ScLRTC: Imputation for Single-Cell RNA-Seq Data Via Low-Rank Tensor Completion

Zhong Li (lizhong@zstu.edu.cn)
Zhejiang Sci-Tech University
Xiutao Pan
Zhejiang Sci-Tech University
Shengwei Qin
Zhejiang Sci-Tech University
Minzhe Yu
Zhejiang Sci-Tech University
Hang Hu
Zhejiang Sci-Tech University

Research Article

Keywords: single-cell RNA-seq, data imputation, low-rank tensor

Posted Date: August 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-757792/v1

ScLRTC: imputation for single-cell RNA-seq data via low-rank tensor completion

Zhong Li[1,*], Xiutao Pan[1], Shengwei Qin[1], Minzhe Yu[1], Hang Hu[1]

1School of Science, Zhejiang Sci-Tech University, Hangzhou, 310018, China

*Corresponding author. Email: lizhong@zstu.edu.cn

Abstract

Background: With single-cell RNA sequencing (scRNA-seq) methods, gene expression patterns at the single-cell resolution can be revealed. But as impacted by current technical defects, dropout events in scRNA-seq lead to missing data and noise in the gene-cell expression matrix and adversely affect downstream analyses. Accordingly, the true gene expression level should be recovered before the downstream analysis is carried out.

Results: In this paper, a novel low-rank tensor completion-based method, termed as scLRTC, is proposed to impute the dropout entries of a given scRNA-seq expression. It initially exploits the similarity of single cells to build a third-order low-rank tensor and employs the tensor decomposition to denoise the data. Subsequently, it reconstructs the cell expression by adopting the low-rank tensor completion algorithm, which can restore the gene-to-gene and cell-to-cell correlations. ScLRTC is compared with other state-of-the-art methods on simulated datasets and real scRNA-seq datasets with different data sizes. Specific to simulated datasets, scLRTC outperforms other methods in imputing the dropouts closest to the original expression values, which is assessed by both the sum of squared error (SSE) and Pearson correlation coefficient (PCC). In terms of real datasets, scLRTC achieves the most accurate cell classification results in spite of the choice of different clustering methods (e.g., SC3 or t-SNE followed by K-means), which is evaluated by using adjusted rand index (ARI) and normalized mutual information (NMI). Lastly, scLRTC is demonstrated to be also effective in cell visualization and in inferring cell lineage trajectories.

Conclusions: a novel low-rank tensor completion-based method scLRTC gave imputation results better than the state-of-the-art tools. Source code of scLRTC can be accessed at https://github.com/jianghuaijie/scLRTC.

Keywords: single-cell RNA-seq; data imputation; low-rank tensor

Background

Over the past few years, with the explosive growth of scRNA sequence data, important biological discoveries have been progressively conducted. However, as impacted by the picogram level of RNAs in a single cell, RNA transcripts may be missed during the reverse transcription and amplification step, so the transcripts are not detected in the following sequencing, which is termed as the dropout problem [1]. The resulting gene-cell expression matrix will consist of numerous false zeros attributed to dropout events, which will corrupt the biological signal and impede downstream analyses (e.g., cell clustering, data visualization and cell trajectory inference). To reduce the impact of this problem, besides increasing the efficiency of transcription capture, an effective imputation algorithm for scRNA-seq data should be developed to predict missing values attributed to dropout events [2].

Existing single-cell imputation methods have two main types: one complies with the deep learning method. For instance, DeepImpute [3] was designed to impute the scRNA sequence by applying a deep neural network (DNN) with a dropout layer and loss function to learn patterns in the data. DCA [4] established an auto-encoder to model the distribution of genes with a zero-inflated negative binomial prior, and then attempted to predict the mean, standard deviation and dropout probability of genes. ScIGANs [5] adopted the generative adversarial networks (GANs) to learn the dependence of nonlinear genes and genes from complex multi-cell type samples, and then trained the neural network model to generate real expression profiles of defined cell types. However, due to the influence of the training set and the existence of over-fitting problem, these methods may generate the false-positive results in differential expression analyses [6]. Another type for single-cell imputation methods complies with the statistical algorithm. For instance, SAVER [7] exploited information across genes in the identical cell type with a Bayesian approach to recover true expression levels; it also measured the uncertainty of recovered values. MAGIC [8] performed a soft clustering after building a Markov transition matrix, and then replaced a gene's raw expression with its weighted mean expression in a cluster. However, MAGIC also imputes the gene expression values that are not affected by dropout. Therefore, it may introduce the new bias into the data and possibly eliminate the meaningful biological variations. ScImpute [9] initially estimated the probability of an entry to be dropout with the use of a mixture model, and then imputed the potential dropout entries of a cell by employing the information from the gene expression of consistent cells. DrImpute [10] presented a clustering-based method and implemented a consensus strategy which estimated a value with several cluster priors or distance matrices and then imputed the data by aggregation.

CMF-Impute [11] drew upon the similarity of cells and genes to build a collaborative matrix factorization-based model for imputing the dropout entries of a given scRNA-seq expression. ALAR [12] provided a low-rank approximation of the expression matrix using singular vector decomposition (SVD). McImpute [13] used the nuclear norm minimization to realize a matrix completion algorithm for the scRNA data imputation. A study [14] suggested that taking advantage of the presence of low-rank submatrix can improve the imputation performance compared to the traditional low-rank matrix restore methods. For example, PBLR [15] considered the cell grouping information and performed a bounded low-rank completion method for each group. ScLRTD [16] introduced the tensor into the imputation of single-cell datasets, but it is mainly aimed at the completion with single-cell multi-omics sequencing data and the result in the scRNA dataset is not better than MAGIC, because this tensor based method did not fully take advantage of the correlation of single-cell data.

In this paper, a novel low-rank tensor completion based method (scLRTC) is proposed for the scRNA-seq data imputation. Since scRNA-seq data commonly involve single cells from different cell types and single cells with the identical type normally exhibit the similar expression pattern, the underlying true expression matrix is reasonably assumed to be able to be approximated by a low-rank matrix [1]. Based on such an assumption, the similar expression patterns of a single cell are adopted to build a third-order tensor, and then the single-cell gene expression is restored by approximating the tensor rank. This method is applied to nine scRNA-seq datasets and four simulation datasets, and it is compared with several state-of-the-art methods (SAVER [7], MAGIC [8], scImpute[9], DrImpute [10], CMF-Impute [11], PBLR [15], WEDGE [17] and scGNN [18]). As revealed from considerable data analyses, the proposed method is capable of achieving more accurate imputation results and improving the downstream analysis.

Results

The proposed imputation method is employed for nine published scRNA-seq datasets (i.e., Pollen [19], Usoskin [20], Yan [21], Zeisel [22], Mouse [23], PBMC [24], Chen [25], Loh [26] and Petropoulos [27]) and four simulation datasets generated from Splatter package [28], and it is compared with some popular methods (e.g., SAVER[7], MAGIC[8], scImpute[9], DrImpute[10], CMF-Impute[11], PBLR[15], WEDGE[17] and scGNN[18]). The method is largely evaluated from five aspects with cell subpopulation clustering, dimensionality reduction visualization, data masking evaluation, correlation analysis and differential expression analysis, and cell trajectory inference. The parameter setting of the respective dataset is listed in Table 1, and the parameter setting of the simulation dataset is presented in Table 2.

Evaluating imputation accuracy through cell clustering

In the relevant research on the scRNA-seq dataset, cell clustering refers to one of the critical contents. There are many clustering algorithms (e.g., K-means and SC3 [29]). Among the mentioned methods, SC3 is recognized as an accurate unsupervised single-cell clustering tool that does not explicitly address dropout events for the scRNA-seq data. Thus, the proposed method together with other popular scRNA-seq imputation methods was added into the preprocessing step of SC3. Then we used the cell clustering accuracy measured by adjusted rand index (ARI) [30] and normalized mutual information (NMI) [31] to evaluate their performance, namely, the consistency between the inferred cell cluster and the real cell cluster.

After the data imputation by the proposed scLRTC and other methods (DrImpute, SAVER, scImpute, MAGIC, CMF-Impute and PBLR), we used SC3 to cluster 6 published scRNA-seq datasets, including Usoskin, Pollen, Yan, Zeisel, Mouse and PBMC. The clustering accuracy measured by ARI and NMI are plotted in Fig. 1A and Fig. 1B, respectively. Obviously, the proposed method has the best ARI performance in Usoskin, Pollen, Yan, Zeiel, Mouse and PBMC, and the performance of NMI on the Usoskin dataset can be as competitive as CMF-Impute. In summary, the proposed method imputation can improve the clustering accuracy of SC3.

To show that the proposed imputation method does not depend on the clustering method, we further used another popular single-cell clustering method (first using the dimensionality reduction by t-distributed stochastic neighbor embedding (t-SNE) [32], and then applying K-means for clustering) [10] to test the performance of the proposed scLRTC and other methods. Compared with the SC3 algorithm, the K-means algorithm is more affected by the initial values. To compare the clustering results more reasonably, we performed t-SNE + K-means 20 times on the Pollen and Usoskin datasets (The perplexity of t-SNE is set to 10, and other parameters of t-SNE are set as default parameters). In the Pollen dataset, our median is the highest at 0.722, and the maximum value is 0.853, which is better than the maximum value (0.847) by DrImpute (Fig. 2A). In the Usoskin dataset, the median of the proposed method is 0.684 and the maximum value is 0.742, which are both the highest compared to other methods (Fig. 2B). In brief, the proposed imputation achieves a better overall effect than other imputation algorithms.

Furthermore, we compared scLRTC with the latest matrix completion based method WEDGE [17] and deep learning based method scGNN [18] on the Zeisel dataset. We applied the Scanpy's Louvain algorithm [33, 34] for the scRNA-seq data clustering and found scLRTC achieved an ARI of 0.692, which is higher than WEDGE's 0.560 and scGNN's 0.678. Finally, we did the test for a large Chen dataset [25] where the number of cells is more than 10000. We also used the Louvain algorithm to cluster the scRNA-seq data and found the clustering performance index ARI increased from 0.611 (raw data) to 0.673 by scLRTC. Considering the time complexity of scLRTC, the addition of tensor computation makes it slower than other methods. But we can control the size of tensor for various datasets to relief the influence of tensor computation. Fig. 2C illustrates the running time of scLRTC for the mentioned experimental datasets with different sizes of tensor setting. It shows that the time complexity of scLRTC is not quadratic proportional to the number of cells, which makes it applicable for scRNA-seq datasets with different sizes.

Cell Visualization

Visually representing scRNA-seq data involves shrinking the gene expression matrix into a lower space, and then mapping each cell's transcriptome in the reduced low dimensional space. Several dimensionality reduction methods are generally known (e.g., PCA [35], t-SNE and UMAP [36]), where UMAP is suggested to be particularly suitable for the visualization of any dimensional data. Accordingly, UMAP was employed to discuss the dimensionality reduction effect before and after imputation on four expression matrices of Yan, Pollen, Usoskin and Zeisel datasets. To be specific, cells were visualized in a two-dimensional space, and different cells were stained using real labels

before and after imputation. To quantify the grouping of cell transcriptomes, an unsupervised clustering quality measurement was conducted with silhouette coefficient (SC) [37] to evaluate the effect of dimensionality reduction. The higher the silhouette coefficient, the more significant the dimensionality reduction effect will be. The UMAP dimensionality reduction visualization and the average SC of the raw and imputed data (4 published datasets) with different methods are illustrated in Fig. 3 and Fig. 4. According to these figures, the SC values of the proposed scLRTC in these datasets are the highest with 0.884, 0.797, 0.861, 0.639, respectively.

Assessing imputation accuracy through data masking

The data masking evaluation was conducted on the real dataset and simulation datasets. First, 5% of non-zero entries were randomly selected from the Loh dataset, and these values were masked to zeros to generate a new gene expression matrix. Subsequently, seven imputation algorithms were applied for the new gene expression matrix and compared with unmasked data. The sum of squared errors (SSE) and Pearson correlation coefficient (PCC) between the imputation values and the true values were adopted to evaluate the effect of imputation. Fig. 5A presents all the results of the imputation accuracy index of the masked data. The proposed method can recover the missing values with the lowest SSE of 268.8 and the highest PCC of 0.707 in all compared imputation algorithms. Note that the SAVER method persistently underestimates the values, especially among the highly expressed genes. Consistent experimental results were also mentioned in references [3,15]. To prevent the influence attributed to randomness, we performed 5 masking repetitions for the above experiment. The results of 7 methods in 5 repeated experiments only slightly fluctuate (Fig. 5B and Fig. 5C), demonstrating that the randomness slightly impacts the mentioned results.

Moreover, the performance of the proposed model was tested on single-cell simulation data that involves three cell populations. These data were generated using the Splatter package [28]. To simulate dropout events, 40%, 50%, 60%, and 70% of the entries were randomly masked in the expression matrix, corresponding to a shape parameter of dropout logistic function (ds) equaling -0.3, 0, 0.05, and 0.25 respectively. The masked entries were imputed with 7 methods and the imputed results are compared with the real values. Fig. 6 shows the visualization results of t-SNE with dropout, unmasked raw data (Full), and 7 imputation methods (including DrImpute, scImpute, MAGIC, SAVER, CMF-Impute, PBLR and scLRTC) under different dropout rates. It can be seen that the proposed method is most consistent with the original data (Full) under the t-SNE visualization, demonstrating that the proposed imputation has a strong ability to restore real cell clusters. Furthermore, we performed the quantitative analysis on the simulation dataset. Fig. 7 shows the SSE and PCC values under different dropout rates. With the increase in the dropout rate, the accuracy of all imputation methods is affected. However, the proposed scLRTC is suggested to exhibit the optimal performance among 7 methods.

Evaluating imputation accuracy through correlation analysis and differential expression analysis

The ability of the imputation method was evaluated to restore gene-gene and cell-cell relationships in complex tissues. The simulated data were employed with a dropout rate of 40% (ds = - 0.3) to calculate the gene-gene and cell-cell correlation matrix, and $\log_{10}(X+1)$ was set as the result after imputation. In the cell-to-cell correlation heat map (Fig. 8A), the color of MAGIC and the proposed scLRTC is the closest to the heat map of Full. For the heat map of gene-gene correlation (Fig. 8B), scImpute and the proposed scLRTC are the closest ones to the expression heat map of Full in color, while MAGIC deviates the most. And then the violin chart was used to display its expression distribution. We find the violin chart of scLRTC is the closest to the unmasked raw data (Full) in the appearance comparison (Fig. 9A), indicating that the position and the upper quartile comply with Full. It is suggested that the data by the scLRTC complement here achieves the most consistent distribution with that of Full. In summary, the proposed method can effectively restore the true gene-gene and cell-cell relationship.

In addition, it is considered that the imputation method should be capable of recovering true differential expressed (DE) genes and reducing the production of false positive genes. Since gold standard of DE genes has been rarely formulated in real datasets, 6 imputation methods (DrImpute, SAVER, scImpute, MAGIC, CMF-Impute and scLRTC) were compared for their capabilities to recover DE genes in the simulation data. The differential expression analysis was performed by using the MAST [38], and the true DE genes identified from the complete data were considered the reference. In terms of the respective method, the DE genes were extracted, which are considered significant by controlling p-value < 0.01 and comparing them with the true DE genes. Fig. 9B presents the average ROC curves of different imputation methods by considering the indices of recall and precision. ScLRTC is found to achieve the highest score (AUC = 0.971) for detecting DE genes, demonstrating that scLRTC is valid to recover more DE genes and detect less false-positives genes.

Evaluating imputation accuracy through cell trajectory inference

A common task of single cell RNA sequence analysis is to rebuild the lineage trajectory and infer the differentiation and progenitor status of single cells, which is a research hotspot over the past few years. Besides, a wide range of algorithms have been developed in this field. For instance, TSCAN [39] performed the differential expression and time series analysis on single-cell expression data, which classified individual cells according to the progress of biological processes. However, TSCAN did not perform dropout imputation for the data reprocessing. Thus, in this study, the scLRTC imputation was integrated into TSCAN, and its performance was compared in the pseudotime inference of the Petropoulos dataset. The Petropoulos data consists of the single cells from five stages of human preimplantation embryonic development from developmental day (E) 3 to day 7. Notably, though the cells at each time point may not be homogeneous, the time label can be exploited to represent an overall developmental trajectory. Accordingly, the known time label acts as the ground truth, and the performance of pseudotime inference is evaluated with TSCAN, as input by the raw data and the imputed data with 6 different methods (scLRTC, SAVER, scImpute, DrImpute, CMF-Impute and MAGIC). Furthermore, Pseudotime ordering score (POS)[39] and Kendall's rank correlation score (KRCS) were used to measure the consistency of time label and pseudotime order derived from the data. The results are presented in Fig. 10.

It is therefore suggested that the proposed method has improved on both POS and KRCS indicators compared with the original data. For the SAVER method, it has an improvement on KRCS, whereas the POS score decreases. In terms of other methods, the pseudotime trajectory by DrImpute and scImpute starts at E3 and ends at E6, the pseudotime trajectory of CMF-Impute starts at E3 and ends at E5, and the trajectory reconstruction error is relatively large. The accuracy of MAGIC from E3 to E5 is relatively high, whereas at E6 and E7 stages, a big discrepancy is identified with the real label, which introduces errors.

Discussion

Since single-cell RNA has a limited extraction efficiency, the occurrence of dropout events adversely affects the downstream analysis. However, the single-cell data imputation is not explicitly involved in the most used scRNA-seq tools for cell clustering, dimensionality reduction visualization, cell type recognition and lineage reconstruction, so it is of high research significance. In this paper, a novel low-rank tensor completion method (scLRTC) is proposed to impute the scRNA sequence data where dropout is present. ScLRTC, a data-driven method, fully considers the similarity and heterogeneity between cells. It builds a third-order tensor representation and employs a low-rank tensor completion model by adopting the ADMM algorithm to achieve the data imputation. This study also inputs the data with scLRTC imputation into SC3 clustering and carries out the clustering by first conducting t-SNE dimensionality reduction and then implementing K-means. Accordingly, it is reported that scLRTC is capable of increasing the clustering accuracy of real data at different dropout rates, as well as improving the quality of cell visualization. Moreover, by integrating the proposed scLRTC into TSCAN, we find it improves the accuracy of pedigree reconstruction and pseudotime inference.

In general, the proposed imputation method can be regarded as one powerful complement to current scRNA sequence data analysis. Our algorithm can be further improved in the future work. For instance, because of the tensor model in scLRTC is relatively independent, we will develop the single-cell completion based on the parallel computing to improve the time complexity of scLRTC. In addition, we currently only use the similarity between cells to build a low-rank tensor. We can also consider the similarity between genes, and combine the similarity between cells and genes to build a higher-order tensor. Besides, we treat all zero values as potential dropout entries, which may be inappropriate for different cases. One possible solution is to first evaluate the probability of a data dropout and then only impute the data with a higher probability case. Lastly, we embed our algorithm in the existing clustering methods. Because each clustering method has its assumption, this strategy may not be optimal for the analysis. Furthermore, how to build a robust clustering method suitable for different single-cell data imputation methods is also a future research content.

Conclusions

Imputation is an essential step in the use and analysis of scRNA-seq. In this research we introduced a novel low-rank tensor completion-based method, termed as scLRTC. Experiments on simulation data and real data sets showed scLRTC to be highly accurate in the imputation and downstream analysis of scRNA-seq.

Methods

Datasets

Nine scRNA-seq datasets (i.e., Pollen, Usoskin, Yan, Zeisel, Mouse, PBMC, Chen, Loh and Petropoulos) with different data sizes is used to test the validity of the proposed scLRTC in imputing dropout events. Besides, these datasets fall to three levels (i.e., gold, silver and copper) based on the supporting evidence of cell markers. To be specific, Pollen, Loh, Yan, Zeisel and Mouse datasets are defined as gold standard datasets, in which all cell markers are defined by complying with experimental conditions or cell lines. Usoskin, PBMC and Chen datasets are defined as the silver standard, with cell markers are calculated and assigned by drawing upon the authors' knowledge of the underlying biology. Petropoulos is considered the copper standard since the cells involved are in the developmental stage (time labeled). Though single cell populations from different time points usually exhibit different expression patterns and biological characteristics, it remains infeasible to separate different populations at each time point based on time tags alone. Table 3 briefs these scRNA sequence datasets with sizes ranging from 90 (Yan) to 12089 (Chen), and the number of cell clusters ranges from 4 (Usoskin) to 46 (Chen). Note that the first seven datasets are normally used for the cluster analysis. Furthermore, the first three datasets are from the low-throughput data sequencing platform, and the last four datasets originate from the high-throughput data sequencing platform. Loh is employed for the data masking evaluation, and Petropoulos is used for the trajectory reconstruction analysis.

Data preprocessing and normalization

In terms of a given scRNA-seq dataset, its gene expression matrix is recorded as X^c. To reduce the effect of underexpressed genes, the gene expressed in less than or equal to 3 cells is removed [40]. To express the filtered matrix by X^N, a matrix X is then made by taking the $\log_2$ transformation with a pseudo count 1

$$X_{ij} = \log_2\left(X_{ij}^N + 1\right);\ i = 1,2,\dots,M, j = 1,2,\dots,N \quad (1)$$

where M denotes the overall number of genes; N is the total number of cells. The pseudo-count is added to avoid infinite values in the parameter estimation in the subsequent data analysis. The logarithmic transformation has an advantage that it can prevent a small number of large observation values from being significantly affected in the data imputation.

Tensor based model for scRNA-seq data imputation

Single-cell dropout events can be formulated as a missing value estimation problem. The core problem of missing value estimation refers to how to develop the relationship between known elements and unknown elements. The scRNA-seq data usually consist of single cells from different cell types, and single cells exhibiting the identical type have similar expression patterns. For this reason, it is assumed that the basic true expression of scRNA-seq data can be approximatively considered as a low-rank matrix. The low-rank matrix restoration essentially complies with the correlation between the rows and columns of a matrix, therefore creates a direct and effective imputation strategy.

A recent study suggested that taking full advantage of the presence of low-rank submatrix can improve the imputation performance compared to traditional low-rank matrix recovery methods [14,15]. However, the low-rank submatrix constructed by clustering is easily influenced by the clustering effect, and the low-rank tensor can be constructed to capture more correlations of similar single cell compared to the low-rank submatrix form. Based on this motivation, the two-dimensional low-rank matrix is extended to the three-order low-rank tensor with the high correlation of scRNA-seq data. Besides, a novel low-rank tensor model is built for single-cell gene expression data, and the tensor trace norm [41] is employed to approximate the rank of the tensor, finally the missing data are rebuilt and the cell's gene expression is restored.

Tensor construction

We construct the tensor form of single-cell which fully considers the high correlation of scRNA-seq data. Specific to a given cell X_i, the Pearson correlation coefficient between cells is first calculated and sorted in a descending order. Subsequently, the gene expression of cell X_i and its K-1 cells with the highest correlation are adopted to build a matrix $Mat_i \in \mathbb{R}^{K\times M}$, where M denotes the number of genes. Subsequently, the difference between cell X_i and other cells are measured in the matrix Mat_i. To be specific, the Euclidean distance is calculated from all cells in Mat_i to X_i and sorted in an ascending order as $D_i = (d_{1i}, d_{2i}, \dots, d_{Ki})$. Next, the direction (angle) similarity of cell gene expression is measured by calculating the cosine similarity from all cells in Mat_i to X_i, and then it is recorded in an ascending order as $C_i = (c_{1i}, c_{2i}, \dots, c_{Ki})$. Lastly, the similarity between cells is measured according to the absolute value of the difference of cell gene expressions, i.e., the Chebyshev distance from all cells in Mat_i to X_i is calculated and restored in an ascending order as $Q_i = (q_{1i}, q_{2i}, \dots, q_{Ki})$.

Three distance vectors obtained from X_i are combined into a feature vector $Vec_i = \{D_i, C_i, Q_i\}$ with a size of $3K \times 1$. Likewise, the feature vector Vec_j can be obtained from the other cell X_j. By calculating the distance between two feature vectors, the P-1 Vec_j closest to Vec_i is searched, and these matrices are merged to build a third-order tensor $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{K\times M\times P}$ for the cell X_i (as shown in Fig. 11).

Tensor fold and unfold

During the tensor analysis, it is convenient to unfold a tensor into a matrix. The "unfold" operation along the kth mode on a tensor $\boldsymbol{\mathcal{Y}}$ is defined as

$$unfold_k(\boldsymbol{\mathcal{Y}}) = \boldsymbol{\mathcal{Y}}_{(k)} \in \mathbb{R}^{I_k\times(I_1\dots I_{k-1}I_{k+1}\dots I_n)} \tag{2}$$

The opposite operation "fold" is defined as

$$fold_k\big(\boldsymbol{\mathcal{Y}}_{(k)}\big) = \boldsymbol{\mathcal{Y}} \tag{3}$$

Tucker decomposition and denoising

Tensor Tucker decomposition is recognized as a form of high-order principal component analysis. The HOSVD method [42] is available for decomposing a third-order tensor $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I\times J\times K}$ by

$$\boldsymbol{\mathcal{Y}} \approx \boldsymbol{\mathcal{G}} \times_1 A \times_2 B \times_3 C \tag{4}$$

where $A \in \mathbb{R}^{I\times P}$, $B \in \mathbb{R}^{J\times Q}$ and $C \in \mathbb{R}^{K\times R}$ denote factor matrices, which can be considered the main components of the corresponding mode. The tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{P\times Q\times R}$ refers to the core tensor, representing the level of interaction between different components.

Next, a hard threshold function is set for factor matrices A, B, C to eliminate the effect of some low value components after the Tucker decomposition. Subsequently, the third-order tensor $\widehat{\boldsymbol{\mathcal{Y}}}$ is restored by updating $\widehat{\boldsymbol{\mathcal{G}}}$, i.e., the convergence of the current tensor data is ensured by the iterative computation, and the denoising effect of some mutation elements is achieved in the tensor form of scRNA-seq data.

Tensor trace norm

The trace norm of a tensor is defined as [41]

$$||\boldsymbol{\mathcal{Y}}||_* = \sum_{i=1}^{n} \alpha_i ||\boldsymbol{\mathcal{Y}}_{(i)}||_* \tag{5}$$

where $\alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1$, $\boldsymbol{\mathcal{Y}}_{(i)}$ expresses the matrix along the ith mode. In fact, the trace norm of a tensor refers to a convex combination of the trace norms of all matrices expanded along the respective mode. Notably, when n is equal to 2 (the matrix case), the definition of the tensor's trace norm complies with the matrix case.

Tensor completion model

By using M_i to replace $\boldsymbol{\mathcal{Y}}$, the tensor completion model is expressed as

$$\begin{aligned} &min_{\boldsymbol{\mathcal{Y}}, M_1,\dots,M_n} \ \textstyle\sum_{i=1}^{n} \alpha_i \big\|M_{i(i)}\big\|_* \\ &s.t.\ \boldsymbol{\mathcal{Y}}_\Omega = \boldsymbol{\mathcal{T}}_\Omega \\ &\boldsymbol{\mathcal{Y}} = M_i, \quad i = 1, \dots, n. \end{aligned} \tag{6}$$

where α_i denotes the coefficient; $M_{i(i)}$ represents the unfold matrix of the tensor along the ith mode; $\boldsymbol{\mathcal{T}}$ is the known tensor; $\boldsymbol{\mathcal{Y}}$ expresses the reconstructed tensor; Ω in $\boldsymbol{\mathcal{T}}_\Omega$ is the index of non-zero observation value.

Tensor model solution

The mentioned model can be solved by adopting the alternating direction method of multipliers (ADMM) algorithm. The augmented Lagrangian function is defined as

$$\begin{aligned} L_\rho(\boldsymbol{\mathcal{Y}}, M_1, \dots, M_n, y_1, \dots, y_n) = \textstyle\sum_{i=1}^{n} \alpha_i ||M_{i(i)}||_* + \langle \boldsymbol{\mathcal{Y}} - M_i, y_i\rangle + \\ \tfrac{\rho}{2} ||M_i - \boldsymbol{\mathcal{Y}}||_F^2 \end{aligned} \tag{7}$$

where $<\cdot,\cdot>$ denotes the inner product; $||\cdot||_F^2$ represents the F-norm, i.e., the root of the square sum of all elements; y_i is the Lagrange multiplier; ρ expresses the penalty parameter.

According to the framework of ADMM, $M_i, \boldsymbol{\mathcal{Y}}, y_i$ can be iteratively updated as

1) $\{M_1^{k+1}, \dots, M_n^{k+1}\} = argmin_{M_1,\dots,M_n} L_\rho(\boldsymbol{\mathcal{Y}}^k, M_1, \dots, M_n, y_1^k, \dots, y_n^k)$
2) $\boldsymbol{\mathcal{Y}}^{k+1} = argmin_{\boldsymbol{\mathcal{Y}}} L_\rho(\boldsymbol{\mathcal{Y}}, M_1^{k+1}, \dots, M_n^{k+1}, y_1^k, \dots, y_n^k)$
3) $y_i^{k+1} = y_i^k - \rho\big(M_i^{k+1} - \boldsymbol{\mathcal{Y}}^{k+1}\big)$

From the augmented Lagrangian function in 1), $\boldsymbol{\mathcal{Y}}_{(i)}^k, y_i^k$ is fixed and minimized to yield

$$\begin{aligned} M_{i(i)}^{k+1} &= arg\min_{M_{i(i)}} \alpha_i \big\|M_{i(i)}\big\|_* + \frac{\rho}{2}\langle M_{i(i)}, M_{i(i)}\rangle - \rho\,\langle M_{i(i)}, \boldsymbol{\mathcal{Y}}_{(i)}^k + \frac{1}{\rho} y_{i(i)}^k\rangle \\ &= arg\min_{M_{i(i)}} \alpha_i \big\|M_{i(i)}\big\|_* + \frac{\rho}{2}\Big\|M_{i(i)} - \boldsymbol{\mathcal{Y}}_{(i)}^k - \frac{1}{\rho} y_{i(i)}^k\Big\|_F^2 \\ &\quad - \langle \boldsymbol{\mathcal{Y}}_{(i)}^k + \tfrac{1}{\rho} y_i^k, \boldsymbol{\mathcal{Y}}_{(i)}^k + \tfrac{1}{\rho} y_{i(i)}^k\rangle \end{aligned} \tag{8}$$

Thus, the optimal solution of $M_{i(i)}^{k+1}$ is

$$M_{i(i)}^{k+1} = arg\,min_{M_{i(i)}} \alpha_i ||M_{i(i)}||_* + \tfrac{\rho}{2} ||M_{i(i)} - (\boldsymbol{\mathcal{Y}}_{(i)}^k + \tfrac{1}{\rho} y_{i(i)}^k)||_F^2 \tag{9}$$

The above Eq. (9) is proven to generate a closed-form in recent references [43, 44], so it can be solved by calculating the singular value thresholding operator $D_\tau(\cdot)$. In terms of any matrix X, the singular value decomposition (SVD) is performed to obtain $X = U\Sigma V^T$, where U, V are orthogonal singular vectors, and $\Sigma \in \mathrm{R}^{r\times r}$ comprises singular values $\sigma_1, \dots, \sigma_r, r = min\{m, n\}$. The singular value thresholding operator can be defined as $D_\tau(X) = U\Sigma_\tau V^T$, where $\Sigma_\tau = diag(max(\sigma_i - \tau, 0))$. Thus, it yields

$$M_{i(i)} = D_{\frac{\alpha_i}{\rho}}\Big(\boldsymbol{\mathcal{Y}}_{(i)} + \tfrac{1}{\rho} y_{i(i)}\Big) \tag{10}$$

By folding $M_{i(i)}$ to get

$$M_i = fold_i\big(M_{i(i)}\big) \tag{11}$$

From the augmented Lagrangian function in 2), it can be minimized by fixing M_i^k and y_i^k, and the optimal solution is obtained as

$$\boldsymbol{\mathcal{Y}}^{k+1} = arg\,min_{\mathcal{Y}} \sum_{i=1}^{n} \langle \boldsymbol{\mathcal{Y}}, y_i^k \rangle + \frac{\rho}{2} \langle \boldsymbol{\mathcal{Y}} - M_i^{k+1}, \boldsymbol{\mathcal{Y}} - M_i^{k+1} \rangle$$

$$= arg\,min_{\mathcal{Y}} \sum_{i=1}^{n} \frac{\rho}{2} \langle \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}} \rangle + \langle \boldsymbol{\mathcal{Y}}, \rho M_i^{k+1} - y_i^k \rangle \quad (12)$$

Take the derivative of (12) with respect to $\boldsymbol{\mathcal{Y}}$ and set it equal to 0 to yield

$$n\rho \boldsymbol{\mathcal{Y}} - \left(\sum_{i=1}^{n} \rho M_i^{k+1} - y_i^k \right) = 0 \quad (13)$$

So

$$\boldsymbol{\mathcal{Y}}^{k+1} = \frac{1}{n} \left(\sum_{i=1}^{n} M_i^{k+1} - \frac{1}{\rho} y_i \right) \quad (14)$$

For $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{Y}}_{\bar{\Omega}} + \boldsymbol{\mathcal{Y}}_{\Omega}$, $\boldsymbol{\mathcal{Y}}_{\Omega} = \mathcal{T}_{\Omega}$ is known, so only $\boldsymbol{\mathcal{Y}}_{\bar{\Omega}}$ is updated

$$\boldsymbol{\mathcal{Y}}_{\bar{\Omega}}^{k+1} = \frac{1}{n} \left(\sum_{i=1}^{n} M_i^{k+1} - \frac{1}{\rho} y_i^k \right)_{\bar{\Omega}} \quad (15)$$

After the tensor completion reconstruction, the corresponding cell expression $X_i (i = 1,2, \dots, N)$ is selected from the respective tensor model, and the gene information of each cell is restored, and a matrix X^* representing the complete scRNA-seq data is lastly formed. Since the gene expression of the cell is non-negative, the matrix X^P is defined after the imputation

$$X^P = \begin{cases} X_{ij}^P = 0 & if (X_{ij}^* < 0) \\ X_{ij}^P = X_{ij}^* & others \end{cases} \quad (16)$$

In brief, the entire scLRTC algorithm process is expressed as

Input: X, ρ, α *and* K
Output: $\hat{X}$ (X after imputing)
1. **for** $m = 1$ *to* N
2. Construction of low rank tensor T_{Ω} of X_m
3: Set $\boldsymbol{\mathcal{Y}}_{\Omega} = T_{\Omega}$ and $\boldsymbol{\mathcal{Y}}_{\bar{\Omega}} = 0, \boldsymbol{\mathcal{Y}} = M_i$
4: **for** $k = 0$ *to* K **do**
5: **for** $i = 1$ *to* n **do**
6: $M_{i(i)} = \boldsymbol{D}_{\frac{\alpha_i}{\rho}} (\boldsymbol{\mathcal{Y}}_{(i)} + \frac{1}{\rho} y_{i(i)})$
7: $M_i = fold_i [M_{i(i)}]$
8: **end for**
9: $\boldsymbol{\mathcal{Y}}_{\bar{\Omega}} = \frac{1}{n} (\sum_{i=1}^{n} M_i - \frac{1}{\rho} y_i)_{\bar{\Omega}}$
10: $y_i = y_i - \rho (M_i - \boldsymbol{\mathcal{Y}})$
11: **end for**
12: extract $\hat{X}_m$ from $\boldsymbol{\mathcal{Y}}$
13: **end for**
14: $X^* = \left(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_m \right)$ and use Eq. (16) to remove negative values

Evaluation measures

To objectively evaluate the effectiveness of the proposed low-rank tensor completion method for single-cell RNA-seq data, the reconstructed data are used for the cell clustering, and two clustering indicators with the normalized mutual information (NMI) and the adjusted rand index (ARI) are adopted to quantify the consistency between inferred and predefined cell clusters in the respective scRNA-seq data. Subsequently, the silhouette coefficient (SC) is adopted to assess the visual effect of cell dimensionality reduction. Lastly, Pseudotemporal ordering score (POS) Tand KRCS are used to evaluate the accuracy of cell trajectory analysis and imputation.

Denote that $U = \{\mu_1, \mu_2, \dots, \mu_P\}$ is adopted to represent the true partition of P classes, $V = \{v_1, v_2, \dots, v_K\}$ is used to denote the partition given by clustering results, n_i and n_j are represented as the number of the class μ_i and cluster v_j respectively, and n_{ij} is expressed as the number of observations in both class μ_i and cluster v_j.

ARI is then formulated as:

$$\frac{\sum_{i=1}^{P} \sum_{j=1}^{K} \binom{n_{ij}}{2} - [\sum_{i=1}^{P} \binom{n_{i\cdot}}{2} \sum_{j=1}^{K} \binom{n_{\cdot j}}{2}] / \binom{n}{2}}{\frac{1}{2} \left[\sum_{i=1}^{P} \binom{n_{i\cdot}}{2} \sum_{j=1}^{K} \binom{n_{\cdot j}}{2} \right] - \left[\sum_{i=1}^{P} \binom{n_{i\cdot}}{2} \sum_{j=1}^{K} \binom{n_{\cdot j}}{2} \right] / \binom{n}{2}} \quad (17)$$

where $n = \sum_{i=1}^{P} n_{i\cdot} = \sum_{j=1}^{K} n_{\cdot j}$.

NMI is expressed as

$$NMI = \frac{2I(U,V)}{H(U)+H(V)} \quad (18)$$

where $I(U,V)$ expresses the amount of mutual information between U and V

$$I(U,V) = \sum_{i=1}^{P} \sum_{j=1}^{K} \frac{|u_i \cap v_j|}{N} \log \frac{N|u_i \cap v_j|}{|u_i| \times |v_j|} \quad (19)$$

$H(U)$ and $H(V)$ are the entropies of partitions U and V

$$H(U) = -\sum_{i=0}^{k} \frac{u_i}{N} log \frac{u_i}{N}, H(V) = -\sum_{i=0}^{k} \frac{v_i}{N} log \frac{v_i}{N} \quad (20)$$

where N is the total number of cells.

SSE is written as

$$\text{sqrt}(\sum_{i=1}^{n} \sum_{i=1}^{n} (X_{ij} - P_{ij})^2) \quad (21)$$

where X_{ij} denotes the true gene expression; P_{ij} represents the predicted gene expression.

SC is expressed as

$$SC = \text{average} \left(\sum_{i=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \right) \quad (22)$$

where i denotes the ith cell, $a(i) = average$ (i to all other cells in the cluster to which it belongs), $b(i) = min$ (the average distance from i to all cells in the other cluster).

Abbreviations

scRNA-seq: Single-cell RNA-sequencing
PCC: Pearson correlation coefficient
SSE: sum of squares of error
ARI: Adjusted Rand Index
NMI: Normalized Mutual Information
SC: silhouette coefficient
POS: Pseudotemporal ordering score
KRCS: Kendall's rank correlation score

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

There are no new data associated with this article. Published datasets used in this study: Pollen et al.'s dataset ([19], P041736) of human tissues cells is available at https://www.ncbi.nlm.nih.gov/sra?term =SRP041736. Usoskin et al.'s dataset ([20, GSE59739) of mouse lumbar dorsal root ganglion cells is available at https://www. ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59739. Yan et al.'s dataset ([21], GSE36552) of the human embryos cells is available at https://www.ncbi. nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552. Zeisel et al.'s dataset ([22]; GSE60361) of the mouse cortex and hippocampus cells is available at https://www.ncbi.nlm.nih.gov/geo/query/acc. cgi?acc=GSE60361. The mouse bladder cell dataset [23] originates from the mouse cell atlas project is available at https://figshare.com/ s/865e694ad06d5857db4b. From the raw count matrix, 2100 cells are selected from the bladder tissue. PBMC dataset [24] is downloaded from the 10X genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k). Chen et al.'s dataset ([25], GSE87544) of adult mouse hypothalamus cells is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87544. Loh et al.'s dataset ([26], SRP 073808) of human embryonic stem cells is available at https://www.ncbi.nlm.nih.gov/sra/?term=SRP073808. Petropoulos et al.'s dataset ([27], E-MTAB-3929) of human preimplantation embryos cells is available at https://www.ebi.ac.uk/arrayexpress /experiments/E-MTAB-3929. All the source codes and supplementary data are available at https://github.com/jianghuaijie/scLRTC.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China under Grant No. 11671009, Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ19A010002.

Authors ' contributions

Z.L. provided the methodology. X.P., Z.L. and S.Q. designed the algorithm. X.P., M.Y. and H.H. arranged the datasets and performed the analysis. X.P. drafted the manuscript. Z.L. and S.Q. reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to express the sincere gratitude to Prof. Xiaojiang Xu from NIH, USA for the helpful discussion.

References

1. Chen C, Wu C, Wu L, Wang X, Deng M, Xi R. scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition. Bioinformatics. 2020;36:3156-3161.
2. Zhu K, Anastassiou D. 2DImpute: imputation in single-cell RNA-seq data from correlations in two dimensions. Bioinformatics. 2020;36:3588-3589.
3. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol.2019; 20:1-14.
4. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun,2019;10:1-14.
5. Xu Y, Zhang Z, You L, Liu J, Fan Z, Zhou X. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Res. 2020; 48:e85.
6. Andrews TS, Hemberg M. False signals induced by single-cell imputation. *F1000Research.* 2018;7:1740.
7. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15(7):539-542.
8. Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716-729.
9. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9(1):1-9.
10. Gong W, Kwak IY, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018;19(1):1-10.
11. Xu J, Cai L, Liao B, Zhu W, Yang J. CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. Bioinformatics. 2020;36(10):3139-3147.
12. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*,2018;397588.
13. Mongia A, Sengupta D, Majumdar A. McImpute: matrix completion based imputation for single cell RNA-seq data. Front Genet. 2019;10:9.
14. Ruchansky N, Crovella M, Terzi E. Targeted matrix completion. Proc SIAM Int Conf Data Min. 2017;255-263.
15. Zhang L, Zhang S. Imputing single-cell RNA-seq data by considering cell heterogeneity and prior expression of dropouts. J Mol Cell Biol. 2020;doi:10.1093/jmcb/mjaa052.
16. Ni Z, Zheng X, Zheng X, Zou X. scLRTD: A novel low rank tensor decomposition method for imputing missing values in single-cell multi-omics sequencing data. IEEE ACM T Comput Biol. 2020;doi:10.1109/TCBB.2020.3025804.
17. Hu Y, Li B, Zhang W, Liu N,Cai P,Chen F, WEDGE: imputation of gene expression values from single-cell RNA-seq datasets using biased matrix decomposition, Brief Bioinform. 2021;doi:https://doi.org/10.1093/bib/bbab085.
18. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, Wang C, Fu H, Ma Q. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021;doi:https://doi.org/10.1038/s41467-021-22197-x.
19. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32(10):1053.
20. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, Hjerling-Leffler J, Haeggström J, Kharchenko O, Kharchenko PV. et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015;18(1):145-153.
21. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J. et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20(9):1131.
22. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138-1142.
23. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F. et al. Mapping the mouse cell atlas by microwell-seq. Cell 2018;172(5):1091-1107.
24. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):1-12.
25. Chen R, Wu X, Jiang L, Zhang Y, Single-cell RNA-seq reveals hypothalamic cell diversity. Cell Rep. 2017;18(13):3227-3241.
26. Loh KM, Chen A, Koh PW, Deng TZ, Sinha R, Tsai JM, Barkal AA, Shen KY, Jain R, Morganti RM. et al. Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. Cell. 2016; 166(2):451-467.

27. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, Plaza, Reyes A, Linnarsson S, Sandberg R, Lanner F. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell 2016;165(4):1012-1026.
28. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 2017;18(1):1-15.
29. Kiselev VY, Kirschner K, Schaub M T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483-486.
30. Strehl A, Ghosh J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3(12):583-617.
31. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846-850.
32. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-2605.
33. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):1-5.
34. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E, Fast unfolding of communities in large networks. J Stat Mech-theoryE. 2008;2008(10):P10008.
35. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab. 1987;2(1):37-52.
36. Becht E, McInnes L, Healy J, Dutertre C, Kwok I, Ng L, Ginhoux F, Newell E. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2019;37(1):38-44.
37. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65.
38. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):1-13.
39. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016;44(13):e117-e117.
40. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14(4): 414-416.
41. Liu J, Musialski P, Wonka P, Ye J. Tensor completion for estimating missing values in visual data. IEEE T Pattern Anal. 2012;35(1):208-220.
42. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. SIAM J Matrix Anal A. 2000;21(4):1253-1278.
43. Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20(4):1956-1982.
44. Ma S, Goldfarb D, Chen L. Fixed point and Bregman iterative methods for matrix rank minimization. Math Program. 2011;128(1):321-353.

Figures

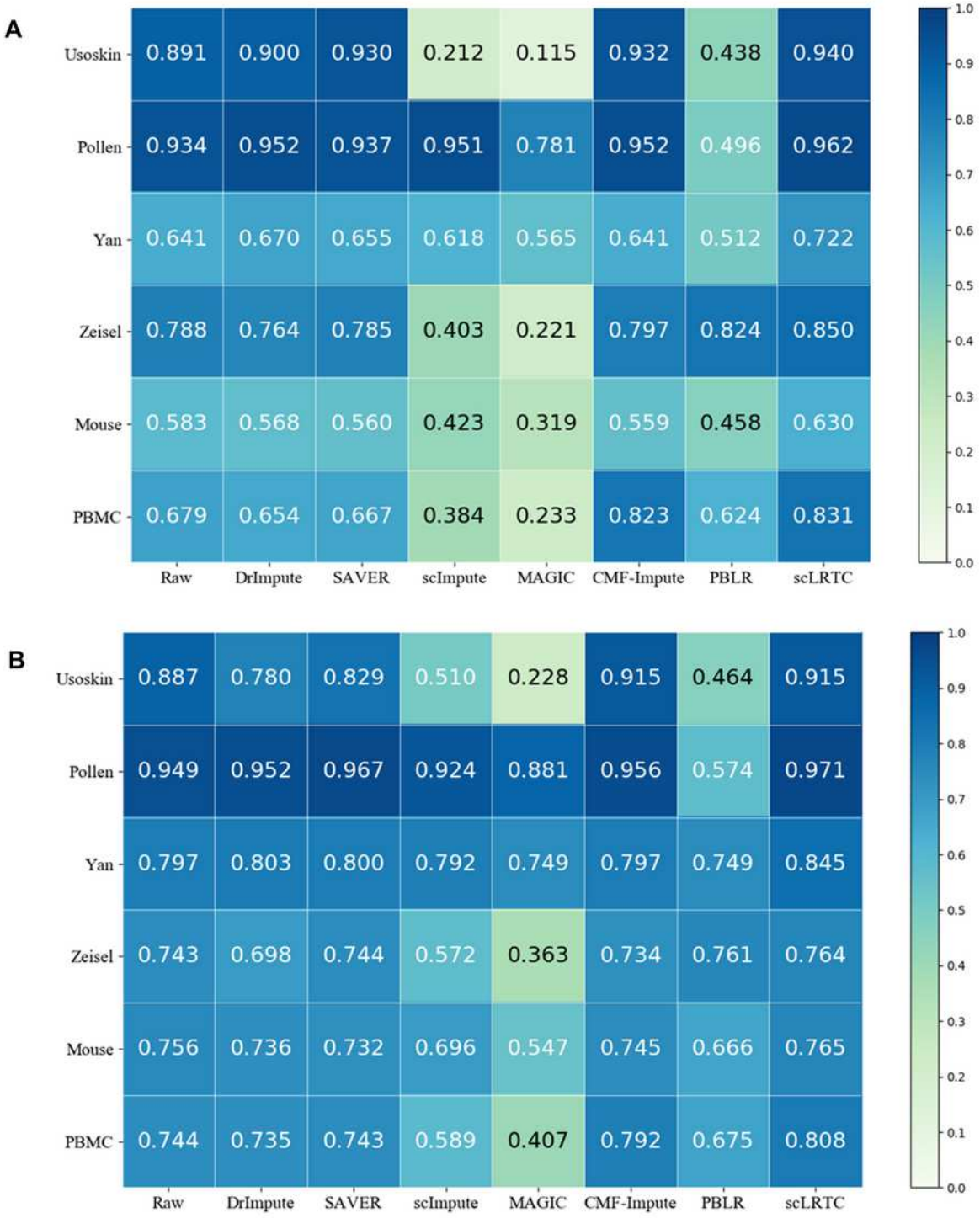

Figure 1

(A) Comparison of NMI indicators obtained by SC3 clustering on 6 scRNA-seq datasets using different imputation algorithms. (B) Comparison of NMI indicators obtained by SC3 clustering on 6 scRNA-seq datasets using different imputation algorithms.

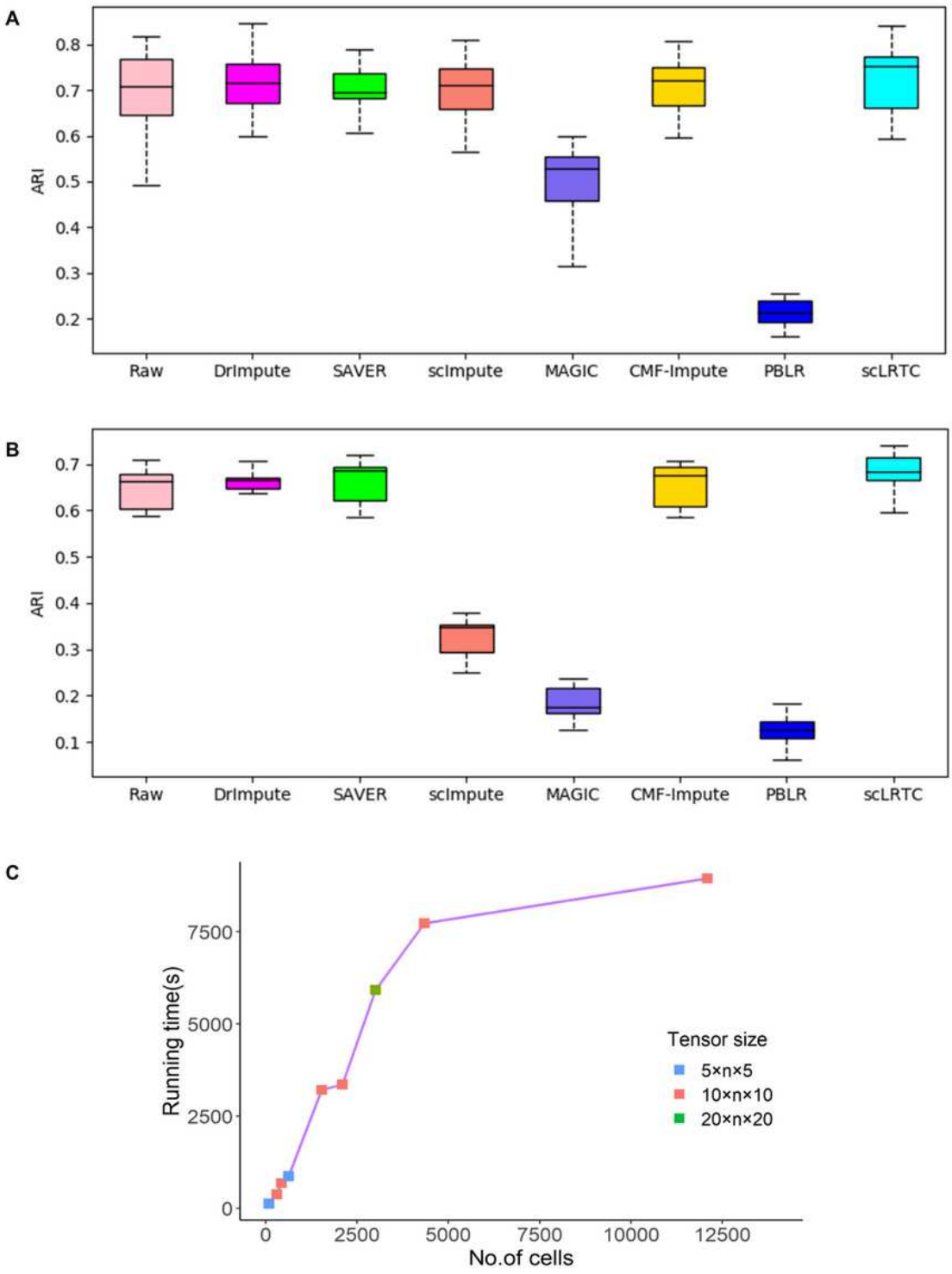

Figure 2

(A) ARI obtained by t-SNE + K-means clustering on the Pollen dataset using different imputation algorithms. (B) ARI obtained by t-SNE + K-means clustering on the Usoskin dataset using different imputation algorithms. (C) Running time of the scLRTC imputation for datasets with different sample sizes. The different tensor size setting is represented by different colors, where n is the number of genes

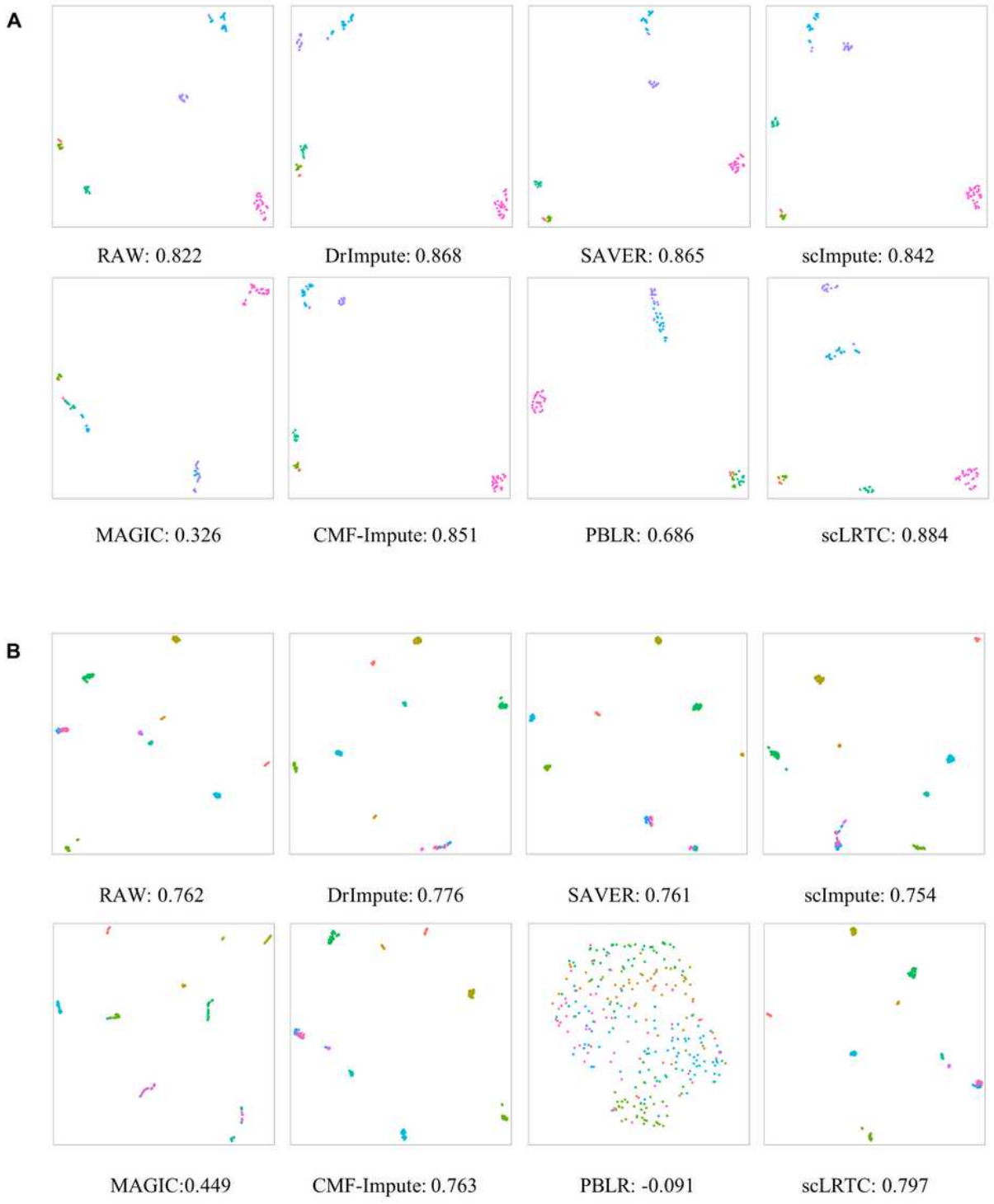

Figure 3

(A) UMAP visualization and SC of the Yan dataset. (B) UMAP visualization and SC of the Pollen dataset.

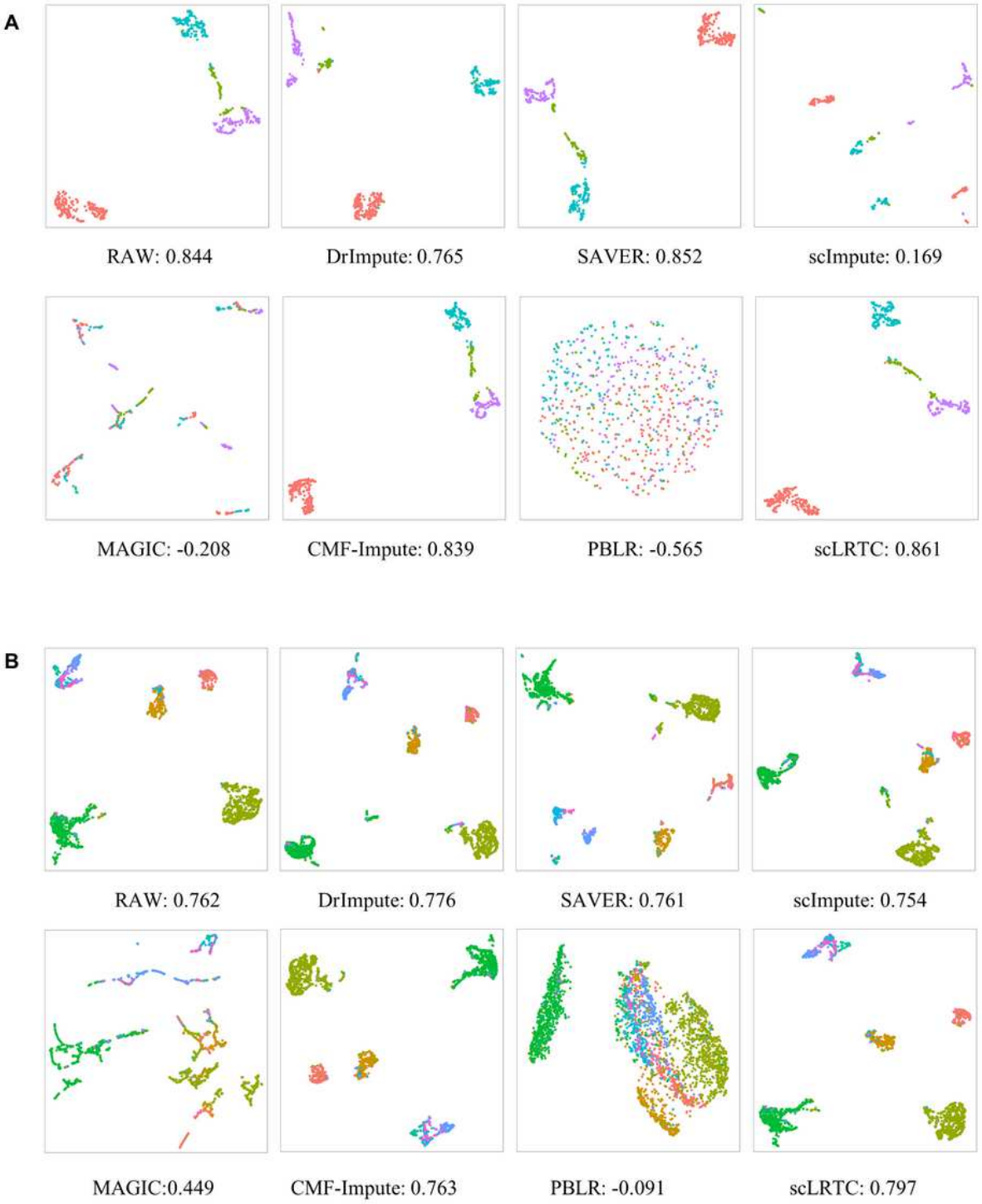

Figure 4

(A) UMAP visualization and SC of the Usoskin dataset. (B) UMAP visualization and SC of the Zeisel dataset.

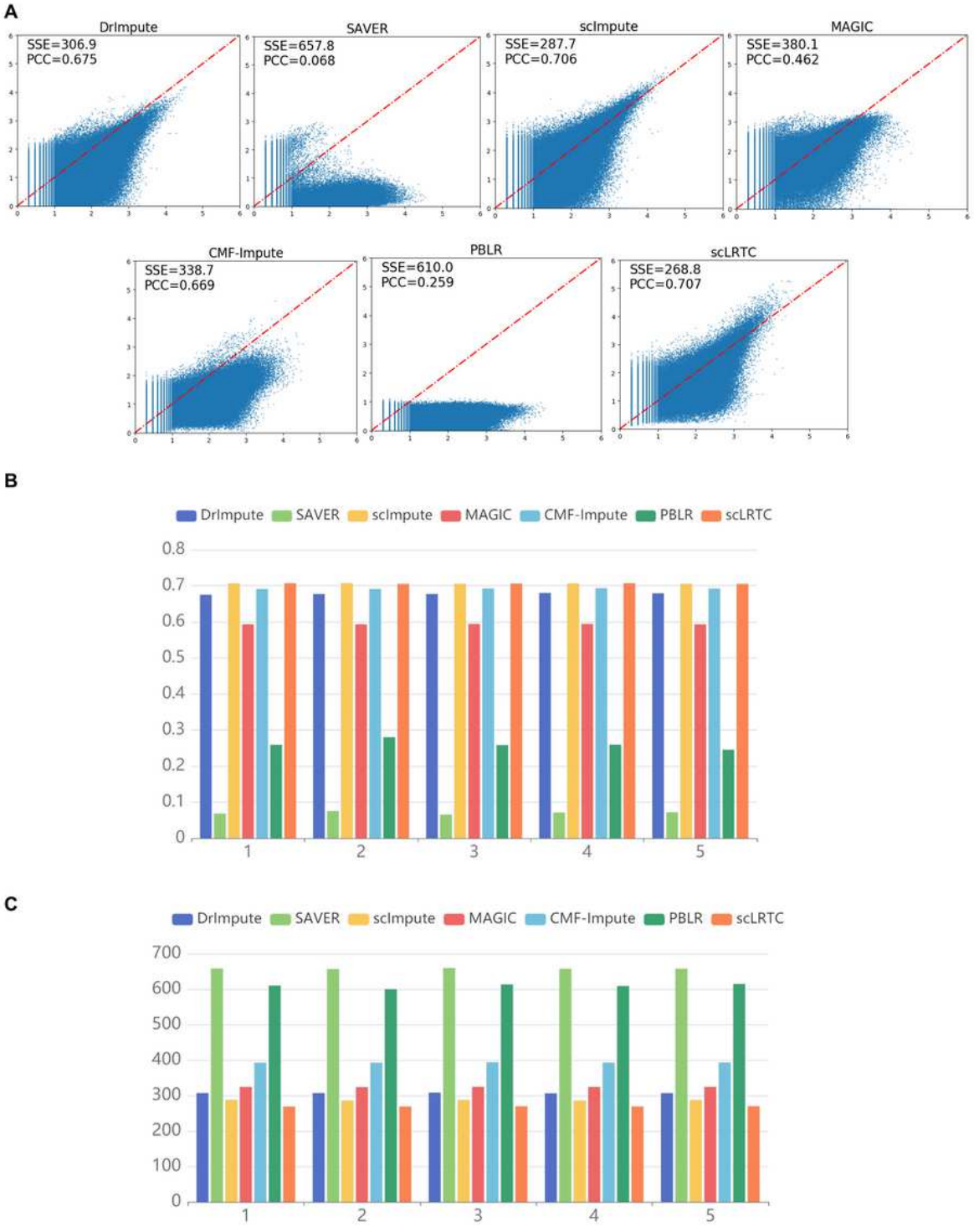

Figure 5

(A) Scatter plots of the masked imputation data and original data. The x-axis corresponds to the true value of the masked data point, and the y-axis represents the imputation value. SSE and PCC are shown in the upper left corner (data is transformed by log10(X +1)). (B) PCC values computed with 7 imputation methods in 5 repeated experiments. (C).

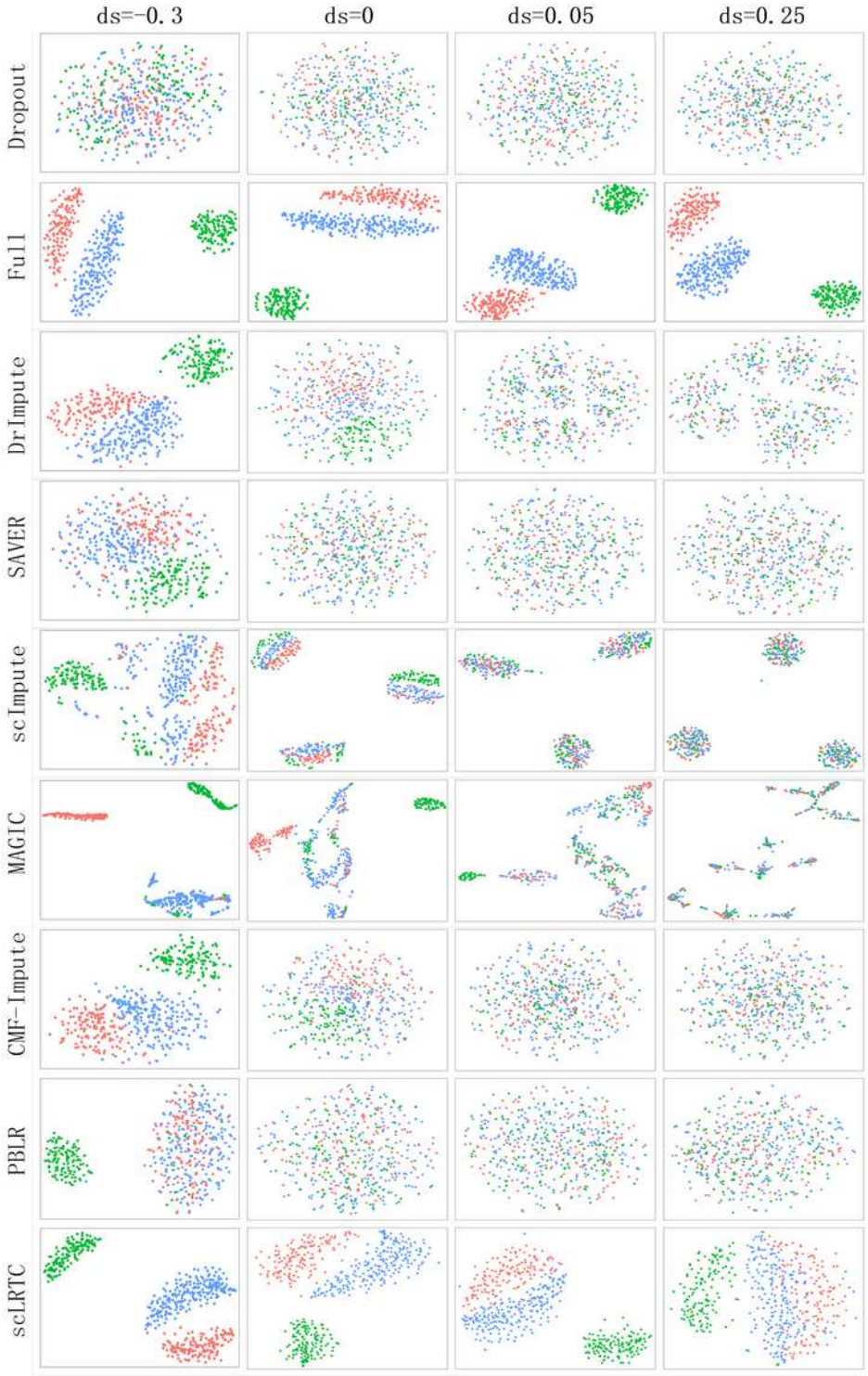

Figure 6

Use t-SNE to visualize the cell gene expression matrix, and use DrImpute, SAVER, scImpute, MAGIC, CMF-Impute, PBLR and scLRTC for imputation. Each column represents a ds, which controls the ratio of zero.

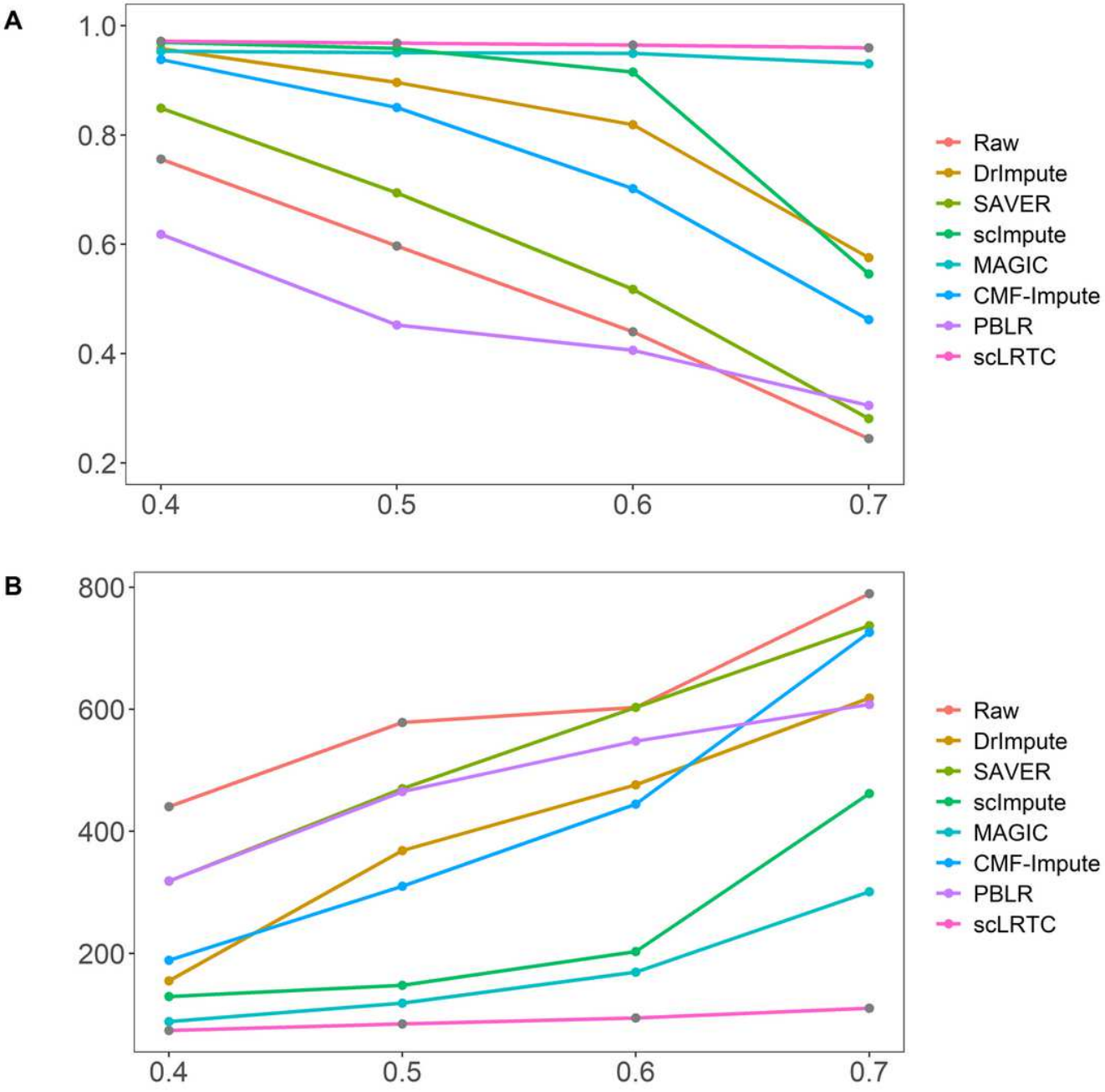

Figure 7

(A) PCC values computed between the Full data and the raw data as well as imputed ones respectively.
(B) SSE values computed between the Full data and the raw data as well as imputed ones respectively.

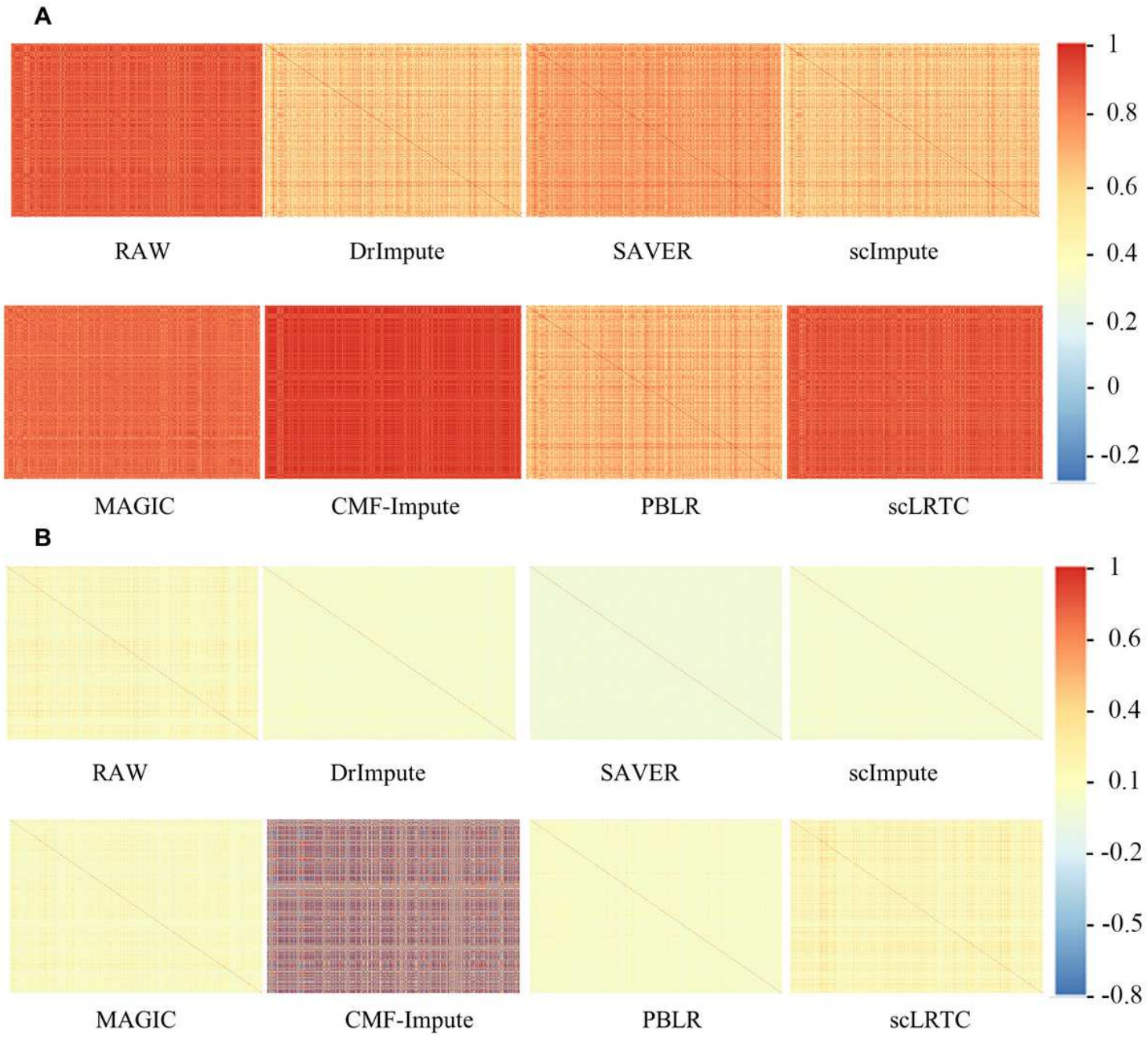

Figure 8

(A) Visualized heat map of cell-cell correlation matrix. (B) Visual-ized heat map of gene-gene correlation matrix.

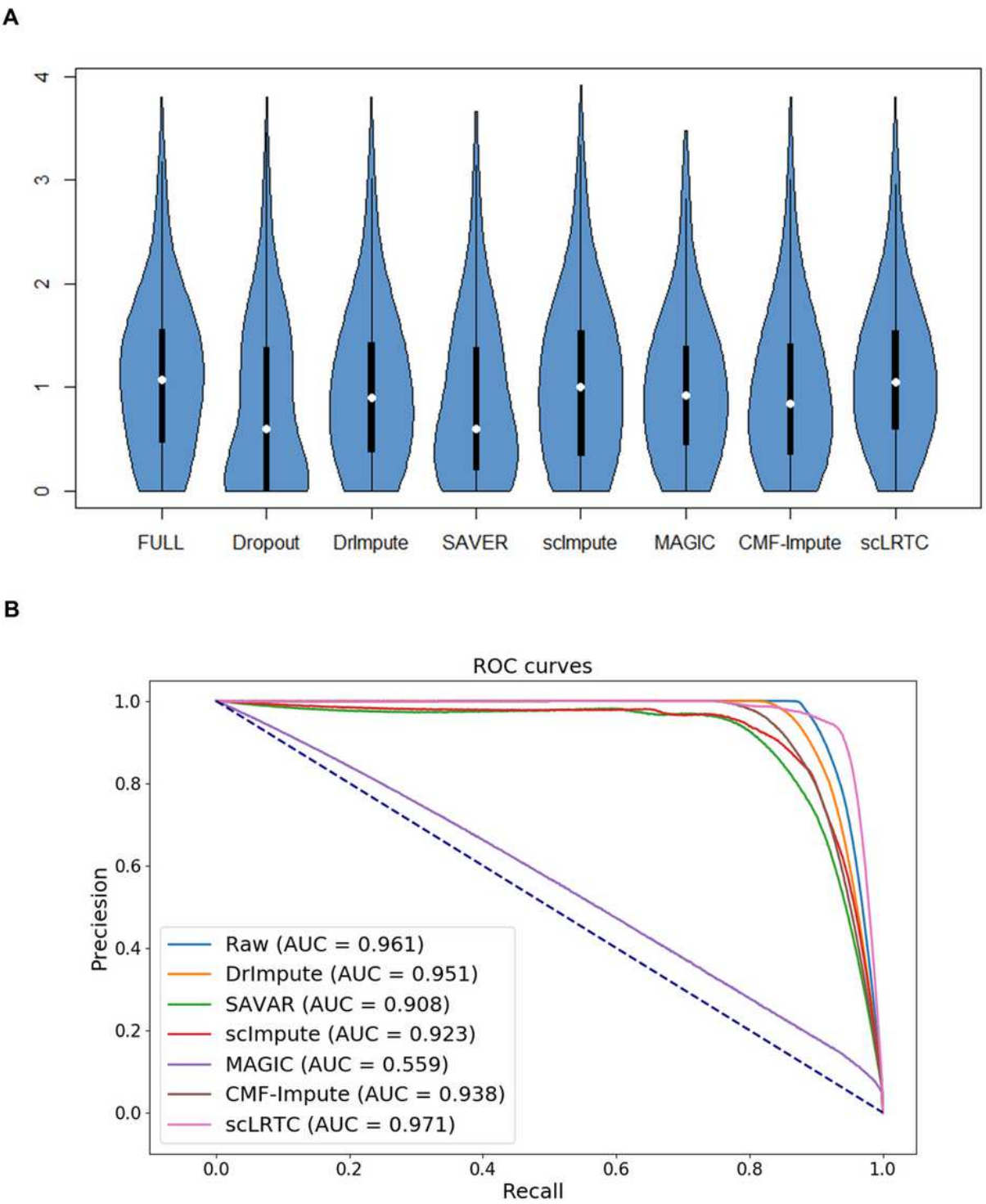

Figure 9

(A) Violin chart of data expression distribution after imputation when the dropout rate is 40% (data is transformed by log10(X +1)). (B) Average ROC of DE genes with different imputation methods.

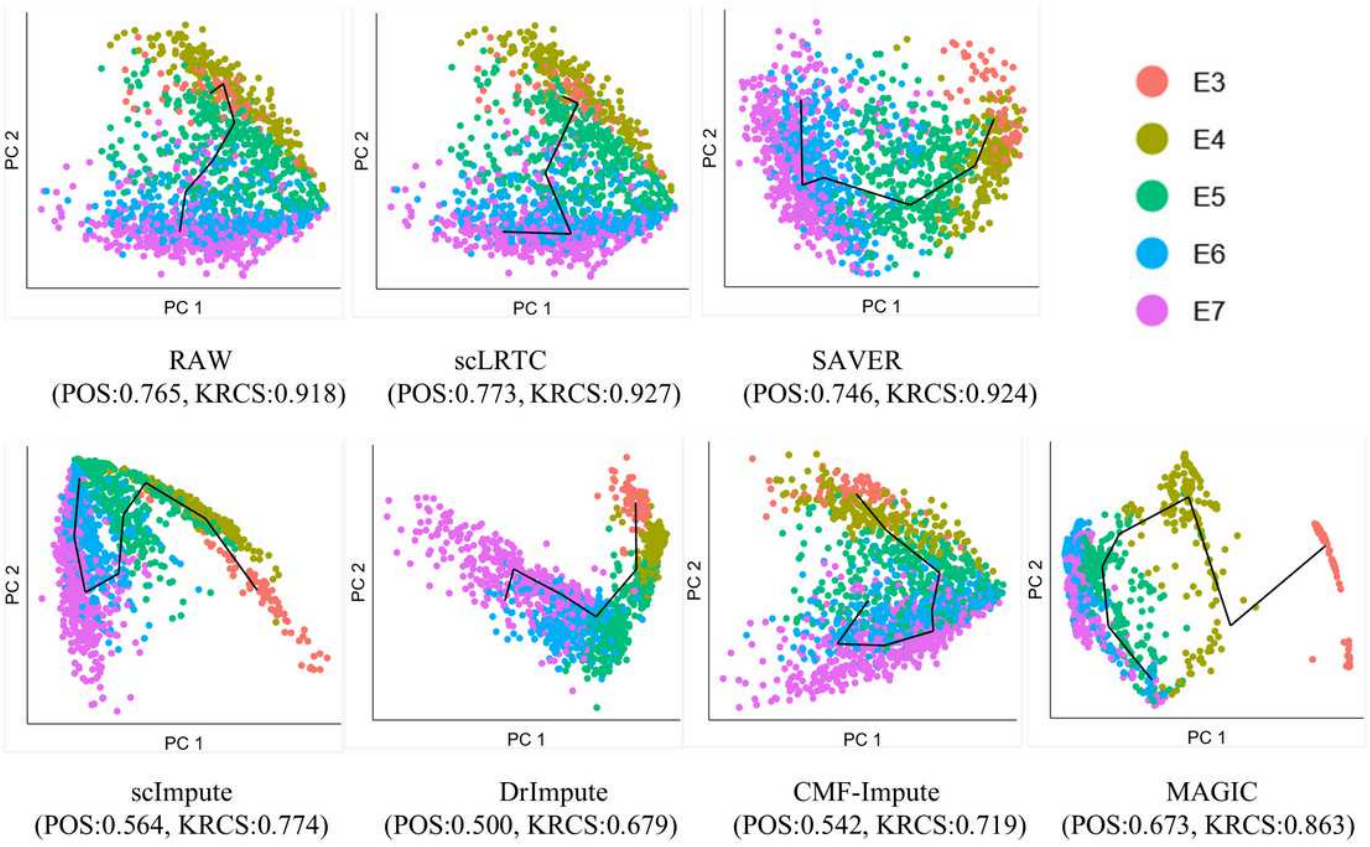

Figure 10

Visualization of lineage reconstruction by TSCAN on the Petropou-los dataset. The POS and Kendall's rank correlation scores of dif-ferent methods are also provided.

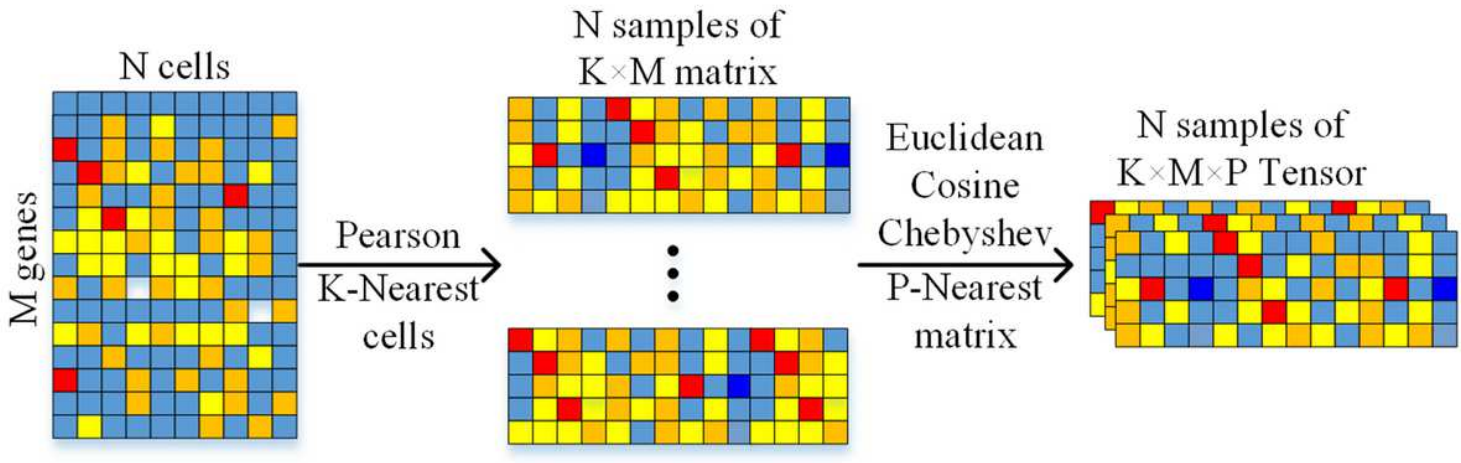

Figure 11

Tensor construction process for the cell Xi

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- table13.pdf