Bioinformatics, 36(10), 2020, 3131–3138
doi: 10.1093/bioinformatics/btaa108
Advance Access Publication Date: 19 February 2020
Original Paper

Gene expression

scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation

Ke Jin[1], Le Ou-Yang[2], Xing-Ming Zhao [3,4], Hong Yan[5] and Xiao-Fei Zhang [1,4,*]

[1]School of Mathematics and Statistics, Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan 430079, China, [2]College of Information Engineering, Shenzhen University, Shenzhen 518060, China, [3]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [4]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai 200433, China and [5]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on September 16, 2019; revised on January 19, 2020; editorial decision on February 10, 2020; accepted on February 12, 2020

Abstract

Motivation: Single-cell RNA sequencing (scRNA-seq) methods make it possible to reveal gene expression patterns at single-cell resolution. Due to technical defects, dropout events in scRNA-seq will add noise to the gene-cell expression matrix and hinder downstream analysis. Therefore, it is important for recovering the true gene expression levels before carrying out downstream analysis.

Results: In this article, we develop an imputation method, called scTSSR, to recover gene expression for scRNA-seq. Unlike most existing methods that impute dropout events by borrowing information across only genes or cells, scTSSR simultaneously leverages information from both similar genes and similar cells using a two-side sparse self-representation model. We demonstrate that scTSSR can effectively capture the Gini coefficients of genes and gene-to-gene correlations observed in single-molecule RNA fluorescence in situ hybridization (smRNA FISH). Down-sampling experiments indicate that scTSSR performs better than existing methods in recovering the true gene expression levels. We also show that scTSSR has a competitive performance in differential expression analysis, cell clustering and cell trajectory inference.

Availability and implementation: The R package is available at https://github.com/Zhangxf-ccnu/scTSSR.

Contact: zhangxf@mail.ccnu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The development of single-cell RNA sequencing (scRNA-seq) technologies provides the measurements of gene expression at single-cell level, which paves the way for studying cellular heterogeneity (Tang *et al.*, 2009). However, dropout events, where a gene is expressed in a cell but not detected, are often observed in scRNA-seq experiments. The resulting gene-cell expression matrix will include many false zeros caused by dropout events, which will corrupt the biological signal and impede downstream analyses, such as cell clustering, data visualization, cell trajectory inference and differential expression analysis. Therefore, it would be useful to impute dropout events in scRNA-seq data before performing downstream analyses.

Several imputation methods designed specifically for scRNA-seq data have been developed recently (Chen *et al.*, 2018; Eraslan *et al.*, 2019; Huang *et al.*, 2018; Kwak *et al.*, 2018; Li and Li, 2018; Linderman *et al.*, 2018; van Dijk *et al.*, 2018; Zhang *et al.*, 2019). The existing methods can be mainly divided into two categories according to how the information from the observed data is used. The first type of methods imputes dropout values by borrowing information from similar genes (Arisdakessian *et al.*, 2019; Eraslan *et al.*, 2019; Huang *et al.*, 2018). That is, the imputed value for a gene is estimated by the observed expression levels of other similar genes in the same cell (Fig. 1A). The second type of methods recovers the expression levels by borrowing information from similar cells (Chen and Zhou, 2018; Kwak *et al.*, 2018; Li and Li, 2018; van Dijk *et al.*, 2018). The predicted expression value of a gene in a cell is obtained by pooling the observed data across similar cells (Fig. 1B). The imputation methods based on similar genes do not make full use of the information shared across cells, while the methods based on similar cells do not take into account similarities among genes. These two types of methods may be suboptimal when the dropout rate is high. For example, the methods based on similar genes cannot make accurate imputation for a cell when the observed

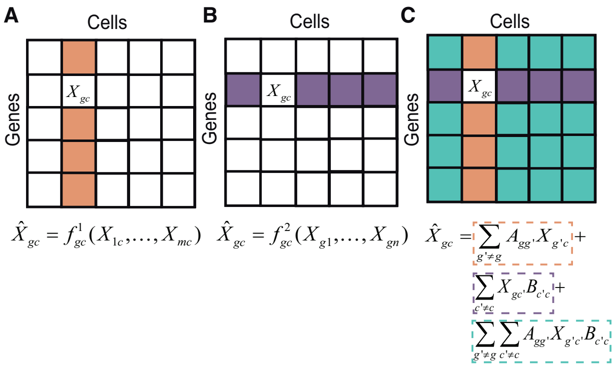
A
Cells
Genes
X_{gc}
$\hat{X}_{gc} = f^1_{gc}(X_{1c},\ldots,X_{mc})$
B
Cells
Genes
X_{gc}
$\hat{X}_{gc} = f^2_{gc}(X_{g1},\ldots,X_{gn})$
C
Cells
Genes
X_{gc}
$\hat{X}_{gc} = \sum_{g'\neq g} A_{gg'}X_{g'c} +$
$\sum_{c'\neq c} X_{gc'}B_{c'c} +$
$\sum_{g'\neq g}\sum_{c'\neq c} A_{gg'}X_{g'c'}B_{c'c}$

Fig. 1. Different strategies for imputing a dropout value, X_{gc}. Here, X_{gc} denotes the expression level of gene g in cell c. The colored blocks represent the expression values that are used to estimate the dropout value by different strategies. (**A**) The first type of methods only uses the expression levels of genes in the same cell (the blocks filled with orange) to impute the dropout value and does not consider expression levels of genes in other cells. The imputed value can be obtained by a linear or non-linear function of the expression levels of genes in the same cell, $\hat{X}_{gc} = f^1_{gc}(X_{1c},\ldots,X_{mc})$. (**B**) The second type of methods only uses the expression levels of gene g in cells similar to the target cell c (the blocks filled with purple) to compute the imputation value and does not take into account expression levels of other genes similar to the target gene g. The imputed value is computed by a linear or non-linear function of the expression levels of gene g in the other cells, $\hat{X}_{gc} = f^2_{gc}(X_{g1},\ldots,X_{gn})$. (**C**) Our method uses the whole expression matrix to compute the imputed value, which simultaneously leverages the expression levels of all other genes in the target cell c (the orange-colored blocks), the expression levels of the target gene g in all other cells (the purple-colored blocks) and the expression levels of all other genes in all other cells (the green-colored blocks). The imputed value can be obtained by a bilinear function of all values in the expression matrix. $A_{gg'}$ is the predictive effect of gene g' on gene g, and $B_{c'c}$ is the predictive effect of cell c' on cell c. (Color version of this figure is available at *Bioinformatics* online.)

expression levels of most similar genes in the cell are zeros. The methods based on similar cells do not perform well if the similar cells contain many dropout values for the target gene of interest.

By assuming that the underlying true gene-cell expression matrix is low rank, several low-rank matrix approximation based imputation methods have been proposed to simultaneously leverage information across cells and genes (Chen *et al.*, 2018; Elyanow *et al.*, 2020; Linderman *et al.*, 2018; Zhang and Zhang, 2018). However, the low-rank assumption is quite strong and does not hold in some situations. If the data consist of discrete cell clusters, it may be low rank. If the clusters are not present and cells are placed on continuous developmental trajectories, the low-rank assumption may not be satisfied and the imputation procedure may produce misleading results (Zhu *et al.*, 2019).

In this study, we propose a new imputation method to recover gene expression for scRNA-seq using a two-side sparse self-representation (scTSSR) model. scTSSR simultaneously learns two non-negative sparse self-representation matrices to capture the gene-to-gene and cell-to-cell similarities. The dropout values are imputed by a bilinear combination of similar genes and cells (Fig. 1C). Following the method of Huang *et al.* (2018), we also couple scTSSR to a Bayesian hierarchical model, where the final imputed value is obtained by a weighted average of the value imputed by scTSSR and the raw read count. To assess the performance of the proposed method, we carry out comprehensive real scRNA-seq data analyses. We first evaluate the accuracy by comparing the imputed data to the data derived from single-molecule RNA fluorescence *in situ* hybridization (smRNA FISH). scTSSR performs better than the compared methods in terms of recovering Gini coefficient of genes and preserving gene–gene correlations. Next, we carry out downsampling experiments and find that scTSSR has a competitive performance in recovering the true expression levels. The comparable performance of scTSSR is also demonstrated in terms of differential expression analysis, cell clustering and cell trajectory inference. These results indicate that scTSSR is a powerful tool for enhancing biological discovery in scRNA-seq data analysis.

2 Materials and methods

2.1 Model

The input to our method is an $m \times n$ normalized expression matrix X with rows and columns representing genes and cells, respectively (In this study, library size normalization and log-transformation are implemented on the raw read count matrix. For details, refer to Supplementary Section S3.1). The goal of this study is to construct an imputed matrix $\hat{X}$ from X so that the dropout values in X can be recovered accurately.

We use the self-representation models to impute the data. Self-representation models aim to find a combination of other data points to represent each target data point in the dataset (Elhamifar and Vidal, 2013). Given the normalized expression matrix X, we can represent the expression level of gene g in cell c using the expression levels of gene g in all other cells following previous studies (Fig. 1B) (Chen and Zhou, 2018; Li and Li, 2018). Mathematically, X_{gc} can be represented as

$$X_{gc} = \sum_{c'\neq c} X_{gc'}B_{c'c} + E_{gc}, \tag{1}$$

where $B_{c'c}$ is the representation coefficient and E_{gc} is noise. $B_{c'c}$ can capture the similarity between cells c' and c and be interpreted as the predictive effect of cell c' on cell c. Let $\hat{B}_{c'c}$ be the estimate of $B_{c'c}$, then we can compute the imputed value for gene g in cell c as $\hat{X}_{gc} = \sum_{c'\neq c} X_{gc'}\hat{B}_{c'c}$. In doing so, only expression levels of gene g in cells similar to the target cell c are leveraged to impute the data, and expression levels of other genes similar to the target gene g are neglected.

To leverage information shared across genes, we can also represent the expression level of gene g in cell c using other genes in cell c (Fig. 1A) (Huang *et al.*, 2018)

$$X_{gc} = \sum_{g'\neq g} A_{gg'}X_{g'c} + E_{gc}, \tag{2}$$

where $A_{gg'}$ is the representation coefficient and can capture the similarity between genes g and g'. After obtaining the estimated representation coefficient $\hat{A}_{gg'}$, we can compute the imputed value for gene g in cell c as $\hat{X}_{gc} = \sum_{g'\neq g} \hat{A}_{gg'}X_{g'c}$. This representation method only uses expression levels of genes in the same cell, and expression levels of genes in other cells are not considered.

Equations (1) and (2) are one-side self-representation models. Equation (1) focuses on representing the data matrix according to cells (the columns of X), and only uses information across cells to impute dropout values. Equation (2) represents the data matrix according to genes (the rows of X) and only uses similar genes to recover the data. To make full use of the information shared across genes and across cells, we propose a two-side sparse self-representation model for recovering scRNA-seq data (scTSSR)

$$X_{gc} = \sum_{g'\neq g} A_{gg'}X_{g'c} + \sum_{c'\neq c} X_{gc'}B_{c'c} \sum_{g'\neq g}\sum_{c'\neq c} A_{gg'}X_{g'c'}B_{c'c} + E_{gc}, \tag{3}$$

where $A_{gg'}$ is the similarity between genes g and g', and $B_{c'c}$ is the similarity between cells c' and c. Here, we use a summation of three terms to represent the expression level of gene g in cell c (Fig. 1C). The first term is a weighted summation of the expression levels of all other genes in the same cell. The second term is a weighted summation of the expression levels of the same gene across all other cells. The third term is a weighted summation of the expression levels of all other genes across all other cells. To reduce the complexity of the model, the similarity between $X_{g'c'}$ and X_{gc} is set to $A_{gg'}B_{c'c}$. Here, we assume that if genes g and g' are similar and cells c and c' are also similar, the expression level of gene g' in cell c' would be similar to the expression level of gene g in cell c. Therefore, we use the product $A_{gg'}B_{c'c}$ to capture the similarity between X_{gc} and $X_{g'c'}$, following the method of previous link prediction studies (Zhao *et al.*, 2017). Here, we prefer scTSSR to be a sparse model since we expect only a small set of cells and genes to be informative for imputing a target dropout value, which means most values of $A_{gg'}$ and $B_{c'c}$

would be zeros. Once we obtain the estimates of $\hat{A}_{gg'}$ and $\hat{B}_{c'c}$, the imputed value can be predicted as

$$\hat{X}_{gc} = \sum_{g' \neq g} \hat{A}_{gg'} X_{g'c} + \sum_{c' \neq c} X_{gc'} \hat{B}_{c'c} + \sum_{g' \neq g} \sum_{c' \neq c} \hat{A}_{gg'} X_{g'c'} \hat{B}_{c'c}. \qquad (4)$$

Note that Huang *et al.* (2018), Chen and Zhou (2018) and Li and Li (2018) also assume that the relationships between genes or cells are linear. For example, by assuming the count of each gene in each cell follows a Poisson-gamma mixture, Huang *et al.* (2018) estimate the prior mean parameters using a Poisson LASSO regression, where the expression levels of other genes are used as predictors, and the posterior mean is used to impute dropout values. Li and Li (2018) first compute the dropout probability of each gene in each cell, then use a penalized linear model to impute the dropout values. Chen and Zhou (2018) first use the lasso regression to select a small set of candidate cells and compute the dropout probabilities, then apply the quadratic programming algorithm for final imputation. When using the regression models to capture the linear relationships between genes or cells, Huang *et al.* (2018) only use the first term of our model to compute the prior mean, and Chen and Zhou (2018) and Li and Li (2018) only use the second term of our model to make predictions. Here, we use a summation of the two terms so that the expression levels of all other genes in the target cell and the expression levels of the target gene in all other cells are used. In addition, we also use the third term to borrow information from all other genes across all other cells $X_{g'c'}$ ($g' \neq g$, $c' \neq c$). The predictive effect of $X_{g'c'}$ on X_{gc} is $\hat{A}_{gg'}\hat{B}_{c'c}$.

We estimate the representation coefficients $\hat{A}_{gg'}$ and $\hat{B}_{c'c}$ from the normalized expression data using a penalized least square method. The optimization problem in the matrix form is

$$\begin{array}{ll} \min_{A,B} & \|X - (AX + XB + AXB)\|_F^2 + \lambda(\|A\|_1 + \|B\|_1) \\ \text{subject to} & A \geq 0,\ B \geq 0,\ \mathrm{diag}(A) = 0,\ \mathrm{diag}(B) = 0. \end{array} \qquad (5)$$

Here, A and B represent the row and column representation coefficient matrices of X, and are used to capture the gene similarities and cell similarities, respectively. $\|\cdot\|_F$ and $\|\cdot\|_1$ are the Frobenious norm and elementwise ℓ_1 norm of a matrix. The first term is the square representation error, and the second term uses the elementwise ℓ_1 norm to promote sparsity of the representation coefficient matrices. λ is a non-negative tuning parameter. We use the constraints $A \geq 0$ and $B \geq 0$ to make sure that the representation coefficients are non-negative so that they can be naturally interpreted as imputation weights (Chen and Zhou, 2018; Li and Li, 2018). Note that negative associations between cells or genes also exist and can be used. Here, we only consider non-negative associations so that the product $A_{gg'}B_{c'c}$ can naturally define the similarity between $X_{g'c'}$ and X_{gc}. diag$(\cdot) \in R^n$ is the vector of the diagonal elements of a matrix. The constraints diag$(A) = 0$ and diag$(B) = 0$ are used to eliminate the trivial solution of representing an expression level as a linear combination of itself. We propose a coordinate descent algorithm to solve the optimization problem (Section 2.2).

After obtaining the solution to (5), we can impute the data according to (4). However, this method may be suboptimal since it does not consider the predictability or provide a measure of uncertainty for the estimated values. Therefore, we couple scTSSR to the Bayesian model used in SAVER (Huang *et al.*, 2018) (Supplementary Section S3.2). SAVER assumes that the observed count expression of each gene in every cell follows a negative binomial distribution. A Bayesian method is used to calculate posterior distribution of every observed count. The mean of the posterior distribution is treated as the final imputation value, which is a weighted average of the scTSSR predicted value and the raw read count. All parameters (expect for the prior mean) in the Bayesian model are estimated using SAVER. The prior mean is estimated as $\hat{\mu}_{gc} = e^{\hat{X}_{gc}}$, where $\hat{X}_{gc}$ is the scTSSR predicted value obtained from Equation (4). We use the function saver in the R package SAVER with setting "mu $=\hat{\mu}$" to compute the final imputed values. Further data analysis steps are then carried out based on the final imputed values.

2.2 Optimization algorithm

We develop a coordinate descent algorithm to solve the optimization problem (5). At each iteration, we minimize the objective function with respect to one representation coefficient matrix while keeping the other one fixed. We first keep A fixed and minimize (5) with respective to B. The optimization problem can be rewritten as

$$\begin{array}{ll} \min_{B} & \|(X - AX) - (A + I)XB\|_F^2 + \lambda\|B\|_1 \\ \text{subject to} & B \geq 0,\ \mathrm{diag}(B) = 0. \end{array} \qquad (6)$$

This is a non-negative constrained sparse learning problem, and Keras is used to solve it (Supplementary Section S3.3). Minimizing (5) with respective to A can be rewritten as

$$\begin{array}{ll} \min_{A} & \|(X - XB) - AX(B + I)\|_F^2 + \lambda\|A\|_1 \\ \text{subject to} & A \geq 0,\ \mathrm{diag}(A) = 0. \end{array} \qquad (7)$$

This optimization problem can also be solved by the algorithm used to solve (6). We cyclically update B and A by solving (6) and (7) until the convergence condition is satisfied. The complete algorithm is presented in Supplementary Section S3.3.

2.3 Tuning parameter selection

Our model (5) has a tuning parameter λ that controls the sparsity level of the estimated coefficient matrices. We determine λ according to a random-matrix-based technology called Gordon's Theorem (Vershynin, 2010). As suggested by this theorem, the proper lasso penalty parameter λ should be set at the order of the standard deviation of the noises (Zhao and Yu, 2006). To choose the proper λ, we need to obtain the noise matrix first. We simply take the mean of the normalized expression matrix X as the imputation estimate of X (Note that the imputation estimate obtained by more complex methods can be also used. Here, we use the simple estimate for the sake of simplicity). So the noise matrix can be calculated by $X - mean(X)$. Then the appropriate λ is set as $\lambda = sd(X - mean(X)) = sd(X)$, where $sd(X)$ is the estimate of the standard deviation of the noise. By selecting the tuning parameter through this way, we can avoid the tedious procedure of trying many different parameter values and to achieve a preferable performance.

3 Results

We compare our method with eleven existing imputation methods to evaluate the performance: ALRA (Linderman *et al.*, 2018), AutoImpute (Talwar *et al.*, 2018), DrImpute (Kwak *et al.*, 2018), MAGIC (van Dijk *et al.*, 2018), SAVER (Huang *et al.*, 2018), scImpute (Li and Li, 2018), SCRABBLE (Peng *et al.*, 2019), scRMD (Chen *et al.*, 2018), scVI (Lopez *et al.*, 2018), VIPER (Chen and Zhou, 2018) and ZINB-WaVE (Risso *et al.*, 2018). Due to the lack of gold standard in evaluating the imputation of scRNA-seq using real data, we conduct the following experiments to assess the accuracy: (i) comparison between imputed scRNA-seq data and smRNA FISH data, (ii) down-sampling experiments, (iii) differential expression analysis, (iv) cell clustering and (v) pseudotime trajectory analysis.

3.1 Evaluating imputation accuracy by comparing to smRNA FISH data

smRNA FISH is a single-cell transcriptomic profiling method that is complementary to scRNA-seq. Compared to scRNA-seq, smRNA FISH has the advantage of measuring gene expression levels with high accuracy. Therefore, the smRNA FISH can be used as a reference to evaluate the imputed scRNA-seq data (Huang *et al.*, 2018). We obtain a Drop-seq dataset and the corresponding smRNA FISH dataset from a melanoma cell line (Torre *et al.*, 2018). We preprocess the data using the method from Huang *et al.* (2018). The preprocessed Drop-seq data consists of 12 241 genes and 8498 cells, and the smRNA FISH data contains 88 040 cells and 26 genes. There are 15 genes common to the smRNA FISH and Drop-seq datasets. Since the cells included in the smRNA FISH and Drop-seq datasets are different, we can only compare the distributions of imputed scRNA-seq data to the distribution of smRNA FISH data. Following Huang *et al.* (2018), we focus on two types of measures. The first one is the Gini coefficient that is a measure of gene's expression variability, which is useful for identifying rare cell types and sporadically expressed genes. The second one is the gene-to-gene correlations, which is important for gene network reconstruction.

We run scTSSR and the 10 individual imputation methods (ALRA, AutoImpute, DrImpute, MAGIC, SAVER, scImpute, scRMD, scVI, VIPER and ZINB-WaVE) on the Drop-seq data and calculate the Gini coefficients for the 15 genes common to the FISH and Drop-seq datasets. We do not run SCRABBLE on the Drop-seq data due to that the computation time is beyond 24 h on a computer with 64GB RAM. Before calculating Gini coefficients, we filter and normalize the data by a GAPDH factor as that performed in Huang *et al.* (2018). We use Pearson correlation coefficient and root-mean-square error (RMSE) between Gini coefficients calculated by the imputed data (or Drop-seq data) and those by the smRNA FISH data as the metrics to assess the recovery of Gini coefficients. Figure 2 shows that ALRA performs best, and our scTSSR performs better than the other nine compared methods and the observed Drop-seq data. We also evaluate the quality of imputed data in terms of the recovery of the gene-to-gene correlations observed in smRNA FISH data. We calculate the gene-to-gene correlation matrix using Pearson correlation coefficient. The Pearson correlation coefficient between gene-to-gene correlation matrix calculated by the imputed data (or Drop-seq data) and that by the smRNA FISH data is used to quantify the performance. As can be seen from Supplementary Figure S1, scTSSR outperforms all other imputation methods with the highest Pearson correlation coefficient, while the two low-rank approximation based methods, ALRA and scRMD, lead to biased estimates of the true correlations. The correlations derived from MAGIC results are much higher than those derived from FISH, while AutoImpute and DrImpute underestimate the gene-to-gene correlations. These results indicate that scTSSR can accurately recover the true distributions observed in smRNA FISH but dampened in Drop-seq.

3.2 Evaluating imputation accuracy through down-sampling experiments

Since it is hard to obtain the gold standard of the true expression levels, we carry out down-sampling experiments to evaluate the performance of different imputation methods. Two down-sampling experiments are conducted. The first one is the same to that conducted by Huang *et al.* (2018), which evaluates the performance using different measures (e.g. correlation with the reference data, cell clustering and t-SNE visualization). The second one is conducted following the method of Chen and Zhou (2018), which evaluates the performance with different down-sampling rates.

We first carry out the down-sampling experiments implemented in Huang *et al.* (2018). We use the four datasets used in Huang *et al.* (2018) to evaluate the accuracy: Baron (Baron *et al.*, 2016), Chen (Chen *et al.*, 2017), La Manno (La Manno *et al.*, 2016), Zeisel (Zeisel *et al.*, 2015). Given a dataset, Huang *et al.* (2018) first selected a subset of cells and genes with high expression to generate a reference dataset, and generated a down-sampled observed dataset from the reference dataset using down-sampling simulation. We download the reference and observed data for the four datasets from https://github.com/mohuangx/SAVER-paper/tree/master/SAVER-data. We run the 12 imputation methods on each observed dataset and evaluate the performance by comparing the imputed data with the reference data using different measures. We first evaluate the performance in terms of correlation with the reference data. The Pearson gene-wise correlation across cells and the Pearson cell-wise correlation across genes between the reference data and the imputed data (or observed data) are calculated. Figure 3A shows that scTSSR achieves the highest gene-wise correlations on the Baron dataset and the La Manno dataset, and performs as the second best on the other datasets. scTSSR also achieves the highest cell-wise correlations on three out of the four datasets (e.g. the Baron dataset, the La Manno dataset and the Zeisel dataset) and performs as the second best on the Chen dataset. Next, we investigate the effect of different imputation methods on cell clustering and visualization. We use the R package Seurat (Butler *et al.*, 2018) to carry out cell clustering and t-SNE visualization, following the method used in SAVER (Huang *et al.*, 2018). We run clustering algorithm and t-SNE on the data imputed by different methods, observed data and reference data. The cell clusters derived from the reference data are treated as the truth. The clustering accuracy is assessed in terms of the Jaccard index and t-SNE visualization. The higher the Jaccard index, the better the clustering performance. scTSSR achieves competitive Jaccard index scores on all datasets (Fig. 3B and Supplementary Fig. S2). From the t-SNE plots, we find that the data imputed by scTSSR can provide a clear representation of the clusters identified from the reference data.

We then carry out the down-sampling experiments implemented in Chen and Zhou (2018) to evaluate the imputation performance with different down-sampling rates. Three datasets used in Chen and Zhou (2018) are considered: Grun (Grun *et al.*, 2014), Cell Type (Chu *et al.*, 2016) and Time Course (Chu *et al.*, 2016). Genes that are expressed in <10% of the cells are filtered out. The down-sampling method adopted here consists of two steps. Firstly, for each gene in turn, we generate initial down-sampled data using a multinomial distribution with smaller library sizes corresponding to different down-sampling rates. Secondly, we set each down-sampled entry of the initially generated data to zero according to its dropout probability computed by a logistic model. After down-sampling, we finally have zero values resulting from low expression values in the original data and the subsequent multinomial down-sampling or the extra dropout events (for detail, refer to Chen and Zhou, 2018). In this experiment, the down-sampling rate, which represents the

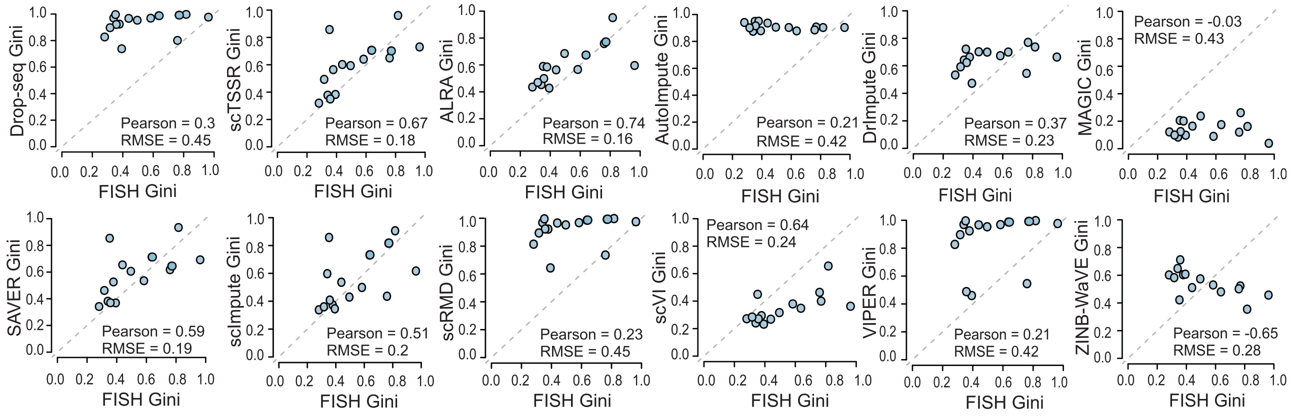

Fig. 2. Comparison of the Gini coefficients computed from the imputed data (or Drop-seq data) to those computed from the smRNA FISH data

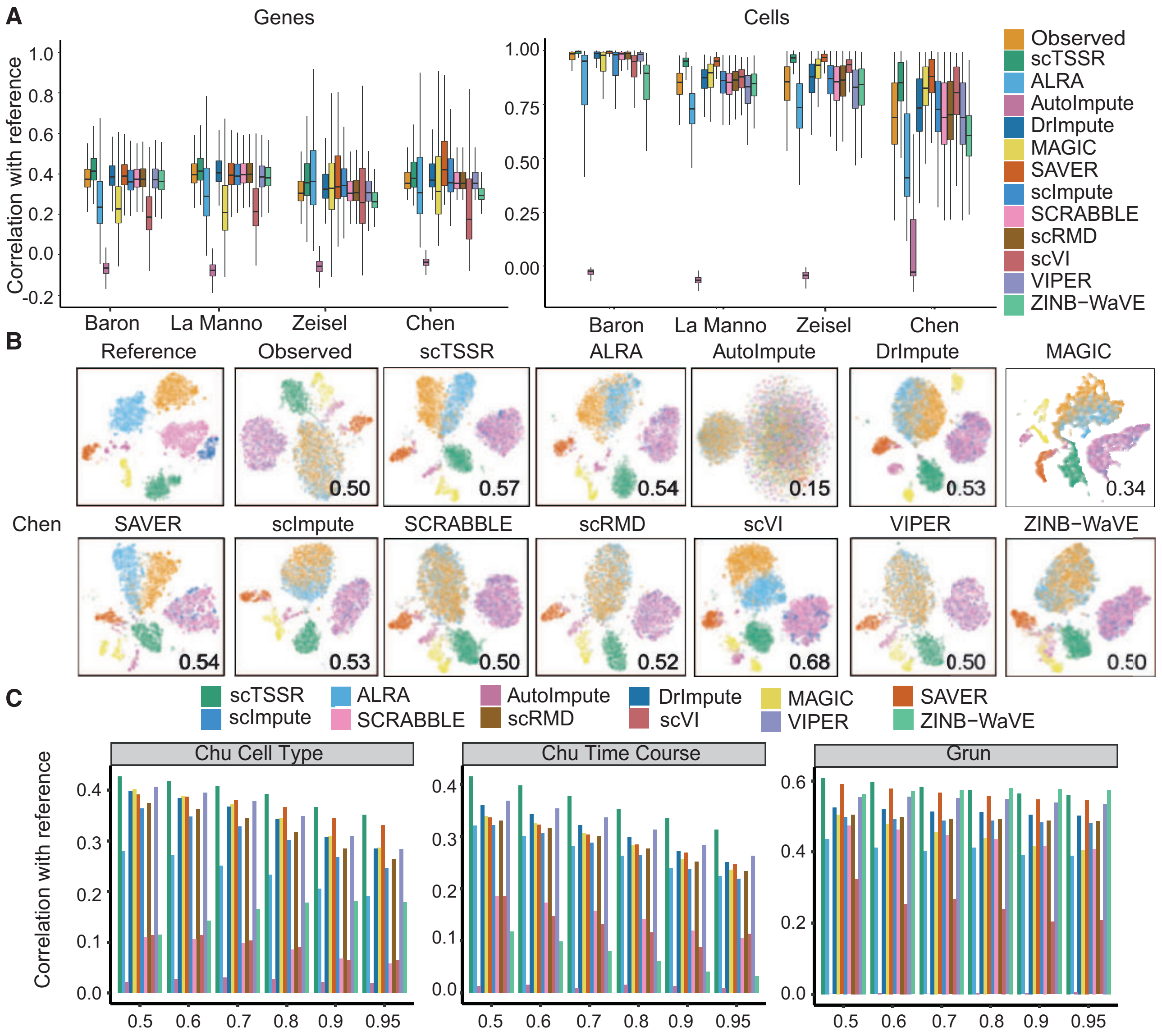

Fig. 3. Evaluation of imputation methods by down-sampling experiments. (**A**) Performance of different imputation methods evaluated by gene-wise correlation with the reference data (left) and cell-wise correlation with the reference data (right). (**B**) Cell clustering and t-SNE visualization of the Chen data. The Jaccard index, used for comparing the similarity between the cell clusters derived from the imputed data (or observed data) and those derived from the reference data, is displayed in the bottom right. The cells are color-coded by the clusters identified from the reference data. (**C**) Imputation accuracy of different methods measured by Pearson correlation between imputed values and the reference with different down-sampling rates. The *X*-axis denotes the down-sampling rates, and the *Y*-axis denotes the correlations. (Color version of this figure is available at *Bioinformatics* online.)

down-sampled proportion of the original read depth, is set to 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95. To assess the performance of each method, we calculate the Pearson correlation across all entries between the original (reference) data and the imputed data. scTSSR outperforms other imputation methods on all the three datasets with all down-sampling rates (Fig. 3C).

3.3 Evaluating imputation accuracy through differential expression analysis

Differential expression analysis, which can be used to identify genes whose expression levels are changed between two conditions or cell types, is useful for characterizing the molecular mechanisms underlying the change. The performance of differential analysis methods will be deteriorated by the dropout events in scRNA-seq. We run the differential analysis methods on the observed and imputed data to illustrate the benefit of imputation. The experiments are conducted following the methods of Eraslan *et al.* (2019), Li and Li (2018) and Chen *et al.* (2018). Due to the lack of gold standard of differentially expressed genes, the differentially expressed genes identified from bulk RNA-seq data are considered as the reference. We consider two real datasets. The first real dataset consists of both bulk RNA-seq and scRNA-seq experiments on human embryonic stem cells (ESC) and definitive endoderm cells (DEC) (Chu *et al.*, 2016). The scRNA-seq data include 138 DEC cells and 212 H1 ESC cells, and the bulk RNA-seq data consists of 4 H1 ESC samples and 2 DEC samples. The second real dataset consists of both bulk RNA-seq and scRNA-seq experiments on human embryonic stem cells (ESC) and endothelial cells (EC) (Chu *et al.*, 2016). The scRNA-seq data includes 105 EC cells and 212 H1 ESC cells, and the bulk RNA-seq data consists of four H1 ESC samples and three EC samples. We use edgeR (Robinson *et al.*, 2010) to identify differentially expressed genes after imputation, and all parameters are set as the default values.

Here, we only consider the 2000 highly variable genes identified from the scRNA-seq data by the function FindVariableFeatures in the R package Seurat (parameters are set to default values). Top 200 genes ranked by adjusted P values from the bulk data are considered as the reference. We use the receiver operating characteristic (ROC) curve to evaluate the performance of each imputation method, where the differentially expressed genes identified from the imputed (and observed) data with different adjusted P value thresholds are compared with the reference. We first carry out the differential expression analysis on the first real dataset (H1 versus DEC). Considering that some of the imputation methods may depend on the random seeds, we run each imputation method 10 times and compute the standard error of the performance metrics, and use black interval to represent the result plus or minus standard error. We plot the ROC curve for each method and calculate the area under curve (AUC) score as well (Fig. 4A and Supplementary Fig. S3). scTSSR outperforms all the other methods. In addition, we also consider the top 400 genes ranked by adjusted P values from the bulk data as the gold standard (Supplementary Fig. S4). scTSSR achieves the highest AUC score among all imputation methods. However, how to set the cutoff to define the reference will influence the ROC analysis results. In order to reduce the uncertainty of the results due to such influence, we calculate the Spearman correlation coefficient between the adjusted P values calculated from the bulk data and those from the imputed data (or observed data). Figure 4B shows that scTSSR performs best on differential expression analysis. We then carry out the differential expression analysis on the second real dataset (H1 versus EC). scTSSR still performs best (Supplementary Fig. S5). The results indicate that the differentially expressed genes identified by scTSSR are consistent with the reference derived from the bulk data.

3.4 Evaluating imputation accuracy through cell clustering

We run different imputation methods on scRNA-seq datasets whose cell labels are well defined through extensive analysis. We consider four datasets: Pollen (Pollen *et al.*, 2014), iPSC (Gong *et al.*, 2018), Guo (Guo *et al.*, 2015) and peripheral blood mononuclear cell (PBMC) (Zheng *et al.*, 2017). The Pollen data includes 23 794 genes and 299 cells from 11 populations. The iPSC data includes 15 724 genes and 315 cells from five different cell types. The Guo data consists of 23 787 genes and 317 cells from 19 different cell types. After filtering out genes that are expressed in <5% of the cells, the Pollen data, the Guo data and the iPSC data include 14 203, 14 898 and 11 960 genes, respectively. The PBMC data consists of 32 738 genes and 5132 cells from five different cell types. We use function FindVariableFeatures in the R package Seurat with default parameter settings to identify 2000 high variable genes of the PBMC data. We first run different imputation methods on each observed dataset, and then use SC3 (Kiselev *et al.*, 2017) which is widely used to carry out clustering analysis on the imputed or observed data. We use R package SC3 to implement SC3, and the number of clusters is chosen as the known number of cell types. Considering that some of the imputation methods may depend on the random seeds, we run each imputation method ten times on the iPSC dataset and compute the standard error of the performance metric adjusted Rand index (ARI), and use black interval to represent the result plus or minus standard error (Fig. 5B).

We use the ARI as the metric to assess the clustering performance. A higher ARI score indicates a better performance. We observe that the performance of different imputation methods depends on the datasets (Fig. 5). We notice that, ALRA, which is based on the low-rank matrix approximation, performs best on the first three datasets, but obtains a lower ARI score on the PBMC dataset than most of the methods, while the performance of scTSSR is robust and always in the top two. Overall, scTSSR can improve the cell clustering through imputing the observed data.

3.5 Evaluating imputation accuracy through cell trajectory inference

Reconstruction of lineage trajectories is important for determining the pattern of a dynamic process. The appearance of dropout events will impair the performance of pseudotime inference algorithms. We compare the pseudotime inference derived with and without imputing the dropout events to evaluate the performance of different imputation methods. The experiments are conducted following the methods of Kwak *et al.* (2018). We consider two published temporal scRNA-seq datasets: Deng (Deng *et al.*, 2014) and Petropoulos (Petropoulos *et al.*, 2016). The Deng data consist of the single cells from ten early mouse developmental stages from zygote, 2-/4-/8-/16- cell stages to blastocyst. The Petropoulos data include the single cells from five stages of human preimplantation embryonic development from developmental Day (E) 3 to Day 7. The reported time labels represent the overall developmental trajectory and are treated as the gold standard to evaluate the performance. We use TSCAN (Ji and Ji, 2016) and Monocle 2 (Qiu *et al.*, 2017) to infer pseudotime from the observed or imputed data, and all parameters are set to the default values. Pseudo-temporal Ordering Score (POS) and Kendall's rank correlation score are considered as the metrics to measure the consistency between the true time labels and pseudotime orderings derived from the data.

Figure 6 shows the visualization of lineages reconstructed by Monocle 2 from the Deng dataset, and the POS and Kendall's rank correlation scores are also provided. DrImpute outperforms all

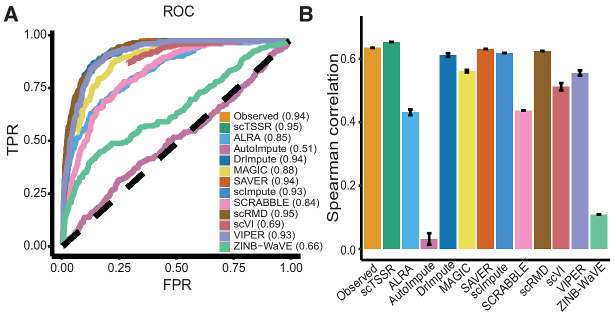

Fig. 4. Evaluation of imputation methods through differential expression analysis on H1 versus DEC cell subpopulations. (**A**) The ROC curve and the calculated AUC score of different methods, where the reference is set as the top 200 genes ranked by adjusted P values from the bulk data. (**B**) The barplots of average Spearman correlation coefficient between the adjusted P values derived from the bulk data and those derived from the imputed data (or observed data). The black interval represents the result plus or minus standard error

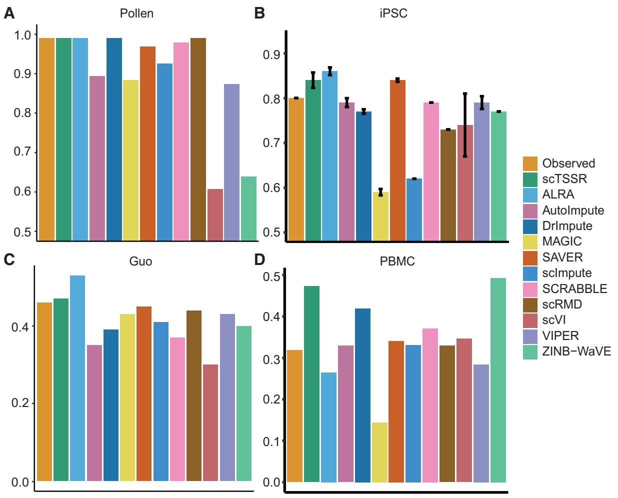

Fig. 5. ARI scores of the clustering results of different imputation methods on four datasets. The Y-axis represents the ARI scores. (**A**) Pollen dataset. (**B**) The barplots of average ARI scores on the iPSC dataset, where the black interval represents the result plus or minus standard error of the category. (**C**) Guo dataset. (**D**) PBMC dataset

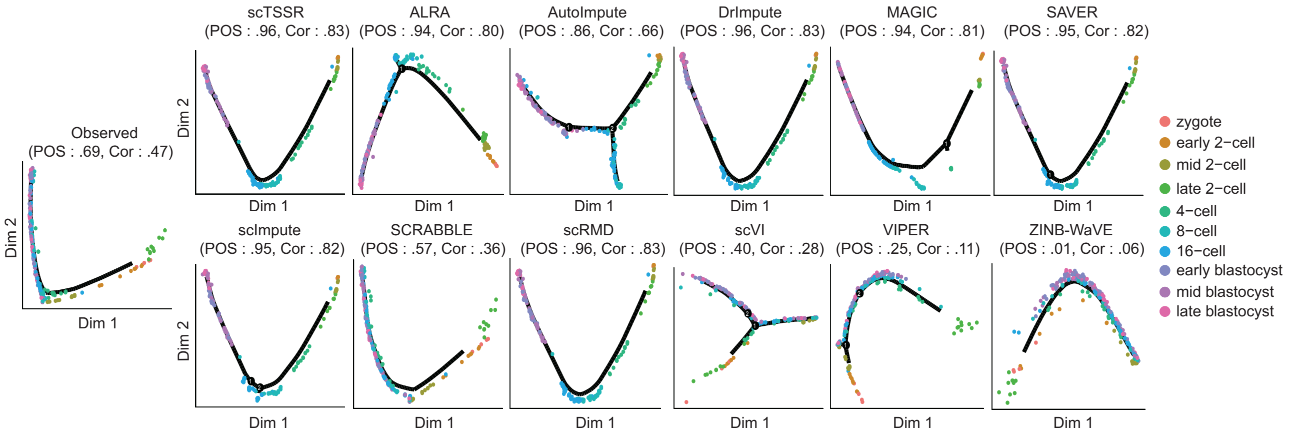

Fig. 6. Visualization of lineages reconstructed by Monocle 2 from the observed and imputed data on the Deng dataset. Cells are embedded into two-dimensional space using reversed graph embedding. The POS and Kendall's rank correlation scores of different methods are also provided

methods on the Deng dataset, while scTSSR is always in top two (Fig. 6 and Supplementary Fig. S6). Also, considering that some of the imputation methods may depend on the random seeds, we run each imputation method ten times on the Deng dataset and compute the standard error of the performance metrics (POS and Kendall's rank correlation score) (Supplementary Fig. S7). Due to that the imputation result obtained by AutoImpute contains many duplicate columns, which results in inevitable errors in the process of using Monocle 2, we do not run AutoImpute on the Petropoulos dataset. scTSSR obtains the top three POS and Kendall's rank correlation score among all methods on the Petropoulos dataset, which are competitive to the results derived from SAVER and scImpute (Supplementary Fig. S8). It is interesting to observe that the two low-rank matrix approximation based methods, ALRA and scRMD, do not perform well on the Petropoulos dataset. This may be explained by the fact that the cells are placed on a continuous trajectory (Supplementary Fig. S8B), and the low-rank assumption of the two methods is not satisfied. The pseudotime trajectory derived from ALRA starts from E3 and ends at E6 and shows fluctuations up and down in the tail of the pseudotime trajectory, which is not consistent with the true time labels. The pseudotime trajectory derived from scRMD starts from E3 and ends at E5, which is also not accurate (Supplementary Fig. S8B). These results indicate that the low-rank matrix approximation based method, ALRA, may have competitive performance in cell clustering where the data matrix is low rank, but it may not perform well in cell trajectory inference when the rank of the data matrix is high. Since our scTSSR does not require the data matrix to be low rank, it performs well in both cell clustering and trajectory inference.

4 Discussion

We have developed a new method to impute dropout events in scRNA-seq data. Five experiments are conducted to evaluate the performance of the proposed method on different real scRNA-seq datasets in terms of different evaluation measures. We count the rankings of all the twelve imputation methods in each experiment (Supplementary Tables S1–S5), and plot the frequency of each method ranking in the top three in five experiments (Supplementary Fig. S9). Experiment results show that our method outperforms other imputation methods in terms of recovery of the Gini coefficients of genes and gene-to-gene correlations, recovery of the true expression levels, differential expression analysis, cell clustering and pseudotime trajectory analysis.

We now discuss the time complexity of the proposed algorithm. In this study, we use Keras (TensorFlow R package) to solve the non-negative constrained sparse learning problem (Equations (6) and (7)). In Keras, the stochastic gradient descent algorithm is used. For m genes and n cells, solving Equations (6) and (7) involves $O(k_1 mn)$ operations, where k_1 is the number of epochs. The total time complexity is $O(k_1 k_2 mn)$, where k_2 is the number of outer iterations. To compare the running time, we run different imputation methods on a workstation with Intel 1 CPU (3.40 GH) and 64 GB RAM. The pbmc33k dataset is used to evaluate the efficiency since it has a large number of cells. The dataset is downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc33k. We filter out genes expressed in <20% of cells, leaving 1856 genes. From this dataset, we generate six subsampled datasets ranging in size (100, 500, 1k, 5k, 10k and 20k cells). We run each method three times on each subsampled dataset to compute the average running time. Experiment results show that our method is comparable to the three regression-based methods (SAVER, scImpute and VIPER) in terms of computing time and can deal with a dataset with 20 000 cells within 200 min (Supplementary Fig. S10), which is affordable in practical applications. Note that the space complexity of our method is $O(m^2 + n^2 + mn)$. As the number of cells is increasing, more memory will be required. To reduce the space complexity, we will extend our method by first dividing cells into different groups (dividing randomly or clustering) and run our method on each group to impute dropout values.

The existing self-representation learning based imputation methods can be divided into two types. The first type (e.g. scImpute and VIPER) represents the expression level of a gene in a cell using the expression levels of the same gene across all other cells, and only borrows the information shared across cells to impute a dropout value. The second type (e.g. SAVER) uses the expression levels of all other genes in the same cell to represent the target expression level, and the information shared only across genes can be leveraged. Unlike scImpute, VIPER and SAVER, scTSSR uses a two-side sparse self-representation model (3). To impute the expression level of a target gene in a target cell, scTSSR simultaneously leverages the expression levels of all other genes in the target cell, the expression levels of the target gene in all other cells, and the expression levels of all other genes in all other cells. That is, scTSSR takes advantage of the whole expression matrix (except the entry corresponding to the dropout event) rather than just the row or the column which contains the dropout event. Low-rank matrix approximation based methods can also borrow information from both genes and cells to impute the data (Chen *et al.*, 2018; Linderman *et al.*, 2018; Zhang and Zhang, 2018). However, the low-rank assumption cannot be satisfied when the cells lie on continuous developmental trajectories. scTSSR does not impose a very strong assumption on the underlying data, and can be applied to data including both discrete cell clusters and continuous trajectories.

A zero count in scRNA-seq data can be caused by a dropout event (technical zero) or reflect a true biological non-expression (biological zero). If we can estimate the dropout probability (denoted as P_{gc}) of the zero count of gene g in cell c, we can incorporate this information into our model to improve the accuracy of the estimated representation matrices $\hat{A}$ and $\hat{B}$ and the imputed data $\hat{X}$. We can rewrite the representation error term in Equation (5) as $\|W \odot \left(X - (AX + XB + AXB)\right)\|_F^2$, where $\odot$ is the Hadamard product of matrices, and W is a weight matrix with element $W_{gc} = 1 - P_{gc}$. According to the weighted least squares, only the errors corresponding to non-zero values and biological zeros are considered, and the errors corresponding to technical zeros are excluded. In addition, in the imputation step, we can also impute the technical zeros only and not change the biological zeros. Several studies can be used to estimate the dropout probabilities of zero counts using different probability models (Chen and Zhou, 2018; Li and Li, 2018; Miao *et al.*, 2019). In the future, we will combine our method with these models to improve the accuracy of the imputed data.

Besides the observed scRNA-seq data, scRNA-seq data collected from other conditions, technologies and species (Wang *et al.*, 2019), bulk RNA-seq data obtained from the same cells/tissues (Peng *et al.*, 2019), smRNA FISH data and biological networks (Elyanow *et al.*, 2020) can also provide valuable information for imputing dropout values. We will extend our model so that information from other domains can be incorporated to improve the performance.

Funding

This work was supported by the National Natural Science Foundation of China [11871026, 61402190, 61532008, 61602309, 61932008 and 61772368], Natural Science Foundation of Hubei province [2018CFB521], Fundamental Research Funds for the Central Universities [CCNU18TS026], Shenzhen Research and Development program [JCYJ20170817095210760], Natural Science Foundation of SZU [2017077], National Key R&D Program of China [2018YFC0910500], Shanghai Municipal Science and Technology Major Project [2018SHZDZX01] and ZJLab, Hong Kong Research Grants Council [Projects C1007-15G and 11200818], and City University of Hong Kong [Project 9610034].

Conflict of Interest: none declared.

References

Arisdakessian,C. *et al.* (2019) Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.*, **20**, 1–14.

Baron,M. *et al.* (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, **3**, 346–360.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

Chen,C. *et al.* (2018) scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition. *bioRxiv*, 459404. doi: 10.1101/459404.

Chen,M. and Zhou,X. (2018) Viper: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.*, **19**, 196.

Chen,R. *et al.* (2017) Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep.*, **18**, 3227–3241.

Chu,L. *et al.* (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, **17**, 173.

Deng,Q. *et al.* (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.

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

Elyanow,R. *et al.* (2020) netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res.*, **30**, 195–204.

Eraslan,G. *et al.* (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 390.

Gong,W. *et al.* (2018) TCM visualizes trajectories and cell populations from single cell data. *Nat. Commun.*, **9**, 2749.

Grun,D. *et al.* (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

Guo,F. *et al.* (2015) The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell*, **161**, 1437–1452.

Huang,M. *et al.* (2018) Saver: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.

Ji,Z. and Ji,H. (2016) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nucleic Acids Res.*, **44**, e117.

Kiselev,V.Y. *et al.* (2017) Sc3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.

Kwak,I.-Y. *et al.* (2018) DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinform.*, **19**, 220.

La Manno,G. *et al.* (2016) Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell*, **167**, 566–580.

Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.

Linderman,G.C. *et al.* (2018) Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*, 397588. doi: 10.1101/397588.

Lopez,R. *et al.* (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.

Miao,Z. *et al.* (2019) scRecover: Discriminating true and false zeros in single-cell RNA-seq data for imputation. *bioRxiv*. doi: 10.1101/665323.

Peng,T. *et al.* (2019) Scrabble: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol.*, **20**, 88.

Petropoulos,S. *et al.* (2016) Single-cell RNA-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, **165**, 1012–1026.

Pollen,A.A. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.

Qiu,X. *et al.* (2017) Single-cell mRNA quantification and differential analysis with census. *Nat. Methods*, **14**, 309–315.

Risso,D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, **9**, 284.

Robinson,M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Talwar,D. *et al.* (2018) Autoimpute: autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.*, **8**, 16329.

Tang,F. *et al.* (2009) mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.

Torre,E.A. *et al.* (2018) Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA fish. *Cell Systems*, **6**, 171–179.

van Dijk,D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 716–729.

Vershynin,R. (2010) Introduction to the non-asymptotic analysis of random matrices. *arXiv: Probability*, 210–268.

Wang,J. *et al.* (2019) Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, **16**, 875–878.

Zeisel,A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.

Zhang,L. and Zhang,S. (2018) PBLR: an accurate single cell RNA-seq data imputation tool considering cell heterogeneity and prior expression level of dropouts. *bioRxiv*, 379883. doi: 10.1101/379883.

Zhang,X.F. *et al.* (2019) EnImpute: imputing dropout events in single-cell RNA-sequencing data via ensemble learning. *Bioinformatics*, **35**, 4827–4829.

Zhao,P. and Yu,B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zhao,Y. *et al.* (2017) Link prediction for partially observed networks. *J. Comput. Graph. Stat.*, **26**, 725–733.

Zheng,G.X. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

Zhu,L. *et al.* (2019) Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. USA*, **116**, 466–471.