is much more precise than the others. In detail, ARI of jSRC on Camp dataset is 0.614, whereas that of SC3, CIDR, SOUP, DR+SR, SAME, SHARP and DRjCC are 0.502, 0.402, 0.525, 0.523, 5806, 0.5733 and 0.612, respectively. jSRC is superior to state-of-the-art methods in terms of Acc (Figure 5B), showing that the superiority of jSRC is not co-factored by the measurements. These results demonstrate that joint learning SR model is promising for discriminating mixture of cells in tissues with high heterogeneity.

Finally, we ask whether jSRC can handle large-scale scRNA-seq datasets using the Birey [41], PBMC[8] and Tabula [39] datasets, where the first dataset contains 11 838 cells with 14 cell types, PBMC contains 68 579 cells with 11 cell types and Tabula contains 110 824 cells with 120 cell types. Performance of various algorithms are shown in Figure 5, where jSRC has a similar performance with DR+SR and SHARP on Birey and PBMC datasets, and it outperforms the others on three datasets. Specifically, ARIs of jSRC are 0.830 (Birey), 0.687 (PBMC) and 0.781 (Tabula), whereas those of the best current methods are 0.838 (DR+SR) for Birey, 0.6998 (SHARP) for PBMC and 0.764 for Tabula. These results demonstrate that jSRC is accurate and efficient for the large-scale datasets.

Overall, across 14 datasets, jSRC on average improves ARI by 20.29% and Acc by 14.93% over state-of-the-art methods, and up to 1.49% and 0.65% over the best available method for all datasets. These results further demonstrate that joint SR adaptive clustering model is promising for cell type discovery. jSRC employs matrix factorization to learn features and cell types. Thus, it is natural to check the robustness of algorithms by using the standard deviation of accuracy with multiple runs (Figure 6), where the average standard deviation of jSRC is 0.02. These results demonstrate that joint learning also improves the robustness of algorithms, which is consistent with [47].

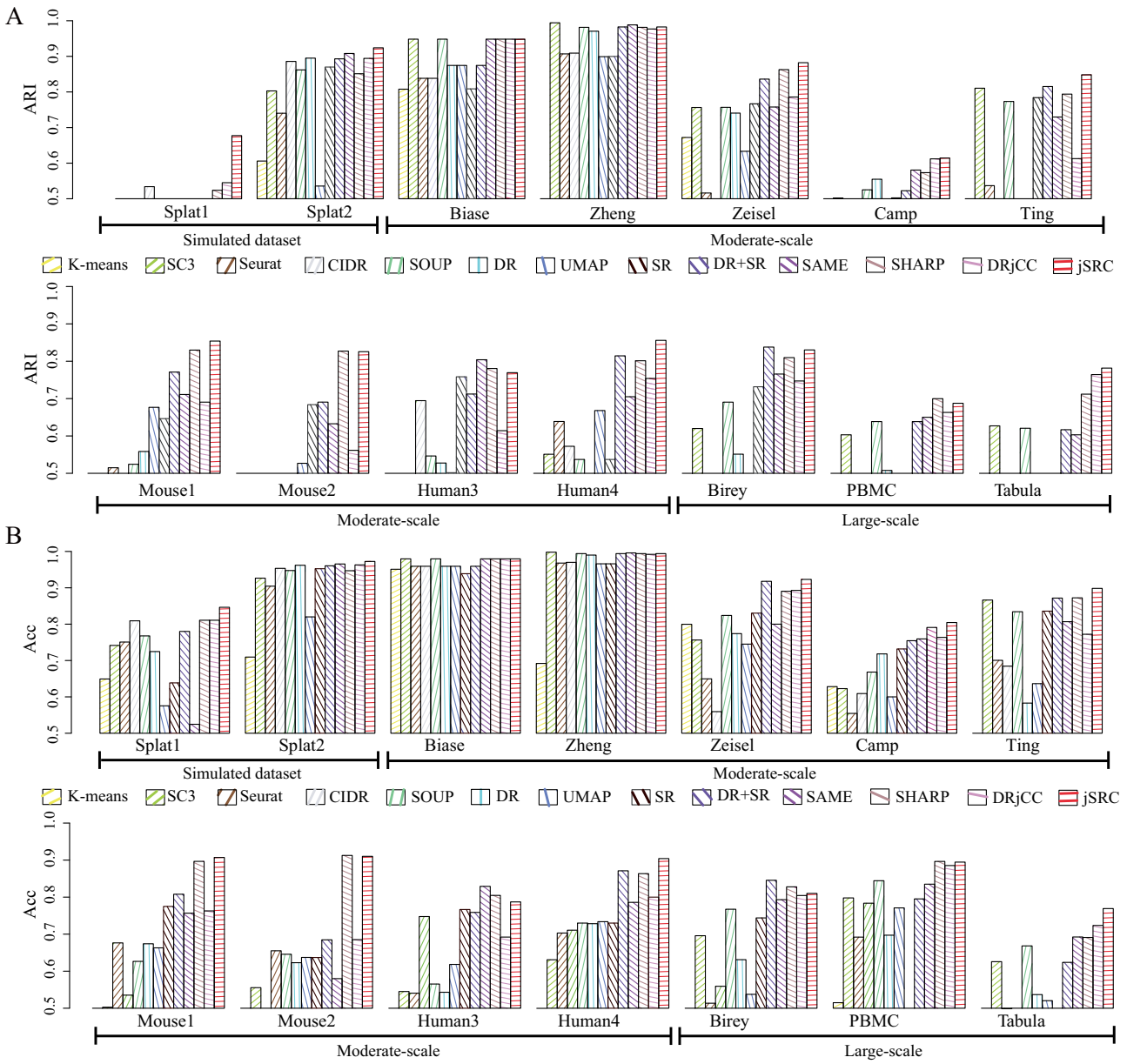

FIG. 5. jSRC significantly improves the accuracy of cell clustering. Performance of various algorithms on 14 scRNA-seq datasets in terms of various measurements: (**A**) ARI and (**B**) Acc, where these algorithms whose values are less than the minimum value of y-axis are missed.

jSRC accurately selects informative genes

Since characteristics and morphology of cells are tightly connected with the expression of genes, it is critically needed to extract informative genes. Here, we propose a module-based strategy to select informative genes (Section Algorithm).

Taking brain cells (Camp dataset) as an example, jSRC identifies seven modules (Figure 7A), where five significant modules are used to select the informative genes. To check whether these modules contain differentially expressed genes (DEGs), we calculate the percentage of DEGs in each module (Figure 7B), where the two nonsignificant modules have no DEG, and the percentage of others is at least 47%. jSRC selects 310 informative genes from the Camp dataset. The heatmap of the top 30 informative genes of gene expression is shown in Figure 7D, where expression profiles are differentially expressed across the top two cell types. The hierarchical clustering is performed on the informative genes as shown in Figure 7E, where these genes are classified into three groups.

Moreover, we find that 214 of the 310 (69.03%) informative genes obtained by jSRC are DEGs across different cell types, while those of DRjCC and SOUP are 59% and 57.50% (Figure 7C), respectively. Evidence shows that biomarker genes are associated with the survival time of patients [48]. Therefore, we assume that informative genes can also serve as biomarkers to predict patients' survival time. By using the gene expression profiles and clinical information of gliomas from TCGA, we use the Kaplan–Meier survival analysis to identify the informative genes that are significantly associated with survival time of patients with a cutoff 0.05. We find that 22 of 30 informative genes (73.3%) with highest deviation are significantly associated with the survival time of patients, while that of DRjCC and SOUP are 53.85% and

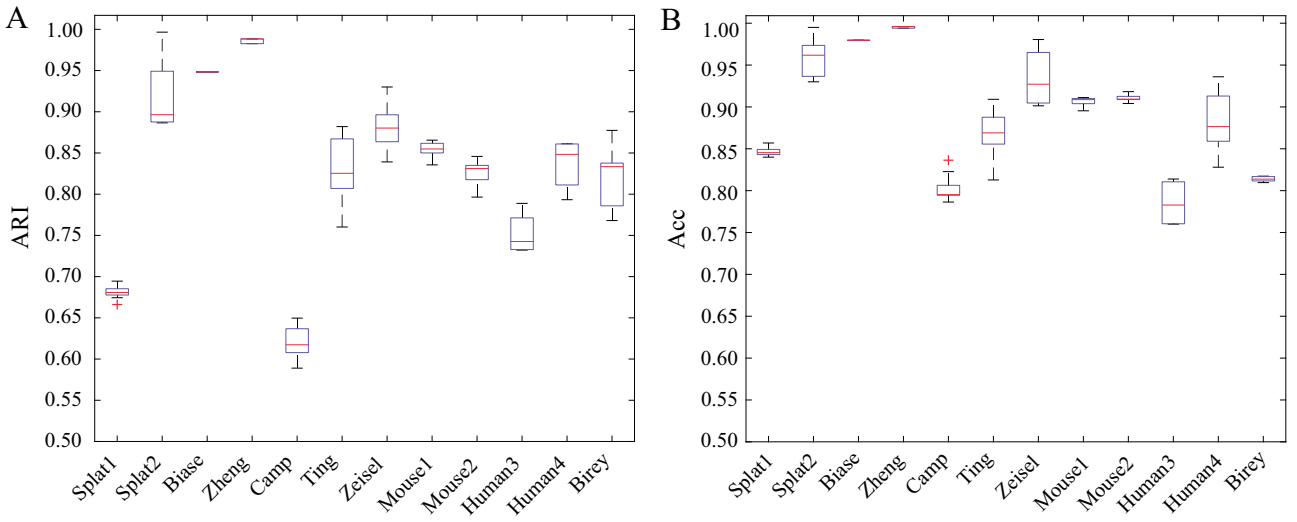

FIG. 6. The distribution of accuracy obtained by jSRC with multiple runs on 12 datasets in terms of various measurements: (**A**) ARI and (**B**) Acc.

53.33%, respectively. For example, *DMD* separates the patients into high and low/mediate expression groups according to gene expression level, where the survival time of these two groups significantly differs with $P = 1.4E\text{-}4$ (Figure 7F). *JPH4* is also significantly associated with the survival time of patient with $P = 2.0E\text{-}3$ (Figure 7G).

To further investigate the biological functions of informative genes, the enrichment analysis is performed using software Metascape [49]. They are significantly enriched by brain cancer related biological processes, such as cell division (P= 1.1E-19, hypergeometric test), brain development ($P = 4.3E\text{-}8$, hypergeometric test) and glial cell differentiation ($P = 1.6E\text{-}9$, hypergeometric test) (Supplementary Figure S6). The network of these enriched biological processes is constructed (Supplementary Figure S7), where each node represents an enriched term (colored by cluster ID), and sizes of nodes are proportional to the P-value of significance. Finally, we ask whether the informative genes are highly associated with brain cell tumors. We find out eight of them are brain-causing genes as shown in Table 2, including ZFHX4 [50], NUF2 [51], PTTG1 [52], DMD [53], CDC27 [54], HMGB2 [55], SATB1 [56] and SCHIP1 [57]. These results may provide biologists with clues for revealing mechanisms of brain tumors.

Then, we apply jSRC to the pancreatic scRNA-seq data and the heatmap of informative genes demonstrates the expression levels are significantly differentially expressed across cell types (Supplementary Figure S8A). The percentage of DEGs in informative genes is 90%, 90% and 88.33% for jSRC, DRjCC and SOUP, respectively (Supplementary Figure S8B). The percentages of specific informative genes significantly associated with patients' survival time is 53.33% for jSRC (Supplementary Figure S8C and D). We also find that five informative genes are disease-causing genes for pancreatic cancer as shown in Table 2, such as CA12 [58], MICB [59], MICA [60, 61], CEL [62–64], IL22RA1 [65]. The enrichment analysis indicates the informative genes are significantly enriched by fatty acid metabolism ($P = 6.9E\text{-}22$, hypergeometric test) and adenylate cyclase-activating G protein-coupled receptor signaling pathway ($P = 7.1E\text{-}16$, hypergeometric test) (Supplementary Figures S9 and S10).

jSRC precisely identifies dynamic cell types associated with progression of COVID-19

Since the emergence of COVID-19 in December 2019, it has been becoming one of the prominent research directions. Immune characteristics associated with COVID-19 severity are currently unclear. Liao *et al.* [42] characterize the BALF immune cells from patients with varying severity of COVID-19 and healthy controls (HCs) using scRNA-seq, where 63 103 cells are selected, including HCs were 19 173, moderate (M) were 7311 and severe (S) were 36 619.

We apply jSRC to scRNA-seq of BALF with an immediate purpose to validate whether it can accurately identify the dynamic cell types associated with progression of COVID-19. The tSNE presentation of cells at various stages are shown in Figure 8, where panel A1 is for the mixed cells in all stages, A2 for HC, A3 for M and A4 for S stage, respectively. In all, jSRC extracts 10 major cell types in BALF, including macrophages, myeloid dendritic cells (mDCs), T cells, natural killer (NK) cells, B cells, epithelial cells, plasmacytoid dendritic cells (pDCs), plasma cells, mast cells and neutrophils. These panels in Figure 8A show that cell types differ greatly. Macrophage cells dominates in the HCs, whereas macrophage and T cells are two major types in the moderate stage and all 10 cell types occur in the severe stage. Specifically, patients with moderate COVID-19 infection has a higher proportion of T, NK, B, epithelial, pDC cells and a lower proportion of macrophages than those of HCs. Patients with severe/critical infection have a much higher proportion of macrophages, epithelial, plasmas and neutrophil cell and a lower proportion of mDCs, T, NK, B and pDCs cells than those with moderate infection. Consistent with previous studies [42, 66, 67], jSRC suggests that cytokine storm is associated with the severity of COVID-19.

The accuracy of various algorithms for clustering of cells in terms of different measurements is shown in Figure 8B, where panel B1 is for ARI and B2 is for accuracy, respectively. Figure 8B1 shows that jSRC outperforms state-of-the-art on all three stages of COVID-19, where ARI is 0.786 (HC), 0.907 (M) and 0.857 (S), respectively. jSRC also achieves the best performance in terms of accuracy (Figure 8B2), where accuracy is 0.966 (HC), 0.942 (M) and

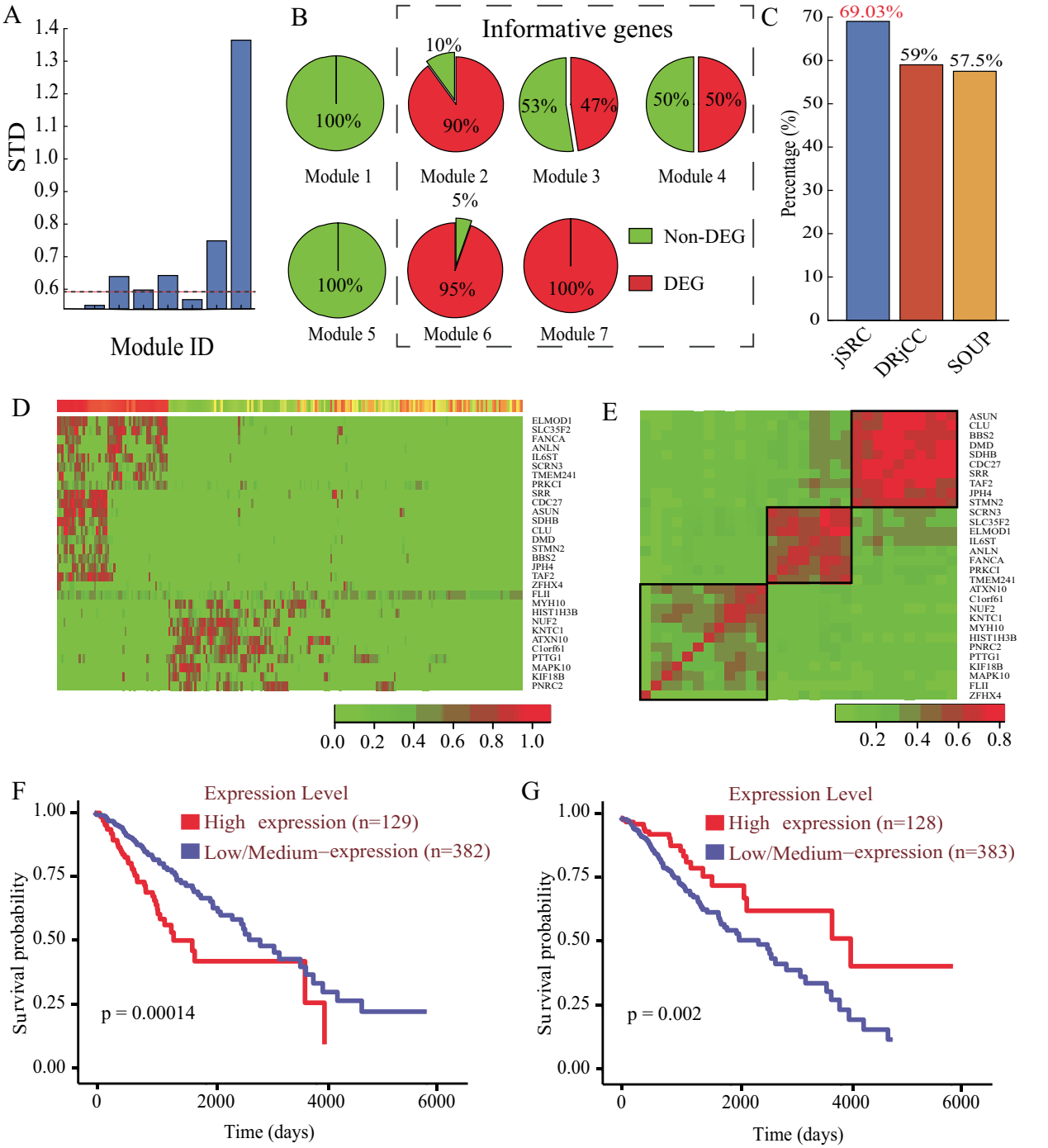

FIG. 7. Analysis of informative genes selected by jSRC: (**A**) the standard deviation (STD) of gene modules obtained by jSRC; (**B**) percentage of DEGs in each module, where the informative genes selected by jSRC in the dotted box; (**C**) percentage of DEGs of informative genes obtained various algorithms; (**D**) heatmap of the expression profile of 30 informative genes selected by jSRC on the Camp dataset, where columns are ordered according to the true cell types; (**E**) clustering of informative genes using hierarchical clustering with correlation; and (**F–G**) Kaplan–Meier survival analysis of informative gene DMD (**F**) and JPH4 (**G**).

0.936 (S), respectively. The similarity of cells using the Euclidean distance of features obtained by jSRC shows the number of cell types and clusters of cells at all stages (Supplementary Figure S11). These results demonstrate that jSRC can precisely capture the characteristics of scRNA-seq of COVID-19, implying the superiority of joint learning.

jSRC identifies 800 informative genes, where 77.03% informative genes are significantly DEGs in HC, while those for the moderate and severe are 94.92% and 98.98% (Supplementary Figure S12). The GO functions of the informative genes are investigated using the enrichment analysis, and they are significantly enriched by the leukocyte activation involved in immune response (P = 8.9E-8, hypergeometric test), T-cell activation (P = 4.5E-7, hypergeometric test) and apoptotic signaling pathway (P = 2.2E-5, hypergeometric test) (Supplementary Figure S13).

Table 2. The biomarker genes obtained by jSRC are highly associated with corresponding tumors

Organ	Gene	Description
	ZFHX4	Regulated the glioblastoma tumor-initiating cell state
	NUF2	Overexpressed in glioma tissues and differentially expressed in a series of glioma cell lines
	PTTG1	PTTG1 axis promotes the proliferation of glioma through stabilizing E2F1
	DMD	Decreased Dp71 expression is associated with cancer proliferation and poor prognosis in glioblastoma, where Dp71 is DMD gene product in the nervous system
Brain	CDC27	mir-218-2 promotes glioblastomas growth, invasion and drug resistance by targeting CDC27
	HMGB2	The oncogene HMGB2 as a downstream target of miR-130a by using luciferase and western blot assays
	SATB1	SATB1 may represent a promising target molecule in glioblastoma therapy
	SCHIP1	Homozygous nonsense mutation in SCHIP1 causes a neurodevelopmental brain malformation syndrome
	CA12	Knockdown of AE2 and of CA12 inhibited pancreatic and salivary gland ductal AE2 activity and fluid secretion, explain the disease
	MICB	MICB shedding in PANC-1 pancreatic cancer cells
	MICA	Pancreatic tumor cells may avoid immune surveillance by releasing the transmembrane MICA protein in soluble form (s-MICA)
Pancreatic	CEL	Rare mutations in the CEL gene cause a syndrome of pancreatic exocrine and endocrine dysfunction, whereas a recombined CEL allele increases the risk for chronic pancreatitis
	IL22RA1	IL22RA1/STAT3 signaling enhances stemness and tumorigenicity in pancreatic cancer

Finally, we tracks the dynamics of cell types by counting the number of, and percentage of, cells in each type obtained by jSRC across various stages as shown in Figure 8C (Supplementary Figure S14). For example, the number of macrophages cells is 17 907 (HC), 4345 (M) and 26 525 (S), respectively. And the percentage of macrophages in HC (93.4%) is much higher than that in stage M (59.4%) and S (72.4%). The number of T cells increases from HC (643) to stage S (4356), while the percentage of T cell reaches the peak at stage M. Interestingly, the number of neutrophil cells is zero in HC and M and dramatically soars up to 1063 in stage S. These dynamic patterns of cell types may serve as potential biomarkers for the diagnosis and therapy of COVID-19.

Discussion

scRNA-seq enables the possibility to explore the expression of genomic at the single cell level, and the accumulated scRNA-seq data provides a great opportunity to reveal the underlying mechanisms of diseases. It also poses a great challenge on designing effective and efficient algorithms for scRNA-seq data. Although great efforts have been devoted to this issue, there are many unsolved problems, such as complexity, accuracy and interpretability.

To solve these problem, we propose an accurate and flexible algorithm jSRC for scRNA-seq data, where DR and SR are simultaneously learned. It leverages information from cells and genes, where DR extracts features of genes under the guidance of cell clustering. We demonstrate that the proposed method improves the performance of visualization through DR (Figures 3A and 4). We benchmark jSRC along with 12 state-of-the-art methods on 15 scRNA-seq datasets, where jSRC significantly outperforms them in terms of various measurements (Figure 5). jSRC not only improves the robustness and accuracy of algorithms but also enhances the interpretability of patterns (Figure 3 and Supplementary Figure S4). Furthermore, the number of cell type parameters of jSRC are automatically determined, which are more accurate than others (Figure 2D). jSRC is also precise to select the informative genes associated with cell types (Figure 7). Finally, we also testify jSRC using scRNA-seq data of COVID-19. The experimental results demonstrate that jSRC precisely identifies dynamic cell types in each stage of COVID-19 (Figure 8A and B), and it also extracts the dynamic patterns of cells associated with disease progression (Figure 8C).

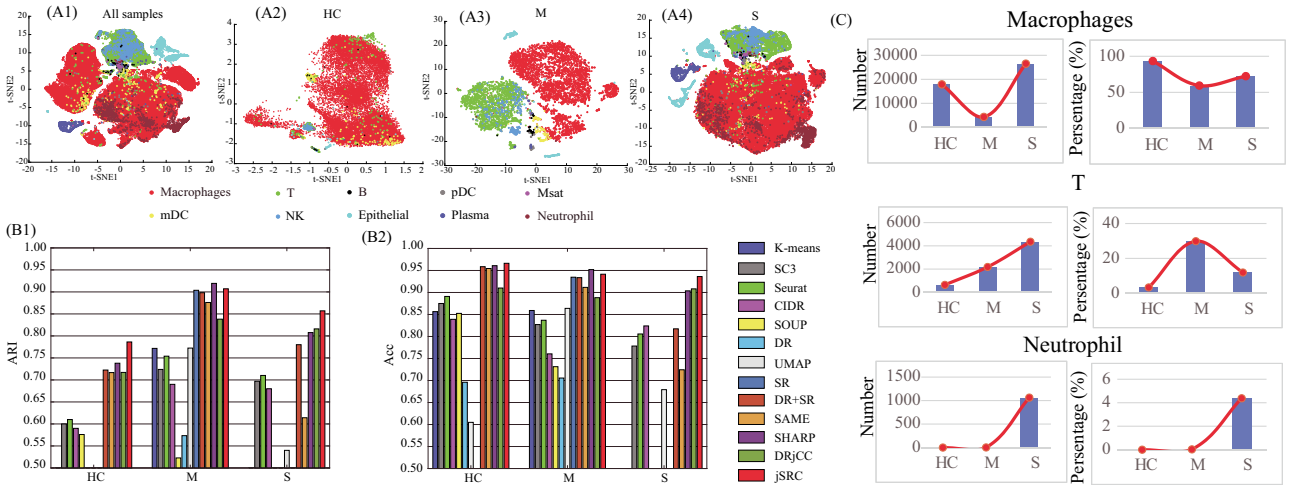

Fig. 8. The application of jSRC to the COVID-19 scRNA-seq data: (**A**) tSNE presentation of cell types based on the coefficient matrix learned by jSRC with various stages, where panel A1 is for all samples, A2 for HCs, A3 for moderate stage and A4 for severe stage; (**B**) performance of various algorithms on the dynamic cell types of COVID-19 in terms of ARI (B1) and Acc (B2), where these algorithms whose values are less than the minimum value of y-axis are missed; (**C**) the dynamics of the major cell types (macrophages, T, neutrophil) across various stages (HC, M, S) of COVID-19 in terms of the number of cells (left) and proportion of cells (right).

The novelty of jSRC is summarized as follows.

- How to determine the number of cell types is fundamental for the downstream analysis of scRNA-seq, and the available methods empirically select it, which is criticized for the accuracy since they ignore intrinsic properties of scRNA-seq data. jSRC overcomes this issue by automatically select the number of cell types by learning the intrinsic features in scRNA-seq data by imposing structure constraint on the representation, thereby facilitating and extending the applications of algorithms (Figure 2).
- How to interpret the patterns with biological backgrounds is one of major concerns of machine learning algorithms and the current algorithms fail to address this issue. To address this problem, jSRC learns features of cells by employing the SR strategy and imposes the sparse constraint on the model, where each cell can be efficiently represented by other cells (Figure 3).
- The current algorithms independently execute procedures for scRNA-seq analysis, such as DR, cell type discovery and feature extraction. However, the independence assumption ignores the connections among these procedures. jSRC jointly learns the features of cells, clustering of cell types and representation, where features are extracted under the guidance of clustering, thereby improving the quality of features. Furthermore, feature extraction and representation learning facilitate the clustering of cells, thereby improving the accuracy of jSRC (Figure 5).
- The informative gene selection is the foundation for the cell proliferation, and many current methods address this issue without considering the interactions among genes. In this study, we select the informative genes by integrating the topological structure and expression profile of genes, where the non-differentially expressed, but critical genes (transcriptional factors) are identified, which provide biologists with clues for further studies.

The experimental results demonstrate that the novelty of jSRC not only provides an efficient tool for analyzing the scRNA-seq data but also suggests the promising directions for designing algorithms for scRNA-seq data.

We see ample opportunities to improve on the function of jSRC in future work. For example, rare cells are critical for transition [3]. These cells are very similar to the originated cells and proliferated cells, where jSRC fails to identify them because SR learning. How to extract rare cell types is promising (Supplementary Figure S2). How to incorporate prior knowledge facilitating the discovery of rare cell types is fundamental. Furthermore, scRNA-seq is insufficient to fully characterize the structure and functions of complex biological systems. How to integrate omic data to further improve the performance of algorithms is also promising.

Key Points

- To improve the scalability, the dimension reduction is employed as a component of algorithm. jSRC jointly learns sparse representation clustering, where features are selected under the guidance of sparse representation clustering of cells, thereby improving the accuracy of clustering.
- To improve the interpretability, the sparse representation constraint is imposed on features based on the hypothesis that cells within the same type have the similar expression patterns on the biomarker genes. The cell types and the number of cell types are automatically determined by the learned coefficient matrix.
- We apply jSRC to 15 scRNA-seq datasets with the number of single cells varies from 49 to 110 824 from various tissues and organisms. The experimental results demonstrate that jSRC significantly outperforms 12 state-of-the-art methods with 20.29% improvement in adjusted Rand index on average.
- And jSRC automatically and precisely selects the number of cell types. Furthermore, it also improves the interpretability and robustness.
- By using scRNA-seq of COVID-19, jSRC accurately identifies cell types and their dynamics associated with disease progression.

Acknowledgments

The authors thank the members of the Ma laboratory for helpful discussion and appreciate the researchers who provide us with source code for a comparison. The authors thank reviewers for their time and suggestions.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) (61772394).

References

1. Huh R, Yang Y, Jiang Y, *et al.* SAME-clustering: single-cell aggregated clustering via mixture model ensemble. *Nucleic Acids Res* 2020; **48**(1): 86–95.
2. Goolam M, Scialdone A, Graham SJ, *et al.* Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* 2016; **165**:61–74.
3. Grun D, Lyubimova A, Kester L, *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 2015; **525**:251–5.
4. Zeisel A, Munoz-Manchado A B, Codeluppi S, *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015; **347**(6226): 1138–42.
5. Kiselev VY, Andrews TS, Hemberg M, *et al.* Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019; **20**:273–82.
6. Han X, Wang R, Zhou Y, *et al.* Mapping the mouse cell atlas by microwell-seq. *Cell* 2018; **172**:1091–107.
7. Cusanovich DA, Reddington JP, Garfield DA, *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* 2018; **555**:538–42.
8. Zheng GX Y, Terry JM, Belgrader P, *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017; **8**:14049.
9. Satija R, Farrell JA, Gennert D, *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015; **33**:495–502.

10. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018; **19**:15.
11. Guo M, Wang H, Potter SS, *et al*. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol* 2015; **11**:e1004575.
12. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* 2019; **16**:983–6.
13. Brennecke P, Anders S, Kim JK, *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013; **10**:1093–5.
14. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017; **14**:483–6.
15. Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform* 2016; **17**:140.
16. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 2017; **18**(1): 59.
17. Yang Y, Huh R, Culpepper HW, *et al*. SAFE-clustering: single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data. *Bioinformatics* 2019; **35**(8): 1269–77.
18. Zhang AW, O'Flanagan C, Chavez EA, *et al*. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 2017; **16**:1007–15.
19. Sun Z, Wang T, Deng K, *et al*. DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* 2018; **34**(1):139–46.
20. Wang B, Zhu J, Pierson E, *et al*. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 2017; **14**:414.
21. Wang S, Karikomi M, MacLean AL, *et al*. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res* 2019; **47**(11): e66.
22. Kim J, Stanescu DE, Won KJ. CellBIC: bimodality-based top-down clustering of single-cell RNA sequencing data reveals hierarchical structure of the cell type. *Nucleic Acids Res* 2018; **46**(21): e124.
23. Wang X, Sun Z, Zhang Y, *et al*. BREM-SC: a Bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Res* 2020; **48**(11): 5814–24.
24. Cheng C, Easton J, Rosencrance C, *et al*. Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell RNA-seq data. *Nucleic Acids Res* 2019; **47**(22): e143.
25. Li L, Dai H, Fang ZY, *et al*. CCSN: single cell RNA sequencing data analysis by conditional cell-specific network. *bioRxiv* 2020. https://doi.org/10.1101/2020.01.25.919829.
26. Zhu L, Lei J, Klei L, *et al*. Semisoft clustering of single-cell data. *Proc Natl Acad Sci USA* 2019; **116**(2): 466–71.
27. Wu WM, Ma XK. Joint learning dimension reduction and clustering of single-cell RNA-sequencing data. *Bioinformatics* 2020; **36**(12): 3825–32.
28. d'Aspremont A, El Ghaoui L, Jordan MI, *et al*. A direct formulation of sparse PCA using semidefinite programming. *SIAM Rev* 2007; **49**:434–48.
29. Wright J, Yang AY, Ganesh A, *et al*. Robust face recognition via sparse representation. *IEEE Trans Pattern Mach Intell* 2009; **31**(2): 210–27.
30. Liu Y, Liu S, Wang Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inform Fusion* 2015; **24**:147–64.
31. Ding RX, Wang X, Shang K, *et al*. Sparse representation-based intuitionistic fuzzy clustering approach to find the group intra-relations and group leaders for large-scale decision making. *IEEE Trans Fuzzy Syst* 2019; **27**(3): 559–73.
32. Boyd S, Parikh N, Chu E. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* 2010; **3**:1–122.
33. Wu S, Joseph A, Hammonds AS, *et al*. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc Natl Acad Sci USA* 2016; **113**:4290–5.
34. Hubert L, Arabie P. Comparing partitions. *J Classif* 1985; **2**:193–218.
35. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017; **18**(1): 174.
36. Biase FH, Cao X, Zhong S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res* 2014; **24**:1787–96.
37. Ting DT, Wittner BS, Ligorio M, *et al*. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep* 2014; **8**:1905–18.
38. Baron M, Veres A, Wolock SL, *et al*. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 2016; **3**:346–60.
39. Schaum N, Karkanias J, Neff NF, *et al*. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 2018; **3**:367–72.
40. Camp JG, Badsha F, Florio M, *et al*. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci USA* 2015; **112**(51): 15672–7.
41. Birey F, Andersen J, Makinson CD, *et al*. Assembly of functionally integrated human forebrain spheroids. *Nature* 2017; **545**(7652): 54–9.
42. Liao MF, Liu Y, Yuan J, *et al*. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* 2020; **26**:842–4.
43. Becht E, McInnes L, Healy J, *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019; **37**(1): 38–44.
44. Maaten L, Hinton G. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008; **9**(2): 579–2605.
45. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models[J]. *Nature communications* 2018; **9**(1): 1–13.
46. Wan S, Kim J, Won KJ. SHARP: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. *Genome Res* 2020; **30**(2): 205–13.
47. Shah SA, Koltun V. Robust continuous clustering. *Proc Natl Acad Sci USA* 2017; **114**(37): 9814–9.
48. Zeng Q, Michael IP, Zhang P, *et al*. Synaptic proximity enables NMDAR signalling to promote brain metastasis. *Nature* 2019; **573**:526–31.
49. Zhou Y, Zhou B, Pache L, *et al*. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019; **10**(1): 1523.
50. Chudnovsky Y, Kim D, Zheng S, *et al*. ZFHX4 interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state. *Cell Rep* 2014; **6**(2): 313–24.
51. Huang SK, Qian JX, Yuan BQ, *et al*. SiRNA-mediated knockdown against NUF2 suppresses tumor growth and induces cell apoptosis in human glioma cells. *Cell Mol Biol (Noisy-le-Grand)* 2014; **60**(4): 30–6.
52. Zhi T, Jiang K, Xu X, *et al*. ECT2/PSMD14/PTTG1 axis promotes the proliferation of glioma through stabilizing E2F1. *Neuro Oncol* 2019; **21**(4): 462–73.

53. Ruggieri S, De Giorgis M, Annese T, *et al*. Dp71 expression in human glioblastoma. *Int J Mol Sci* 2019; **20**(21): 5429.
54. Feng Z, Zhang L, Zhou J, *et al*. mir-218-2 promotes glioblastomas growth, invasion and drug resistance by targeting CDC27. *Oncotarget* 2017; **8**(4): 6304–18.
55. Tang C, Yang Z, Chen D, *et al*. Downregulation of miR-130a promotes cell growth and epithelial to mesenchymal transition by activating HMGB2 in glioma. *Int J Biochem Cell Biol* 2017; **93**:25–31.
56. Frömberg A, Rabe M, Oppermann H, *et al*. Analysis of cellular and molecular antitumor effects upon inhibition of SATB1 in glioblastoma cells. *BMC Cancer* 2017; **17**(1): 3.
57. Elsaid MF, Chalhoub N, Ben-Omran T, *et al*. Omozygous nonsense mutation in SCHIP1/IQCJ-SCHIP1 causes a neurodevelopmental brain malformation syndrome. *Clin Genet* 2018; **193**(2): 387–91.
58. Hong JH, Muhammad E, Zheng C, *et al*. Essential role of carbonic anhydrase XII in secretory gland fluid and HCO3 (-) secretion revealed by disease causing human mutation. *J Physiol* 2015; **593**(24): 5299–312.
59. Duan X, Mao X, Sun W. ADAM15 is involved in MICB shedding and mediates the effects of gemcitabine on MICB shedding in PANC-1 pancreatic cancer cells. *Mol Med Rep* 2013; **7**(3): 991–7.
60. Onyeaghala G, Lane J, Pankratz N, *et al*. Association between MICA polymorphisms, s-MICA levels, and pancreatic cancer risk in a population-based case-control study. *PLoS One* 2019; **14**(6): e0217868.
61. Michita RT, Chies JAB, Schramm S, *et al*. A valine mismatch at position 129 of MICA is an independent predictor of cytomegalovirus infection and acute kidney rejection in simultaneous pancreas–kidney transplantation recipients. *Int J Mol Sci* 2018; **19**(9): 2618.
62. El Jellas K, Johansson BB, Fjeld K, *et al*. The mucinous domain of pancreatic carboxyl-ester lipase (CEL) contains core 1/core 2 O-glycans that can be modified by ABO blood group determinants. *J Biol Chem* 2018; **293**(50): 19476–91.
63. Dalva M, El Jellas K, Steine SJ, *et al*. Copy number variants and VNTR length polymorphisms of the carboxyl-ester lipase (CEL) gene as risk factors in pancreatic cancer. *Pancreatology* 2017; **17**(1): 83–8.
64. Fjeld K, Weiss FU, Lasher D, *et al*. A recombined allele of the lipase gene CEL and its pseudogene CELP confers susceptibility to chronic pancreatitis. *Nat Genet* 2015; **47**(5): 518–22.
65. He W, Wu J, Shi J, *et al*. IL22RA1/STAT3 signaling promotes stemness and tumorigenicity in pancreatic cancer. *Cancer Res* 2015; **78**(12): 3293–305.
66. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan. *Lancet* 2020; **395**:49706.
67. Zhou C, Gao C, Xie Y, *et al*. COVID-19 with spontaneous pneumomediastinum. *Lancet* 2020; **20**:384–510.