scImpute: accurate and robust imputation for single cell RNA-seq data

Wei Vivian Li [1], Jingyi Jessica Li [1,2,*]

Abstract

The analysis of single-cell RNA-seq (scRNA-seq) data is complicated and biased by excess zero or near zero counts, the so-called dropouts due to the low amounts of mRNA sequenced within individual cells. We introduce scImpute, a statistical method to accurately and robustly impute the dropouts in scRNA-seq data. scImpute is shown as an effective tool to enhance the clustering of cell populations, improve the accuracy of differential expression analysis, and aid the study of gene expression dynamics in time series scRNA-seq experiments.

1 Introduction

Bulk cell RNA-sequencing (RNA-seq) technology has been widely used for transcriptome profiling to study transcriptional structures, splicing patterns, and expression levels of transcripts and genes [1]. However, it is important to account for cell-specific transcriptome landscapes in order to address biological questions such as the cell heterogeneity and the gene expression stochasticity. The single-cell RNA sequencing (scRNA-seq) technologies are now emerging as a powerful tool to capture transcriptome-wide cell-to-cell variability [2, 3, 4]. The scRNA-seq enables the quantification of intra-population heterogeneity at a much higher resolution, potentially revealing dynamics in heterogeneous cell populations and complex tissues [5]. An important characteristic of scRNA-seq data is the "dropout" phenomenon where a gene is observed at a moderate expression level in one cell but is not detected in another cell [6]. Usually these events occur due to the low amounts of mRNA in individual cells, and thus a truly expressed transcript may not be detected by sequencing in some samples. This dropout phenomenon is also protocol-dependent. The number of cells that can be analyzed with one chip is usually no more than hundreds of cells on the Fluidigm C1 platform, with around 1-2 million reads per cell. On the other hand, protocols based on droplet microfluidics can parallelly profile more than $10,000$ cells, but with only 100-200k reads per cell [7]. As a consequence, there is a much higher dropout rate with droplet microfluidics technologies.

[1] Department of Statistics, University of California, Los Angeles, CA 90095-1554
[2] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[*] To whom correspondence should be addressed. Email: jli@stat.ucla.edu

Methods for analyzing scRNA-seq data have been developed from different perspectives, such as clustering, cell type identification, and dimension reduction [8, 9, 10, 11, 12]. Some of these methods address the dropout events by implicit imputation while others do not. CIDR is the first clustering method that incorporates imputation of dropout values, but the imputed expression of a particular gene in a cell changes each time when the cell is paired up with a different cell [10]. Seurat is a computational strategy for spatial reconstruction of cells from single-cell gene expression data [11]. It includes an imputation step to impute the expression of landmark genes based on highly variable or so-called structured genes. ZIFA is a dimensionality reduction model specifically designed for zero-inflated single-cell gene expression analysis [12]. The model is built upon an empirical observation: dropout rate for a gene depends on its mean expression level in the population.

Since most downstream analyses on scRNA-seq, such as differential expression analysis, identification of cell-type-specific genes, and all the analyses mentioned earlier, rely on the accuracy of gene expression measurements, it is important to correct the false zero expression due to dropout events by model-based imputation methods. To our knowledge, MAGIC is the only available method for explicit and genome-wide imputation of single-cell gene expression profiles [13]. A key step in this method is to create a Markov transition matrix. In the imputation of a single cell, the weights of the other cells are determined through this matrix. However, imputing all counts including those not affected by dropout would introduce new bias into the data and possibly eliminate meaningful biological variation. It is also inappropriate to treat all zero counts as missing values, since some of them may reflect true biological non-expression. Therefore, we propose a new imputation method for scRNA-seq data, scImpute, to simultaneously determine which values are affected by dropout events in data and perform imputation only on dropout entries. To achieve this goal, scImpute first learns each gene's dropout probability in each cell by fitting a mixture model. Next, scImpute imputes the (highly probable) dropout values in a cell by borrowing information of the same gene in other similar cells, which are selected based on the genes not severely affected by dropout events.

2 Results

2.1 scImpute improves the identification of cell populations

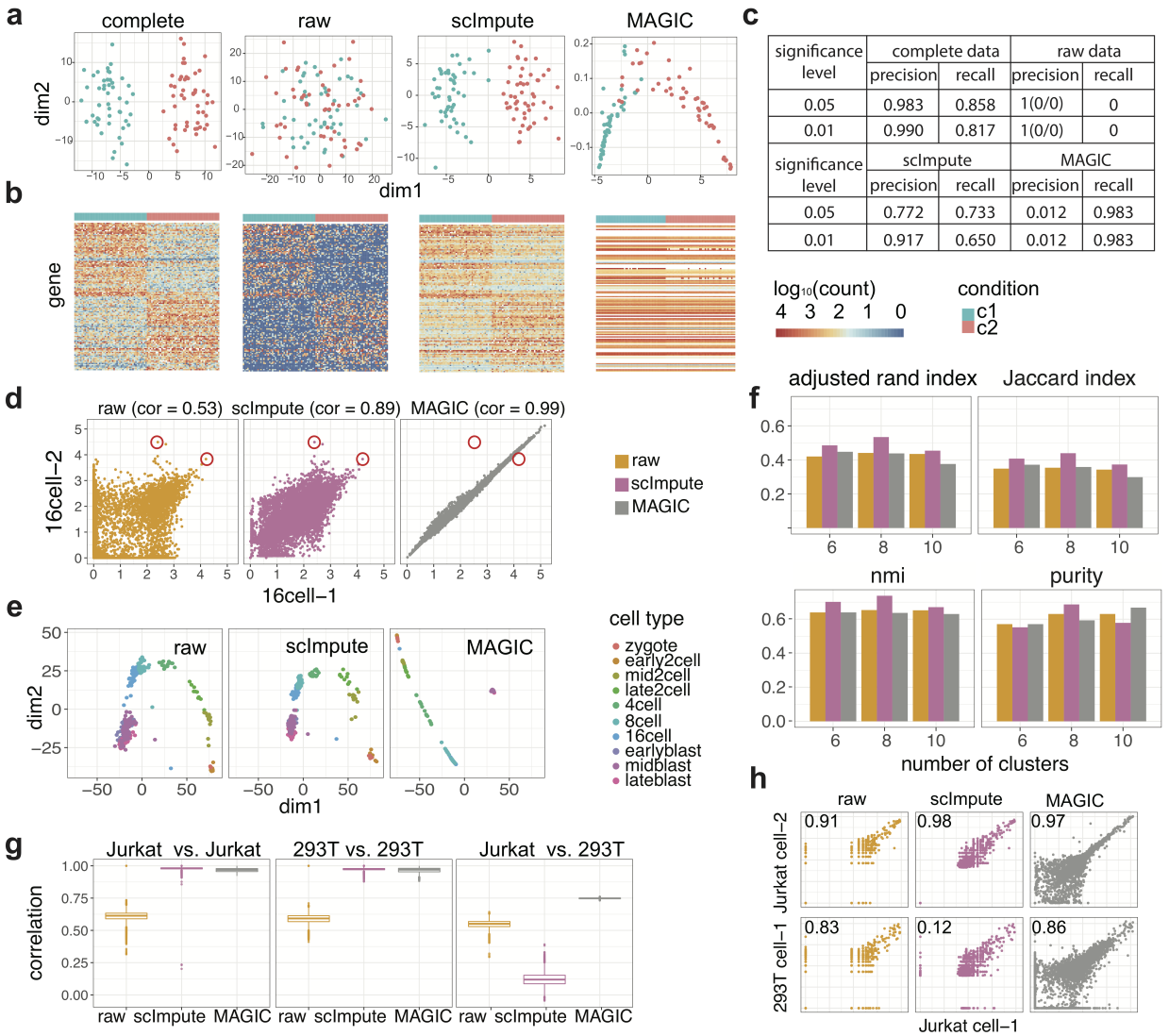
We first use a simulation study to illustrate scImpute's efficacy and robustness in imputing gene expressions for better identification of multiple cell types (without true class labels supplied). We suppose there are two cell types c_1 and c_2, each with 50 cells, and 120 among $10,000$ genes are truly differentially expressed (DE) (details in Methods). scImpute is able to unveil the cell identities masked by dropout events (Figure 1a), and recover the missing gene expression to a good extent to make the imputed matrix resemble the complete matrix (Figure 1b). Imputed data by scImpute lead to accurate identification of DE genes (Figure 1c). MAGIC also improves grouping of cells, but

it introduces artificial signals that largely alter the data (Figure 1b) and thus the PCA result (Figure 1a). We further investigate a case with with three cell types c_1, c_2, and c_3 each with 50 cells, and 180 among $10,000$ genes are differentially expressed. Again, scImpute leads to expression profiles that mostly resemble the complete data. (Supplementary Figure S1). In addition, we assess how the prevalence of dropout values influences the performance of scImpute. The differential expression analysis based on the imputed data has increased accuracy as the dropout proportion in the raw data decreases. But scImpute is able to achieve $> 80.0\%$ AUC when the proportion of zero count in raw data is less than 75.0% (Supplementary Figure S3). The only tuning parameter in scImpute is a threshold t, and the imputation is only applied to the genes with dropout probabilities larger than t in a cell (see Methods). A sensitivity analysis suggests that scImpute is robust to varying threshold values (Supplementary Figure S2).

We also use two real data sets to illustrate scImpute's capability in identifying cell types or subpopulations. The first one is a smaller data set of mouse preimplantation embryos [15]. It contains RNA-seq profiles of 268 single cells from 10 developmental stages (see Methods). 70.0% entries in the raw read count matrix are 0. To illustrate the the dropout phenomenon, we plot the $\log_{10}$ read counts of two 16-cell stage cells as an example in Figure 1d. Even though the two cells come from the same stage, many expressed genes have zero counts in only one cell. This problem is alleviated in the imputed data by scImpute, and the Pearson correlation between the two cells increases from 0.53 to 0.89 (Figure 1d). MAGIC achieves an even higher correlation (0.99) but also introduces very large counts that do not exist in the raw data.

Although it is possible to differentiate the major developmental stages from the raw data, the imputed data by scImpute output more compact clusters (Figure 1e). MAGIC gives a very clean pattern of developmental stages, but it has a high risk of removing biologically meaningful variation, given that many cells of the same stage have almost identical scores in the first two PCs. We then compare the clustering results using the k-means algorithm ($k = 6, 8,$ or 10) on the first two PCs. The results are evaluated by four different measures: adjusted rand index [16], Jaccard index [17], normalized mutual information (nmi) [18], and purity (see Supplementary Note 1). The four measures are all between 0 and 1, with 1 indicating perfect match between the clustering result and the truth. All the four measures indicate that scImpute leads to best agreement with the true labels (Figure 1f). This comparison suggests that scImpute could improve clustering among cell subpopulations by imputing dropout values in scRNA-seq data.

We also apply scImpute to a larger data set generated by the high-throughput droplet-based system [19]. It contains scRNA-seq data of $2,885$ 293T cells and $3,258$ Jurkat cells. After we remove the genes that have zero expression in all the cells, the proportion of zero expression is 84.6% among the Jurkat cells (Supplementary Figure S7) and 83.5% among the 293T cells. Among the Jurkat cells, the median of pairwise Pearson correlations is 0.61 in the raw data, and increases to 0.98 and 0.97 in the imputed data by scImpute and MAGIC, respectively. Among the 293T cells, the median of pairwise Pearson correlations is 0.59 in the raw data, and increases to 0.98 and 0.97 in the imputed data by scImpute and MAGIC, respectively. The increased correlations among



significance level	complete data		raw data	
	precision	recall	precision	recall
0.05	0.983	0.858	1(0/0)	0
0.01	0.990	0.817	1(0/0)	0

significance level	scImpute		MAGIC	
	precision	recall	precision	recall
0.05	0.772	0.733	0.012	0.983
0.01	0.917	0.650	0.012	0.983

Figure 1: scImpute helps defining cellular identity in simulated and real data. **a-c**: Comparison of imputation methods based on simulated data with two cell types (100 cells). **a**: The first two dimensions of PCA results. **b**: The expression profile of the 120 true DE genes in the two cell types in complete, raw, and imputed data. Vertical axis represents genes and horizontal axis represents cells. **c**: The precision and recall rates of DE gene identification. The p-values from the two-sided t-test are adjusted using the Benjamini-Hochberg procedure [14] and compared against significance level 0.01 or 0.05. **d-f**: Comparison of imputation methods based on the mouse embryo data (268 single cells). **d:** Scatter plots of $\log_{10}$(count) in two example cells of 16cell stage before and after imputation. Largest raw count in the two cells are marked by red circles. **e:** The first two dimensions of PCA results. **f:** The adjusted rand index, Jaccard index, nmi, and purity scores of clustering results based on the raw and imputed data. Clustering is performed by the k-means algorithm on the single cells' scores in the first two PCs. **g-h**: Comparison of imputation methods based on the Jurkat cell line ($3,258$ cells) and the 293T cell line ($2,885$ cells). **g**: Boxplots of Pearson correlation calculated for pairwise Jurkat vs. Jurkat, 293T vs. 293T, or 293T vs. Jurkat cells. **h**: Scattor plots showing two examples of gene expression patterns. Pearson correlation scores are marked on the top-left of each scattor plot.

cells within the same cell line (Figure 1g-h) suggest that the imputation strengthens the similarities of cells. Among all the pairs of Jurkat and 293T cells, the median Pearson correlation is 0.55. scImpute decreases the median to 0.12, while MAGIC increases the median to 0.75. This result

shows that scImpute is able to reveal trasncriptomic differences of different cell types, while MAGIC makes different cell types more similar. Therefore, scImpute is the better method to differentiate cell types by increasing both within-cell-type similarities and between-cell-type differences.

2.2 scImpute assists differential gene expression analysis on scRNA-seq data

An effective imputation method should lead to a better agreement between scRNA-seq and bulk RNA-seq data of the same biological condition, since bulk data have higher signal-to-noise ratios compared with scRNA-seq data. To evaluate whether the DE genes identified from single-cell or bulk data have larger overlap after imputation, we utilize a real data set with both single-cell and bulk RNA-seq experiments on human embryonic stem cells (ESC) and definitive endorderm cells (DEC) [20] This data set includes 6 samples of bulk RNA-seq (4 in H1 ESC and 2 in DEC) and 350 samples of scRNA-seq (212 in H1 ESC and 138 in DEC). The percentage of zero gene expression is 14.8% in bulk data and 49.1% in single-cell data.

We apply both scImpute and MAGIC to impute the gene expression values, and then compare the results of differential expression analysis with the results on raw data. Since the ground truth is unknown, we take the DE genes detected from bulk data as the standard to be compared against. Figure 2a illustrates the expression profile of the top 200 DE genes identified in the bulk data. We compare the DE gene lists obtained from the bulk and single-cell data to calculate precision and recall rates and F scores (Figure 2b, Supplementary Note 2). The imputed data lead to more similar DE gene lists (around 20% higher F score) to the DE genes from the bulk data, implying that differential analysis on imputed data has better accuracy compared with raw data. scImpute has a higher precision rate, while MAGIC has a higher recall rate, thus scImpute is preferred when users have a priority on the precision of the identified DE genes. We also assess the consistency of the identified DE genes from the raw and imputed data with the bulk data by comparing the overlap of top DE genes (Figure 2c). From this perspective, scImpute leads to the best agreement with the bulk data.

2.3 scImpute assists pattern recognition in time course scRNA-seq data

Chu et al. [20] also generated bulk and single-cell time-course RNA-seq data profiled at 0, 12, 24, 36, 72, and 96 h of differentiation during DEC emergence (Supplementary Table S2). We utilize this data set to show that scImpute can help recover DE signals difficult to identify in the raw time-course data, and reduce false discoveries resulted from dropouts. We first apply scImpute to the raw scRNA-seq data, and then use the R package `SCPattern` [21] to identify genes with differential expression along the time points. `SCPattern` identifies $3,247$ genes from the raw data and $5,861$ from the imputed data, and $2,587$ of them are overlapping. We infer that for the 660 genes that are only identified from the raw data, dropout events have introduced false down-regulating signals. The four example genes in Supplementary Figure S4a are only identified in the raw data, but neither the imputed data nor the bulk data provide strong evidence of differential

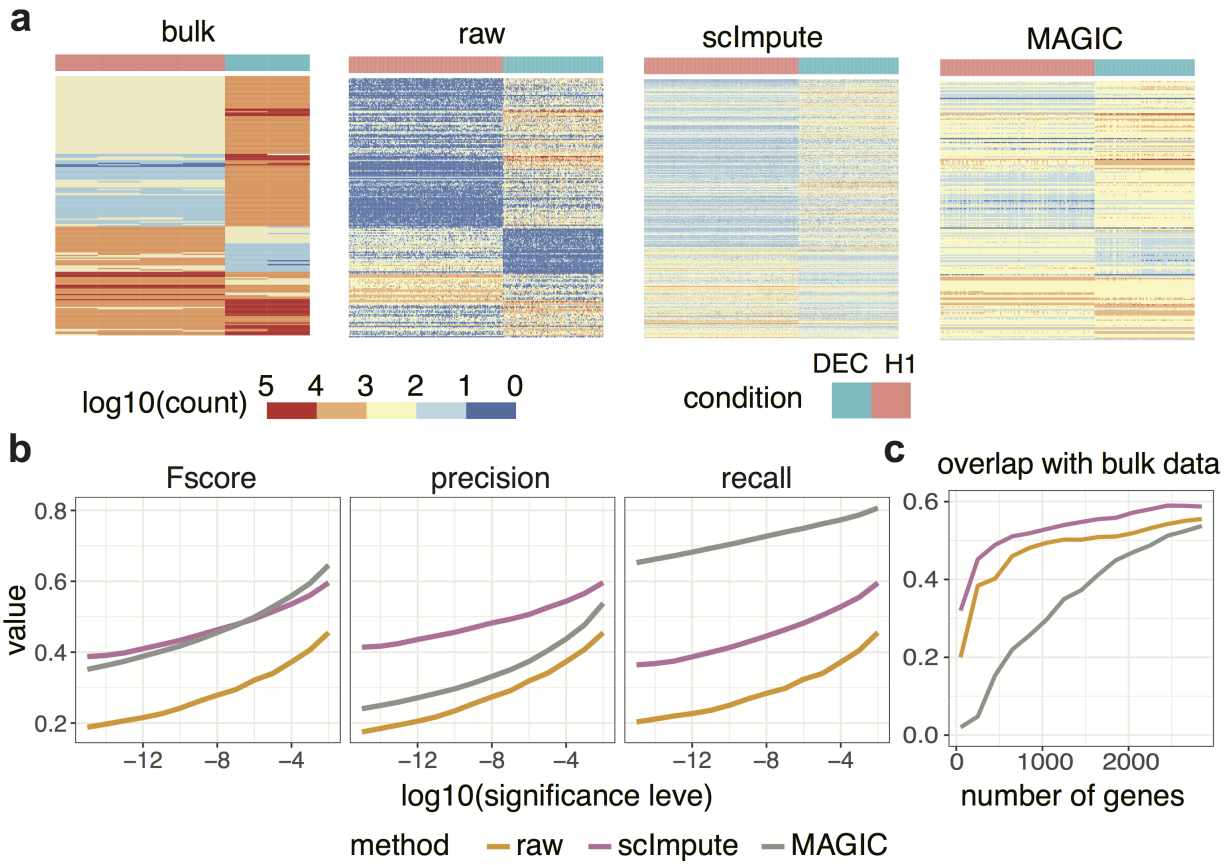


Figure 2: scImpute improves differential gene expression analysis in human embryonic cells. **a:** Heatmaps of the top 200 DE genes found in bulk RNA-seq data (4 samples in H1 and 2 samples in DEC). Gene expression values of each data set are displayed in the corresponding heatmap. **b-c** Performance of DE gene identification based on the raw and imputed single-cell data. The p-values are calculated by `DESeq2`. **b:** The F score, precision rate, and recall rate are calculated by comparing the identified genes from each single-cell data set with those identified from the bulk data under the same significance level (horizontal-axis). **c:** The overlap with bulk data is calculated as (number of common genes in the top n identified genes from single-cell or bulk data)/n.

expression. In these examples, the differential patterns are mainly due to varying dropout rates at different time points. As for the genes that are only identified in the imputed data, we suspect that dropout events of some genes make it difficult to separate biological signals from technical noises. But after imputation, the underlying expression patterns are more easily detected by statistical testing. The four genes in Supplementary Figure S4b illustrate how the imputed data by scImpute present up or down regulation patterns that are consistent with the bulk data. For a genome wide comparison, the imputed data have significantly higher Pearson correlations with the bulk data (Supplementary Figure S5).

3 Discussion

We propose a statistical method scImpute to address the dropout events prevalent in scRNA-seq data. scImpute focuses on imputing the missing expression values of dropout genes, while retaining the expression levels of genes that are largely unaffected by dropout events. Hence, scImpute can reduce technical variation resulted from scRNA-seq and better represent cell-to-cell biological variation, while it also avoids introducing excessive bias during its imputation process. An attractive advantage of scImpute is that it can be incorporated into any existing pipelines or downstream analysis of scRNA-seq data. scImpute takes the raw read count matrix as input and outputs an imputed count matrix of the same dimensions, so it can be seamlessly combined with other computational tools without data reformatting or transformation. Another important feature of scImpute is that it does not require manual tuning of parameters. The only parameter in the method is a threshold t on dropout probabilities. We show in a sensitivity analysis that scImpute is robust to the threshold value (Supplementary Figure S2), and a default threshold value 0.5 is sufficient for most scRNA-seq data. Moreover, cell type information is not necessary for the scImpute method. When cell type information is available, separate imputation on each cell type is expected to produce more accurate results. But as illustrated by simulation and real data results, scImpute is able to infer cell-type-specific expression even when the true labels are not supplied.

scImpute scales up well when the number of cells increases, and the computation efficiency can be largely improved if a filtering step on genes can be performed based on biological knowledge. Aside from computational complexity, another future direction is to further improve imputation efficiency when dropout rates in raw data are severely high, as with the droplet-based technologies. Imputation task becomes more difficult when proportion of missing values increases. More complicated models that account for gene similarities may yield more accurate imputation results, but the prevalence of dropout events may require additional prior knowledge on similar genes to assist modeling.

Methods

The scImpute model

The input of our method is a count matrix $\boldsymbol{X}^C$ with rows representing genes and columns representing cells, and our eventual goal is to construct an imputed count matrix with the same dimensions. We start by normalizing the count matrix by the library size of each sample (cell) so that all samples have one million reads. Denote the normalized matrix by $\boldsymbol{X}^N$, we then make a matrix $\boldsymbol{X}$ by taking $\log_{10}$ transformation with a pseudo count 1.01:

$$X_{ij} = \log_{10}(X_{ij}^N + 1.01); \quad i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J,$$

where I is the total number of genes and J is the total number of cells. The pseudo count is added to avoid infinite values in parameter estimation in a later step.

Once we obtain the transformed gene expression matrix $\boldsymbol{X}$, the first step is to infer which genes are affected by the dropout events in which cells. Instead of treating all zero values as dropout events, we construct a statistical model to systematically determine whether a zero value comes from a dropout event or not. With the existence of dropout events, most genes have a bimodal expression pattern that can be described by a mixture model of two components (Supplementary Figure S6). The first component is a Gamma distribution used to account for the dropouts, while the second component is a Normal distribution to represent the actual gene expression levels. Thus for each gene i, its expression can be considered as a random variable with density function

$$f_{X_i}(x) = \lambda_i \text{Gamma}(x; \alpha_i, \beta_i) + (1 - \lambda_i)\text{Normal}(x; \mu_i, \sigma_i)\,, \tag{3.1}$$

where λ_i is gene i's dropout rate, α_i, β_i are shape and rate parameters of Gamma distribution, and μ_i, σ_i are the mean and standard deviation of Normal distribution. The intuition behind this mixture model is that if a gene has high expression and low variation in most cells, then a zero count is more likely to be a dropout value; on the other hand, if a gene has constantly low expression or medium expression with high variation, then a zero count may reflect real biological variability. An advantage of this model is that it does not assume an empirical relationship between dropout rates and genes' mean expression, as did in [6], allowing more flexibility in the model estimation. The parameters in the mixture model can be estimated by the Expectation-Maximization (EM) algorithm and we denote their estimates as $\hat{\lambda}_i$, $\hat{\alpha}_i$, $\hat{\beta}_i$, $\hat{\mu}_i$, and $\hat{\sigma}'_i$. It follows that the dropout probability of gene i in cell j can be estimated as

$$d_{ij} = \frac{\hat{\lambda}_i \text{Gamma}(X_{ij}; \hat{\alpha}_i, \hat{\beta}_i)}{\hat{\lambda}_i \text{Gamma}(X_{ij}; \hat{\alpha}_i, \hat{\beta}_i) + (1 - \hat{\lambda}_i)\text{Normal}(X_{ij}; \hat{\mu}_i, \hat{\sigma}_i)}\,.$$

Therefore, each gene i has an overall dropout rate $\hat{\lambda}_i$, which does not depend on individual cells, and also dropout probabilities $d_{ij}, j = 1, 2, \ldots, J$, which may vary among different cells.

Now we impute the gene expressions cell by cell. For each cell j, we select a gene set A_j in need of imputation based on the genes' dropout probabilities in cell j: $A_j = \{i : d_{ij} \geq t\}$, where t is a threshold on dropout probabilities. Similarly we also have a gene set $B_j = \{i : d_{ij} < t\}$ that have accurate gene expression with high confidence and do not need imputation. We learn cells' similarities through gene set B_j to impute the expression of genes in set A_j by borrowing information from their expression in other similar cells. To do this, we construct a weighted Lasso model [22]

$$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}^{(j)}}{\text{argmin}} \frac{1}{|B_j|} ||\boldsymbol{w}_{B_j}^T (\boldsymbol{X}_{B_j,j} - \boldsymbol{X}_{B_j,-j}\boldsymbol{\beta}^{(j)})||_2^2 + \theta ||\boldsymbol{\beta}^{(j)}||_1,$$

where $\theta > 0$ is the regularization parameter and $\boldsymbol{w}_{B_j}$ is the weight vector of gene set B_j. Note that the response $\boldsymbol{X}_{B_j,j}$ is a vector representing the B_j rows in the $j-$th column of $\boldsymbol{X}$, and the

design matrix $\boldsymbol{X}_{B_j,-j}$ is a sub-matrix of $\boldsymbol{X}$ with dimensions $|B_j| \times (j-1)$. The weight of gene i is determined by its dropout rate: $w_i = \frac{1-\hat{\lambda}_i}{\sum_{i' \in B_j}(1-\hat{\lambda}'_i)}$. The indices of cells selected by the Lasso are denoted as $S_j = \left\{j' : \hat{\beta}^{(j)}_{j'} > 0\right\}$. The strengths of this model are two-folds. First, the Lasso ensures sparsity in variable selection, thus only cells that are informative for imputation in cell j would be selected. Second, incorporating weights on genes can give priority to genes with high expression, as these genes are more informative in learning cells' relationships. Finally, the estimated coefficients $\hat{\boldsymbol{\beta}}^{(j)}$ from the Ordinary Least Squares (OLS) linear regression are used to impute the expression of gene set A_j in cell j:

$$\hat{\boldsymbol{\beta}}^{(j)}_{\text{OLS}} = \underset{\boldsymbol{\beta}^{(j)}_{S_j}}{\operatorname{argmin}} ||\boldsymbol{w}^T_{B_j}(\boldsymbol{X}_{B_j,j} - \boldsymbol{X}_{B_j,S_j}\boldsymbol{\beta}^{(j)}_{S_j})||^2_2$$

$$\hat{X}_{ij} = \begin{cases} X_{ij}, & i \in B_j, \\ X_{i,S_j}\hat{\boldsymbol{\beta}}^{(j)}_{\text{OLS}}, & i \in A_j. \end{cases}$$

We construct a separate lasso model for each cell to impute the expression of genes with high dropout probabilities. This method simultaneously determines the values that need imputation, and would not introduce bias to the high expressions of accurately measured genes. Supplementary Figure S7 give some real data examples of distribution of zero count proportion, showing the reasonability of dividing genes into set A and B. Supplementary Figure S8 illustrates the imputation step in scImpute with a toy example.

Generation and analysis of simulated data

In the simulation study with two cell types, we suppose there are two cell types c_1 and c_2, each with 50 cells , and there are $10,000$ genes in total. In this gene population, only 120 genes are truly differentially expressed, with one half having higher expression in c_1 and the other half having higher expression in c_2. We directly generate genes' $\log_{10}$-based read counts to construct the simulated data set. First, mean expressions of the $10,000$ genes are randomly drawn from a Normal distribution with mean 1.8 and standard deviation 0.5. Similarly, standard deviations of gene expressions are randomly drawn from a Normal distribution with mean 0.6 and standard deviation 0.1. These parameters are estimated from the real data set of mouse embryo cells. Second, we randomly draw 60 genes and shift their mean expression in cell type c_1 by multiplying it with an integer randomly sampled from $\{2, 3, \ldots, 10\}$; we also draw another 60 genes and shift their mean expression in cell type c_2 in the same way. Next, the expression of each gene in the 100 cells can be simulated from Normal distributions defined by the parameters obtained in the first two steps. We refer to the resulting gene expression data as the *complete data*. Finally, we suppose the dropout rate of each gene follows a double exponential function $\exp(-0.1 *$mean expression$^2)$, as assumed in [12]. Zero values are then introduced into the simulated data for each gene based on a Bernoulli distribution defined by its dropout rate, resulting in a gene expression matrix with

excess zeros and in need of imputation. We refer to the gene expression data after introducing zero values as the *raw data*. Please note that the generation of gene expression values does not directly follow the mixture model used in scImpute, so that we can investigate the efficacy and robustness of scImpute in a fair way using this simulation. (Data with three cell types is simulated in a similar procedure.)

Based on the parameters estimated from real data, we simulate a $10,000 \times 100$ gene expression matrix (*raw data*) with 71.8% of zero values, while the *complete data* only have 1.3% of zero values. Even though the two cell types are clearly distinguishable when we apply principal component analysis (PCA) to the complete data, they become indistingusiable when dropout events exist (Figure 1a). However, when we apply scImpute and MAGIC to impute the dropout values, the relationships between the 100 cells are unveiled. Both methods are able to distinguish type c_1 from c_2, but MAGIC introduces artificial signals that largely alter the data and thus the PCA result (Figure 1a). In the gene expression matrices, we observe that the dropout events largely mask the expression dynamics of the 120 DE genes. scImpute is able to recover the missing gene expression to a good extent and make the imputed matrix resemble the complete matrix, while MAGIC would largely distort the expression matrix and make it even more deviated from the complete data than the raw data (Figure 1b). This difference between scImpute and MAGIC can be more easily observed by directly comparing the accuracy of DE genes' identification (Figure 1c). The DE genes are found by the t-test with a significance level of 0.01 or 0.05 on the simulated gene expression. The 120 true DE genes cannot be detected at all from the raw data, but scImpute is able to recover the differential pattern.

Data availability

The scRNA-seq data of mouse preimplantation embryos [15] is publicly available through GEO Series accession number GSE45719. It contains RNA-seq profiles of 268 single cells from zygote (4), early 2-cell stage (8), middle 2-cell stage (12), late 2-cell stage (10), 4-cell stage (14), 8-cell stage (37), 16-cell stage (50), early blastocyst (43), middle blastocyst (60), and late blastocyst (30) stages.

The scRNA-seq data of Jurkat and 293T cells [19] are available at `https://www.10xgenomics.com/single-cell/`.

The RNA-seq data on human embryonic cells [20] is publicly available under the GEO Series accession number GSE75748.

Code availability

The scImpute method is implemented in the `R` package `scImpute`. The package and the source code are publicly available at Github: `https://github.com/Vivianstats/scImpute`.

References

[1] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[2] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

[3] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6):e1004333, 2015.

[4] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[5] Serena Liu and Cole Trapnell. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research*, 5, 2016.

[6] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[7] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1):44–73, 2017.

[8] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, page btv088, 2015.

[9] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

[10] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biology*, 18(1):59, 2017.

[11] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

[12] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241, 2015.

[13] David van Dijk, Juozas Nainys, Roshan Sharma, Pooja Kathail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *bioRxiv*, page 111591, 2017.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[15] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[16] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[17] Glenn W Milligan and Martha C Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458, 1986.

[18] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[19] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8, 2017.

[20] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.

[21] Ning Leng, Li-Fang Chu, Jeea Choi, Christina Kendziorski, James Thomson, and Ron Stewart. Scpattern: A statistical approach to identify and classify expression changes in single cell rna-seq experiments with ordered conditions. *bioRxiv*, page 046110, 2016.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Supplementary Materials

Supplementary Note

1 Four evaluation measures of clustering results

The four measures (adjusted rand index, Jaccard index, nmi, and purity) focus on different properties of the clustering results. The adjusted rand index penalizes both false positive and false negative decisions, where a positive decision means that two cells are clustered into one cluster, while a negative decision means that two cells are clustered into different clusters. The Jaccard index is similar to the adjusted rand index, but it does not account for true negatives. The normalized mutual information (nmi) measures the similarity from the perspective of information theory. The purity score simply calculates the percentage of the total number of samples that are from the same true class and clustered together correctly, and it does not penalize on splitting a true class into multiple clusters.

We choose a to represent the number of observation pairs which are corrected grouped into the same class by the clustering method. b represents the number of observation pairs which are grouped into the same cluster but actually belong to different classes. c represents the number of observation pairs which are grouped into different clusters but actually belong to the same class. d represent the number of observation pairs which are correctly grouped into different clusters (See Supplementary table S1).

We use $U = \{u_1, \ldots, u_P\}$ to denote the true partition of P class and $V = \{v_1, \ldots, v_K\}$ to denote the partition given by K-means clustering results. Let $n_{i\cdot}$ and $n_{\cdot j}$ be the number of observations in class u_i and cluster v_j respectively, and n is the total number of observations.

The adjusted rand index is calculated as

$$\frac{a + d - n_c}{a + b + c + d - n_c},$$

where $n_c = \left(n(n^2+1) - (n+1)\sum_i n_{i\cdot}^2 - (n+1)\sum_j n_{\cdot j}^2 + 2\sum_{i,j} n_{i\cdot}^2 n_{\cdot j}^2/n\right) / \left(2(n-1)\right)$.

The Jaccard index is calculated as

$$\frac{a}{a + b + c}.$$

The normalized mutual information is calculated as

$$\frac{2I(U,V)}{H(U) + H(V)},$$

where $I(U,V)$ is mutual information, and $H(U)$ and $H(V)$ are the entroy of partition U and V.

The purity score is calculated as

$$\frac{1}{n}\sum_i \max_j |V_i \cap U_j|.$$

2 Procedures in the differential gene expression analysis

Since the ground truth is unknown, we take the DE genes detected from bulk RNA-seq data as the standard to be compared against, given that the bulk data have better signal-to-noise ratios and thus more information for differential gene expression analysis. Figure 2a illustrates the expression profile of the top 200 DE genes identified in the bulk data. The identification of DE genes is performed using the R package `DESeq2`. In the bulk data, we apply different thresholds to the Benjamini-Hochberg adjusted p-values of genes to obtain a DE gene list for every threshold as the standard. The same thresholds are applied to the p-values obtained from the raw and imputed single-cell data to obtain DE gene lists. Then we compare the DE gene lists obtained from the bulk and single-cell data to calculate precision and recall rates and F scores (Figure 2b).

Supplementary tables

Table S1: Confusion matrix used to calculate the clustering measures.

		clustered classes		
		# pairs in the same class	# pairs in different classes	Total
true classes	# pairs in the same class	a	c	$a+c$
	# pairs in different classes	b	d	$b+d$
	Total	$a+c$	$b+d$	n

Table S2: Sample information in timecourse RNA-seq data.

time point	00h	12h	24h	36h	72h	96h	total
scRNA-seq (cells)	92	102	66	172	138	188	758
bulk RNA-seq (replicates)	0	3	3	3	3	3	15

Supplementary Figures

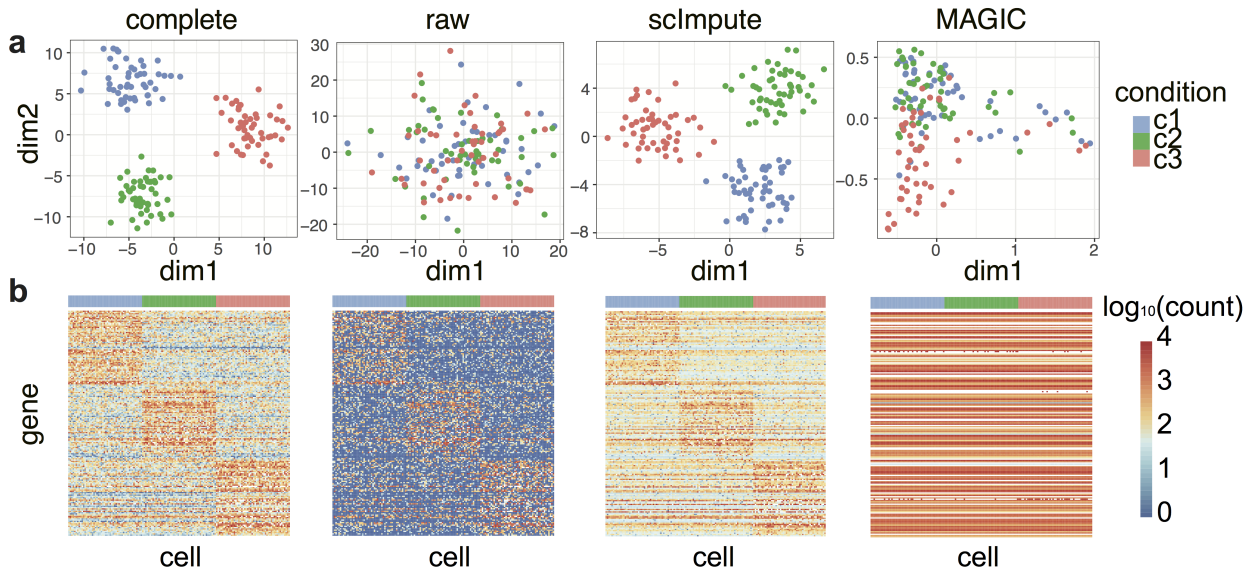


Figure S1: Comparison of imputation methods based on simulated data with three cell types (150 cells). **a**: The first two dimensions of PCA results. **b**: The expression profile of the 180 true DE genes in the three cell types in complete, raw, and imputed data. Vertical axis represents genes and horizontal axis represents cells.

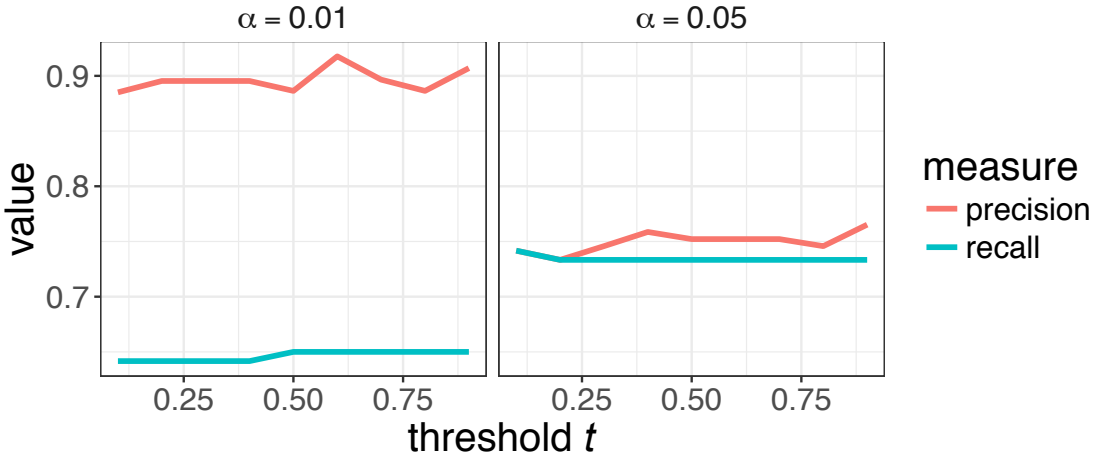
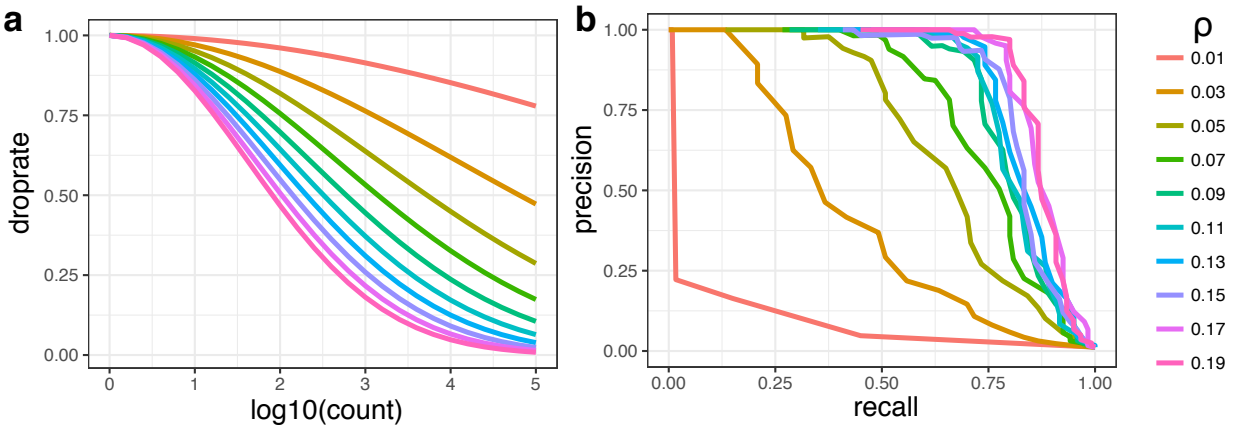


Figure S2: Sensitivity analysis on the threshold t used to select gene set A_j and B_j in cell j. Precision and recall rates are calculated by comparing the DE genes identified from the imputed data by scImpute with the truth. α refers to the significance level used in the two-sided t-test for selecting differentially expressed genes.



c

ρ	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19
zero proportion	0.97	0.90	0.84	0.79	0.74	0.70	0.65	0.62	0.58	0.55
AUC	0.07	0.42	0.66	0.75	0.81	0.82	0.84	0.83	0.88	0.88

Figure S3: Performance of scImpute given different dropout rates in raw simulated data. **a**: The theoretical dropout rates determined by the double exponential function $\exp(-\rho \times \log_{10}(\text{count})^2)$, with ρ varying from 0.01 to 0.19 by a step of 0.02. **b**: The precision-recall curves for the identification of differentially expressed genes from the imputed data. The raw data are simulated with different values of ρ, leading to varying proportions of dropout expression values. **c**: The proportion of zero count and the area under the precision-recall curve (AUC) for each ρ value.

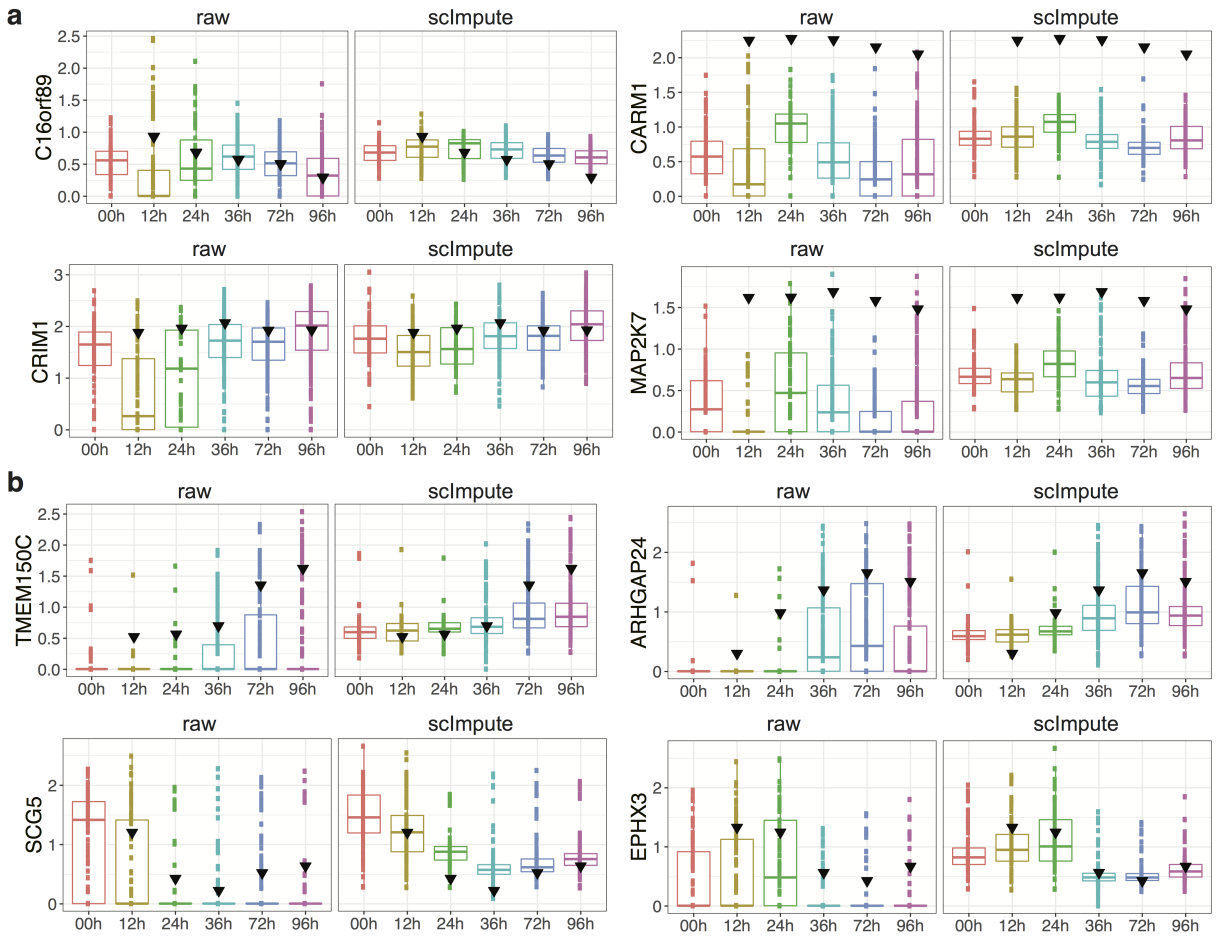


Figure S4: Comparison of the raw and the imputed data by scImpute based on the time-course scRNA-seq human ESC data. **a:** Four example genes that are called by `SCPattern` in the raw data, but not in the imputed data. **b:** Four example genes that are called by `SCPattern` in the imputed data, but not in the raw data. The boxplots show the $\log_{10}$(count) of the corresponding gene in single cells along the 6 time points, while the triangles mark the average expression of the same gene in the three replicates of bulk data.

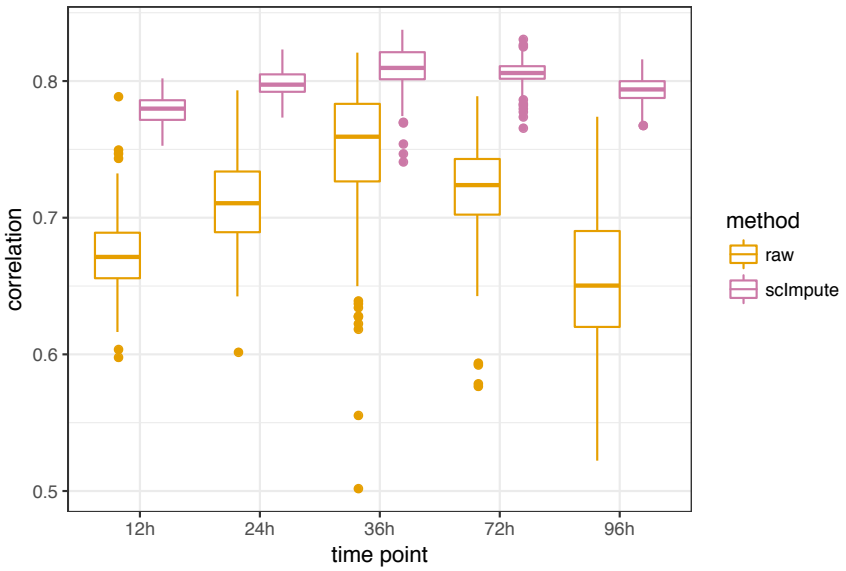


Figure S5: Correlation between gene expression in single-cell and bulk data. The Pearson correlation coefficients are calculated between each individual cell and averaged bulk data, at each time point. The correlations based on the imputed data are significantly increased compared with the raw data.

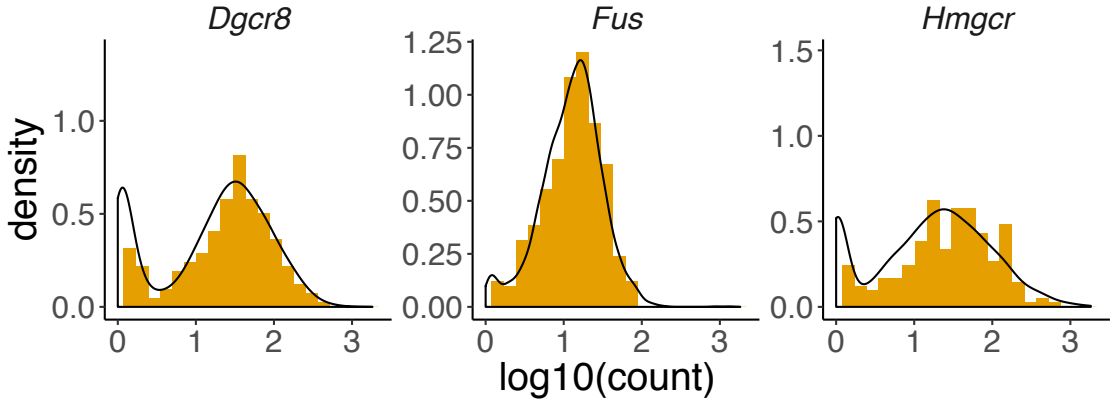


Figure S6: Three example genes (*Dgcr8, Fus, and Hmgcr*) for comparison of observed and fitted expression distribution. The results are based on the mouse embryo data. Yellow histogram represents observed distribution and black line represents fitted distribution by the Gamma-Gaussian mixture model (3.1).

Figure S7: Three examples in real data showing the number of genes with various proportion of zero expression values in the raw and imputed data. The proportions of zero values in the imputed data are significantly reduced compared with the raw data. **a**: raw data; **b**: imputed data generated by scImpute.

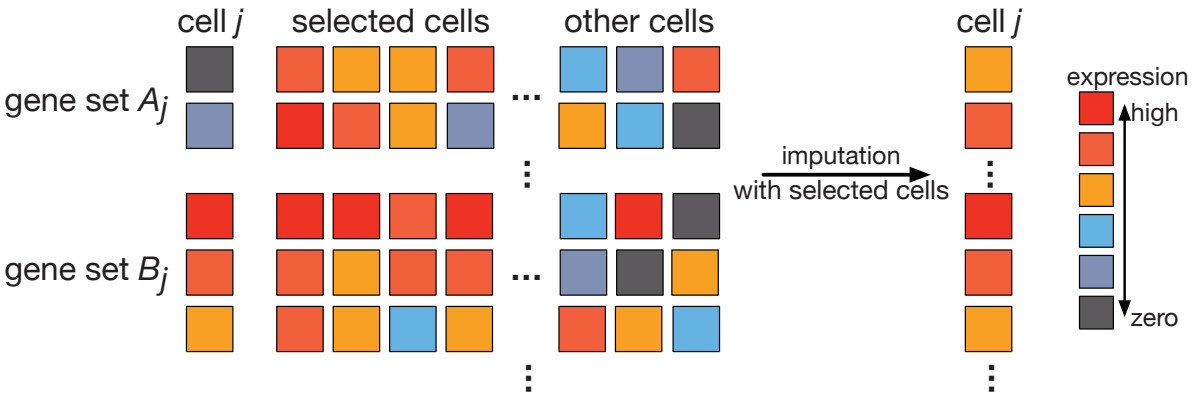


Figure S8: A toy example illustrating the workflow in the imputation step of scImpute method.