frontiers in Genetics

METHODS
published: 28 April 2020
doi: 10.3389/fgene.2020.00407

An Adaptive Sparse Subspace Clustering for Cell Type Identification

Ruiqing Zheng[1], *Zhenlan Liang*[1], *Xiang Chen*[1], *Yu Tian*[1], *Chen Cao*[2] *and Min Li*[1]*

[1] *School of Computer Science and Engineering, Central South University, Changsha, China,* [2] *Departments of Biochemistry & Molecular Biology and Medical Genetics, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada*

The rapid development of single-cell transcriptome sequencing technology has provided us with a cell-level perspective to study biological problems. Identification of cell types is one of the fundamental issues in computational analysis of single-cell data. Due to the large amount of noise from single-cell technologies and high dimension of expression profiles, traditional clustering methods are not so applicable to solve it. To address the problem, we have designed an adaptive sparse subspace clustering method, called AdaptiveSSC, to identify cell types. AdaptiveSSC is based on the assumption that the expression of cells with the same type lies in the same subspace; one cell can be expressed as a linear combination of the other cells. Moreover, it uses a data-driven adaptive sparse constraint to construct the similarity matrix. The comparison results of 10 scRNA-seq datasets show that AdaptiveSSC outperforms original subspace clustering and other state-of-art methods in most cases. Moreover, the learned similarity matrix can also be integrated with a modified t-SNE to obtain an improved visualization result.

Keywords: single cell RNA-seq, subspace clustering, adaptive sparse strategy, similarity learning, visualization

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Chunhou Zheng,
Anhui University, China
Tao Zeng,
Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, China

Correspondence:
Min Li
limin@mail.csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics

Received: *23 October 2019*
Accepted: *31 March 2020*
Published: *28 April 2020*

Citation:
Zheng R, Liang Z, Chen X, Tian Y, Cao C and Li M (2020) An Adaptive Sparse Subspace Clustering for Cell Type Identification. Front. Genet. 11:407. doi: 10.3389/fgene.2020.00407

1. INTRODUCTION

Cells are the basic functional unit all organisms are made of and play significant roles in the different stages of life. Through various DNA and RNA sequencing data, researchers have a comprehensive and deep understanding of cell biology. However, traditional sequencing data is obtained from bulks of cells, and these are composed of the mixed effect of numerous cells and ignore cell heterogeneity. These bulk-seq data will lead to deviations in downstream analysis if a specific type of cell is expected. Recently, single-cell sequencing techniques have developed rapidly and make up the defect of bulk sequencing data. Although the single-cell sequencing technique cannot capture all cell information, it provides a great opportunity to reveal the characteristics of an individual cell.

The fundamental step of analyzing the single-cell data is to identify the cell types. Utilizing single-cell RNA-seq (scRNA-seq) data to obtain the cell clusters is one of the most efficient methods available. The amount of clustering methods on the basis of scRNA-seq data have been proposed. A group of methods are focused on calculating more accurate and robust similarity scores between cells. SNN-cliq (Xu and Su, 2015) constructed the distance matrix and counted the number of common neighbor cells for each pair of cells as the similarity scores and then incorporated these within a clique-based clustering method. Seurat (V3.0) was inspired by an SNN-cliq and applied the SNN graph with a louvain algorithm (Butler et al., 2018; Stuart et al., 2019). Seurat is one of the most widely used methods. SIMLR (Wang et al., 2017) and SC3 (Kiselev et al., 2017) adopted multiple similarity metrics from different aspects. In SIMILR, we could learn the inherent similarity matrix from a different resolution of Gaussian kernels, while SC3 combined multiple sub-clustering results together to build up a consensus matrix. Random forest (Pouyan and Kostka, 2018) was another way to calculate the similarity. The correlation coefficient has been proven to be effective when estimating the pairwise similarity of cells, and a high-order correlation coefficient was also applied in the scRNA-seq data analysis (Jiang et al., 2018; Tang et al., 2019). Compared to the methods based on pair-wise distance or correlation measurement, SinNLRR (Zheng et al., 2019b) considered the subspace characteristics of cells' expression and assumed the low rank and non-negative properties of the similarity matrix. Besides, several methods, including nonnegative matrix factorization (NMF) (Shao and Höfer, 2017; Zhu et al., 2017), imputation, and dimensionality reduction-based methods (Yau et al., 2016; Lin et al., 2017), have been used widely in assessing cellular heterogeneity. In the other aspect, the increasing number of well-learned scRNA-seq datasets also drives the appearance of supervised methods. These methods depended on labeled training datasets or some prior biological knowledge, such as gene markers (Wagner and Yanai, 2018; Pliner et al., 2019). According to the latest study (Abdelaal et al., 2019), most of the supervised methods are sensitive to prior knowledge, dataset complexity, or input features. Moreover, this kind of method has a fixed resolution and cannot find the detailed subtypes from a rough cell group. In this study, we have focused on the unsupervised clustering methods to identify the cell types. Inspired by previous methods, calculating the distance or similarity matrix of cells is a critical step. To recognize more accurate similarities of cells from high dimensional expression profiles, we have proposed an adaptive sparse subspace clustering method called AdaptiveSSC. AdaptiveSSC follows the subspace assumption and remains the nearest neighbors of a cell by a data-driven adaptive sparse constraint. The derived similarity matrix is used to obtain the clustering result and visualization. AdaptiveSSC obtains an improved performance on multiple experimental datasets.

2. MATERIALS AND METHODS

The pipeline of AdaptiveSSC is shown in **Figure 1**. Taking the scRNA-seq expression matrix as the input, AdaptiveSSC constructs the sparse cell-to-cell similarity matrix by keeping the most similar cells for each cell before then applying it to spectral clustering and modified t-distributed stochastic neighbor embedding (t-SNE) to obtain cell groups and the visualization result.

2.1. Data Pre-processing

The quantified scRNA-seq data contain thousands of genes, and the sparsity of gene expression is usually high. Therefore, AdaptiveSSC filters the genes expressed in <10% of the cells (the maximum number is 100), which are not regarded as informative genes. AdaptiveSSC investigates the linear effect of other cells on the target cell. To remove the scale of cells' expression, the L_2 normalization is carried on the original gene expression matrix.

$$X_{ij} = G_{ij} / \sqrt{\sum_{k=1}^{M} G_{kj}^2} \tag{1}$$

where G is the original expression matrix with M genes and N cells. The normalized matrix X is used in the following calculation.

2.2. Adaptive Sparse Subspace Clustering

Most clustering methods depend on the calculation of the similarity or distance matrix. The most popular similarity measurements include Euclidean distance, Pearson or Spearman correlations, and cosine similarity, which are all based on a pairwise estimation. The scRNA-seq data usually contains thousands of genes; however, only a part of a gene determines the cell type, which corresponds to a low-dimensional manifold surface. According to the common strategy in manifold learning, only the local measurement of similarity or distance is reliable, so previous scRNA-seq clustering methods (Xu and Su, 2015; Wang et al., 2017) usually apply k-nearest neighbors (KNN) to keep the locality. However, the KNN is used arbitrary to select the same number of neighbors for each cell, and the selection of k would have a great influence on the final result in some situations. In order to overcome these shortcomings, we propose an adaptive sparse subspace clustering method, which we have called AdaptiveSSC.

AdaptiveSSC is developed from sparse subspace clustering (SSC) methods. SSC is proposed to solve the motion segmentation and face clustering problems (Elhamifar and Vidal, 2013). SSC assumes that the feature vector of a sample can be expressed as the linear combination of other samples in the same subspace or type. Based on the assumption, the expression of a cell $X_i = c_1X_1 + c_2X_2 + \cdots + c_{i-1}X_{i-1} + c_{i+1}X_{i+1} + \cdots + c_NX_N$ and c_k is the subspace coefficient denoting the similarity score between cells. If the cell i and k are the same type, $c_k > 0$, otherwise it is 0. By adding l_1 term, the most similar cells lying in the same subspace are retained. Extending it to all cells, the calculation of the subspace coefficient matrix is defined as Equation (2):

$$min|C|_1 \quad s.t, \quad X = XC \quad \text{and} \quad diag(C) = 0 \tag{2}$$

where X is the normalized expression matrix. C is the coefficient matrix and C_{ij} denotes similarity between cell i and j. $|\cdot|_1$ denotes l_1 norm. The larger values in C mean the more similar cells. The relaxation formula of the optimization problem is shown:

$$min\frac{1}{2}||X - XC||_F^2 + \lambda|J|_1 \quad s.t, \quad diag(C) = 0 \quad \text{and} \quad C - J = 0 \tag{3}$$

where $||\cdot||_F^2$ means the Fresenius norm and λ is the l_1 penalty factor, which controls the sparsity of the coefficient matrix. J is an auxiliary matrix.

In the Equation (3), the coefficient matrix C is sensitive to the selection of the l_1 penalty factor. Another problem is that the same penalty factor for all coefficients will lead to the loss

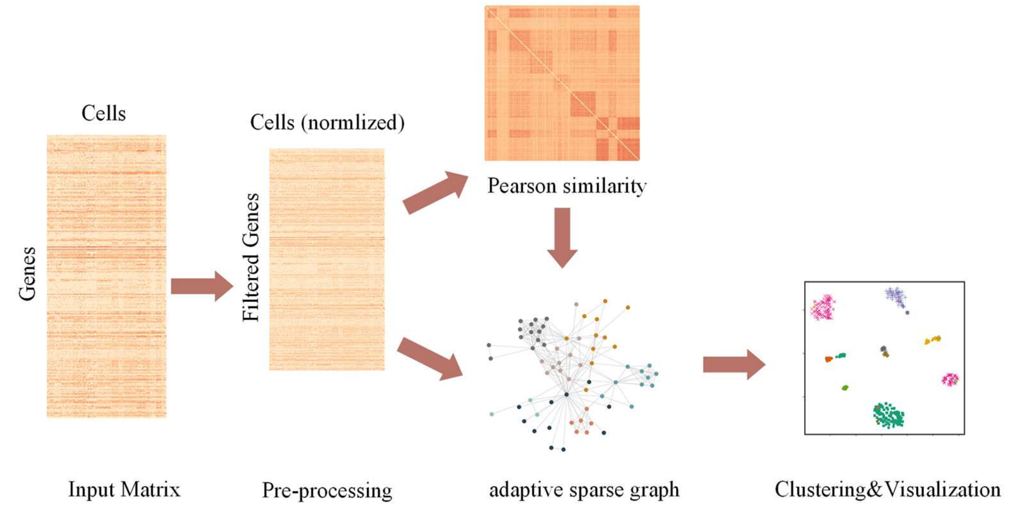

FIGURE 1 | The pipeline of AdaptiveSSC to identify and visualize cell types from scRNA-seq data.

of consistency between estimation and variable selection (Zou, 2006). Therefore, we have introduced a data-driven adaptive strategy to solve these problems. As a Pearson correlation has been proven to be effective when measuring the similarity in previous studies (Kiselev et al., 2017; Wang et al., 2017), we utilized it to adjust the penalty factor for each coefficient. If the correlation of two cells is high, the penalty factor is decreased and vice versa. The modified optimization problem is therefore defined:

$$min\frac{1}{2}||X-XC||_F^2+\lambda|\frac{J}{W}|_1\ s.t,\quad diag(C)=0\quad \text{and}\quad C-J=0$$
$$\text{where, } W_{ij}=\begin{cases}pearson(X_i,X_j)\ \text{If}\quad pearson(X_i,X_j)>0\\ 0\qquad\qquad\qquad\qquad \text{otherwise}\end{cases}\tag{4}$$

where $\frac{J}{W}$ means element division of matrix J and W. We set the negative value of the Pearson correlation to 0. Because only the trend of the expression of two cells are positively correlated, we regard them as similar cells. Some zero values in W would lead to zero values in J during the optimization.

Alternating direction method of multipliers (ADMM) (Boyd et al., 2011) is an efficient method to solve Equation (4). According to ADMM, the augmented Lagrangian formula is defined:

$$\iota_{\gamma,\lambda}\left(C,J,Y\right)=\frac{1}{2}||X-XC||_F^2+\lambda|\frac{J}{W}|_1+tr(Y^T\left(C-J\right))$$
$$+\frac{1}{2\gamma}||C-J||^2\quad \text{and}\quad diag(C)=0\tag{5}$$

where Y is a dual variable, γ is an augmented Lagrangian penalty parameter, and tr means the trace of the matrix. ADMM updates C, Y, or J by fixing others. In iteration $k+1$, the optimized form of C^{k+1}, J^{k+1}, and Y^{k+1} is shown in Equations (6–8):

$$C^{k+1}=\left(X^TX+\frac{1}{\gamma}I\right)^{-1}\left(X^TX+\frac{1}{\gamma}\left(J^k-Y^k\right)\right)$$
$$C^{k+1}=C^{k+1}-diag\left(C^{k+1}\right)\tag{6}$$

$$J^{k+1}=threshold_{\frac{\lambda}{W},\gamma}\left(C^{k+1}+Y^k\right)$$
$$=sign\left(C^{k+1}+Y^k\right)\cdot max\left(|C^{k+1}+Y^k|-\frac{\lambda}{\gamma W},0\right)\tag{7}$$
$$J^{k+1}=J^{k+1}-diag\left(J^{k+1}\right)$$

$$Y^{k+1}=Y^k+\frac{1}{\gamma}\left(C^{k+1}-J^{k+1}\right)\tag{8}$$

where $sign()$ means the sign function. The convergence of ADMM mainly includes primal residuals and dual residuals. On the basis of updating process, the penalty parameter γ affects the speed of convergence. In AdaptiveSSC, we apply a balance strategy (Boyd et al., 2011) between primal residuals and dual residuals to adjust γ. The setting of γ is shown:

$$\gamma_{k+1}=\begin{cases}\gamma_k/2, & \text{when}\quad ||r^k||_2>\mu||s^k||_2,\\ 2\gamma_k, & \text{when}\quad ||s^k||_2>\mu||r^k||_2,\\ \gamma_k, & \text{others.}\end{cases}\tag{9}$$

where $r^k=C^k-J^k$ is the primal residual and $s^k=\frac{1}{\gamma}\left(J^k-J^{k-1}\right)$ is the dual residual. The μ is set to 50 as default. To reduce the computational complexity, γ is updated by 10 iterations. When $max(abs(C-J))<0.0001$ or the number of iteration is larger

TABLE 1 | Single cell RNA-seq datasets.

Datasets	Cell number	Gene number	Techniques
Darmanis (Darmanis et al., 2015)	420	22,085	SMARTer
Kolod (Kolodziejczyk et al., 2015)	704	10,685	Smart-Seq2
Treutlein (Treutlein et al., 2014)	80	959	SMARTer
Yan (Yan et al., 2013)	90	20,214	Tang et al., 2011
Ting (Ting et al., 2014)	114	14,405	Single CTC RNA-Seq
Engel (Engel et al., 2016)	203	23,337	Smart-seq2
Kumar (Kumar et al., 2014)	361	11,497	SMARTer
Vento (Vento-Tormo et al., 2018)	5,418	33,693	Smart-seq2
Baron (Baron et al., 2016)	8,569	20,125	inDrop
Shekhar (Shekhar et al., 2016)	26,830	13,166	Drop-seq

than 200, this update process is finished. To keep the symmetry of the similarity matrix, the final similarity matrix $S = C^T + C$.

Finally, the spectral clustering (SC) (Von Luxburg, 2007) is applied on the learned similarity matrix. The SC is based on the point of graph cut and utilizes the characteristic of the corresponding Laplacian matrix to divide the graph into several clusters. In AdaptiveSSC, we use the normalized Laplacian matrix $L^{norm} = I - D^{-\frac{1}{2}} SD^{-\frac{1}{2}}$, where D is the degree matrix, to obtain its k eigenvectors corresponding to the smallest k eigenvalues. Then, k-means is used to obtain the final clusters.

3. RESULTS AND DISCUSSION

3.1. scRNA-seq Datasets

We collected 10 scRNA-seq datasets to evaluate the performance of AdaptiveSSC. These datasets are based on different single-cell techniques or protocols, such as Smart-seq, SMARTer, and Drop-seq based methods. Meanwhile, the scale of these datasets ranges from the tens to the tens of thousands. The variety of the datasets could indicate the generalization ability of AdaptiveSSC comprehensively. The details of these datasets are shown in **Table 1**. All datasets contain the real cell types from the original researches.

3.2. Evaluation Metrics

In order to compare the performance of different clustering methods, we selected two popular metrics: normalized mutual information (NMI) and adjusted rand index (ARI). Both NMI and ARI can quantify the consistency between the clustering results and the real labels. The definition of NMI and ARI is shown:

$$NMI\,(T,P) = \frac{I(T,P)}{\left[H(T) + H(P)\right]} \tag{10}$$

$$ARI\,(T,P) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}} \tag{11}$$

Where T and P mean the real labels and clustering labels, respectively. In Equation (11), n_{ij} denotes the number of cells belonging to i group in real labels and j group in clustering labels; n_i denotes the number of cells belonging to the i group in real labels, while n_j denotes the number of cells belonging to the j group in clustering labels.

3.3. Parameter Analysis

Although the adaptive strategy is used in AdaptiveSSC, there are still some hyperparameters to be set. The most important hyperparameter is the l_1 penalty factor λ. By the adaptive adjustment, the learned similarity matrix is not so sensitive to it. We evaluated the NMI and ARI of AdaptiveSSC on eight small datasets (smaller than 5,000 cells) with λ ranging from 0.01 to 0.19 and the interval set to 0.02. The results for eight small datasets are shown in **Figure 2**. Based on the result, when the λ was in the 0.01–0.05, both NMI and ARI were in the best range and were more stable. Therefore, we used $\lambda = 0.03$ as a default in AdaptiveSSC. During the experiment, we also found the optimal λ was not consistent for big datasets (in Baron is 0.01 and in Shekhar and Vento is 0.007). We recommend that users select the proper λ by grid searching with the following rule. If the corresponding sparsity of C is between 0.02 and 0.05, the λ should be selected. In Baron and Shekhar, we selected the corresponding λ with the sparsity of C is 0.03.

3.4. Comparison Analysis of Clustering Methods

To validate the effectiveness of AdaptiveSSC, we selected seven competitive methods, including SIMLR (Wang et al., 2017), MPSSC (Park and Zhao, 2018), SNN-cliq (Xu and Su, 2015), RAFSIL (Pouyan and Kostka, 2018), Seurat(V3.0) (Butler et al., 2018; Stuart et al., 2019), SinNLRR (Zheng et al., 2019b), and sparse subspace clustering (SSC) (Elhamifar and Vidal, 2013). All these methods are based on the construction of similarity matrix. SNN-cliq and Seurat recalculate the similarities based on their shared neighbors. SIMILR and MPSSC focus on the different resolution of Gaussian kernels, while RAFSIL applies random forest. SinNLRR is based on the subspace assumption with low rank constraint. The original SSC was selected as the baseline method. The results of NMI and ARI on 10 datasets are shown

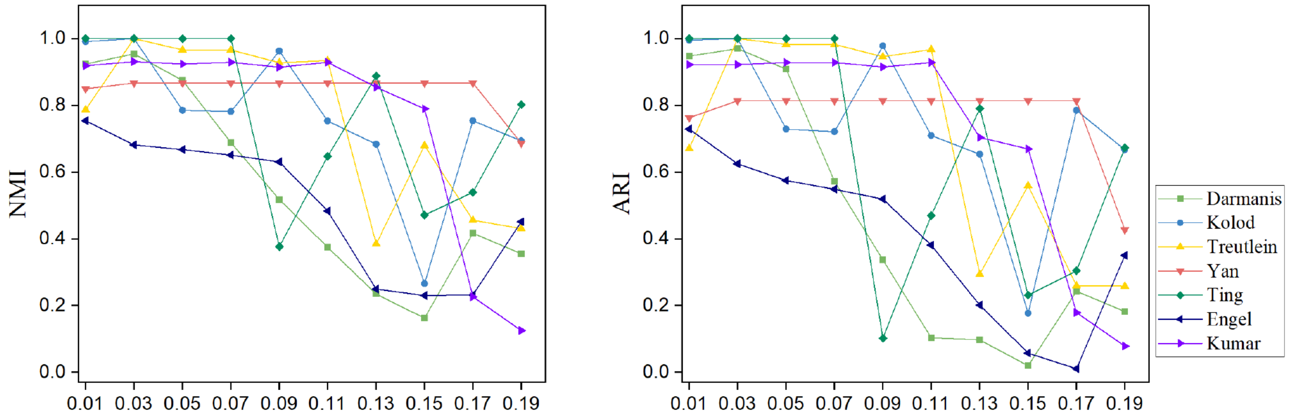

FIGURE 2 | The corresponding NMI and ARI with different values of λ on eight datasets.

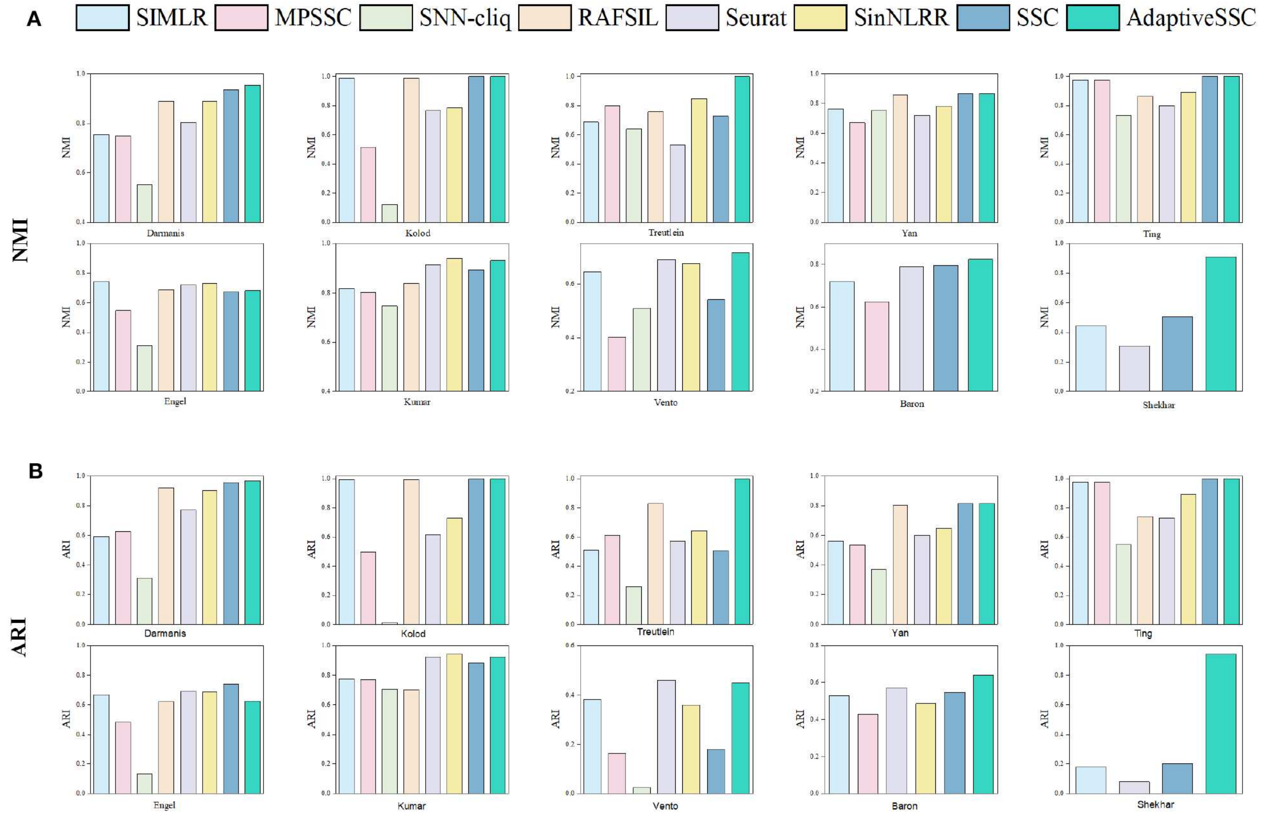

FIGURE 3 | The corresponding **(A)** NMI and **(B)** ARI of SIMLR, MPSSC, SNN-cliq, RAFSIL, Seurat, SinNLRR, SSC, and AdaptiveSSC on 10 datasets.

in **Figure 3**. Compared to SSC, AdaptiveSSC improved NMI and ARI in six datasets. Especially in Treutelin, Kumar, Vento, and Shekhar, AdaptiveSSC exhibited a significant improvement, more so than SSC, which means the adaptive penalty factor leads to the more accurate similarity matrix. In Kolod and Ting, AdaptiveSSC achieved the same performance with SSC. Overall, AdaptiveSCC exhibited a better performance than SSC in most cases. Besides, AdaptiveSSC achieved the best (or a tie for first place) performance in seven datasets upon NMI and eight datasets upon ARI compared with other six state-of-the-art methods. It is worth noting that only AdaptiveSSC obtains the perfect result on Treutelin. The results in Baron and Shekhar also verify AdaptiveSSC's effectiveness in large datasets. Estimation of the number of cell types is another important aspect in

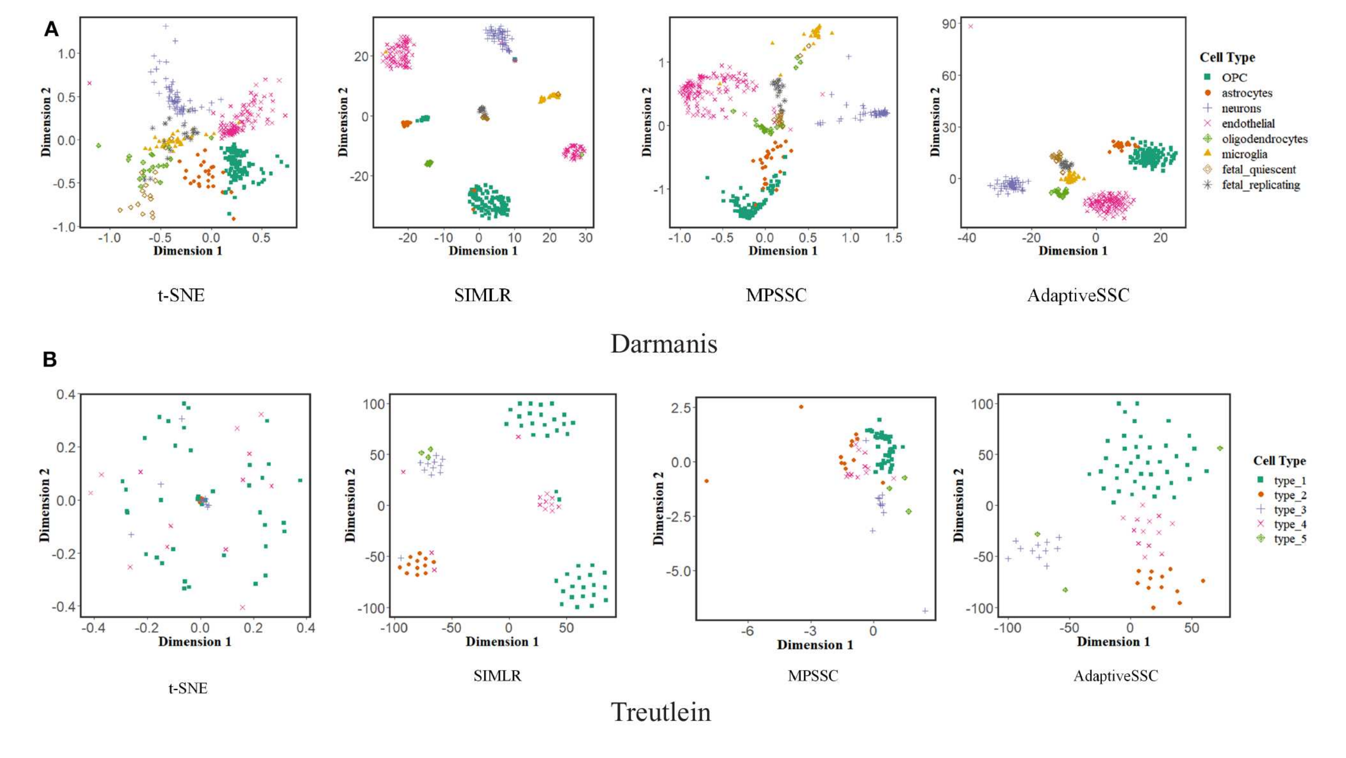

FIGURE 4 | The visualization of t-SNE, SIMLR, MPSSC, and AdaptiveSSC on **(A)** Darmanis and **(B)** Treutelin.

application. In AdaptiveSSC, we also used *eigengap* to determine the number of clusters, which was popular in previous studies. The results can be found in the **Supplementary Material**. As shown in the results, none of the methods predict the correct number of clusters in all datasets. However, AdaptiveSSC obtains the correct number of clusters in three datasets and gets the closest number in five datasets, which is a better selection overall. Moreover, we select five different scale datasets to evaluate the computational efficiency of these methods. The running time can be found in the **Supplementary Material**. AdaptiveSSC has a faster speed than SSC but is still time-consuming in large datasets compared with SIMLR and Seurat. All the experiments run on the server with 24 cores and 512 GB memory. The methods with running time more than 36 h are excluded, such as RAFSIL, SNN-cliq, and SinNLRR in large scale datasets, and MPSSC gets out of memory error on Shekhar.

3.5. Comparison Analysis of Visualization

Visualization of scRNA-seq is another important issue. Previous study (Wang et al., 2017) proposed a modified t-distributed stochastic neighbor embedding (t-SNE) to validate the performance of learned similarity. We also adopted this evaluation to AdaptiveSSC and generate 2D-embedding images on Darmanis and Treutelin with the learned similarity matrix of t-SNE, SIMLR, MPSSC, and AdaptiveSSC, respectively. The result is shown in **Figure 4**. The points with the same color mean they have the same cell type. Compared to other methods, AdaptiveSSC could group the same cells together and exhibits good silhouettes. Although SIMILR and MPSSC contain more dense parts, they divide cells with same type into different cliques, which are usually far away from each other. This will give the researchers a misconception that they are belong to exactly different types. Therefore, AdaptiveSSC has a better performance and potential in the visualization of scRNA-seq data.

3.6. Discussion and Conclusion

The identification of cell types is a fundamental problem is scRNA-seq data analysis. In recent years, a lot of clustering methods have been proposed to solve it. However, most of these methods do not exhibit a good generalization on different datasets. In this study, we proposed a subspace clustering with an adaptive sparse constraint, called AdaptiveSSC. AdaptiveSSC regards the expression of a cell can be expressed as a linear combination of other cell's expression from the same type. A data-driven adaptive sparse strategy is applied to keep the locality of cells in the original dimension and decrease the sensitivity to the penalty factor. Eight scRNA-seq datasets were used to evaluate the performance of AdaptiveSSC. By comparing with SSC, AdaptiveSSC improves the clustering results significantly in some cases, which indicates the effectiveness of our strategy. Moreover, six state-of-the-art methods were selected as comparison. From the NMI and ARI, AdaptiveSSC achieves the best performance in most of datasets. Finally, we integrated the learned similarity with modified t-SNE further, which also shows the powerful potential of AdaptiveSSC in visualization.

However, the computational efficiency of AdaptiveSSC is still low for large datasets and should be improved in the future. Some strategies used in the fast clustering method could be considered to make AdaptiveSSC more efficient (Ren et al., 2019). Moreover, AdaptiveSSC explores the cell heterogeneity from a gene level, but it is also important to study the different biological functions of cells. Regulatory modules (Aibar et al., 2017) have been proved effective when showing the functional heterogeneity of cells. It is possible to identify the cell type from the whole gene regulatory network perspective (Li et al., 2017; Zheng et al., 2018, 2019a). Besides, motivated by previous studies (Lan et al., 2018; Chen et al., 2019; Shi et al., 2019), multi-view learning and integrating with prior knowledge are promising directions to improve the accuracy of clustering and give a higher resolution of cell types.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/zrq0123/AdaptiveSSC.

AUTHOR CONTRIBUTIONS

RZ and CC designed the methodology. RZ, ZL, XC, and YT run the comparison experiments on datasets. RZ and ML wrote the paper. All authors revised and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61732009) and the Fundamental Research Funds for the Central Universities of Central South University (Nos. 2018zzts028, 2019zzts592).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00407/full#supplementary-material

REFERENCES

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J., et al. (2019). A comparison of automatic cell identification methods for single-cell RNA-sequencing data. *Genome Biol.* 20:194. doi: 10.1186/s13059-019-1795-z

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., et al. (2017). Scenic: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., 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. doi: 10.1016/j.cels.2016.08.011

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

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

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y.-P. P., et al. (2019). ILDMSF: Inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinformatics.* doi: 10.1109/TCBB.2019.2936476

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., 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. doi: 10.1073/pnas.1507125112

Elhamifar, E., and Vidal, R. (2013). Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2765–2781. doi: 10.1109/TPAMI.2013.57

Engel, I., Seumois, G., Chavez, L., Samaniego-Castruita, D., White, B., Chawla, A., et al. (2016). Innate-like functions of natural killer t cell subsets result from highly divergent gene programs. *Nat. Immunol.* 17, 728–739. doi: 10.1038/ni.3437

Jiang, H., Sohn, L. L., Huang, H., and Chen, L. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics* 34, 3684–3694. doi: 10.1093/bioinformatics/bty390

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-Seq data. *Nat. Methods* 14:483. doi: 10.1038/nmeth.4236

Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Ilicic, T., Henriksson, J., Natarajan, K. N., et al. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17, 471–485. doi: 10.1016/j.stem.2015.09.011

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., DaleyKeyser, A. J., Li, H., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516, 56–61. doi: 10.1038/nature13920

Lan, W., Wang, J., Li, M., Liu, J., Wu, F.-X., and Pan, Y. (2018). Predicting microRNA-disease associations based on improved microRNA and disease similarities. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 1774–1782. doi: 10.1109/TCBB.2016.2586190

Li, M., Zheng, R., Li, Y., Wu, F.-X., and Wang, J. (2017). MGT-SM: a method for constructing cellular signal transduction networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 417–424. doi: 10.1109/TCBB.2017.2705143

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

Park, S., and Zhao, H. (2018). Spectral clustering based on learning similarity matrix. *Bioinformatics* 34, 2069–2076. doi: 10.1093/bioinformatics/bty050

Pliner, H. A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* 16, 983–986. doi: 10.1038/s41592-019-0535-3

Pouyan, M. B., and Kostka, D. (2018). Random forest based similarity learning for single cell RNA sequencing data. *Bioinformatics* 34, i79–i88. doi: 10.1093/bioinformatics/bty260

Ren, X., Zheng, L., and Zhang, Z. (2019). SSCC: a novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data. *Genomics Proteomics Bioinformatics* 17, 201–210. doi: 10.1016/j.gpb.2018.10.003

Shao, C., and Höfer, T. (2017). Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* 33, 235–242. doi: 10.1093/bioinformatics/btw607

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166, 1308–1323. doi: 10.1016/j.cell.2016.07.054

Shi, Q., Zhang, C., Hu, B., and Zeng, T. (2019). Multi-view subspace clustering analysis for aggregating multiple heterogeneous omics data. *Front. Genet.* 10:744. doi: 10.3389/fgene.2019.00744

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. III, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. doi: 10.1016/j.cell.2019.05.031

Tang, F., Lao, K., and Surani, M. A. (2011). Development and applications of single-cell transcriptome analysis. *Nat. Methods* 8, S6–S11. doi: 10.1038/nmeth.1557

Tang, H., Zeng, T., and Chen, L. (2019). High-order correlation integration for single-cell or bulk RNA-seq data analysis. *Front. Genet.* 10:371. doi: 10.3389/fgene.2019.00371

Ting, D. T., Wittner, B. S., Ligorio, M., Jordan, N. V., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi: 10.1016/j.celrep.2014.08.029

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375. doi: 10.1038/nature 13173

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353. doi: 10.1038/s41586-018-0698-6

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. doi: 10.1007/s11222-007-9033-z

Wagner, F., and Yanai, I. (2018). Moana: A robust and scalable cell type classification framework for single-cell RNA-seq data. *bioRxiv [preprint]* 456129. doi: 10.1101/456129

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

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. *Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660

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

Zheng, R., Li, M., Chen, X., Wu, F.-X., Pan, Y., and Wang, J. (2018). Bixgboost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics* 35, 1893–1900. doi: 10.1093/bioinformatics/bty908

Zheng, R., Li, M., Chen, X., Zhao, S., Wu, F., Pan, Y., et al. (2019a). An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. doi: 10.1109/TCBB.2019.2900614

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

Zhu, X., Ching, T., Pan, X., Weissman, S. M., and Garmire, L. (2017). Detecting heterogeneity in single-cell RNA-seq data by non-negative matrix factorization. *PeerJ* 5:e2888. doi: 10.7717/peerj.2888

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101, 1418–1429. doi: 10.1198/016214506000000735

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.