G3, 2021, 11(6), jkab098
DOI: 10.1093/g3journal/jkab098
Advance Access Publication Date: 31 March 2021
Software and Data Resources

Single-cell data clustering based on sparse optimization and low-rank matrix factorization

Yinlei Hu,[1,†] Bin Li,[2,†] Falai Chen,[1,3,*] and Kun Qu [2,3,4,*]

[1]School of Mathematical Sciences, University of Science and Technology of China, 230026 Hefei, Anhui, China
[2]Department of Oncology, The First Affiliated Hospital of USTC, Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, School of Basic Medical Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, 230021 Hefei, Anhui, China
[3]School of Data Science, University of Science and Technology of China, 230026 Hefei, Anhui, China
[4]CAS Center for Excellence in Molecular Cell Sciences, the CAS Key Laboratory of Innate Immunity and Chronic Disease, University of Science and Technology of China, 230027 Hefei, Anhui, China

*Corresponding author: Department of Oncology, The First Affiliated Hospital of USTC, Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, School of Basic Medical Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, 230021 Hefei, Anhui, China. qukun@ustc.edu.cn (K.Q.); School of Mathematical Sciences, University of Science and Technology of China, 230026 Hefei, Anhui, China. chenfl@ustc.edu.cn (F.C.)
†These authors contributed equally to this work.

Abstract

Unsupervised clustering is a fundamental step of single-cell RNA-sequencing (scRNA-seq) data analysis. This issue has inspired several clustering methods to classify cells in scRNA-seq data. However, accurate prediction of the cell clusters remains a substantial challenge. In this study, we propose a new algorithm for scRNA-seq data clustering based on Sparse Optimization and low-rank matrix factorization (scSO). We applied our scSO algorithm to analyze multiple benchmark datasets and showed that the cluster number predicted by scSO was close to the number of reference cell types and that most cells were correctly classified. Our scSO algorithm is available at https://github.com/QuKunLab/scSO. Overall, this study demonstrates a potent cell clustering approach that can help researchers distinguish cell types in single- scRNA-seq data.

Keywords: scSO; single-cell cluster; spectral cluster; sparse optimization; scRNA-seq

Introduction

Single-cell RNA-sequencing (scRNA-seq) technology has been widely used in many biological investigations, including the elucidation of cell subtype heterogeneity (Zeisel *et al.* 2015; Goolam *et al.* 2016), construction of gene regulatory networks (Darmanis *et al.* 2015), profiling of cell development and differentiation (Deng *et al.* 2014; Liu *et al.* 2017), and depiction of disease in an immunoresponsive environment (Guo *et al.* 2018; Zhang *et al.* 2018). The analysis of scRNA-seq data contains, but is not limited to, quality control (Chen *et al.* 2016), data normalization (Cole *et al.* 2019), unsupervised clustering (Kiselev *et al.* 2017; Wang *et al.* 2017; Wolf *et al.* 2018; Yang and Wang 2020), trajectory construction (Wolf *et al.* 2019), and differentially expressed gene identification (Soneson and Robinson 2018). As a fundamental step of scRNA-seq data analysis, cell clustering determines the results of subsequent downstream analyses to a certain extent, but is often inaccurate and misconstrues analyses. In recent years, various clustering methods emerged to address this problem, and they have been widely used in single-cell data analysis (Kiselev *et al.* 2019). For example, in Seurat (Butler *et al.* 2018; Stuart *et al.* 2019), Butler and Stuart *et al.* employed K-nearest-neighbor graphs to obtain cell–cell similarity and used the community detection algorithm to cluster cells. To estimate cell-cell correlation, Wang *et al.* (2017) proposed a multi-kernel learning method in SIMLR, and Kiselev *et al.* (2017) presented a consensus clustering algorithm in SC3. However, the clustering accuracy of the currently established algorithms is limited, and as such, algorithms need to be further improved for the accurate prediction of cell clusters (Kiselev *et al.* 2019).

In this work, by assuming that the expression vectors of cells in the same cluster are approximately linearly correlated, we proposed the use of Sparse Nonnegative Matrix Factorization (SNMF) and a Gaussian mixture model (GMM) to calculate cell-cell similarity. After assembling the cell–cell similarity matrix, we introduced a novel, unsupervised algorithm to predict cell clusters based on spectral methods and sparse optimization techniques (see Materials and Methods). As shown by our experimental results derived from 12 benchmark datasets whose cell types have been biologically verified (Yan *et al.* 2013; Biase *et al.* 2014; Deng *et al.* 2014; Pollen *et al.* 2014; Treutlein *et al.* 2014; Klein *et al.* 2015; Kolodziejczyk *et al.* 2015; Usoskin *et al.* 2015; Zeisel *et al.* 2015; Baron *et al.* 2016; Goolam *et al.* 2016; Li *et al.* 2017), the clustering accuracy of our scRNA-seq data clustering based on Sparse Optimization and low-rank matrix factorization (scSO) method outperforms the previously established, state-of-the-art single-cell clustering methods. Furthermore, our scSO algorithm can be used to generate a visual representation of cell–cell similarity (see Figure 4).

Received: February 15, 2021. **Accepted:** March 20, 2021

Materials and methods

Methods

As an input, scSO takes an expression matrix $\mathbf{A}$ whose element $\mathbf{A}_{ij}$ represents the expression of the ith gene in the jth cell. There are four elementary procedures in scSO, including preprocessing, SNMF dimensionality reduction, cell-to-cell similarity matrix construction, and unsupervised clustering based on sparse optimization. The pipeline is illustrated in Figure 1. For the convenience of other researchers, we have uploaded scSO to GitHub (https://github.com/QuKunLab/scSO).

***Preprocessing*:** The preprocessing procedure includes two steps. In the paper of SC3 method (Kiselev *et al.* 2017), Kiselev *et al.* pointed out that the ubiquitous genes and rare genes usually cannot help clustering, and filtering out these genes can significantly improve the efficiency of calculations. Moreover, in other recently published studies, researchers also removed ubiquitous genes in the preprocessing stage (Gan *et al.* 2018, 2020; Vans *et al.* 2019; Lu *et al.* 2020; Ye *et al.* 2020). Therefore, we first removed genes that were not detected in any cell (uncaptured genes), and genes detected in all cells (ubiquitous genes). Second, the genes used for classification were filtered by their average expression among cells. We defined a function $f(\bar{\imath})$, as:

$$f(\bar{\imath}) = \begin{cases} 1, & \bar{\imath} \in [\eta_1, \eta_2] \\ 0, & otherwise \end{cases},$$

to determine whether the ith gene should be considered for cluster analysis, where $\bar{\imath} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{A}_{ij}$ is the average expression of the ith gene in all cells. We set $\eta_1 = 0.1\rho$ and $\eta_2 = 8.5\rho$ for all datasets tested in this article, where $\rho = \frac{1}{m}\sum_{i}^{m}\bar{\imath}$, and retain the genes with $f(\bar{\imath}) = 1$. Finally, the scSO algorithm normalizes the total expression of each cell to 10,000, and updates the expression values by performing a logarithmic translation on them that is, $\mathbf{A}_{ij} = log_{10}\left(\frac{10{,}000\times\mathbf{A}_{ij}}{\sum_i \mathbf{A}_{ij}} + 1\right)$.

***SNMF dimensionality reduction*:** We used SNMF for dimensionality reduction [a detailed introduction to NMF can refer to (Kim and Park 2007)] that is,

$$min_{\mathbf{W},\mathbf{H}\geq 0} \parallel \mathbf{A} - \mathbf{WH} \parallel_F^2 + \alpha^2 \sum_{i=1}^{n} \parallel \mathbf{H}(:,i) \parallel_1^2 + \beta^2 \sum_{j=1}^{m} \parallel \mathbf{W}(j,:) \parallel_1^2 (1)$$

where $\mathbf{W}(j,:)$ is the jth row of $\mathbf{W}$ and $\boldsymbol{H}(:,i)$ is the ith column of $\boldsymbol{H}$, and α and β are nonnegative tuning parameters. $\mathbf{W}$ and $\boldsymbol{H}$ are the basis matrix and coefficient matrix of $\mathbf{A}$, respectively. Each column of $\boldsymbol{H}$ is a low-dimensional representation of the corresponding cell. Because the model (1) is a convex problem after fixing $\mathbf{W}$ or $\boldsymbol{H}$, it can be solved by the alternating iteration algorithm. The detailed process is described in Supplementary Note S1.

***Constructing the cell-to-cell similarity matrix*:** After obtaining the low-dimensional expression matrix $\boldsymbol{H}$ of cells, we used the GMM (Supplementary Note S2) to fit the distribution of cells in space spanned by $\mathbf{W}$. We adopted the MATLAB function *fitgmdist*() to build the GMM ($P(x) = \sum_{i=1}^{k}\pi_i N(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$), and then calculated the probability that the ith cell belongs to the jth component through Bayes' rule [*i.e.*, $\gamma(i,j) = \frac{\pi_i N(\boldsymbol{x}_i|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{P(x_i)}$]. Finally, the similarity matrix $\boldsymbol{S}$ can be derived by calculating the probability that any two cells belong to the same component that is, $\boldsymbol{S} = \boldsymbol{\gamma}\boldsymbol{\gamma}^T$, here $\boldsymbol{\gamma} = (\gamma(i,j))$.

***Unsupervised clustering based on sparse optimization*:** Assuming that a dataset $\mathbf{A}$ can be divided into c clusters ($C_1, C_2, C_3, \ldots, C_c$), and the perfect similarity matrix should be:

$$S_{pq} = \begin{cases} 1, & if\ p,q \in C_i \\ 0, & otherwise \end{cases} \quad (p,q = 1 \sim n).$$

The zero eigenvalue of $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$ here, $D = diag(d_1, d_2, \ldots, d_n)$ and $d_i = \sum_{j=1}^{n}\mathbf{S}_{ij}$) has multiplicity c, and its eigenspace is spanned by the indicator vectors $1_{C_1}, 1_{C_2}, \ldots, 1_{C_c}$,

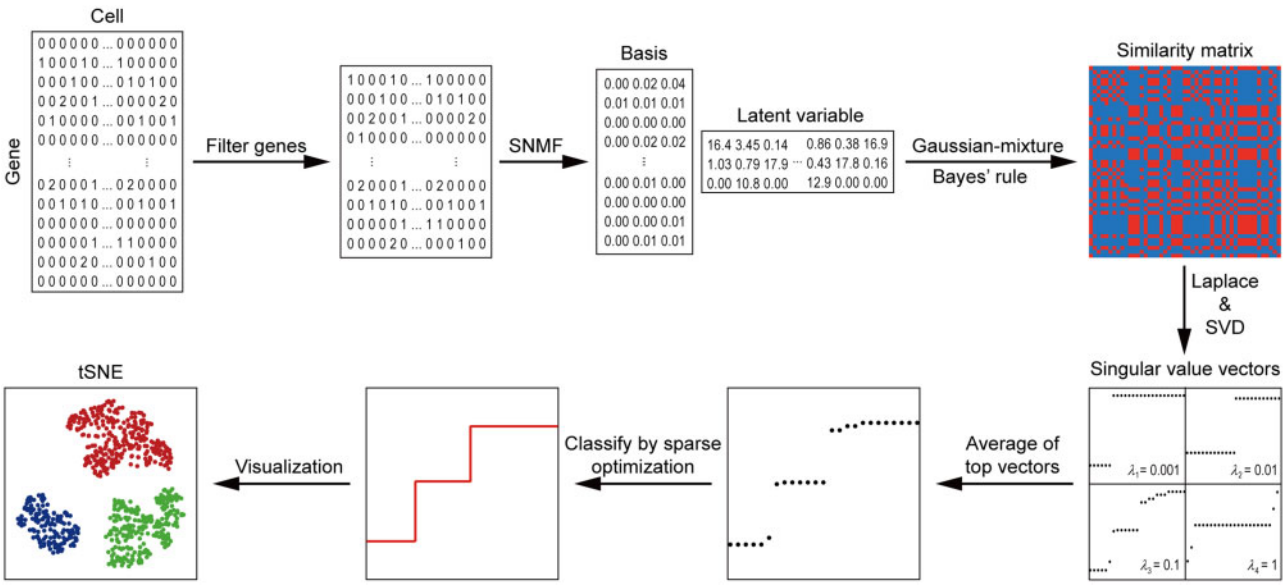


Figure 1 A conceptual overview of the scSO workflow.

here $1_{C_i} \in \mathbb{R}^n$ *and* $(1_{C_i})_j = \begin{cases} 1, & j \in C_i \\ 0, & otherwise \end{cases}$. Moreover, each nonzero eigenvalue l_i of $\boldsymbol{L}$ has multiplicity $l_i - 1$ $(i = 1 \sim c)$, where l_i is the number of cells in C_i (Supplementary Note S3). For the eigenvector $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^T$ corresponding to a zero eigenvalue of $\boldsymbol{L}$, if we sort the entries of $\boldsymbol{u}$ in ascending order, as:

$$u'_1 = \cdots = u'_{l_1} < u'_{l_1+1} = \cdots = u'_{l_2} < \cdots < u'_{l_{c-1}+1} = \cdots = u'_{l_c} = u'_n ,$$

the entries of $\boldsymbol{u}$ with the same values belong to the same class and they can be used to segment C into c classes.

After obtaining the similarity matrix $\boldsymbol{S} = (\boldsymbol{S_{ij}})_{\boldsymbol{n\times n}}$ (see previous section), we constructed the Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, and computed the eigenvalues $l_1 \le l_2 \le \ldots \le l_n$ and the eigenvector $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$ of $\boldsymbol{L}$. It is easy to show that $\boldsymbol{L}$ is positive semidefinite. Because the similarity matrix $\boldsymbol{S}$ generated from previous section contains noise, the minimum eigenvalue l_{min} of $\boldsymbol{L}$ is not exactly zero, and the multiplicity l_{min} is usually equal to 1. However, the vector $\boldsymbol{l} = (l_1, l_2, \ldots, l_n)$ is a piecewise constant vector if $\boldsymbol{S}$ is "perfect" (Supplementary Note S3). Thus, we calculated the piecewise constant approximation of $\boldsymbol{l}$ as $\boldsymbol{l}'$ and defined the multiplicity of eigenvalue zero (termed c′) as the number of the minimum elements in $\boldsymbol{l}'$. As shown in Supplementary Figure S1, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ cannot separate zygote, two-, and three-cell, but the center of $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ can distinguish them well. Therefore, we used the center of $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{c'}$ as an initial guess $\boldsymbol{u}$ of cell clustering (blue points in Figure 2A).

After obtaining the $\boldsymbol{u}$, scSO takes the following process to predict cell clusters: (1) sort the elements of $\boldsymbol{u}$ in ascending order to obtain $\tilde{\boldsymbol{u}}$ (blue points in Figure 2A); (2) calculate the piecewise constant approximate $\hat{\boldsymbol{u}}$ of $\tilde{\boldsymbol{u}}$ to remove noise (solid red line in Figure 2B); (3) group cells corresponding to a piecewise constant (*i.e.*, a step in Figure 2C) into a cluster, and the number of piecewise constants is the number of cell clusters.

In this work, we obtain the piecewise constant approximation of a vector $\mathbf{b}$ by the weighted sparse optimization model (Tong *et al.* 2020) that is:

$$\min_{\mathbf{x},\mathbf{z}} [(1-\lambda)||\mathbf{b}-\mathrm{x}||_2^2 + \lambda||z||_1], \mathrm{s.t.}\mathbf{z} = D\mathbf{x} \qquad (2)$$

where the entries of $\mathbf{b}$ are sorted in ascending order, $\mathbf{x}$ is the piecewise constant approximation of $\mathbf{b}$, $\lambda \in (0,1)$ is the balance parameter, and $\boldsymbol{D}_{n\times(n-1)}$ is a sparse matrix. All entries of $\boldsymbol{D}$ are zeros, except for $\boldsymbol{D}_{i,i} = -1$ and $\boldsymbol{D}_{i,i+1} = 1$. As the model (2) is a convex quadratic programming problem, we solved it by the convex optimization package CVXPY (https://www.cvxpy.org/; Diamond and Boyd 2016; Agrawal *et al.* 2018).

Parameters for other tools

Seurat (version 3.0.0) was downloaded from https://github.com/satijalab/seurat/. For all datasets in the paper, Seurat was performed with default parameters. Specially, we set the number of neighbors to 20, the cluster resolution to 0.8, and used the *ScoreJackStraw()* function and 0.05 (the bound of *P*-value) to determine the number of principal components. (2) Scanpy (version 1.4.0) was downloaded from https://github.com/theislab/scanpy. In addition to using *n_pcs* = 20 in the function *pp.neighbors ()*, we ran Scanpy with default parameters (*n_neighbors* = 15 and *resolution* = 1.0). (3) We downloaded SC3 (version 1.10.1) from https://github.com/hemberg-lab/SC3. SC3 was run with default parameters, for example, *gene_filter* = *FALSE*, *pct_dropout_min* = 10, *pct_dropout_max* = 90, *d_region_min* = 0.04, and *d_region_max* = 0.07. (4) SIMLR (version 1.8.1) was downloaded from https://github.com/BatzoglouLabSU/SIMLR. We use the default parameters to perform SIMLR except that we set *kk* = 30 in the *SIMLR_Large_Scale ()* function and used *NUMC* = 2:15 to calculate the number of clusters.

Data availability

There is no new data associated with this article. Published datasets used in this study: nine datasets are available at GEO with the accession numbers: GSE36552 (Yan *et al.* 2013), GSE57249 (Biase *et al.* 2014), GSE45719 (Deng *et al.* 2014), GSE65525 (Klein *et al.* 2015), GSE60361 (Zeisel *et al.* 2015), GSE81861(Rca; Li *et al.* 2017), GSE52583 (Treutlein *et al.* 2014), GSE84133 (Baron *et al.* 2016), and GSE59739 (Usoskin *et al.* 2015); the rest of three datasets can be available with the accession number: E-MTAB-2600 (Kolodziejczyk *et al.* 2015), E-MTAB-3321 (Goolam *et al.* 2016), and SRP041736 (Pollen *et al.* 2014) (Supplementary Table S1). Supplementary Figures S1–4, Supplementary Tables S1 and S2, and Supplementary Notes S1–3 are available at figshare: https://doi.org/10.25387/g3.14256641. The source code of scSO is released at https://github.com/QuKunLab/scSO, and all results in this article are obtained by using python version of scSO.

Results

To assess the performance of scSO, we applied our approach to 12 benchmark scRNA-seq datasets (Yan *et al.* 2013; Biase *et al.* 2014; Deng *et al.* 2014; Pollen *et al.* 2014; Treutlein *et al.* 2014; Klein

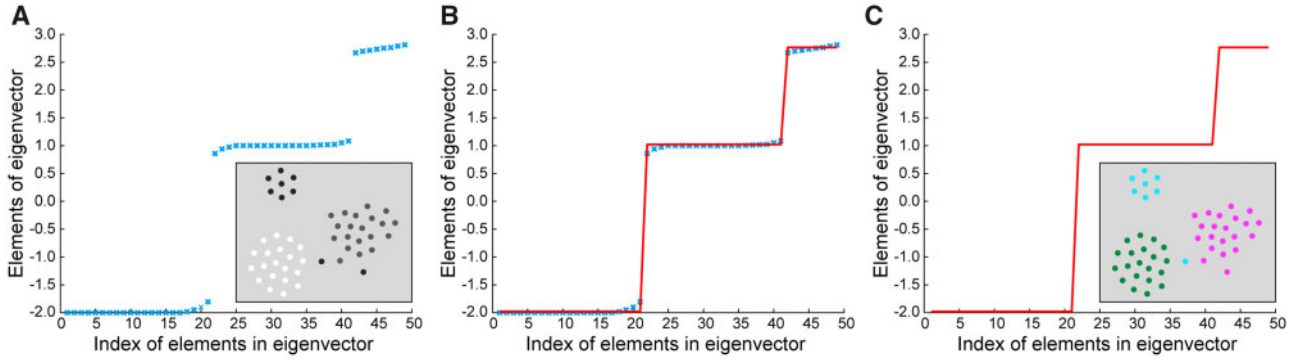


Figure 2 An example of scSO determine cell clusters. (A) The blue points represent the sorted elements of the eigenvector corresponding to the zero eigenvalue of the Laplacian matrix. (B) The solid red line is the piecewise constant approximation of the sorted eigenvector calculated by sparse optimization. (C) The result of clustering generated by the piecewise constant vector.

et al. 2015; Kolodziejczyk et al. 2015; Usoskin et al. 2015; Zeisel et al. 2015; Baron et al. 2016; Goolam et al. 2016; Li et al. 2017), Among 12 benchmark datasets, the cells in Biase (Biase et al. 2014), Yan (Yan et al. 2013), Goolam (Goolam et al. 2016), and Deng (Deng et al. 2014) came from different cell stages (those cell types represent different cell stages), and the cells in Kolodziejczyk (Kolodziejczyk et al. 2015) and Pollen (Pollen et al. 2014) were generated under different experimental conditions (those cell labels represent different experimental conditions). Kiselev et al. (2017) believed that these cell labels in the six datasets have high-confidence and call them "gold standard" data. The other six datasets (such as the Zeisel (Zeisel et al. 2015) dataset) are considered as "silver standard" data. Their cell types were assigned by computational methods and experimentally verified by differential expressed mark genes of each cell group. Those benchmark datasets also were used in other recent studies (Kiselev et al. 2017; Qi et al. 2020).

Currently there are >50 single-cell clustering algorithms (Zappia et al. 2018); however, the recently published benchmark article from *Briefings in Bioinformatics* (Qi et al. 2020), Qi et al. tested five representative clustering methods (SC3, SNN-Cliq, SINCERA, SEURAT, and pcaReduce) of the most advanced scRNA-seq tools currently available through 12 public benchmark datasets and showed that SC3 had the highest clustering accuracy under default parameters, while Seurat performed well in the mixture control experiment reported by the recently published benchmark article in *Nature Methods* (Tian et al. 2019) Tian et al. believed that RaceD3, Seurat, clusterExperiment, RCA, and SC3 were representative, and tested them in their mixture control experiment). Scanpy is a widely used python packages for single-cell analysis (Li et al. 2020a,b). SIMLR measures the cell–cell similarity based on a multi-kernel learning method, which is a representative method in single-cell clustering algorithms (Kiselev et al. 2019). Therefore, we only compared our algorithm with SC3, Scanpy, Seurat, and SIMLR. To ensure that comparisons between algorithms were based on the same criteria, we used the same gene-filtering and normalization steps for all these algorithms.

We used the following three metrics to quantify the performance of scSO in predicting the number of cell types: (1) the difference between the predicted and reference cell type numbers of each dataset ($\Delta_m = N_{pred,m} - N_{ref,m}$ for dataset m), (2) the Pearson correlation coefficient between the predicted and reference cell type numbers ($P_{corr} = corr(\bar{N}_{pred}, \bar{N}_{ref})$), and (3) the number of datasets with Δ_m less than or equal to 2 ($N_{\Delta\leq 2}$). Specifically, Δ_m represents the error of the predicted number of cell types, and the average Δ_m of the prediction results of scSO on the 12 datasets was 1.58, which was the lowest of the 5 methods (Figure 3A). For the P_{corr} parameter that represents the prediction accuracy of the cell type number, the scSO result was 0.79, and the results of SC3, Scanpy, Seurat, and SIMLR were 0.41, 0.51, 0.47, and 0.23, respectively (Figure 3B). Moreover, $N_{\Delta\leq 2}$ indicates the number of datasets for which the cell type number was accurately predicted, and the $N_{\Delta\leq 2}$ for the scSO results was 10, whereas for the other four methods $N_{\Delta\leq 2}$ were 6 (SC3), 5 (Scanpy), 4 (Seurat), and 6 (SIMLR; Figure 3C).

To further evaluate the clustering performance of scSO, we calculated the adjusted rand index (ARI; Kiselev et al. 2017; Huang et al. 2018) between its clustering results and reference labels of cells. A higher ARI value indicates higher consistency between the clustering result and the reported cell types of each dataset. In the scSO clustering results of all the 12 datasets, the ARI values of 8 datasets were greater than or equal to 0.8 (Figure 3D). On the contrary, in the clustering results of SC3, the ARI values of three datasets were greater than 0.8, which was the second-highest among all the five methods (The results of different ARI thresholds can refer to Supplementary Table S2). We also tested these state-of-the-art methods with their filtering and normalization processes, which did not significantly change their performance on these datasets (Supplementary Figure S2).

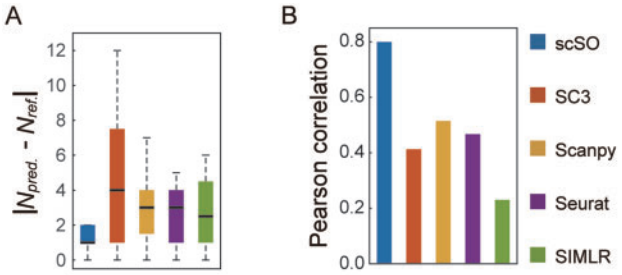
Specifically, scSO classified all the oligodendrocytes in the Zeisel dataset (Zeisel et al. 2015) into one cell cluster, and the d0 mouse embryonic stem cells in the Klein dataset. In contrast, the other four methods divided these cell types into multiple subclusters, and the subclusters classified by different methods were not identical (Supplementary Figure S3).

Furthermore, we used scSO to sort cells by its corresponding element in $\boldsymbol{u}$ (the eigenvector corresponding to zero eigenvalue, Figure 4). The cells with the same height belong to one cluster, and the height of a "step" indicates the difference between two clusters. The $\boldsymbol{u}$ can be used to visualize the similarity between cells. Also, the users can tune the balance parameter λ to adjust cluster number, according to the heights of "steps" in $\boldsymbol{u}$, to obtain a better clustering result.

Discussion

In this article, we propose a new clustering algorithm, scSO, to classify cell types from scRNA-seq data. In the construction of the scSO algorithm, we assumed that the gene expression of cells of the same type was approximately linearly correlated, and we adopted NMF method and GMM to construct the similarity matrix. Then, a new spectral clustering algorithm based on sparse optimization (scSO) was proposed to predict the final cell clusters. The eigenvector that we used to obtain cell identity (Figure 4) can be used to visualize the cell–cell similarity determined by scSO. Finally, we applied scSO to 12 benchmark datasets to assess its performance. Our results indicate that scSO outperformed other state-of-the-art methods in terms of clustering accuracy, implying its advantage in processing single-cell sequencing data.

In scSO, there are several parameters for setting, such as the gene filtering parameter η, the number of reduced dimensions r in NMF, the number of clusters and covariance structure in GMM and the balance parameter λ in the piecewise constant approximation. All these parameters are set as default values in this study. The gene filtering parameter η is used to screen out the genes that do not contribute to scSO clustering. For all the datasets shown in Figure 3, we empirically set $\eta_1 = 0.1\rho$ and $\eta_2 = 8.5\rho$, where ρ is the average expression of all the genes. Experiments show that such setting outperforms other gene filtering methods. For the rank r of gene expression matrix, we use a heuristic algorithm to automatically estimate r in scSO. The heuristic algorithm is based on the observation that the ratios $\frac{\sigma_i}{\sigma_{i+1}}$ of singular values ($\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_{min(m,\,n)}$), has a large jump at $i = r$, and σ_{r+1} is small (Hu et al. 2020). For the covariance structure for GMM, since the r features obtained by SNMF are approximately linearly independent, we set the covariance matrix structure to be diagonal for the 12 benchmark datasets. For the number of clusters in GMM, under the assumption that cells of different types are approximately linearly independent, and in order to better fit the distribution of cells in the low-dimensional space, the number of clusters in the GMM should not be less than r. Meanwhile, in order to avoid over-fitting caused by setting the cluster number in GMM to be too large, the number of clusters should be close to r. We found that $r + 1$ is a good choice for the number of clusters of GMM by experiment. For the balance parameter λ, we tested the performance of scSO with different λ



C

Estimation of the number of cell types

	Ref.	scSO	SC3	Scanpy	Seurat	SIMLR
Biase	3	3	3	2	2	7
Kolodz.	3	4	10	8	7	9
Yan	7	5	6	4	3	11
Goolam	5	6	6	5	4	7
Pollen	11	10	11	8	8	9
Deng	10	6	9	7	5	13
Klein	4	5	24	11	13	5
Zeisel	9	9	21	13	12	10
Rca	7	9	13	9	8	7
Treutlein	5	4	3	3	2	6
Baron	14	9	22	13	14	9
Usoskin	4	5	11	8	7	15

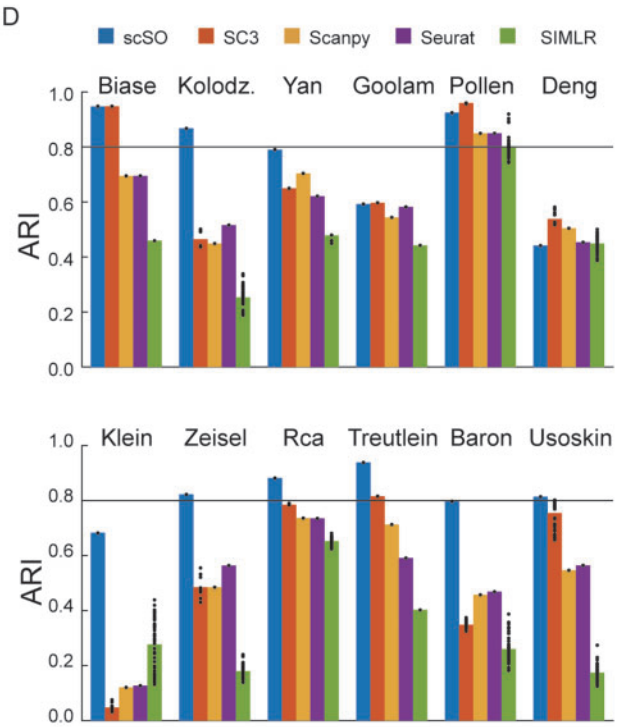


Figure 3 Performance of scSO on 12 benchmark datasets. (A) Difference (Δ_m) between the reference and predicted cell type numbers. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. (B) Pearson correlation coefficient between the reference and predicted cell type numbers. (C) A table listing the reference cell type numbers reported by the data source papers, and the predicted cell type numbers generated by different methods. (D) ARI values between the reference cell types and predicted cell clusters. Bar, average ARI; Dots, ARI values for multiple runs. SC3 and SIMLR were performed 50 times for each dataset.

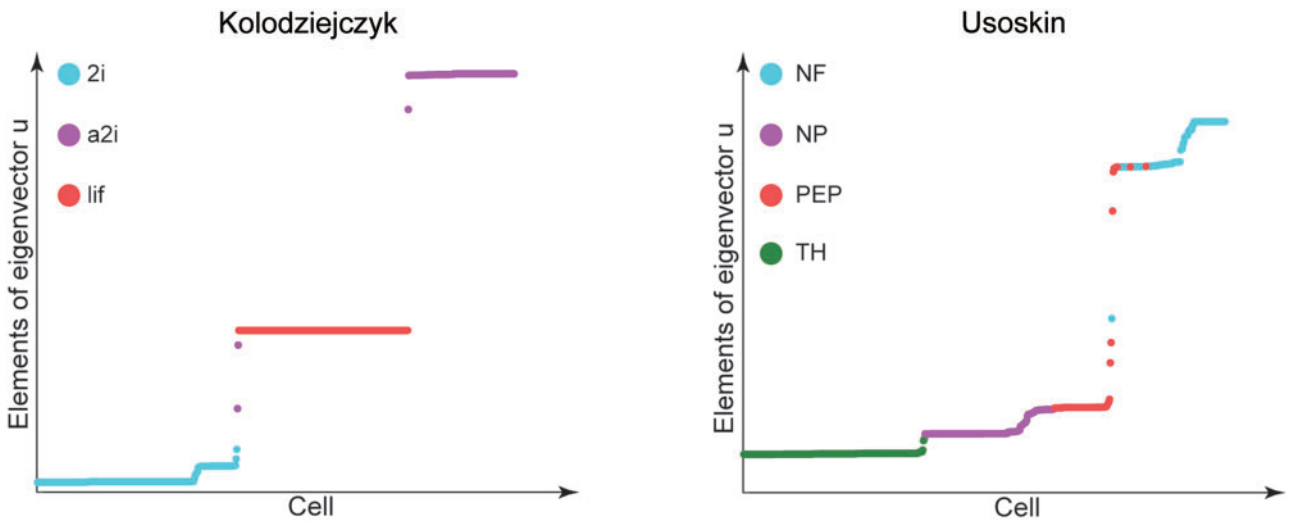


Figure 4 Eigenvector *u* generated by scSO for Kolodziejczyk's and Usoskin's datasets. Each point denotes a cell, and colors denote cell types. "lif," "2i," and "a2i" represent single-cell RNA-sequencing of Mouse embryonic stem cells (mESCs) cultured under three different conditions (Usoskin *et al.* 2015). PEP, NP, NF, and TH are the abbreviations for peptidergic nociceptor, non-peptidergic nociceptor, neurofilament, and tyrosine hydroxylase, respectively (Kolodziejczyk *et al.* 2015).

through 12 benchmark datasets (Supplementary Figure S4). The results show that for λ values between 0.15 and 0.2, the ARI values of the scSO were stable (except the ARI values of Klein data near $\lambda = 0.172$), which implies scSO is robust to λ (Supplementary Figure S4). Therefore, we set $\lambda = 0.15$ as the default value.

Regarding future work, there are three possible directions to improve the clustering effect of scSO. First, to reduce the number of free parameters in scSO, we can try to use a hierarchical version of the Expectation–Maximization algorithm to automatically estimate the number of cell clusters in the future version of scSO. Second, recent studies (Huang *et al.* 2018; Hu *et al.* 2020) indicated that using an appropriate imputation method for single-cell data can improve the profiling of cell types. As such, we will progress the performance of scSO by enhancing the resistance of scSO to dropout events (due to the low capture and sequencing efficiency of single-cell sequencing technology, most genes are represented by zero values in scRNA-seq data). Finally, batch effects hinder scRNA-seq data analyses, and Stuart *et al.* (2019) showed that a suitable batch effect elimination algorithm can effectively improve the accuracy of clustering. Therefore, in future work, we will improve the method of measuring the similarity among cells to make scSO more robust for batch effects.

Acknowledgments

We thank the USTC supercomputing center and the School of Life Science Bioinformatics Center for providing supercomputing resources for this project. We thank the CAS interdisciplinary innovation team for helpful discussion.

Funding

This work was supported by the National Key R&D Program of China (grant numbers 2020YFA0112200 and 2017YFA0102900 to K.Q.), the National Natural Science Foundation of China grants (grant numbers 91940306, 81788101, 31970858, 31771428, and 91640113 to K.Q.; grant numbers 61972368 and 11571338 to F.C.), and the Fundamental Research Funds for the Central Universities (grant number YD2070002019, WK2070000158, and WK9110000141 to K.Q.). It was also supported by Anhui Provincial Natural Science Foundation (grant number BJ2070000097 to B.L.).

Conflicts of interest

There is no conflict of interest.

Literature cited

Agrawal A, Verschueren R, Diamond S, Boyd S. 2018. A rewriting system for convex optimization problems. J Control Decis. 5:42–60.

Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, *et al.* 2016. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst. 3: 346–360.e4.

Biase FH, Cao X, Zhong S. 2014. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Res. 24:1787–1796.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 36:411–420.

Chen H-IH, Jin Y, Huang Y, Chen Y. 2016. Detection of high variability in gene expression from single-cell RNA-seq profiling. BMC Genomics. 17:508.

Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, *et al.* 2019. Performance assessment and selection of normalization procedures for single-cell RNA-seq. Cell Syst. 8:315–328.e8.

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, *et al.* 2015. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A. 112:7285–7290.

Deng Q, Ramsköld D, Reinius B, Sandberg R. 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 343:193–196.

Diamond S, Boyd S. 2016. CVXPY: A Python-embedded modeling language for convex optimization. J Mach Learn Res. 17:1–5.

Gan Y, Li N, Zou G, Xin Y, Guan J. 2018. Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method. BMC Med Genomics. 11: 117.

Gan Y, Liang S, Wei Q, Zou G. 2020. Identification of differential gene groups from single-cell transcriptomes using network entropy. Front Cell Dev Biol. 8:588041.

Goolam M, Scialdone A, Graham SJL, MacAulay IC, Jedrusik A, *et al.* 2016. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. Cell. 165:61–74.,

Guo X, Zhang Y, Zheng L, Zheng C, Song J, *et al.* 2018. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med. 24:978–985.

Hu Y, Li B, Liu N, Cai P, Chen F, *et al.* 2021. WEDGE: recovery of gene expression values for sparse single-cell RNA-seq datasets using matrix decomposition. Brief. Bioinform. bbab085.

Huang M, Wang J, Torre E, Dueck H, Shaffer S, *et al.* 2018. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 15:539–542.

Kim H, Park H. 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 23:1495–1502.

Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 20: 273–282.

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, *et al.* 2017. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 14:483–486.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, *et al.* 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 161:1187–1201.

Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, *et al.* 2015. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 17:471–485.

Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, *et al.* 2017. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 49:708–718.

Li B, Gould J, Yang Y, Sarkizova S, Tabaka M, *et al.* 2020a. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat Methods. 17:793–798.

Li J, Yu C, Ma L, Wang J, Guo G. 2020b. Comparison of Scanpy-based algorithms to remove the batch effect from single-cell RNA-seq data. Cell Regen. 9:10.

Liu Z, Lou H, Xie K, Wang H, Chen N, *et al.* 2017. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nat Commun. 8:1–9.

Lu X, Gao Y, Li J, He K, Chen G, *et al.* 2020. Identification of cell types from single-cell transcriptomes using a novel clustering framework BT. In: D-S Huang, V Bevilacqua, A Hussain, editors.

Intelligent Computing Theories and Application. Cham: Springer International Publishing, p. 17–27.

Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, *et al.* 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 32:1053–1058.

Qi R, Ma A, Ma Q, Zou Q. 2020. Clustering and classification methods for single-cell RNA-sequencing data. Brief Bioinform. 21: 1196–1208.

Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 15:255–261.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, *et al.* 2019. Comprehensive integration of single-cell data. Cell. 177: 1888–1902.e21.

Tian L, Dong X, Freytag S, Cao KAL, Su S, *et al.* 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 16:479–487.

Tong W, Yang X, Pan M, Chen F. 2020. Spectral mesh segmentation via l0 gradient minimization. IEEE Trans Vis Comput Graph. 26: 1807–1820.

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, *et al.* 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 509:371–375.

Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, *et al.* 2015. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 18:145–153.

Vans E, Sharma A, Patil A, Shigemizu D, Tsunoda T. 2019. Clustering of small-sample single-cell rna-seq data via feature clustering and selection BT. In: AC Nayak, A Sharma, editors. PRICAI 2019: Trends in Artificial Intelligence. Cham: Springer International Publishing, p. 445–456.

Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. 2017. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 14:414–416.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19:15.

Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, *et al.* 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 20:59.

Yan L, Yang M, Guo H, Yang L, Wu J, *et al.* 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 20:1131–1139.

Yang F, Wang M. 2020. A review of systematic evaluation and improvement in the big data environment. Front Eng Manag. 7: 27–46.

Ye X, Zhang W, Futamura Y, Sakurai T. 2020. Detecting interactive gene groups for single-cell RNA-seq data based on co-expression network analysis and subgraph learning. Cells. 9:1938.

Zappia L, Phipson B, Oshlack A. 2018. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol. 14:e1006245.

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, Manno GL, *et al.* 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 347:1138–1142.

Zhang L, Yu X, Zheng L, Zhang Y, Li Y, *et al.* 2018. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature. 564:268–272.