CNLLRR: A Novel Low-Rank Representation Method for Single-cell RNA-seq Data Analysis

Na Yu¹, Jin-Xing Liu^{1*}, Ying-Lian Gao², Chun-Hou Zheng³, Junliang Shang¹, Hongmin Cai⁴

1 School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong, China

2 Qufu Normal University Library, Qufu Normal University, Rizhao, Shandong, China

3 Co-Innovation Center for Information Supply and Assurance Technology, Anhui University, Hefei, Anhui, China

4 School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China

* sdcavell@126.com

Abstract

The development of single-cell RNA-sequencing (scRNA-seq) technology has enabled the measurement of gene expression in individual cells. This provides an unprecedented opportunity to explore the biological mechanisms at the cellular level. However, existing scRNA-seq analysis methods are susceptible to noise and outliers or ignore the manifold structure inherent in the data. In this paper, a novel method called Cauchy non-negative Laplacian regularized low-rank representation (CNLLRR) is proposed to alleviate the above problem. Specifically, we employ the Cauchy loss function (CLF) instead of the conventional norm constraints in the noise matrix of CNLLRR, which will enhance the robustness of the method. In addition, graph regularization term is applied to the objective function, which can capture the paired geometric relationships between cells. Then, alternating direction method of multipliers (ADMM) is adopted to solve the optimization problem of CNLLRR. Finally, extensive experiments on scRNA-seq data reveal that the proposed CNLLRR method outperforms other state-of-the-art methods for cell clustering, cell visualization and prioritization of gene markers. CNLLRR contributes to understand the heterogeneity between cell populations in complex biological systems.

Author summary

Analysis of single-cell data can help to further study the heterogeneity and complexity of cell populations. The current analysis methods are mainly to learn the similarity between cells and cells. Then they use the clustering algorithm to perform cell clustering or downstream analysis on the obtained similarity matrix. Therefore, constructing accurate cell-to-cell similarity is crucial for single-cell data analysis. In this paper, we design a novel Cauchy non-negative Laplacian regularized low-rank representation (CNLLRR) method to get a better similarity matrix. Specifically, Cauchy loss function (CLF) constraint is applied to punish noise matrix, which will improve the robustness of CNLLRR to noise and outliers. Moreover, graph regularization term is applied to the objective function, which will effectively encode the

local manifold information of the data. Further, these will guarantee the quality of the cell-to-cell similarity matrix learned. Finally, single-cell data analysis experiments show that our method is superior to other representative methods.

Introduction

In recent years, a large number of single-cell RNA sequencing (scRNA-seq) data have been generated due to the development of next-generation sequencing technologies [1]. At the same time, scRNA-seq data contain a wealth of information on biological function and gene regulation. Analysis and research on these information pave us a way to observe individual cells unprecedentedly [2]. It thus provides us possibilities to explore the heterogeneity and complexity of cell population. It also offers us an unprecedented opportunity to learn the biological mechanisms and functional diversity at the cellular level.

With the help of scRNA sequencing, identification of subpopulation of cells [3] is now possible. The identification can be considered as a clustering problem through the similarities of gene expression. Those cells with high similarity will be categorized into the same subgroup. Cell clustering contributes to the targeted therapy of disease and the discovery of new cell subtypes. By using visualization methods, such as *t*-distributed stochastic neighbor embedding (t-SNE) [5], biologists can intuitively understand the distribution of the cells. Through prioritizing the genes that define the subgroup of cells, a new direction is provided for us to study cell differences.

Currently, many methods for single-cell clustering have been proposed [6, 7]. Generally speaking, these methods are mainly to learn the similarity between cells and cells. Then they use the clustering algorithm to perform cell clustering or downstream analysis on the obtained similarity matrix. Therefore, constructing accurate cell-to-cell similarity is crucial for scRNA-seq data analysis. Up to now, many efforts have been taken, for example, Xu et al. used the idea of shared nearest neighbors to measure the similarity between cells [8]. To enhance clustering accuracy, Wang et al. proposed the single-cell interpretation via multi-kernel learning (SIMLR) [9], which obtains a similarity matrix by learning the appropriate distance metric. Then Kim et al. designed a new framework based on improved SIMLR by using Pearson correlation as the measure of similarity [10]. Influenced by the assumption that neighbor information can measure the similarity, Jiang et al. proposed a method called Corr, which is based on differentiability correlation of cell pairs [11].

The aforementioned method relied on the accurate pairwise similarities among the cells. However, the similarity prones to be corrupted by noises or outliers. The high dimensional characteristics of the scRNA data further deteriorates the accuracy of the similarity calculation by involving an averaging operation. Alternatively, recent advances focus on capturing the global structure of the data. Zheng et al. proposed a subspace clustering algorithm called SinNLRR based on low-rank representation (LRR) to identify cell types [4]. Although the SinNLRR method achieves satisfactory experimental results, two issues still need to be considered. One is that single-cell data inevitably contain noise (for example, in the process of measuring and collecting scRNA-seq data) and outliers. However, the noise matrix of the SinNLRR method is constrained by the Frobenius norm, in which case its performance is affected. In addition, the common L_1 -norm, $L_{2,1}$ -norm or Frobenius norm is applied to the noise term, which can only model specific noise. In general, L_1 -norm can model the noise that matches the Laplacian distribution, $L_{2,1}$ -norm constraint is imposed to tackle sample-specific outliers, and Frobenius norm can well characterize Gaussian noise [12]. Unfortunately, noise and outliers are generally unknown and have complex statistical distributions. Therefore, a reasonable new strategy is needed to effectively deal with

noise and outliers. Another issue is that there are low-dimensional manifold structures in the high-dimensional sample space of data. If the manifold information is not considered, the similarity matrix learned will not better reflect the similarity among cells. This will seriously affect the results of single-cell analysis.

To overcome the above limitations, we design a novel Cauchy non-negative Laplacian regularized low-rank representation (CNLLRR) method for clustering cells, visualizing cells and prioritizing gene markers. Specifically, Cauchy loss function (CLF) constraint is applied to punish noise matrix. It can effectively reduce the influence of a single sample containing noise and outliers and avoid to be affected pronouncedly when the sample contains large noise. In addition, the influence function of CLF has an upper limit when estimating residuals, so it seldom relies on the distribution of noise and outliers [13]. Therefore, the CNLLRR method is robust to noise and outliers. Moreover, to preserve the intrinsic geometric local information in scRNA-seq data, the graph regularization term is applied to the objective function. The main contributions of this paper are as follows:

1. A novel method called CNLLRR is presented to capture the global structure of the data. CNLLRR uses the CLF instead of the conventional norm constraints on the noise term, which effectively reduces the influence of a single sample. Therefore, the robustness of the CNLLRR method to noise and outliers is improved. Moreover, the CNLLRR method also applies graph regularization terms to encode manifold information of the data. These will improve the quality of the learned similarity matrix.
2. The alternating direction method of multipliers (ADMM) is employed to optimize the CNLLRR method. And the robustness of our method is proved in theory and experiments.
3. We perform comprehensive experiments to validate the CNLLRR method on the scRNA-seq data. The experimental results show that our method is meaningful and superior to several representative methods.

The rest of this paper is organized as follows. In Materials and methods Section, LRR, CLF, and graph regularization are simply reviewed. The CNLLRR method and its optimization are also described in detail. In Results and discussion Section, the experimental results on scRNA-seq data are demonstrated. The paper is concluded in Conclusion Section.

Materials and methods

Low-rank representation

LRR is a widely used data analysis technique for exploring multiple low-dimensional subspace structures [14, 15]. In the field of bioinformatics, a scRNA-seq dataset can be represented by a matrix. The columns and rows of the matrix represent cells and the level of gene expression in different cells, respectively.

Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in R^{m \times n}$, where $\mathbf{x}_j$ can be considered as a linear combination of the bases in the dictionary [16]. Therefore, the data matrix plays the same role as the dictionary which helps to discover the potential affinity between data samples. LRR decomposes the data matrix into the low-rank representation of the data and the noise, the purpose of which is to search the lowest rank that can represent all the data [17]. Therefore, the objective function of the LRR can be expressed as the following rank minimization problem:

$$\min_{\mathbf{C}, \mathbf{E}} \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_l, \text{ s.t. } \mathbf{X} = \mathbf{XC} + \mathbf{E}, \quad (1)$$

where $\mathbf{C} \in R^{n \times n}$ and $\mathbf{E} \in R^{m \times n}$ represent the coefficient and the noise matrices, respectively. $\|\cdot\|_*$ is the nuclear norm of the matrix, which is defined as the sum of the singular values of the matrix. And $\|\cdot\|_l$ is some sort of regularization strategy, such as $\|\cdot\|_1$, $\|\cdot\|_{2,1}$, and $\|\cdot\|_F$. $\lambda \geq 0$ denotes the weighting parameter which is used to balance the two terms.

Recently, Zheng et al. proposed a modified version of LRR to detect cell types in scRNA-seq data [4]. If the expressions of cells lie in the same subspace, it is implied that these cells are most likely of the same type. Substituting $\mathbf{X} - \mathbf{XC}$ for $\mathbf{E}$, the objective function of SinNLRR is defined as

$$\min \frac{1}{2} \|\mathbf{X} - \mathbf{XC}\|_F^2 + \lambda \|\mathbf{C}\|_*, \text{ s.t. } \mathbf{C} \geq \mathbf{0}, \quad (2)$$

where $\|\cdot\|_F$ represents the Frobenius norm of matrix. $\mathbf{C}$ can be seen as a representation of $\mathbf{X}$, with $\mathbf{C} \geq \mathbf{0}$ meaning the non-negative similarity between cells. The penalty is adaptive due to the selection of λ .

Cauchy loss function (CLF)

In real-world applications, noise and outliers are common in the data. What is more, they are usually unknown and complex. Therefore, a robust loss function is urgently needed to deal with this problem. Recently, CLF, as a robust estimator, has been successfully used in face recognition [18] and image clustering [13]. CLF can also be applied in solving the minimum issue in the form of the following optimization problem:

$$\min \sum_j \rho(r_j), \quad (3)$$

where r_j is the residual between the j -th data and its estimated value. $\rho(\cdot)$ represents a symmetric and positive-definite function where there is a unique minimum at zero. The $\rho(\cdot)$ function in CLF is $\log(1 + (x/a)^2)$, hence the influence function of it can be calculated as

$$\phi(x) = \frac{\partial \rho(x)}{\partial (x)} = \frac{2x}{x^2 + a^2}, \quad (4)$$

where a is a constant. $\phi(\cdot)$ reflects the impact of sample points changes on the parameter estimation. It can be seen from Eq (4) that as the residual x increases, the influence function of CLF changes slightly. In other words, the influence function of the CLF has an upper bound when estimating the residual [13]. Therefore, using CLF as a constraint on the noise matrix in the model helps to improve the robustness of the method.

Graph regularization

Due to the rapid development of manifold learning, people pay more and more attention to the inherent nonlinear structure of data. Graph regularization is based on local invariance assumption that if two sample points are close to each other in the high-dimensional data space, they are also close in the new low-dimensional space [19]. Hence graph regularization can capture the potential local manifold structure in the original data. For a given data matrix $\mathbf{X}$, we use the data points as the vertices to construct a k -nearest neighbor (k -NN) graph [20]. Then, a symmetric weight matrix $\mathbf{W}$ is defined [21], whose element $\mathbf{W}_{ij}$ represents the weight of edge that joins the vertex $\mathbf{x}_i$ and the vertex $\mathbf{x}_j$. It can be calculated as

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathbf{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathbf{N}_k(\mathbf{x}_i), \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $\mathbf{N}_k(\mathbf{x}_i)$ represents the set of k nearest neighbors of $\mathbf{x}_i$. In mathematics, the graph regularization can be measured by

$$\begin{aligned} & \min_{\mathbf{C}} \sum_{i,j} \mathbf{W}_{ij} \|\mathbf{c}_i - \mathbf{c}_j\|^2 \\ &= \min_{\mathbf{C}} \text{tr}(\mathbf{C}^T (\mathbf{D} - \mathbf{W}) \mathbf{C}) \\ &= \min_{\mathbf{C}} \text{tr}(\mathbf{C}^T \mathbf{L} \mathbf{C}), \end{aligned} \quad (6)$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ denote the mappings of $\mathbf{x}_i$ and $\mathbf{x}_j$ under some transformations [22]. The degree matrix $\mathbf{D}$ is a diagonal matrix whose non-zero element $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. $\mathbf{L}$ is the graph Laplacian matrix of $\mathbf{X}$ [23].

Cauchy non-negative Laplacian regularized low-rank representation

In this section, we introduce the Cauchy non-negative Laplacian regularized low-rank representation (CNLLRR) method in detail. The CNLLRR algorithm is summarized in Algorithm 1.

Objective function

LRR has been successfully applied in many fields because it can effectively explore the global structure of data [22]. However, data are inevitably contaminated by noise and outliers in real world applications. To alleviate this issue, some methods are proposed as the variants of LRR by adding a regularization on the noise matrix. Nevertheless, noise and outliers are usually unknown and have complex statistical distributions. The L_1 -norm, the $L_{2,1}$ -norm and the Frobenius norm constraints are no longer suitable for most real world situations. Meanwhile, it is necessary to consider the low-dimensional manifold structure embedded in high-dimensional space in LRR-based methods.

To overcome the above limitations, a method called CNLLRR is proposed. CLF rather than conventional norm constraints is applied to penalize the noise matrix. CLF can effectively reduce the influence of a single sample containing noise and outliers. Moreover, it seldom relies on the distribution of noise and outliers [13] which helps to improve the robustness of the method. In addition, a graph regularization term is applied to encode the potential manifold information of data. In conclusion, the formulation of CNLLRR can be described as follows:

$$\min \sum_{i=1}^n \log \left(1 + \frac{\|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_F^2}{a^2} \right) + \lambda \|\mathbf{C}\|_* + \beta \text{tr}(\mathbf{C}^T \mathbf{L} \mathbf{C}), \quad s.t. \mathbf{C} \geq \mathbf{0}, \quad (7)$$

where λ and β denote penalty factors for balancing the regularization terms. $\mathbf{c}_i$ is a representation of the i -th data $\mathbf{x}_i$. Similar to the SinNLR method, we replace $\mathbf{X} - \mathbf{X}\mathbf{C}$ with the noise matrix $\mathbf{E}$. The CLF constraint is applied on it. To achieve the computational separability of Eq (7), an auxiliary matrix $\mathbf{J} \in R^{n \times n}$ is introduced. Then CNLLRR can be rewritten as the following optimization problem:

$$\min \log \left(1 + \frac{\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2}{a^2} \right) + \lambda \|\mathbf{J}\|_* + \beta \text{tr}(\mathbf{C}^T \mathbf{L} \mathbf{C}), \quad s.t. \mathbf{C} \geq \mathbf{0} \text{ and } \mathbf{J} - \mathbf{C} = \mathbf{0}. \quad (8)$$

Optimization of CNLLRR

Alternating direction method of multipliers (ADMM) is used to optimize the CNLLRR method. Then the augmented Lagrangian function of Eq (8) can be defined as follows:

$$L_{\mu}(\mathbf{C}, \mathbf{J}, \mathbf{Y}) = \log \left(1 + \frac{\|\mathbf{X} - \mathbf{XC}\|_F^2}{a^2} \right) + \lambda \|\mathbf{J}\|_* + \beta \text{tr}(\mathbf{C}^T \mathbf{LC}) + \langle \mathbf{Y}, \mathbf{J} - \mathbf{C} \rangle + \frac{\mu}{2} \|\mathbf{J} - \mathbf{C}\|_F^2. \quad (9)$$

By using $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T \mathbf{B})$, Eq (9) can be further rewritten as

$$L_{\mu}(\mathbf{C}, \mathbf{J}, \mathbf{Y}) = \log \left(1 + \frac{\|\mathbf{X} - \mathbf{XC}\|_F^2}{a^2} \right) + \lambda \|\mathbf{J}\|_* + \beta \text{tr}(\mathbf{C}^T \mathbf{LC}) + \frac{\mu}{2} \left\| \mathbf{J} - \mathbf{C} + \frac{\mathbf{Y}}{\mu} \right\|_F^2, \quad (10)$$

where $\mathbf{Y} \in R^{n \times n}$ represents the Lagrange multiplier. μ denotes the user-defined parameter [4]. We can easily solve the problem Eq (10) by alternately updating one variable while fixing other variables.

First, $\mathbf{J}^K$ and $\mathbf{Y}^K$ are fixed when solving $\mathbf{C}^{K+1}$. And $L_{\mu}(\mathbf{C}, \mathbf{J}^K, \mathbf{Y}^K)$ is optimized, we get

$$\mathbf{C}^{K+1} = \arg \min_{\mathbf{C}} \log \left(1 + \frac{\|\mathbf{X} - \mathbf{XC}\|_F^2}{a^2} \right) + \beta \text{tr}(\mathbf{C}^T \mathbf{LC}) + \frac{\mu}{2} \left\| \mathbf{J} - \mathbf{C} + \frac{\mathbf{Y}}{\mu} \right\|_F^2. \quad (11)$$

Through algebraic calculations, there exists an explicit solution of $\mathbf{C}$,

$$\mathbf{C}^{K+1} = (2Q^{K+1} \mathbf{X}^T \mathbf{X} + \mu \mathbf{I} + 2\beta \mathbf{L})^{-1} (2Q^{K+1} \mathbf{X}^T \mathbf{X} + \mu \mathbf{J}^K + \mathbf{Y}^K), \quad (12)$$

where $Q^{K+1} = (\|\mathbf{R}^{K+1}\|_F^2 + a^2)^{-1}$, $\mathbf{R}^{K+1} = \mathbf{X} - \mathbf{XC}^K$ and $\mathbf{I}$ denotes an identity matrix.

Similarly, the subproblem used to update $\mathbf{J}$ can be calculated as

$$\mathbf{J}^{K+1} = \arg \min_{\mathbf{J}} \lambda \|\mathbf{J}\|_* + \frac{\mu}{2} \left\| \mathbf{J} - \mathbf{C} + \frac{\mathbf{Y}}{\mu} \right\|_F^2. \quad (13)$$

It is a quadratic optimization with low-rank constraint. It has an explicit solution by soft-thresholding,

$$\mathbf{J}^{K+1} = \text{soft}_{\lambda, 1/\mu} \left(\mathbf{C}^{K+1} - \frac{\mathbf{Y}^K}{\mu} \right), \quad (14)$$

where $\text{soft}_{\lambda, 1/\mu}(\cdot)$ represents the soft-thresholding operator [24]. And $\text{soft}_{\lambda, 1/\mu}(\mathbf{A}) = \mathbf{U} D_{\lambda, 1/\mu}(\mathbf{\Sigma}) \mathbf{V}^T$, $\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$. The element on the diagonal of matrix $\mathbf{\Sigma}$ is σ_{ii} , $D_{\lambda, 1/\mu}(\mathbf{\Sigma}) = \text{diag}(\max(\sigma_{ii} - \lambda/\mu, 0))$. Then the solution of $\mathbf{Y}$ is given by

$$\mathbf{Y}^{K+1} = \mathbf{Y}^K + \mu(\mathbf{C}^{K+1} - \mathbf{J}^{K+1}). \quad (15)$$

Finally, the optimization procedure for solving the proposed CNLLRR method is presented in Algorithm 1.

Algorithm 1: CNLLRR

Data Input: scRNA-seq data matrix $\mathbf{X} \in R^{m \times n}$

Parameters: λ, β, a and μ

Initialization: $\mathbf{C}^0 = \mathbf{J}^0 = \mathbf{Y}^0 = \mathbf{0}$, $\max_{iter} = 200$, $tol = 10^{-4}$

Set $K = 0$

1) The graph Laplacian matrix is constructed by Eq (6).

2) While not converged and $K \leq \max_{iter}$

2.1) Update $\mathbf{R}^{K+1}$: $\mathbf{R}^{K+1} = \mathbf{X} - \mathbf{X}\mathbf{C}^K$;

2.2) Update Q^{K+1} : $Q^{K+1} = \left(\|\mathbf{R}^{K+1}\|_F^2 + a^2 \right)^{-1}$;

2.3) Update $\mathbf{C}^{K+1}$ by Eq (12);

2.4) Update $\mathbf{J}^{K+1}$ by Eq (14);

2.5) Update $\mathbf{Y}^{K+1}$ by Eq (15);

2.6) $K = K + 1$;

2.7) If the convergence condition is satisfied:

$\max |\mathbf{J}_{ij}^{K+1} - \mathbf{C}_{ij}^{K+1}| < tol$ and $\max |\mathbf{J}_{ij}^{K+1} - \mathbf{J}_{ij}^K| < tol$ and $\max |\mathbf{C}_{ij}^{K+1} - \mathbf{C}_{ij}^K| < tol$

End while

Output: $\mathbf{C}$, $\mathbf{J}$ and $\mathbf{Y}$

Robustness analysis

In this subsection, we analyze the robustness of the proposed CNLLRR method. From the viewpoint of statistics, using the robust estimator to measure errors is beneficial to enhance the robustness of the algorithm. Fig 1 shows the different estimators and corresponding influence functions. L_2 estimator, L_1 estimator and CLF are defined as $\rho_2(x) = x^2$, $\rho_1(x) = |x|$ and $\rho_c(x) = \log(1 + (x/a)^2)$, respectively. Their influence functions are represented as $\phi_2(x) = 2x$, $\phi_1(x) = \begin{cases} 1, x > 0 \\ -1, x < 0 \end{cases}$ and $\phi_c(x) = 2x/x^2 + a^2$, respectively. The influence function can reflect the impact of sample points changes on parameter estimation. In Fig 1(b), the influence of a sample point on the parameter estimation increases linearly with respect to its error for L_2 estimator. Therefore, L_2 estimator is sensitive to noise and outliers. From Fig 1, we can also observe that L_1 estimator can reduce the impact of large error on parameter estimation. However, with the error increasing, the influence function does not cut off. In Fig 1(b), for CLF, as the error increases, its influence function value tends to be zero. That is to say, the influence function of CLF has an upper limit. This is the reason why the CNLLRR method is robust to noise and outliers.

Fig 1. Explanation of the different estimators and corresponding influence functions. (a) Estimators. (b) Influence functions.

Next, we design an experiment to further illustrate the robustness of CNLLRR. Another method SinNLRR is also used for comparison. The experiment is conducted on a synthetic dataset consisting of 200 two-dimensional data points and the results are plotted in Fig 2. The subspace learning ability of the two methods on clean synthetic dataset can be revealed by Fig 2(a). From Fig 2(b) to 2(d), there are 20, 120, 150 randomly contaminated data points in the synthetic dataset, respectively. It can be seen that the subspace recovering ability of SinNLRR is severely limited with the number of contaminated points increasing. In contrast, CNLLRR can still explore the subspace structure successfully, even in the case of extreme noise points. In conclusion, our proposed CNLLRR method is more robust than SinNLRR.

Fig 2. Robustness comparison of CNLLRR and SinNLRR on the synthetic dataset. (a) Clean dataset. (b) 20 contaminated data points. (c) 120 contaminated data points. (d) 150 contaminated data points.

Convergence and complexity analysis

We use the ADMM method to optimize the proposed CNLLRR method. As shown in Algorithm 1, $\mathbf{C}$, $\mathbf{J}$, and $\mathbf{Y}$ are iteratively updated until the convergence condition of CNLLRR is satisfied or the number of iterations exceeds the maximum we set. Therefore, the convergence of our method can be guaranteed. We also performed experiments to verify the convergence of the CNLLRR method. In Fig 3, the x-axis and y-axis represents the number of iterations and the error value, respectively. We can observe that as the number of iterations increases, the error value decreases gradually. This proves that the CNLLRR method has the property of fast convergence on all datasets. In addition, O notation is used to characterize the computational complexity. The computational cost of CNLLRR is mainly produced in the procedure of updating $\mathbf{C}$ and $\mathbf{J}$. Furthermore, each time $\mathbf{C}$ is updated, $\mathbf{R}$ and $\mathbf{Q}$ also need to be calculated. Hence the time complexity of updating $\mathbf{C}$ can be calculated as $O(n^3 + mn^2)$ according to Eq (12). As shown in Eq (14), singular value decomposition (SVD) needs to be employed when updating $\mathbf{J}$ whose computational cost is $O(n^3)$. Moreover, data graph also requires $O(mn^2)$ to be built. Assuming that k is the number of iterations of the algorithm, the total computational complexity of CNLLRR is $O(kn^3 + kmn^2 + mn^2)$.

Fig 3. Convergence curves of CNLLRR on ten datasets.

Framework of CNLLRR

In this subsection, we use Fig 4 to describe the entire framework of CNLLRR in detail. Fig 4 can be seen as consisting of two main steps.

Fig 4. The entire framework of CNLLRR.

The first step is the adoption of the CNLLRR method. Specifically, preprocessing is needed for the input data matrix. The preprocessing includes gene filtering [25] and L_2 -norm normalization [26]. Gene filtering is commonly used in single-cell analysis, which aims to filter out the 5% genes that expresses less than all cells. L_2 -norm normalization normalizes each column vector of the matrix to 1, ie., $\mathbf{X}_{ij} = \mathbf{X}_{ij} / \sqrt{\sum_i \mathbf{X}_{ij}^2}$. It can eliminate the scale differences between samples. After preprocessing, we use data matrix as the input of CNLLRR method to get the corresponding coefficient matrix $\mathbf{C}$. Then we obtain the similarity matrix which has a better block diagonal structure in the same way as SinNLRR [4]. The localized similarity matrix is defined as

$$\mathbf{P}_{ij} = \begin{cases} \mathbf{C}_{ij} & \text{if } \mathbf{C}_{ij} > (\mathbf{C}_{ii}/f), \\ 0 & \text{otherwise,} \end{cases} \quad (16)$$

where f denotes the relaxation coefficient. Like SinNLRR, the value of f is also 1.5 in our method. The similarity matrix can be easily obtained through $\mathbf{S} = \mathbf{P}^T + \mathbf{P}$.

The second step is the single-cell analysis. We perform spectral clustering on the learned similarity matrix. In addition, the similarity matrix is visualized and gene markers are prioritized.

Results and discussion

In this section, clustering cells, visualizing cells and prioritizing gene markers are performed to analyze the performance of the CNLLRR method. In addition, t-SNE [5], K-means, PCA [27], sparse subspace clustering (SSC) [28], SIMLR [9], Corr [11], and SinNLRR [4] are used as comparison methods to verify the effectiveness of the proposed method.

Datasets

With the launch of the Human Cell Atlas program, high-throughput single-cell sequencing technology has developed rapidly. This allows us to study gene expression in individual cells due to the large amount of valuable and unpredictable information from Single-cell data generated by sequencing technology. Studies in these information offer us a new insight into the learning of biological mechanisms at cellular level. In this paper, we analyze the performance of CNLLRR on scRNA-seq data, including Goolam [29], Pollen [30], Treutlein [31], Deng [32], Ting [33], Breton [34], Engel3 [35], Grover [36], Engel4 [35], and Darmanis [37]. The Engel dataset can be divided into 3 or 4 categories based on cell types, and they are represented as Engel3 and Engel4, respectively. In the experiment, gene filtering and L_2 -norm normalization are utilized to preprocess the data. Specific information on scRNA-seq data is listed in Table 1.

Table 1. Summary of ten scRNA-seq data.

Datasets	Cells	Genes	Cell types	Species
Goolam	124	40315	5	Mus musculus
Pollen	249	14805	11	Homo sapiens
Treutlein	80	959	5	Mus musculus
Deng	135	12548	7	Mus musculus
Ting	114	14405	5	Mus musculus
Breton	957	20689	4	Homo sapiens
Engel3	203	21690	3	Homo sapiens
Grover	135	14739	2	Mus musculus
Engel4	203	23337	4	Homo sapiens
Darmanis	420	22085	8	Homo sapiens

Parameter setting

In fact, the appropriate parameter setting will contribute to the performance of our method. There are four parameters in the proposed CNLLRR, which are the penalty factors λ and β , the constant a , the user-defined parameter μ . We use the grid search algorithm to select the optimal parameters. The values of λ and β are set in intervals $[0.1, 1.3]$ and $[0.001, 100]$, respectively. It should be noted that the value of β was changed in the Breton dataset from $\{3, 4, 5, 6, 7\}$. As for a and μ , the variation ranges of them are $[0.001, 1]$ and $[6, 11]$, respectively. Fig 5, Fig 6, and Fig 7 demonstrate the performance of CNLLRR with different λ , β , a , and μ on ten scRNA-seq datasets, respectively. From Fig 5 and Fig 6, we can observe that the proposed method is sensitive to parameters λ and β . Fortunately, λ and β can be selected to achieve better performance in the aforementioned interval. We can find in Fig 7(a) and 7(b) that as the values of a and μ increase, the behavior of CNLLRR is relatively stable. That is to say, the proposed method is robust to these two parameters. From Fig 7(b), we can observe that all ten datasets can achieve satisfactory clustering results. For convenience,

we set the value of μ to 10 in the following experiment. In conclusion, reasonable parameter values can be selected from Fig 5 to Fig 7.

Fig 5. Performance of the CNLLRR set with different values of λ .

Fig 6. Performance of the CNLLRR set with different values of β .

Fig 7. Performance of the CNLLRR versus parameters a and μ . (a) The different values of a . (b) The different values of μ .

Clustering results

In this section, we perform clustering experiments to illustrate the effectiveness of the proposed method on scRNA-seq data. The spectral clustering algorithm is performed on the similarity matrix, which can divide cells into different subgroups according to cell types.

Evaluation metrics

To quantify the clustering results, we employ two widely used evaluation metrics: Adjusted Rand Index (ARI) [38] and Normalized Mutual Information (NMI) [39]. ARI is more suitable for sample-balanced datasets compared with NMI [40]. Let $T = \{T_1, T_2, \dots, T_K\}$ and $P = \{P_1, P_2, \dots, P_K\}$ represent the ground truth cluster set and the cluster set predicted by the clustering method, respectively. The ARI can be computed by

$$ARI(T, P) = \frac{\sum_{ij} \binom{N_{ij}}{2} - \left[\sum_i \binom{N_{i.}}{2} \sum_j \binom{N_{.j}}{2} \right] / \binom{N}{2}}{1/2 \left[\sum_i \binom{N_{i.}}{2} + \sum_j \binom{N_{.j}}{2} \right] - \left[\sum_i \binom{N_{i.}}{2} \sum_j \binom{N_{.j}}{2} \right] / \binom{N}{2}}, \quad (17)$$

where N_{ij} represents the number of sample points in T_i and P_j . The number of sample points in T_i is represented by $N_{i.}$. And the number of sample points in P_j is denoted by $N_{.j}$. $\binom{a}{b}$ means the binomial coefficient.

NMI is another metric that represents the similarity of cluster sets T and P . It is formulated as:

$$NMI(T, P) = \frac{2MI(T, P)}{H(T) + H(P)}, \quad (18)$$

where $H(\cdot)$ denote the entropy of the cluster set and $MI(\cdot, \cdot)$ represents mutual information between cluster sets. Both ARI and NMI can be used to measure the consistency between the two data distributions. The domains of values of ARI and NMI are $[-1, 1]$ and $[0, 1]$, respectively. The larger the value, the more consistent the clustering results with the real situation.

Comparison of clustering performance

To evidence the effectiveness of CNLLRR, we perform clustering experiments on the scRNA-seq data. To make a fair comparison, CNLLRR and the other seven methods

are provided with real cluster sets, even though the Corr itself can automatically estimate the number of clusters. In our experiment, ARI and NMI are adopted to evaluate the clustering performance. The experimental results are illustrated in Table 2.

Table 2. The clustering performance on the scRNA-seq data.

Methods	Adjusted Rand Index (%)							
	t-SNE	PCA	K-means	SSC	SIMLR	Corr	SinNLRR	CNLLRR
Goolam	52.55	45.94	41.82	44.41	29.91	30.46	90.97	97.22
Pollen	80.55	88.86	83.78	92.92	94.15	48.03	90.51	93.89
Treutlein	54.73	57.27	61.72	52.42	51.14	68.22	64.19	79.79
Deng	53.01	48.15	49.14	38.04	45.65	47.53	47.06	65.33
Ting	63.84	87.61	85.67	97.84	98.03	63.02	89.43	94.46
Breton	9.02	3.96	4.11	3.97	4.42	1.61	14.32	17.32
Engel3	27.04	50.57	47.32	54.05	27.79	11.15	59.44	67.49
Grover	27.12	27.12	27.12	18.49	9.46	10.55	28.71	35.52
Engel4	57.25	32.78	34.53	52.02	66.82	43.77	65.33	77.01
Darmanis	57.25	32.78	34.53	52.02	39.82	61.83	60.57	69.90

Methods	Normalized Mutual Information (%)							
	t-SNE	PCA	K-means	SSC	SIMLR	Corr	SinNLRR	CNLLRR
Goolam	70.43	62.53	56.86	58.40	56.93	57.29	88.85	93.94
Pollen	91.90	92.34	91.42	94.46	94.83	88.24	93.15	94.94
Treutlein	73.46	75.30	71.57	70.22	68.81	74.46	85.08	85.85
Deng	71.55	72.70	71.74	65.79	74.22	68.90	73.89	78.64
Ting	77.68	89.44	88.13	97.55	97.54	79.95	89.05	92.91
Breton	23.70	14.67	14.86	16.72	20.97	30.36	32.73	25.27
Engel3	34.45	53.19	49.16	56.79	40.38	24.34	60.26	68.02
Grover	21.97	21.25	20.80	13.82	7.11	16.34	22.40	30.16
Engel4	60.21	44.45	46.54	28.71	74.13	52.06	69.32	78.92
Darmanis	60.21	44.45	46.54	58.71	60.71	78.24	75.13	77.68

Based on Table 2, we can draw the following conclusions:

- 1) In the ten single-cell datasets, SinNLRR outperforms SIMLR and Corr by about 14% and 9% in terms of ARI and NMI. The CNLLRR is also superior to the SIMLR and Corr by around 23% and 13% on the metrics of ARI and NMI, respectively. The reason is that the LRR-based variants SinNLRR and CNLLRR capture the global structure of the data, while SIMLR and Corr only consider the similarity of the pair-wise data samples. This suggests that the LRR-based approach is more suitable for single-cell clustering due to its ability to capture complex relationships among samples.
- 2) From Table 2, we can observe that CNLLRR method exceeds SinNLRR method by approximately 9% and 4% on the AC and NMI metrics, respectively. There are mainly two reasons. First, using CLF instead of conventional norms on the error term in CNLLRR enable it with better robustness. Second, the graph regularization term applied in the proposed method explores local manifold information from data. On the account, CNLLRR has satisfactory clustering performance.
- 3) Basic simple clustering methods such as t-SNE, PCA, K-means, and SSC can achieve acceptable performances. For example, the average ARI score of SSC is 4% higher than of SIMLR and 12% higher than of Corr on ten single-cell data sets. This implies that improvements on traditional methods can not always promote

clustering performance. For instance, the useful information may loss when naively modifying these methods, which can affect the clustering results.

- 4) It can be seen from Table 2 that our CNLLRR method can achieve best performance on most of datasets. Compared with other methods, the average scores of ARI and NMI of CNLLRR have approximately 9% and 4% improvements. We can conclude that our method is reasonable and effective in suppressing noise, outliers and preserving potential geometric information for the data.

In real-world applications, the number of clusters is usually unknown. To tackle this issue, we design an experiment to confirm the number of clusters in the scRNA-seq dataset. It is well known that SIMLR, Corr and SinNLRR are clustering methods proposed for the specific purpose of single-cell cluster analysis. Hence we use the three methods as comparison methods. In the experiment, we first construct the normalized Laplacian matrices $\mathbf{L}_{norm}$ which are based on the corresponding similarity matrices $\mathbf{S}$ obtained by each method. $\mathbf{L}$ is defined as

$$\mathbf{L}_{norm} = \mathbf{I} - \mathbf{B}^{-\frac{1}{2}} \mathbf{S} \mathbf{B}^{\frac{1}{2}}, \quad (19)$$

where $\mathbf{B}$ represents the diagonal matrix and the elements on its diagonal are $\mathbf{B}_{ii} = \sum_j \mathbf{S}_{ij}$. $\mathbf{I}$ is the identity matrix. The eigengap [41] method is then applied on $\mathbf{L}_{norm}$, which can estimate the number of clusters by maximizing the eigenvalues gap $|\lambda_k - \lambda_{k-1}|$ of $\mathbf{L}_{norm}$. It should be mentioned that Corr decides the cluster number by itself automatically. Table 3 lists the comparison results of the four methods. We can see that CNLLRR concludes the same cluster number as the real cell types on the Treutlein, Ting and Engel3 datasets. In the remaining seven single-cell datasets, it obtains approximate cluster numbers as the number of real clusters. Although the performance of determining cluster numbers by all these methods is not satisfactory, the CNLLRR method is generally superior. It provides another perspective to validate the better clustering performance of our method.

Table 3. Comparison of the number of clusters estimated by the four methods.

Datasets	Cell types	SIMLR	Corr	SinNLRR	CNLLRR
Goolam	5	3	3	5	3
Pollen	11	19	3	10	10
Treutlein	5	12	3	4	5
Deng	7	6	4	5	5
Ting	5	51	5	5	5
Breton	4	2	3	6	2
Engel3	3	2	3	5	3
Grover	2	4	6	4	4
Engel4	4	2	2	6	2
Darmanis	8	2	6	10	7

Visualization and gene markers

Visualize cells using t-SNE

In scRNA-seq data analysis, data visualization is also commonly used in identifying cell groups [42]. As a dimensionality reduction method, t-SNE has become a mainstream tool of visualization. According to [9], the modified t-SNE algorithm projects the input similarity matrix obtained by CNLLRR into a two-dimensional space, which shows the

distinctions between cell subpopulations more intuitively and help biologists understand the diversity of cells more simply and directly.

Due to the limitation of space, we choose Engel4 and Darmanis datasets as the instances for data visualization to illustrate the performance of our method. Engel4 data are obtained from the thymus natural killer T (NTK) cells of female mice at 5 weeks [35]. It consists of 45 NKT0 cells, 46 NKT1 cells, 44 NKT17 cells and 68 NKT2 cells. Darmanis data contain 420 cells from adult and fetal cerebral cortical tissue [37]. Specifically, it has 110 fetal_quiescent neurons, 25 fetal_replicating neurons, 62 astrocytes, 131 neurons, 20 endothelial, 38 oligodendrocytes 16 microglia and 18 oligodendrocyte precursor (OPC) cells. Fig 8 shows the two-dimensional cell visualization results for t-SNE, SIMLR, SinNLRR, and CNLLRR on the Engel4 and Darmanis datasets. It can be seen from Fig 8(a) and 8(b) that t-SNE has the worst visualization effect. The clustering sets of the SIMLR method are more compact on the Engel4 and Darmanis datasets. This is because it requires the real cell labels to get the similarity matrix, whereas other methods do not. SIMLR always visualizes the same class of cells as different ones. For example, in Engel4 dataset, NTK0, NKT1, and NKT2 are not well distinguished. In contrast, as Fig 8 shows, the visualization result of CNLLRR is superior to SinNLRR and CNLLRR has better grouping results than most other methods. This above confirms the effectiveness of the CNLLRR method.

Fig 8. Comparison of cell visualization of the four methods. (a) Engel4 dataset. (b) Darmanis dataset.

Gene markers prioritization

The prioritization of gene markers has received intensive attentions since it was proposed. Cell gene markers which contain a wealth of valuable biological information are helpful in distinguishing cell subgroups and reveal cell complexity. The in-depth mining of biological information is of great significance in exploring cellular heterogeneity and life mechanisms. In our method, we use bootstrapped Laplacian score to identify gene markers on the similarity matrix obtained by CNLLRR [9]. The gene markers are then sorted in descending order corresponding to the importance of gene in distinguishing subgroups of cells.

Fig 9 demonstrates the top 10 gene markers in Engel4 and Darmanis datasets, and indicate their expression levels in different colors. The x-axis and y-axis represent the gene name and the cell type, respectively. As can be seen from Fig 9(a), the top 10 genes highly expressed in NKT2. REXO2, HMGB2 and CCT3 can be regarded as marker genes of NKT17, NKT1 and NKT0, respectively. REXO2 is a protein coding gene which may have a role in DNA repair, replication, and recombination, and in RNA processing and degradation. Therefore, the lack of REXO2 leads to a significant decrease in mitochondrial nucleic acid content [43]. HMGB2 encodes a member of non-histone chromosomal high mobility group protein family. The proteins of this family are chromatin-associated and distributed in the nucleus of higher eukaryotic cells ubiquitously [44]. The relevant pathways of CCT3 are cell cycle_role of APC in cell cycle regulation and organelle biogenesis and maintenance. The protein encoded by this gene is a molecular chaperone that is a member of the chaperonin containing TCP1 complex (CCT), also known as the TCP1 ring complex (TRiC) [45]. In Fig 9(b), PLP1, TMEM144 and CLDND1 are the gene markers of oligodendrocytes. MALAT1 can be regarded as a gene marker of neurons. The gene markers of astrocytes are SLC1A2, SLC1A3, AQP4 and SPARCL1. MAP1B and TUBA1A are highly expressed on fetal_quiescent neurons. The protein encoded by PLP1 may play a role in the compaction, stabilization, and maintenance of myelin sheaths, as well as in

oligodendrocyte development and axonal survival. The protein encoded by AQP4 is the predominant aquaporin found in brain and has an important role in brain water homeostasis. Published articles further confirmed that PLP1 and AQP4 are the marker genes of oligodendrocytes and astrocytes, respectively [46]. To save space, we perform a brief analysis of some of the gene markers in the Engel4 and Darmanis datasets. The experiments and analysis provide new insights for researchers into the studies of NTK cells and cerebral cortical tissue.

Fig 9. The top 10 gene markers of the two datasets. (a) Engel4 dataset. (b) Darmanis dataset.

Conclusion

In this paper, we propose a robust LRR method called Cauchy non-negative Laplacian regularized LRR (CNLLRR) for scRNA-seq analysis. The proposed CNLLRR method uses CLF to reduce the negative impact of noise and outliers in clustering. Moreover, the graph regularization was applied to exploit the local manifold structure of data. Extensive experiments compared with other state-of-art methods on scRNA-seq data demonstrate the effectiveness of our proposed CNLLRR method. This method may provide a guiding for the next single-cell analysis and be applied to other relative fields.

For future work, we will make further efforts in model simplifying and application in larger dataset. In our model, the optimal solution is not easy to be obtained owing to the four free parameters. Another shortcoming of CNLLRR is the lack of flexibility when analyzing a large number of cells. Thus, we will continue to develop the Cauchy non-negative Laplacian regularized LRR method and promote its flexibility in the application of scRNA-seq analysis.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61872220, 61873001, and 61972226.

References

1. Shao C, Höfer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics*. 2017;33(2):235–242.
2. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*. 2013;14(9):618.
3. Park S, Zhao H. Spectral clustering based on learning similarity matrix. *Bioinformatics*. 2018;34(12):2069–2076.
4. Zheng R, Li M, Liang Z, Wu F-X, Pan Y, Wang J. SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics*. 2019. doi: 10.1093/bioinformatics/btz139.
5. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*. 2008;9(Nov):2579–2605.

6. Kanter I, Dalerba P, Kalisky T. A cluster robustness score for identifying cell subpopulations in single cell gene expression datasets from heterogeneous tissues and tumors. *Bioinformatics*. 2018;35(6):962–971.
7. Durif G, Modolo L, Mold JE, Lambert-Lacroix S, Picard F. Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis. In: *RECOMB*. 2018; P. 254–255.
8. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*. 2015;31(12):1974–1980.
9. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature methods*. 2017;14(4):414.
10. Kim T, Chen IR, Lin Y, Wang AY-Y, Yang JYH, Yang P. Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in bioinformatics*. 2018. doi: 10.1093/bib/bby076.
11. Jiang H, Sohn LL, Huang H, Chen L. Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics*. 2018;34(21):3684–3694.
12. Wei L, Wang X, Wu A, Zhou R, Zhu C. Robust subspace segmentation by self-representation constrained low-rank representation. *Neural Processing Letters*. 2018;48(3):1671–1691.
13. Li X, Lu Q, Dong Y, Tao D. Robust subspace clustering by cauchy loss function. *IEEE transactions on neural networks and learning systems*. 2019;30(7):2067–2078.
14. Cui G, Li X, Dong Y. Subspace clustering guided convex nonnegative matrix factorization. *Neurocomputing*. 2018;292:38–48.
15. Wang Y-X, Liu J-X, Gao Y-L, Zheng C-H, Shang J-L. Differentially expressed genes selection via Laplacian regularized low-rank representation method. *Computational biology and chemistry*. 2016;65:185–192.
16. Liu G, Lin Z, Yu Y. Robust subspace segmentation by low-rank representation. In: *ICML*. 2010; P. 8.
17. Xu X-X, Gao Y-L, Liu J-X, Wang Y-X, Dai L-Y, Kong X-Z, et al. A novel low-rank representation method for identifying differentially expressed genes. *International Journal of Data Mining and Bioinformatics*. 2017;19(3):185–201.
18. Guan N, Liu T, Zhang Y, Tao D, Davis LS. Truncated Cauchy non-negative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017;41(1):246–259.
19. Yi Y, Wang J, Zhou W, Zheng C, Kong J, Qiao S. Non-Negative matrix factorization with locality constrained adaptive graph. *IEEE Transactions on Circuits and Systems for Video Technology*. 2019. doi: 10.1109/TCSVT.2019.2892971.
20. Li X, Cui G, Dong Y. Refined-graph regularization-based nonnegative matrix factorization. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2017;9(1):1.

21. Cai D, He X, Han J, Huang TS. Graph Regularized Non-Negative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2011;33(8):1548–1560.
22. Yin M, Gao J, Lin Z. Laplacian regularized low-rank representation and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015;38(3):504–517.
23. Liu X, Zhai D, Zhao D, Zhai G, Gao W. Progressive image denoising through hybrid graph Laplacian regularization: A unified framework. *IEEE Transactions on image processing*. 2014;23(4):1491–1503.
24. Cai J-F, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*. 2010;20(4):1956–1982.
25. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods*. 2017;14(5):483.
26. Vidal R, Favaro P. Low rank subspace clustering (LRSC). *Pattern Recognition Letters*. 2014;43:47–61.
27. Abdi H, Williams LJ. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*. 2010;2(4):433–459.
28. Elhamifar E, Vidal R. Sparse subspace clustering. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009; P. 2790–2797.
29. Goolam M, Scialdone A, Graham SJ, Macaulay IC, Jedrusik A, Hupalowska A, et al. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*. 2016;165(1):61–74.
30. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*. 2014;32(10):1053.
31. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*. 2014;509(7500):371.
32. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*. 2014;343(6167):193–196.
33. Ting DT, Wittner BS, Ligorio M, Jordan NV, Shah AM, Miyamoto DT, et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell reports*. 2014;8(6):1905–1918.
34. Breton G, Zheng S, Valieris R, da Silva IT, Satija R, Nussenzweig MC. Human dendritic cells (DCs) are derived from distinct circulating precursors that are precommitted to become CD1c+ or CD141+ DCs. *Journal of Experimental Medicine*. 2016;213(13):2861–2870.
35. Engel I, Seumois G, Chavez L, Samaniego-Castruita D, White B, Chawla A, et al. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nature immunology*. 2016;17(6):728.

36. Grover A, Sanjuan-Pla A, Thongjuea S, Carrelha J, Giustacchini A, Gambardella A, et al. Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nature communications*. 2016;7:11075.
37. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*. 2015;112(23):7285–7290.
38. Zhang S, Wong H-S, Shen Y. Generalized adjusted rand indices for cluster ensembles. *Pattern Recognition*. 2012;45(6):2214–2226.
39. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*. 2002;3(Dec):583–617.
40. Romano S, Vinh NX, Bailey J, Verspoor K. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*. 2016;17(1):4635–4666.
41. Von Luxburg U. A tutorial on spectral clustering. *Statistics and computing*. 2007;17(4):395–416.
42. Zhong S, Zhang S, Fan X, Wu Q, Yan L, Dong J, et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*. 2018;555(7697):524.
43. Bruni F, Gramegna P, Oliveira JM, Lightowlers RN, Chrzanowska-Lightowlers ZM. REXO2 is an oligoribonuclease active in human mitochondria. *PLoS One*. 2013;8(5):e64670.
44. Pallier C, Scaffidi P, Chopineau-Proust S, Agresti A, Nordmann P, Bianchi ME, et al. Association of chromatin proteins high mobility group box (HMGB) 1 and HMGB2 with mitotic chromosomes. *Molecular biology of the cell*. 2003;14(8):3414–3426.
45. Hynes G, Sutton C, U S, Willison K. Peptide mass fingerprinting of chaperonin-containing TCP-1 (CCT) and copurifying proteins. *The FASEB journal*. 1996;10(1):137–147.
46. Numasawa-Kuroiwa Y, Okada Y, Shibata S, Kishi N, Akamatsu W, Shoji M, et al. Involvement of ER stress in dysmyelination of Pelizaeus-Merzbacher Disease with PLP1 missense mutations shown by iPSC-derived oligodendrocytes. *Stem cell reports*. 2014;2(5):648–661.

(a)

(b)

(a)

(b)

(c)

(d)

(a) Engel4

(b) Darmanis

(a) Engel4

(b) Darmanis