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

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

**Abstract**

The emerging single cell RNA sequencing (scRNA-seq) technologies enable the investigation of transcriptomic landscapes at single-cell resolution. The analysis of scRNA-seq data is complicated by excess zero or near zero counts, the so-called dropouts due to the low amounts of mRNA sequenced within individual cells. Downstream analysis of scRNA-seq would be severely biased if the dropout events are not properly corrected. We introduce scImpute, a statistical method to accurately and robustly impute the dropout values in scRNA-seq data. ScImpute automatically identifies gene expression values affected by dropout events, and only perform imputation on these values without introducing new bias to the rest data. ScImpute also detects outlier or rare cells and excludes them from imputation. Evaluation based on both simulated and real scRNA-seq data on mouse embryos, mouse brain cells, human blood cells, and human embryonic stem cells suggests that scImpute is an effective tool to recover transcriptome dynamics masked by dropout events. scImpute is shown to correct false zero counts, enhance the clustering of cell populations and subpopulations, improve the accuracy of differential expression analysis, and aid the study of gene expression dynamics.

## 1 Introduction

Bulk cell RNA-sequencing (RNA-seq) technology has been widely used for transcriptome profiling to study transcriptional structures, splicing patterns, and gene and transcript expression levels [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. Despite its popularity, bulk RNA-seq does not allow people to study cell-to-cell variation in terms of transcriptomic dynamics. In bulk RNA-seq, cellular heterogeneity cannot be addressed since signals of variably expressed genes would be averaged across cells. Fortunately, single-cell RNA sequencing (scRNA-seq) technologies are now emerging as a powerful tool to capture

[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

transcriptome-wide cell-to-cell variability [2, 3, 4]. 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].

One important characteristic of scRNA-seq data is the "dropout" phenomenon where a gene is observed at a moderate expression level in one cell but undetected 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 during sequencing in some cells. This characteristic of scRNA-seq is shown to be protocol-dependent. The number of cells that can be analyzed with one chip is usually no more than a few hundreds 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 $100-200$k reads per cell [7]. Hence, there is usually a much higher dropout rate in scRNA-seq data generated by the droplet microfluidics than the Fluidigm C1 platform. Statistical or computational methods developed for scRNA-seq need to take the dropout issue into consideration, and otherwise they may present varying efficacy when applied to data from different protocols.

Methods for analyzing scRNA-seq data have been developed from different perspectives, such as clustering, cell type identification, and dimension reduction. Some of these methods address the dropout events in scRNA-seq by implicit imputation while others do not. SNN-Cliq is a clustering method that uses scRNA-seq to identify cell types [8]. Instead of using conventional similarity measures, SNN-Cliq uses the ranking of cells/nodes to construct a graph from which clusters are identified. CIDR is the first clustering method that incorporates imputation of dropout values, but the imputed expression value of a particular gene in a cell changes each time when the cell is paired up with a different cell [9]. The pairwise distances between every two cells are later used for clustering. Seurat is a computational strategy for spatial reconstruction of of cells from single-cell gene expression data [10]. It infers the spatial origins of individual cells from the cell expression profiles and a spatial reference map of landmark genes. It also 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 [11]. The model is built upon an empirical observation: dropout rate for a gene depends on its mean expression level in the population and, and ZIFA accounts for dropout events in factor analysis.

Since most downstream analyses on scRNA-seq, such as differential gene expression analysis, identification of cell-type-specific genes, reconstruction of differentiation trajectory, 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 in scRNA-seq data by model-based imputation methods. To our knowledge, MAGIC is the first available method for explicit and genome-wide imputation of single-cell gene expression profiles [12]. MAGIC imputes missing expression values by sharing information across similar cells, based on the idea of heat diffusion. A key step in this method is to create a Markov transition matrix, constructed by normalizing the

similarity matrix of single cells. In the imputation of a single cell, the weights of the other cells are determined through the transition matrix. During the preparation of this manuscript, we also noticed another imputation method SAVER [13], which borrows information across genes using a Bayesian approach to estimate (unobserved) true expression levels of genes. Both MAGIC and SAVER would alter all gene expression levels including those not affected by dropouts, and this would potentially introduce new biases into the data and possibly eliminate meaningful biological variation. We think 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 based on 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 unlikely affected by dropout events (Figure 1).

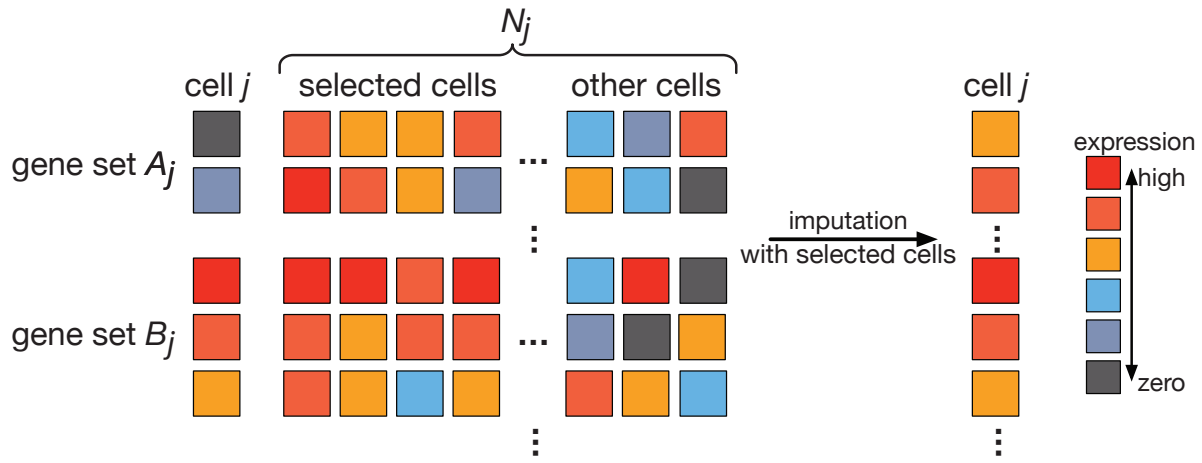


**Figure 1:** A toy example illustrating the workflow in the imputation step of scImpute method, described in Equation (4.2)-(4.3). scImpute first learns each gene's dropout probability in each cell by fitting a mixture model. Next, scImpute imputes the (highly probable) dropout values in cell $j$ (gene set $A_j$) by borrowing information of the same gene in other similar cells, which are selected based on gene set $B_j$ (not severely affected by dropout events).

# 2 Results

## 2.1 scImpute recovers gene expression affected by dropout events

A key reason for performing imputation on scRNA-seq data is to recover biologically meaningful transcriptome dynamics in single cells so that we can determine cell identity and identify differentially expressed (DE) genes among different cell types. We first use three examples to illustrate scImpute's efficacy in imputing gene expressions. (All the imputation results in Section 2 are

obtained without using true cell type information unless otherwise noted.)

First, we show that scImpute recovers the true expression of the ERCC spike-in transcripts [14], especially low abundance transcripts that are impacted by dropout events. The ERCC spike-ins are synthesized RNA molecules with known concentrations, which serve as gold standards of true expression levels, so the read counts can be compared with the true expression for accuracy evaluation. The dataset contains $3,005$ cells from the mouse somatosensory cortex region [15]. After imputation, the median correlation among the $57$ transcripts' read counts and their true concentration increases from $0.92$ to $0.95$, and the minimum correlation increases from $0.81$ to $0.89$ (Figure S1). The read counts and true concentrations also present a stronger linear relationship in each single cell (Figure 2).

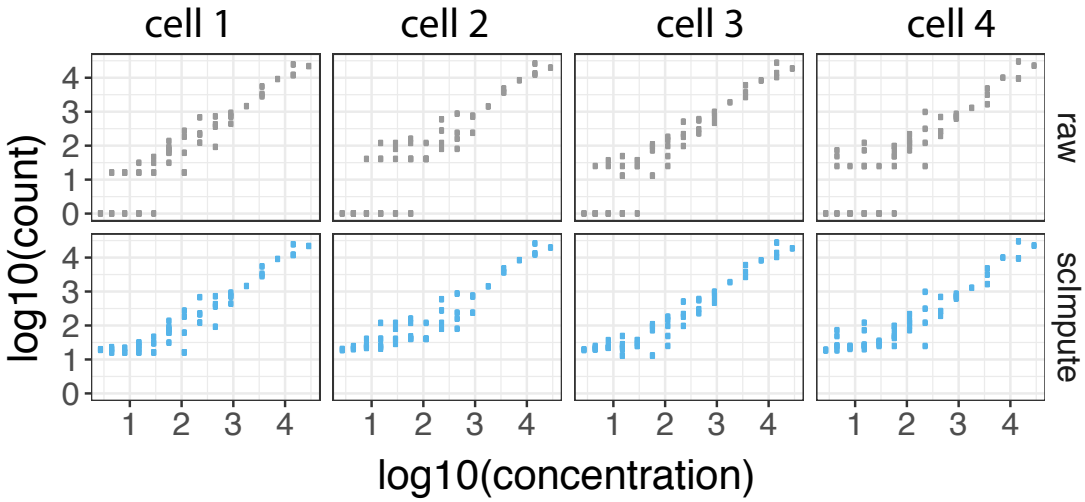


**Figure 2:** The ERCC spike-ins' $\log 10$(counts) and $\log 10$(concentration) in four randomly selected mouse cortex cells. Imputed data presents a better linear relationship between the true concentration and the observed counts.

Second, we show that scImpute correctly imputes the dropout values of $892$ annotated cell-cycle genes in $182$ embryonic stem cells (ESCs) that had been staged for cell-cycle phases (G1, S and G2M) [16]. These genes are known to modulate the cell cycle and are expected to have non-zero expression during different stages of the cell cycle. Before imputation, $22.5\%$ raw counts of the cell-cycle genes are zeros, which are highly likely due to dropouts. After imputation, most of the dropout values are corrected, and true dynamics of these genes in the cell cycle are revealed (Figure S2 and S3). The imputed counts also better represents the true biological variation in these cell-cycle genes (Figure 3).

Third, we use a simulation study to illustrate the efficacy of scImpute in enhancing the identification of cell types. We simulate expression data of three cell types $c_1, c_2$, and $c_3$, each with $50$ cells, and $810$ among $20,000$ genes are truly differentially expressed (DE) (details in Methods). Even though the three cell types are clearly distinguishable when we apply principal component analysis (PCA) to the complete data, they become less separated in the raw data with dropout events. However, the relationships among the $150$ cells are clarified after we apply scImpute. The

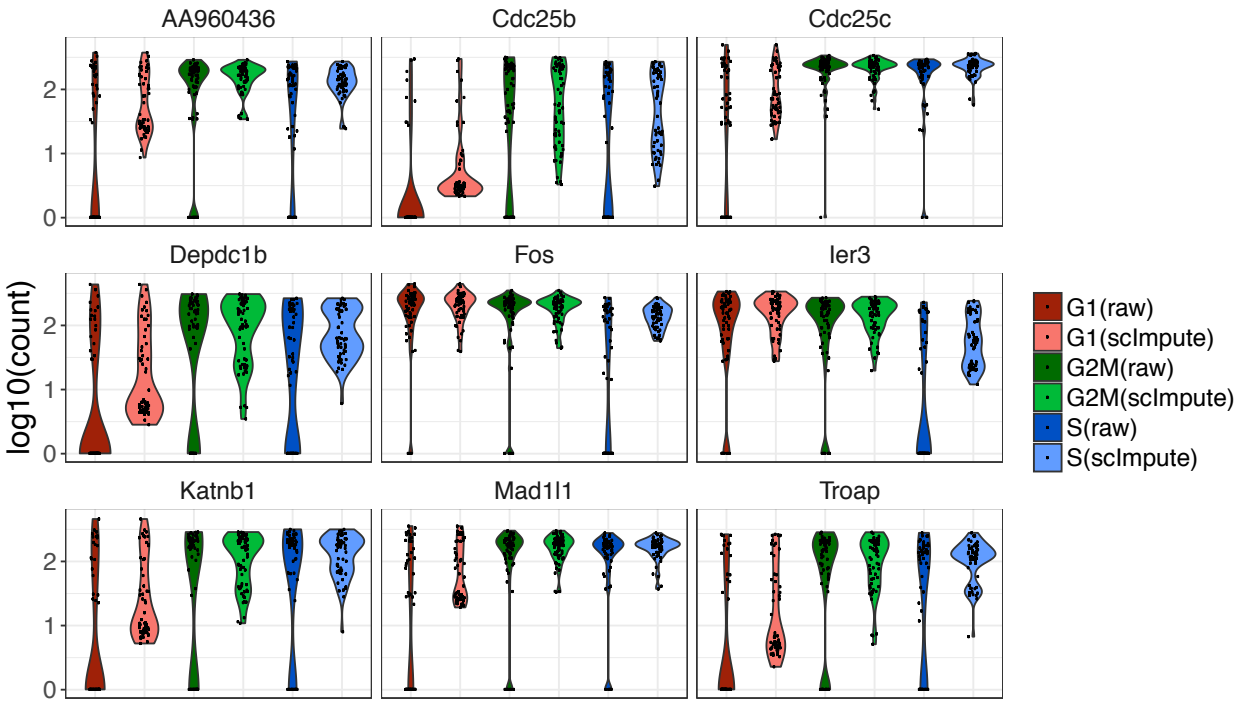


**Figure 3:** Violin plots showing the $\log 10$(count) of nine cell cycle genes in the three phases (G1, S, and G2M). scImpute has corrected the dropout values of cell cycle genes.

other two methods MAGIC and SAVER are also able to distinguish the three cell types, but MAGIC introduces artificial signals that largely alter the data and thus the PCA result, while SAVER only slightly improve the clustering result over that of the raw data (Figure 4). In addition, the dropout events obscure the differential pattern and thus increase the difficulty of detecting DE genes. The imputed data by scImpute lead to a clearer comparison between the up-regulated genes in different cell types, while the imputed data by MAGIC and SAVER fail to recover this pattern (Figure 4). We also assess how the prevalence of dropout values influences the performance of scImpute. As expected, the DE analysis based on the imputed data has increased accuracy as the dropout proportion decreases. Yet scImpute still achieves $> 80.0\%$ area under the curve even when the proportion of zero count in raw data is $75.0\%$ (Figure S4).

## 2.2 ScImpute improves the identification of cell subpopulations in real data

To illustrate scImpute's capacity in aiding the identification of cell types or cell subpopulations, we apply our method to two real scRNA-seq datasets. The first one is a smaller dataset of mouse preimplantation embryos [17]. It contains RNA-seq profiles of $268$ single cells from $10$ developmental stages. Partly due to dropout events, $70.0\%$ of read counts in the raw count matrix are zeros. To illustrate the dropout phenomenon, we plot the $\log 10$-transformed read counts of two 16-cell stage cells as an example in Figure S5. Even though the two cells come from the same

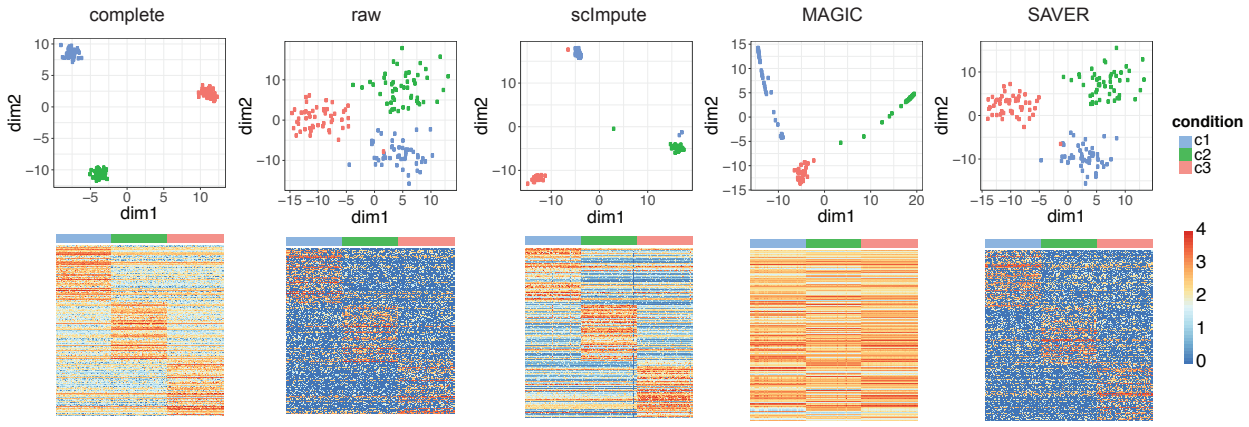


**Figure 4: a**: The first two PCs calculated from the complete data, raw data, and imputed data by scImpute, MAGIC, and SAVER. **b**: The expression profile of the $810$ true DE genes in the complete, raw, and imputed datasets.

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 $0.72$ to $0.82$ (Figure S5), especially due to the decreased number of genes only expressed in one cell. MAGIC achieves an even higher correlation $(0.95)$ but also introduces very large counts that do not exist in the raw data. Biological variation between the two cells is likely lost in the imputation process of MAGIC. On the other hand, SAVER's imputation does not have a clear impact on the data.

We compare the imputation results by investigating the clustering accuracy in the first two principal components (PCs). Although it is possible to differentiate the major developmental stages from the raw data, the imputed data by scImpute output more compact clusters (Figure 5). MAGIC gives a 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. scImpute is the only method that is able to detect outlier cells. We then compare the clustering results of the spectral clustering [18] algorithm on the first two PCs. Since the true cluster labels include several sub-stages in embryonic development, we use different numbers of clusters, $k = 6, 8, 10, 12$ and $14$. The results are evaluated by four different measures: adjusted rand index [19], Jaccard index [20], normalized mutual information (nmi) [21], and purity (Supplementary Text). 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 the best clustering result as compared with no imputation and the imputation by MAGIC or SAVER (Figure S6). This result suggests that scImpute improves the clustering of cell subpopulations by imputing dropout values in scRNA-seq data.

We also apply scImpute to a large dataset generated by the high-throughput droplet-based system [22]. The dataset contains $4,500$ peripheral blood mononuclear cells (PBMCs) of nine immune cell types, with $500$ cells of each type. In the raw data, $92.6\%$ read counts are exactly zeros.

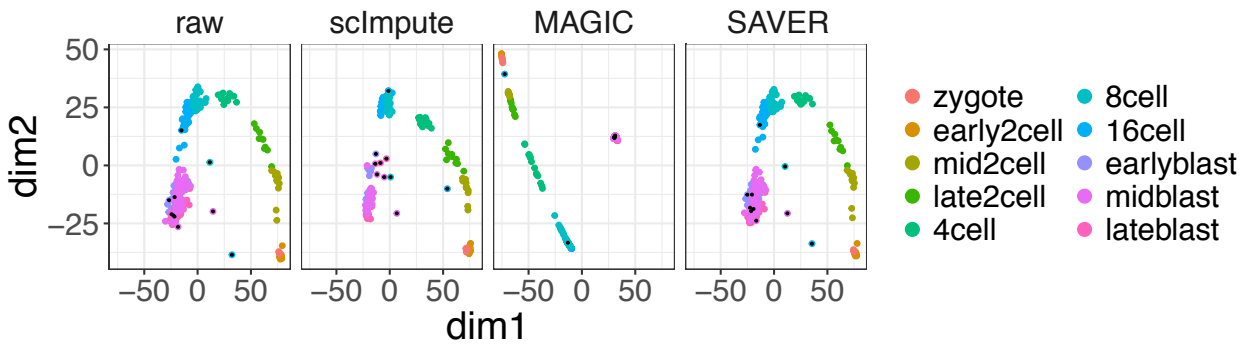


**Figure 5:** The first two PCs obtained from the raw and imputed data of mouse embryonic cells. The black dots mark the outlier cells detected by scImpute.

Given dimension reduction by t-SNE [23], the cytotoxic and naïve cytotoxic T cells are clustered together, and the other four types of T cells are not separated. After scImpute's imputation, the cytotoxic (label 11) and naïve cytotoxic T cells (label 8) are separated into two groups, and the naïve T cells (label 5) and memory T cells (label 3) are now distinguishable from the remaining T cells (Figure 6). This evidence shows the strong ability of scImpute to identify cell subpopulations despite missing cell type information. On the other hand, MAGIC does not improve the clustering of cells in the same type (Figure S7), and we could not obtain SAVER's results after running the program overnight. After the imputation by scImpute, the monocyte cells are grouped into one large and two small clusters, and we find that the three clusters reveal dynamics of two signature genes, *FCER1A*, which accumulates during the dendritic cell differentiation from monocytes [24], and *S100A8*, whose expression differs between subsets of human monocytes [25] (Figure S8 and 6). The large cluster (label 10) is characterized by high expression of *S100A8* and moderate expression of *FCER1A*; one of the small clusters (label 1) presents high expression of both *S100A8* and *FCER1A*, while in the other small cluster (label 2) *FCER1A* is largely non-expressed. We also investigate the three clusters (labels 6, 9 and 12) of regulatory/memory/helper T cells. The three clusters are supported by the expression of eight potential marker genes: cells in the same cluster have a similar expression pattern (Figure S9). This example shows that scImpute provides an opportunity to discover new cell subpopulations and their marker genes.

## 2.3 ScImpute assists differential gene expression analysis on scRNA-seq data

ScRNA-seq data provide insights into the stochastic nature of gene expression in single cells, but suffer from a relatively low signal-to-noise ratio compared with bulk RNA-seq data. Thus an effective imputation method should lead to a better agreement between scRNA-seq and bulk RNA-seq data of the same biological condition on genes known to have little cell-to-cell heterogeneity. To evaluate whether the DE genes identified from single-cell data are more accurate after imputation, we utilize a real dataset with both bulk and single-cell RNA-seq experiments on human embryonic stem cells (ESC) and definitive endorderm cells (DEC) [26]. This dataset includes $6$ samples of

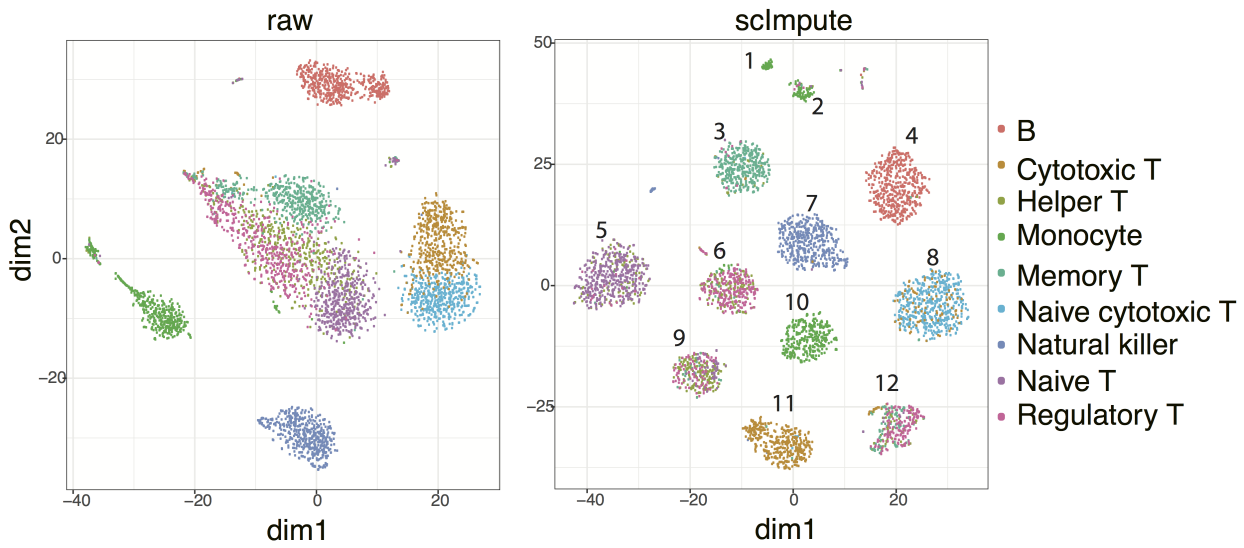


**Figure 6:** The first two dimensions of the t-SNE results calculated from raw and imputed PBMC dataset. Numbers marked on imputed data are cluster labels.

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 percentages of zero gene expression are $14.8\%$ in bulk data and $49.1\%$ in single-cell data.

We apply scImpute, MAGIC, and SAVER to impute the gene expression for each cell type respectively, and then perform DE analysis on the raw data and the imputed data by each method, respectively. We use the R package `DESeq2` [27] to identify DE genes from the bulk data, and the R packages `DESeq2` and `MAST` [28] to identify DE genes from the scRNA-seq data. Inspecting the top $200$ DE genes from the bulk data, we find that their expression profiles in the scRNA-seq data have stronger concordance with those in the bulk data after imputation by scImpute (Figure S10). We apply different thresholds to false discovery rates (FDRs) of genes in the bulk data to obtain a DE gene list for every threshold. The same thresholds are applied to the FDRs of genes calculated from the raw and imputed scRNA-seq datasets to obtain DE gene lists respectively. Then we compare the DE gene lists obtained from the scRNA-seq data with those from the bulk data (i.e., the standard) to calculate precision and recall rates and $F$ scores (Figure S11). ScImpute leads to more similar DE gene lists to those from the bulk data, and achieves around $10\%$ higher $F$ scores compared with results on the raw data. We find that scImpute makes a good balance between the precision and recall rate, while MAGIC has low precision, and SAVER has low recall rate and is barely distinguishable from no imputation. We conclude that scImpute is preferred when users have a priority on the overall accuracy of the DE genes.

A comparison between the expression profiles of DEC and ESC marker genes [26, 29, 30] shows that the imputed data by scImpute best reflect the gene expression signatures by removing undesirable technical variation resulted from dropouts (Figure 7a and S13). To determine if the DE genes identified in scRNA-seq data are biologically meaningful, we performed gene ontology

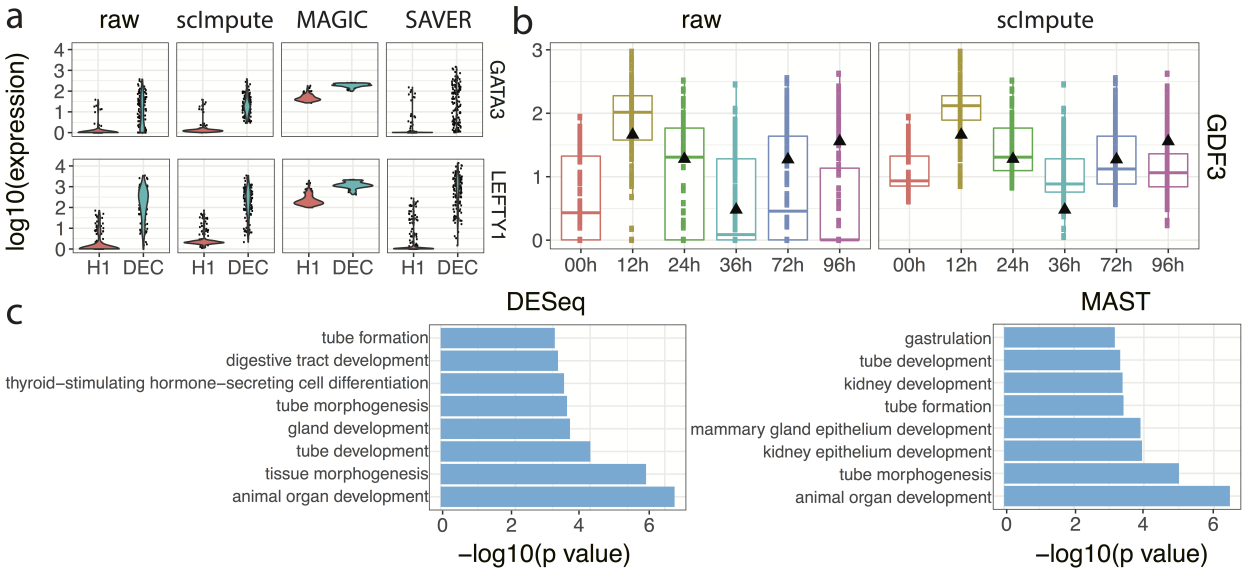


**Figure 7:** ScImpute improves differential gene expression analysis and reveals expression dynamics in time-course experimments. **a**: Raw and imputed expression levels of two marker genes of DEC. **b**: Time-course expression patterns of the gene *GDF3*, which is annotated with the GO term "endoderm development". Black triangles mark the gene's expression in bulk data. **c**: Selected GO terms enriched in the DEC up-regulated genes that can be only detected (by `DESeq2` or `MAST`) in the imputed data by scImpute, but not in the raw data.

(GO) enrichment analysis [31]. In the $\sim 300$ DEC up-regulated genes that are only detected in the imputed data by scImpute but not in the raw data, enriched GO terms are highly relevant to the functions of DECs (Figures 7c, S12, S14, and S15). However, in the $\sim 300$ DEC up-regulated genes that are only detected in the raw data, enriched GO terms are general and not characteristic to DECs (Figures S16 and S17). These results also demonstrate that scImpute can facilitate the usage of DE method that were not designed for single-cell data.

## 2.4 ScImpute recovers gene expression dynamics in time course scRNA-seq data

Aside from the data used in Section 2.3, Chu et al. [26] 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 dataset to show that scImpute can help recover the DE signals that are 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 with true cell type labels, and then study how the time-course expression patterns change in imputed data. The imputed data better distinguish cells of different time points (Figure S18), suggesting that imputed read counts reflect more accurate transcriptome dynamics along the time course. Even though the scRNA-seq data present more biological variation than the bulk data, it is reasonable to expect that the average gene expression signal across cells in scRNA-seq should correlate with the signal in bulk RNA-seq. For a genome wide comparison, the imputed data have significantly higher Pearson correlations with the bulk data (Figure S19). We study $70$ genes associated with

the GO term "endoderm development" [32] and found that a subset of these genes that are likely affected by dropout events show higher expression and better consistency with the bulk data after the imputation by scImpute (Figure 7b and S20). Similarly, we also study the marker genes (e.g., *FOXA2, HHEX,* and *CXCR4*) of DECs [26, 29, 30] and these genes' expression levels at time point 96h are recovered by scImpute even though they have a median read count of zero in the raw data (Figure S21).

## 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 excess bias during its imputation process. To achieve the above goal, scImpute first learns each gene's dropout probability in each cell by fitting a mixture model for each cell type. Next, scImpute imputes the (highly probable) dropout values of genes 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. Comprehensive studies on both simulated and real data suggest that compared with the raw scRNA-seq data, the imputed data by scImpute better present cell type identity and lead to more accurate DE analysis results.

An attractive advantage of scImpute is that it can be incorporated into any existing pipelines or downstream analysis of scRNA-seq data, such as normalization [3, 33], differential expression analysis [6], clustering and classification [8, 9], etc. 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 only involves two parameters that can be easily understood and selected. The first parameter $K$ denotes the potential number of cell populations. It can be selected based on clustering of raw data and the resolution level desired by the users. The second parameter is a threshold $t$ on dropout probabilities. We show in a sensitivity analysis that scImpute is robust to the different parameters (Figure S24), 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 cells 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. Despite the availability of computational methods that directly model zero-inflation in data [6, 28], scImpute takes the imputation perspective to improve the data quality, and its applicability is not restricted to a specific task. Hence, scImpute is an useful tool that benefits all types of scRNA-seq downstream analyses.

# 4 Methods

## 4.1 Data processing and normalization

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 $1.01$:

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

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.

## 4.2 Dectection of cell subpopulations and outliers

Since scImpute borrows information of the same gene from similar cells to impute the dropout values, a critical step is to first determine which cells are from the same subpopulation. Due to excess zero counts in scRNA-seq data, it is difficult to accurately cluster cells into true cell types. Hence, the goal of this step is to find a candidate pool of "neighbors" for each cell. ScImpute will select similar cells from the candidate neighbors in a subsequent imputation step. Suppose that scImpute clusters the cells in a dataset into $K$ subpopulations in this step. For each cell, its candidate neighbors are the other cells in the same cluster.

1. PCA is performed on matrix $\boldsymbol{X}$ for dimension reduction and the resulting matrix is denoted as $\boldsymbol{Z}$, where columns representing cells and rows representing principal components (PCs). The purpose of dimension reduction is to reduce the impact of large portions of dropout values. The PCs are selected such that at least $40\%$ of the variance in data could be explained.

2. Based on the PCA-transformed data $\boldsymbol{Z}$, the distance matrix $\boldsymbol{D}_{J\times J}$ between the cells could be calculated. For each cell $j$, we denote its distance to the nearest neighbor as $l_j$. For the set $\boldsymbol{L} = \{l_1, \ldots, l_J\}$, we denote its first quartile as $Q_1$, and third quartile as $Q_3$. The outlier

cells are those cells which do not have close neighbors:

$$O = \{j : l_j > Q_3 + 1.5(Q_3 - Q_1)\}.$$

For each outlier cell, we set its candidate neighbor set $N_j = \emptyset$. Please note that the outlier cells could be a result of experimental/technical error or bias, but they may also represent real biological variation as rare cell types. ScImpute would not impute gene expression values in outlier cells, nor use them to impute gene expression values in other cells.

3. The remaining cells $\{1, \ldots, J\} \backslash O$ are clustered into $K$ groups by spectral clustering [18]. We denote $g_j = k$ if cell $j$ is assigned to cluster $k$ $(k = 1, \ldots, K)$. Hence, cell $j$ has the candidate neighbor set $N_j = \{j' : g_{j'} = g_j, j' \neq j\}$.

### 4.3 Identification of dropout values

Once we obtain the transformed gene expression matrix $\boldsymbol{X}$ and the candidate neighbors of each cell $N_j$, the next 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 across similar cells, and that pattern can be described by a mixture model of two components (Supplementary Figure S22). 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. For each gene, the proportions and parameters of the two components could be different in various cell types, so we construct separate mixture models for different cell subpopulations.

For each gene $i$, its expression in cell subpopulation $k$ is modeled as a random variable $X_i^{(k)}$ with density function

$$f_{X_i^{(k)}}(x) = \lambda_i^{(k)} \text{Gamma}\left(x; \alpha_i^{(k)}, \beta_i^{(k)}\right) + \left(1 - \lambda_i^{(k)}\right) \text{Normal}\left(x; \mu_i^{(k)}, \sigma_i^{(k)}\right), \tag{4.1}$$

where $\lambda_i^{(k)}$ is gene $i$'s dropout rate in cell subpopulation $k$, $\alpha_i^{(k)}, \beta_i^{(k)}$ are the shape and rate parameters of Gamma distribution, and $\mu_i^{(k)}, \sigma_i^{(k)}$ 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 within a cell subpopulation, a zero count is more likely to be a dropout value; on the other hand, if a gene has constantly low 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 mean expression levels of genes, as [6] did, 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 $\hat{\lambda}_i^{(k)}$, $\hat{\alpha}_i^{(k)}$, $\hat{\beta}_i^{(k)}$, $\hat{\mu}_i^{(k)}$, and $\hat{\sigma}_i^{(k)}$. It follows that the *dropout probability* of gene $i$ in cell $j$, which

belongs to subpopulation $k$, can be estimated as

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

Therefore, each gene $i$ has an overall *dropout rate* $\hat{\lambda}_i^{(k)}$ in cell subpopulation $k$, which does not depend on individual cells within the subpopulation. Gene $i$ also has *dropout probabilities* $d_{ij}$ $(j = 1, 2, \ldots, J)$, which may vary among different cells. During the preparation of this manuscript, it came to our attention that Ghazanfar et al. also used the Gamma-Normal mixture model to analyze scRNA-seq data but only applied it to categorize non-zero expression values into low expression or high expression values [34].

### 4.4 Imputation of dropout values

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. 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 $B_j$. Then we impute the expression of genes in set $A_j$ by borrowing information from the same gene's expression in other similar cells learned from $B_j$. Supplementary Figures S23 and S24c give some real data examples of zero count proportions in genes and dropout probabilities in cells, showing that it is reasonable to divide genes into two sets. To learn the cells similar to cell $j$ from $B_j$, we use the non-negative least squares (NNLS) regression:

$$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}^{(j)}}{\text{argmin}} \, ||\boldsymbol{X}_{B_j,j} - \boldsymbol{X}_{B_j,N_j}\boldsymbol{\beta}^{(j)}||_2^2, \text{ subject to } \boldsymbol{\beta}^{(j)} \geq \mathbf{0}. \tag{4.2}$$

Recall that $N_j$ represents the indices of cells that are candidate neighbors of cell $j$. The response $\boldsymbol{X}_{B_j,j}$ is a vector representing the $B_j$ rows in the $j$-th column of $\boldsymbol{X}$, the design matrix $\boldsymbol{X}_{B_j,N_j}$ is a sub-matrix of $\boldsymbol{X}$ with dimensions $|B_j| \times |N_j|$, and the coefficients $\boldsymbol{\beta}^{(j)}$ is a vector of length $|N_j|$. Note that NNLS itself has the property of leading to a sparse estimate $\hat{\boldsymbol{\beta}}^{(j)}$, whose components may have exact zeros [35], so NNLS can be used to select similar cells of cell $j$ from its neighbors $N_j$. Finally, the estimated coefficients $\hat{\boldsymbol{\beta}}^{(j)}$ from the set $B_j$ are used to impute the expression of gene set $A_j$ in cell $j$:

$$\hat{X}_{ij} = \begin{cases} X_{ij}, & i \in B_j, \\ X_{i,N_j}\hat{\boldsymbol{\beta}}^{(j)}, & i \in A_j. \end{cases} \tag{4.3}$$

We construct a separate regression model for each cell to impute the expression of genes with high dropout probabilities (Figure 1). This method simultaneously determines the values that need imputation, and would not introduce bias to the high expressions of accurately measured genes.

The application of scImpute involves two parameters. The first parameter is $K$, which determines the number of initial clusters to help identify candidate neighbors of each cell. The imputation results does not heavily rely on the choice of $K$ is since scImpute uses a model-based method to select similar cells in a later stage. However, setting $K$ to a value close to the true number of cell subpopulations can assist the selection of similar cells. The second parameter is a threshold $t$, and the imputation is only applied to the genes with dropout probabilities larger than $t$ in a cell to avoid over-imputation. Please note that the threshold is set on the dropout probability (the probability that a gene being a dropout in a cell), not on the dropout rate (the proportion of cells in which the gene is affected by dropout events). The sensitivity analysis based on the mouse embryo data [17] suggests that scImpute is robust to varying parameter values (Supplementary Figure S24a-b). Especially, the choice of parameter $t$ only affects a minute fraction of genes (Supplementary Figure S24c).

## 4.5 Generation of simulated scRNA-seq data

We suppose there are three cell types $c_1, c_2,$ and $c_3$, each with $50$ cells, and there are $20,000$ genes in total. In the gene population, only $810$ genes are truly differentially expressed, with one third having higher expression in each cell type respectively. We directly generate genes' $\log 10$-transformed read counts as expression values. First, mean expressions of the $20,000$ 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 dataset of mouse embryo cells. Second, we randomly draw $270$ 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 create $270$ highly expression genes for each of cell types $c_2$ and $c_3$ in the same way. Next, the expression values of each gene in the $150$ cells are simulated from Normal distributions defined by the mean and standard deviation 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 $\exp(-0.1 * \text{mean expression}^2)$, as assumed in [11]. Zero values are then introduced into the simulated data for each gene based on a Bernoulli distribution defined by the dropout rate of the gene, 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 use this simulation to investigate the efficacy and robustness of scImpute in a fair way.

## 4.6 Availability of data and software

The scRNA-seq data used in this manuscript are all publicly available and their sources are summarized in Supplementary Table S3. The `R` package `scImpute` is freely available at: `https:`

//github.com/Vivianstats/scImpute.

## 5 Funding

This work was supported by the PhRMA Foundation Research Starter Grant in Informatics, NIH/NIGMS grant R01GM120507, and NSF grant DMS-1613338.

## 6 Acknowledgements

We are grateful to Douglas Arneson, Feiyang Ma, Dr. Robert Modlin, Dr. Matteo Pellegrini, and Dr. Xia Yang at University of California, Los Angeles, for providing insightful discussions. We thank Dr. Mark Biggin at Lawrence Berkeley National Laboratory for his suggestions on this manuscript. We also thank Dr. Daria Merkurjev for assisting the data collection.

# 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] 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.

[10] 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.

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

[12] 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.

[13] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. Gene expression recovery for single cell rna sequencing. *bioRxiv*, 2017.

[14] Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for rna-seq experiments. *Genome research*, 21(9):1543–1551, 2011.

[15] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[16] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[17] 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.

[18] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

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

[20] 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.

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

[22] 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.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[24] Natalija Novak, Carmen Tepel, Susanne Koch, Klaudia Brix, Thomas Bieber, and Stefan Kraft. Evidence for a differential expression of the fc$\varepsilon$ri$\gamma$ chain in dendritic cells of atopic and nonatopic donors. *Journal of Clinical Investigation*, 111(7):1047, 2003.

[25] Alexandru Schiopu and Ovidiu S Cotoi. S100a8 and s100a9: Damps at the crossroads between innate immunity, traditional risk factors, and cardiovascular disease. *Mediators of inflammation*, 2013, 2013.

[26] 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.

[27] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[28] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):278, 2015.

[29] Pei Wang, Ryan T Rodriguez, Jing Wang, Amar Ghodasara, and Seung K Kim. Targeting sox17 in human embryonic stem cells creates unique strategies for isolating and analyzing developing endoderm. *Cell stem cell*, 8(3):335–346, 2011.

[30] Pei Wang, Kristen D McKnight, David J Wong, Ryan T Rodriguez, Takuya Sugiyama, Xueying Gu, Amar Ghodasara, Kun Qu, Howard Y Chang, and Seung K Kim. A molecular signature for purified definitive endoderm guides differentiation and isolation of endoderm from mouse and human embryonic stem cells. *Stem cells and development*, 21(12):2273–2287, 2012.

[31] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, 2009.

[32] Judith A Blake, Janan T Eppig, James A Kadin, Joel E Richardson, Cynthia L Smith, and Carol J Bult. Mouse genome database (mgd)-2017: community knowledge resource for the laboratory mouse. *Nucleic acids research*, 45(D1):D723–D729, 2017.

[33] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. Scnorm: robust normalization of single-cell rna-seq data. *Nature Methods*, 2017.

[34] Shila Ghazanfar, Adam J Bisogni, John T Ormerod, David M Lin, and Jean YH Yang. Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. *BMC Systems Biology*, 10(5):11, 2016.

[35] Martin Slawski, Matthias Hein, et al. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.

# Supplementary Materials

### 6.0.1 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

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

## 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 |

**Table S3:** Summary of scRNA-seq data used in this manuscript.

| data | source | % zero count in raw data | % zero count in imputed data |
|---|---|---|---|
| ERCC spike-ins | [14] | 27.1 | 0.1 |
| cell cycle | [16] | 22.6 | 0.5 |
| mouse embryo | [17] | 61.0 | 52.2 |
| PBMC | [22] | 98.6 | 28.3 |
| H1 vs DEC | [26] | 49.1 | 20.9 |
| H1 vs DEC (time-course) | [26] | 54.6 | 23.6 |

## Supplementary Figures

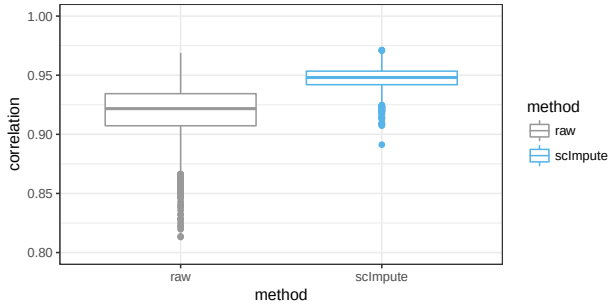


**Figure S1:** Correlation between the ERCC spike-ins' $\log 10$(counts) and $\log 10$(concentration) in the $3,005$ mouse cortex cells.

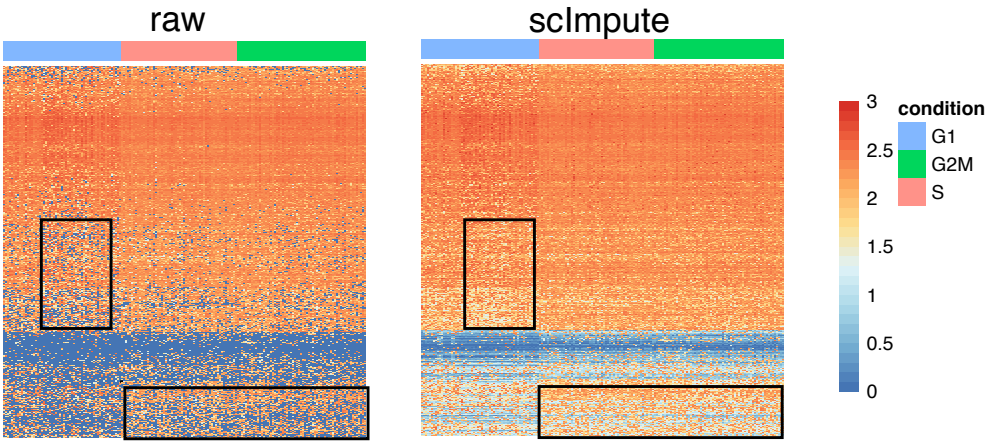


**Figure S2:** Heatmaps showing the $\log 10$(counts) of the $892$ cell cycle genes before and after imputation.

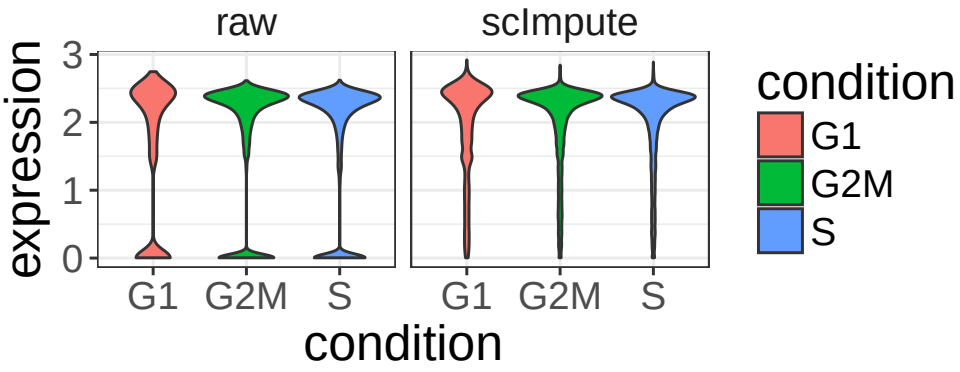


**Figure S3:** Violin plots showing the $\log 10$(counts) of the $892$ cell cycle genes in the three phases (G1, S, and G2M). scImpute has corrected the dropout values of cell cycle genes.

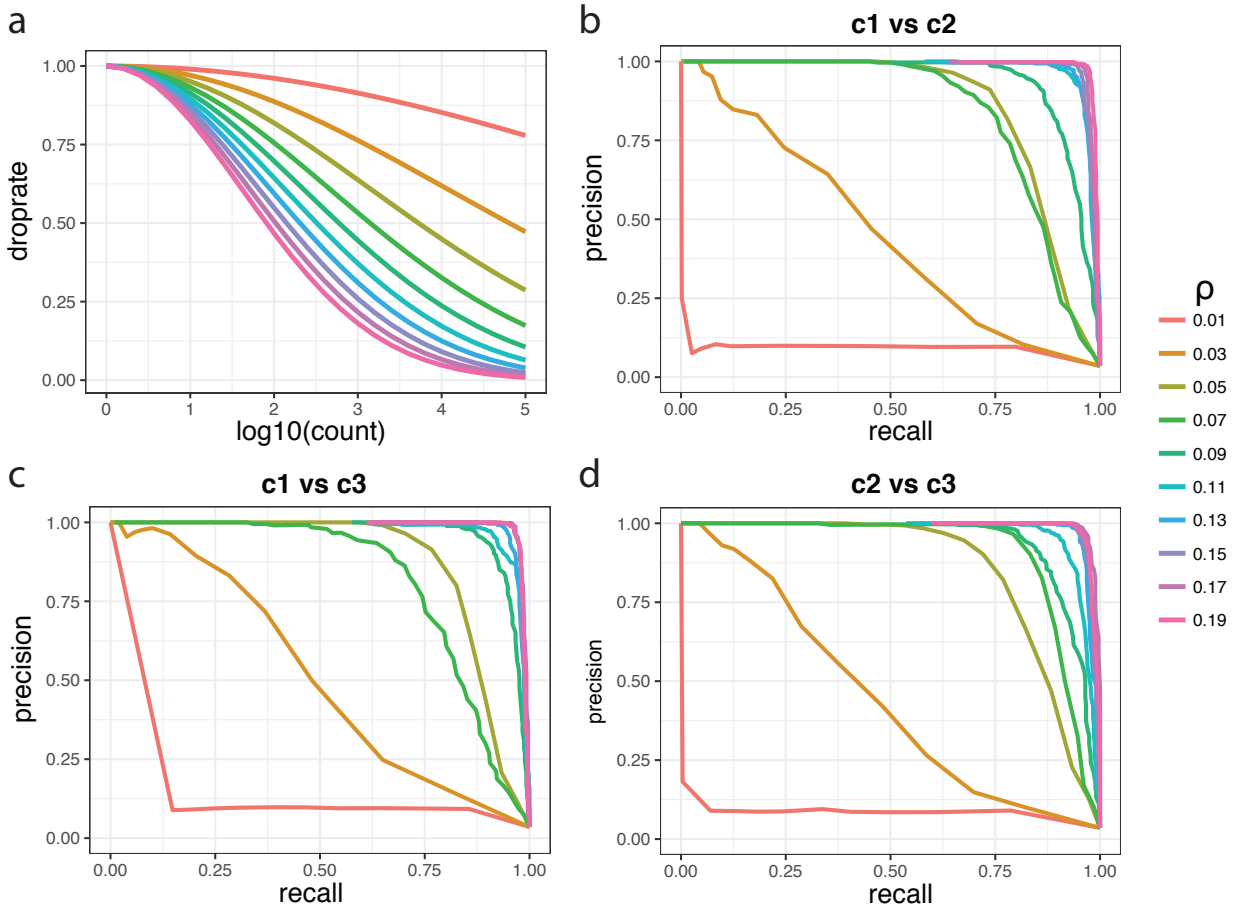


**Figure S4:** Performance of scImpute given different dropout rates in raw simulated data. **a**: The theoretical dropout rates determined by the double exponential function $\exp(-\rho \times \log 10(\text{count})^2)$, with $\rho$ varying from $0.01$ to $0.19$ by a step of $0.02$. **b-d**: The precision-recall curves for the identification of differentially expressed genes from the imputed data.

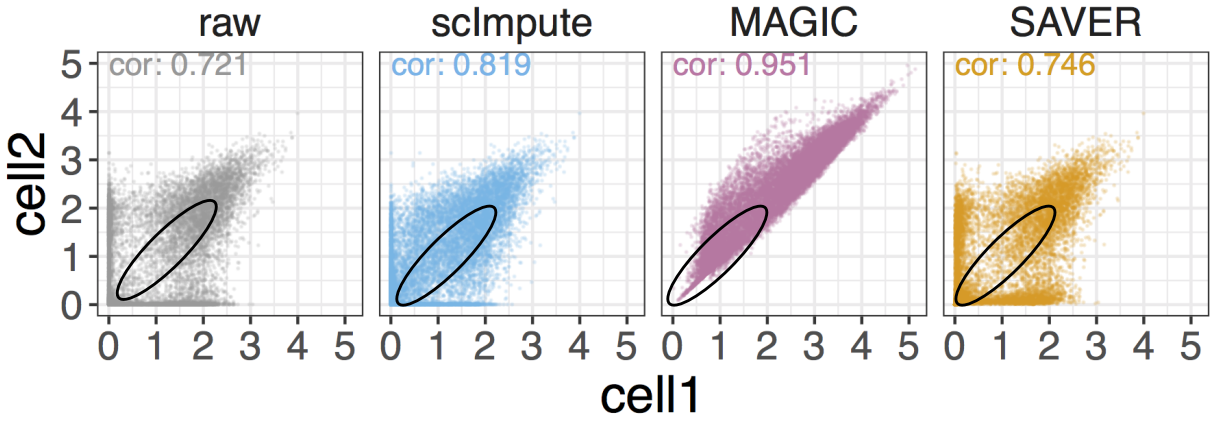


**Figure S5:** The first two PCs obtained from the raw and imputed data of mouse embryonic cells. The black dots mark the outlier cells detected by scImpute.

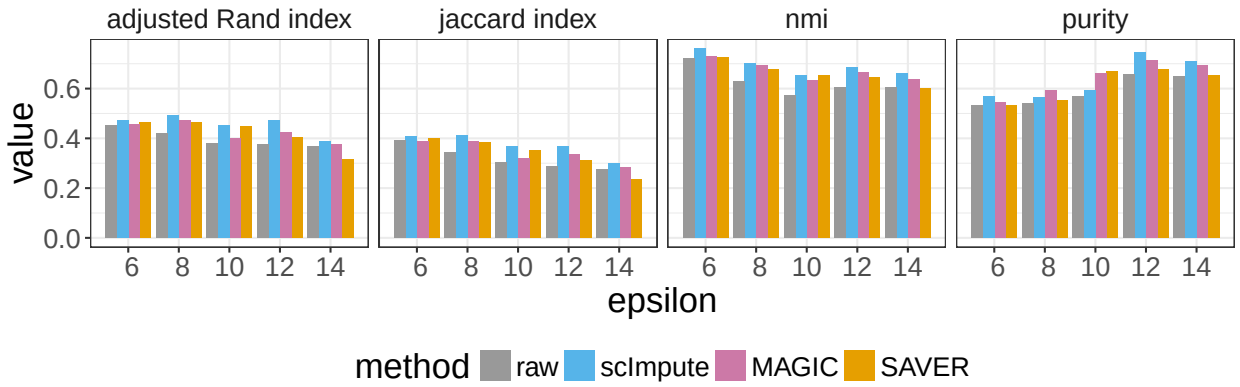


**Figure S6:** The adjusted rand index, Jaccard index, nmi, and purity scores of clustering results based on the raw and imputed data. Clustering is performed by the spectral clustering algorithm on the single cells' scores in the first two PCs.

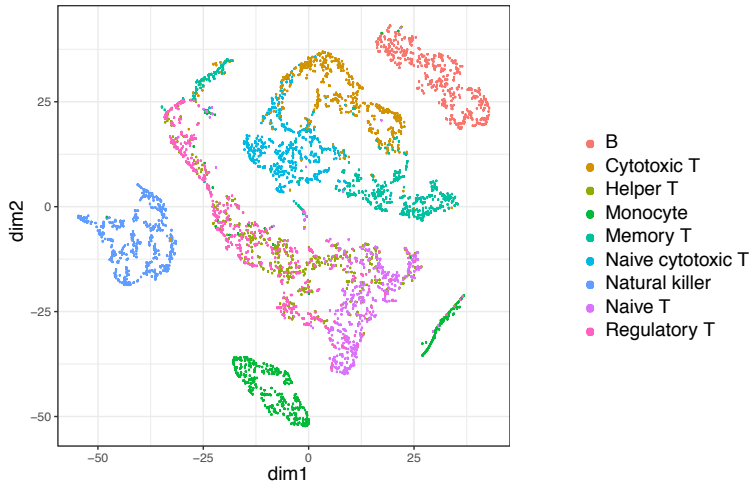


**Figure S7:** The first two dimensions of the t-SNE results calculated from imputed PBMC data by MAGIC.

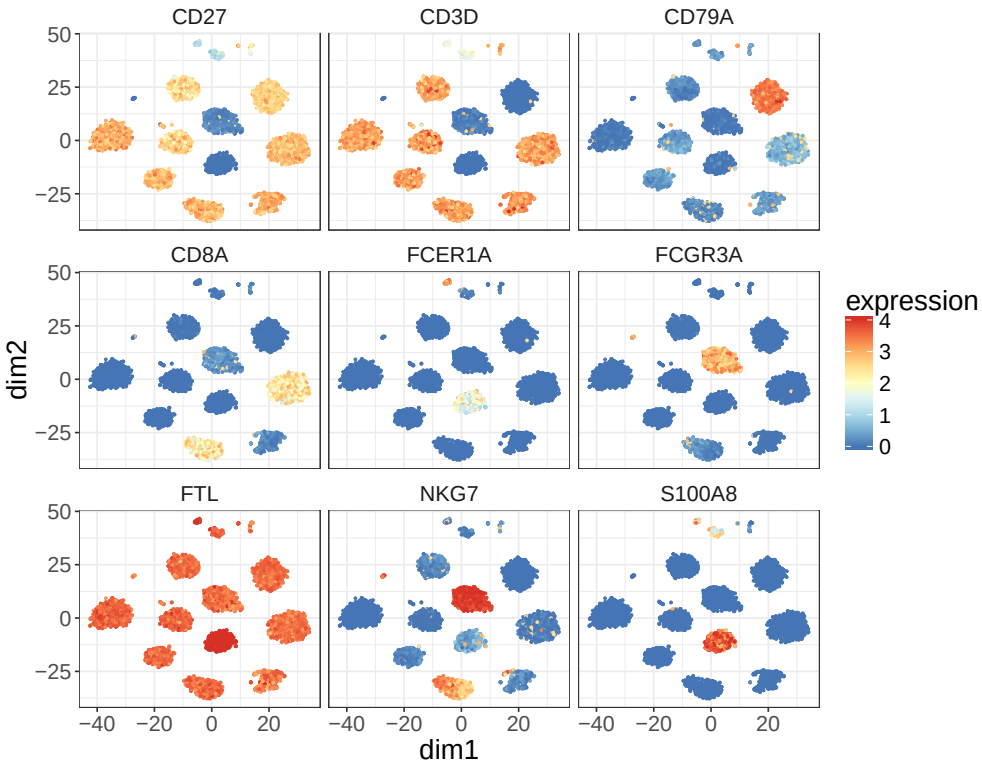


**Figure S8:** The $\log 10$ expression levels of nine known marker genes shown in clusters obtained from imputed data by scImpute.

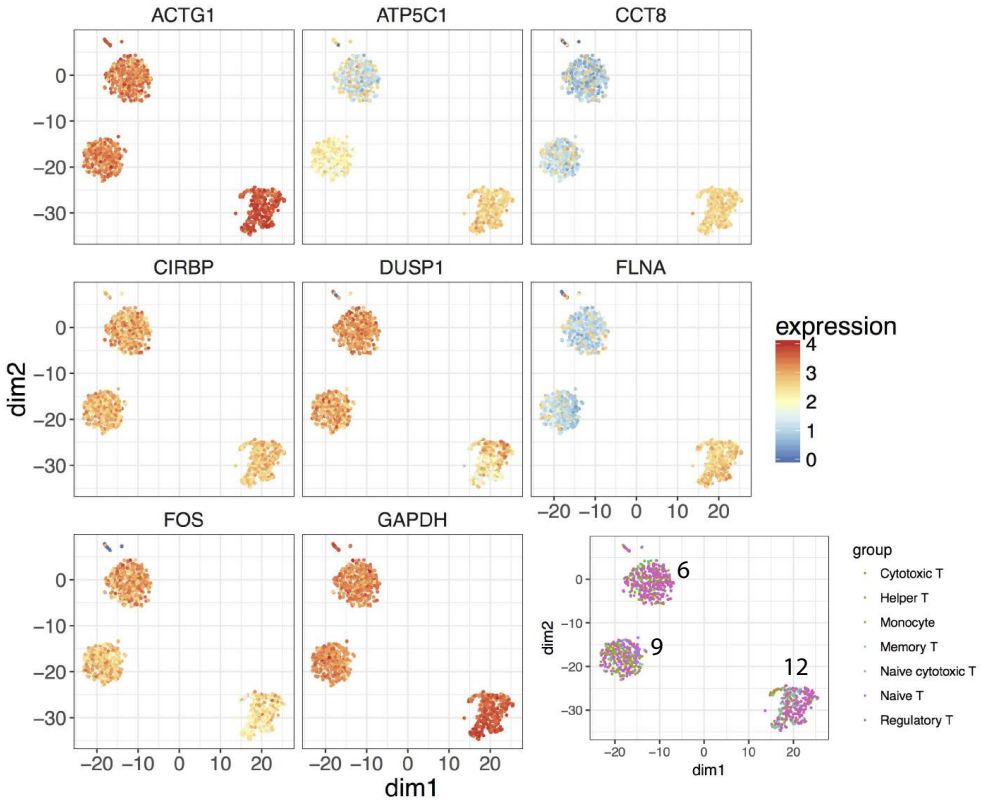


**Figure S9:** The $\log 10$ expression levels of eight potential marker genes to distinguish subpopulations of T cells.

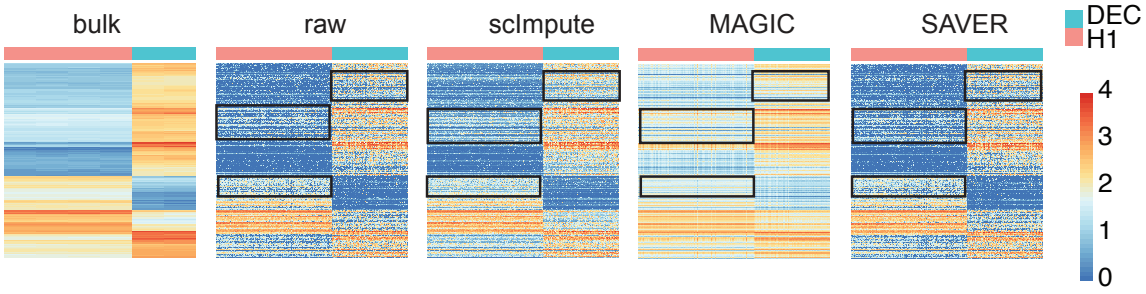


**Figure S10:** The $\log 10$ expression profiles of the top $200$ DE genes detected in the bulk data by DESeq.

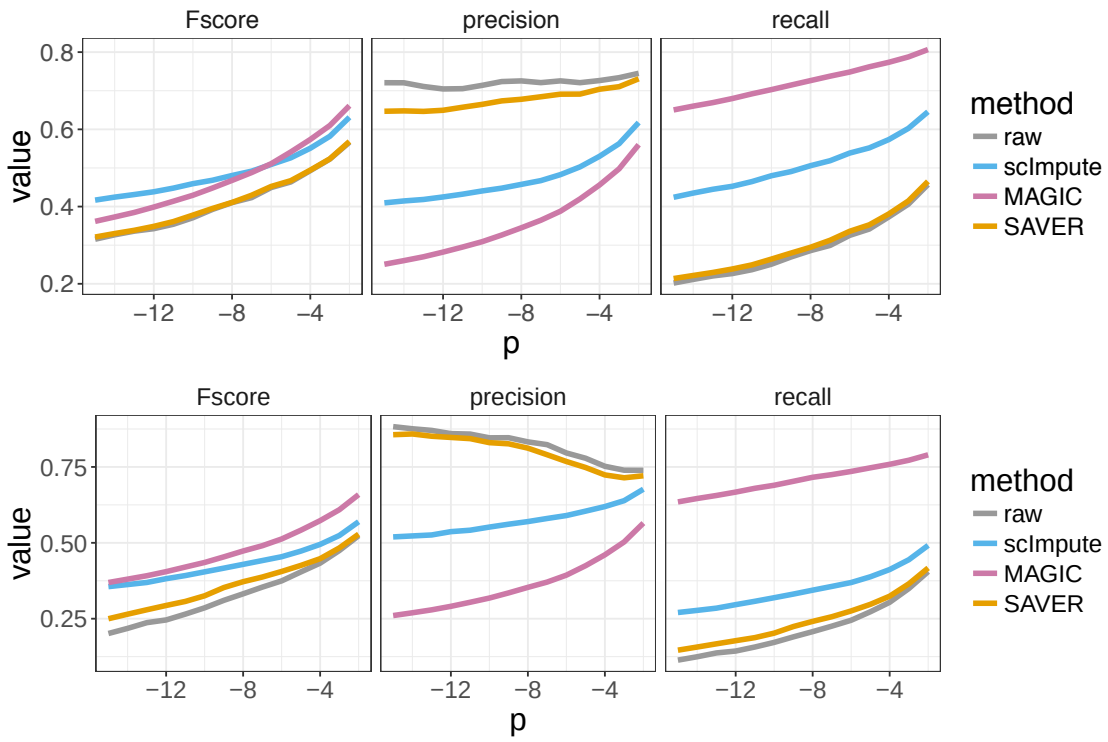


**Figure S11:** Comparison of DE analysis between bulk and single-cell data. **a**: $p$-values for both bulk and single-cell data are calculated using DESeq. **b**: $p$-values for bulk data are calculated using DESeq and $p$-values for single-cell data are calculated using MAST.

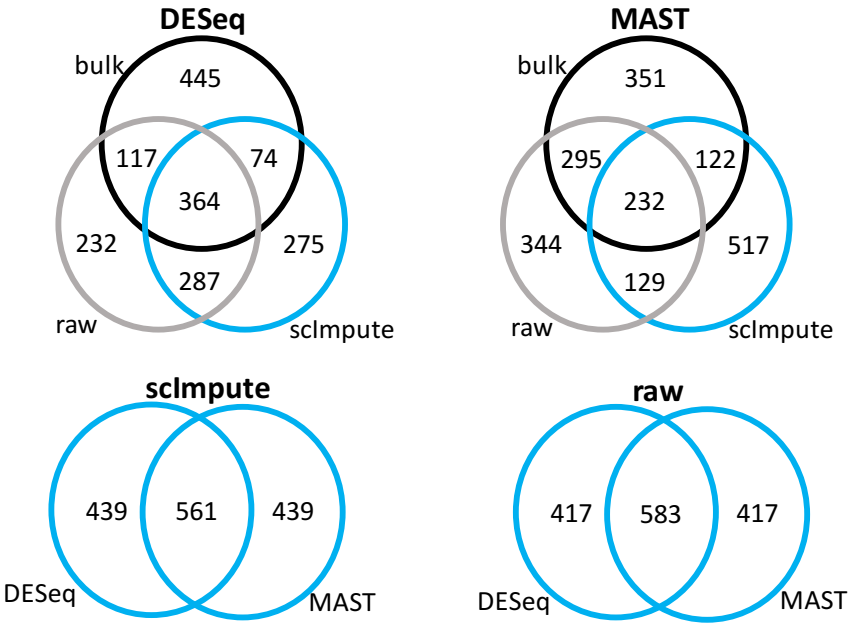


**Figure S12:** Intersection of the top $1,000$ DE genes detected in bulk, raw, and imputed data by DESeq or MAST. **a**: DE genes are detected using DESeq. **b**: DE genes in bulk data are detected using DESeq; DE genes in single-cell data are detected using MAST. **c**: Top $1,000$ genes detected in scImpute's imputed data by DESeq and MAST have $561$ in common. **d**: Top $1,000$ genes detected in raw data by DESeq and MAST have $583$ in common.

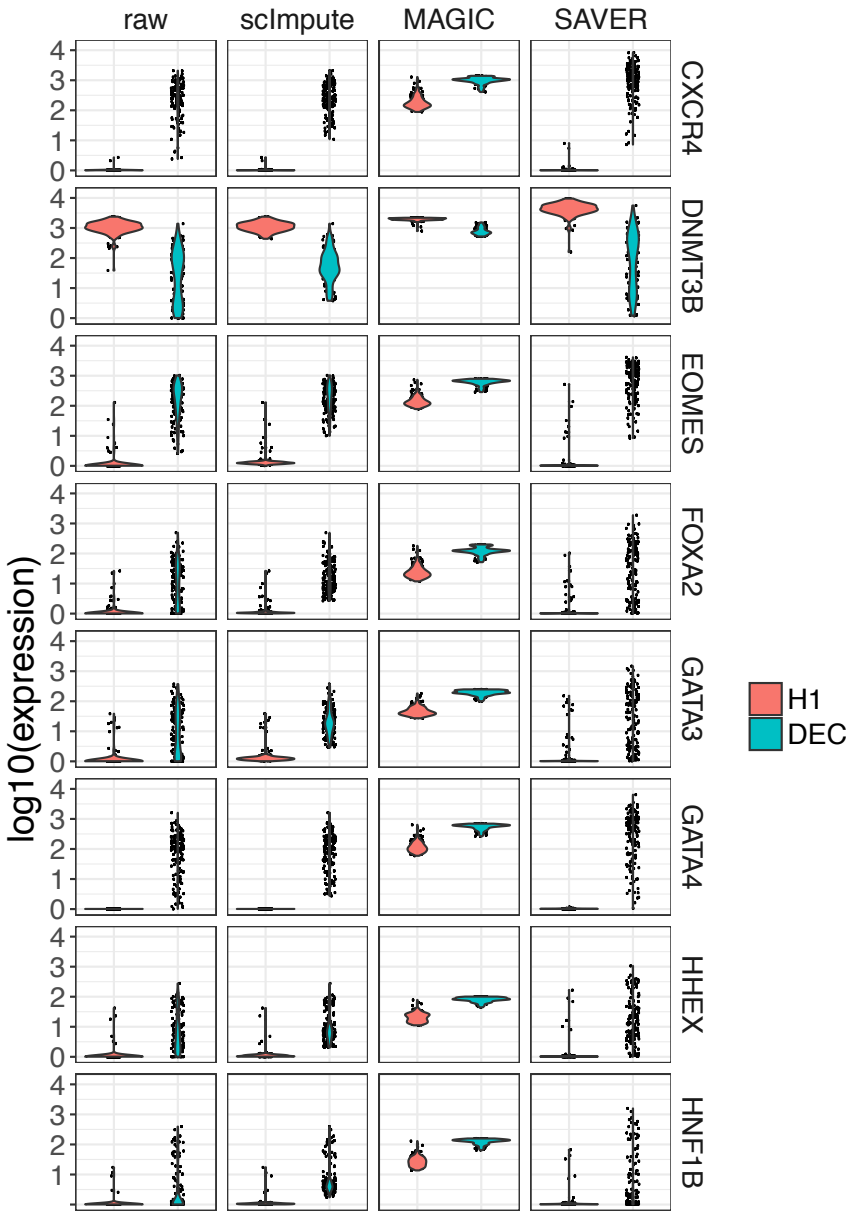


**Figure S13:** Violin plots showing $\log 10$ gene expression of nine genes before and after imputation.

| GO Term | Description | P-value | GO Term | Description | P-value |
|---|---|---|---|---|---|
| GO:0048646 | anatomical structure formation involved in morphogenesis | 4.86E-11 | GO:1902680 | positive regulation of RNA biosynthetic process | 1.44E-04 |
| GO:0032502 | developmental process | 2.53E-09 | GO:0070169 | positive regulation of biomineral tissue development | 1.60E-04 |
| GO:2000026 | regulation of multicellular organismal development | 9.38E-09 | GO:0030198 | extracellular matrix organization | 1.62E-04 |
| GO:0045597 | positive regulation of cell differentiation | 2.51E-08 | GO:0001837 | epithelial to mesenchymal transition | 1.64E-04 |
| GO:0045595 | regulation of cell differentiation | 3.38E-08 | GO:0043062 | extracellular structure organization | 1.67E-04 |
| GO:0050793 | regulation of developmental process | 4.96E-08 | GO:0060363 | cranial suture morphogenesis | 1.71E-04 |
| GO:0051094 | positive regulation of developmental process | 9.31E-08 | GO:0048562 | embryonic organ morphogenesis | 1.72E-04 |
| GO:0051239 | regulation of multicellular organismal process | 2.46E-07 | GO:0051241 | negative regulation of multicellular organismal process | 1.85E-04 |
| GO:0048856 | anatomical structure development | 2.81E-07 | GO:0045944 | positive regulation of transcription from RNA polymerase II promoter | 1.90E-04 |
| GO:0048869 | cellular developmental process | 3.60E-07 | GO:0010646 | regulation of cell communication | 1.98E-04 |
| GO:0030154 | cell differentiation | 3.91E-07 | GO:0042221 | response to chemical | 2.14E-04 |
| GO:0048513 | animal organ development | 5.47E-07 | GO:0006928 | movement of cell or subcellular component | 2.22E-04 |
| GO:0071310 | cellular response to organic substance | 6.98E-07 | GO:0022603 | regulation of anatomical structure morphogenesis | 2.69E-04 |
| GO:0098609 | cell-cell adhesion | 8.42E-07 | GO:0051093 | negative regulation of developmental process | 2.83E-04 |
| GO:0051240 | positive regulation of multicellular organismal process | 8.54E-07 | GO:0023051 | regulation of signaling | 2.88E-04 |
| GO:0045778 | positive regulation of ossification | 1.15E-06 | GO:0050679 | positive regulation of epithelial cell proliferation | 2.96E-04 |
| GO:0001525 | angiogenesis | 1.21E-06 | GO:0048523 | negative regulation of cellular process | 3.00E-04 |
| GO:0070887 | cellular response to chemical stimulus | 2.00E-06 | GO:0007568 | aging | 3.22E-04 |
| GO:0048729 | tissue morphogenesis | 3.18E-06 | GO:0022407 | regulation of cell-cell adhesion | 3.23E-04 |
| GO:0070848 | response to growth factor | 4.21E-06 | GO:0042482 | positive regulation of odontogenesis | 3.29E-04 |
| GO:0048598 | embryonic morphogenesis | 5.49E-06 | GO:0007178 | transmembrane receptor protein serine/threonine kinase signaling pathway | 3.39E-04 |
| GO:0071560 | cellular response to transforming growth factor beta stimulus | 5.96E-06 | GO:0048732 | gland development | 3.45E-04 |
| GO:0045669 | positive regulation of osteoblast differentiation | 8.61E-06 | GO:0048585 | negative regulation of response to stimulus | 3.48E-04 |
| GO:0009653 | anatomical structure morphogenesis | 1.21E-05 | GO:0051254 | positive regulation of RNA metabolic process | 3.61E-04 |
| GO:0003198 | epithelial to mesenchymal transition involved in endocardial cushion formation | 1.26E-05 | GO:0060317 | cardiac epithelial to mesenchymal transition | 3.83E-04 |
| GO:0071495 | cellular response to endogenous stimulus | 1.29E-05 | GO:0030155 | regulation of cell adhesion | 3.90E-04 |
| GO:0061138 | morphogenesis of a branching epithelium | 1.47E-05 | GO:1903707 | negative regulation of hemopoiesis | 4.03E-04 |
| GO:0071559 | response to transforming growth factor beta | 1.52E-05 | GO:0035239 | tube morphogenesis | 4.14E-04 |
| GO:0042481 | regulation of odontogenesis | 1.85E-05 | GO:0045165 | cell fate commitment | 4.23E-04 |
| GO:0010038 | response to metal ion | 1.99E-05 | GO:0048519 | negative regulation of biological process | 4.34E-04 |
| GO:0071363 | cellular response to growth factor stimulus | 2.25E-05 | GO:0071241 | cellular response to inorganic substance | 4.38E-04 |
| GO:0030500 | regulation of bone mineralization | 2.30E-05 | GO:0043392 | negative regulation of DNA binding | 4.63E-04 |
| GO:0030278 | regulation of ossification | 2.39E-05 | GO:1904381 | Golgi apparatus mannose trimming | 4.98E-04 |
| GO:0048754 | branching morphogenesis of an epithelial tube | 2.50E-05 | GO:0060129 | thyroid-stimulating hormone-secreting cell differentiation | 4.98E-04 |
| GO:0001763 | morphogenesis of a branching structure | 2.52E-05 | GO:0051270 | regulation of cellular component movement | 5.01E-04 |
| GO:0035116 | embryonic hindlimb morphogenesis | 2.71E-05 | GO:0060429 | epithelium development | 5.38E-04 |
| GO:2000677 | regulation of transcription regulatory region DNA binding | 3.08E-05 | GO:0014070 | response to organic cyclic compound | 5.44E-04 |
| GO:0048583 | regulation of response to stimulus | 3.43E-05 | GO:0061029 | eyelid development in camera-type eye | 5.60E-04 |
| GO:0048518 | positive regulation of biological process | 3.45E-05 | GO:0048762 | mesenchymal cell differentiation | 6.03E-04 |
| GO:0048468 | cell development | 3.83E-05 | GO:0071407 | cellular response to organic cyclic compound | 6.14E-04 |
| GO:0007166 | cell surface receptor signaling pathway | 4.11E-05 | GO:0003006 | developmental process involved in reproduction | 6.38E-04 |
| GO:0009966 | regulation of signal transduction | 4.46E-05 | GO:0030326 | embryonic limb morphogenesis | 6.45E-04 |
| GO:0035115 | embryonic forelimb morphogenesis | 4.54E-05 | GO:0035113 | embryonic appendage morphogenesis | 6.45E-04 |
| GO:0045785 | positive regulation of cell adhesion | 4.68E-05 | GO:0051101 | regulation of DNA binding | 6.46E-04 |
| GO:0070167 | regulation of biomineral tissue development | 5.30E-05 | GO:1904018 | positive regulation of vasculature development | 6.76E-04 |
| GO:0048522 | positive regulation of cellular process | 5.64E-05 | GO:0008284 | positive regulation of cell proliferation | 6.84E-04 |
| GO:0010033 | response to organic substance | 6.09E-05 | GO:2001212 | regulation of vasculogenesis | 7.06E-04 |
| GO:0002009 | morphogenesis of an epithelium | 7.04E-05 | GO:0071248 | cellular response to metal ion | 7.06E-04 |
| GO:0009887 | animal organ morphogenesis | 7.67E-05 | GO:0007165 | signal transduction | 7.11E-04 |
| GO:0030501 | positive regulation of bone mineralization | 8.32E-05 | GO:0030855 | epithelial cell differentiation | 7.21E-04 |
| GO:0032501 | multicellular organismal process | 8.47E-05 | GO:0048565 | digestive tract development | 7.22E-04 |
| GO:0045667 | regulation of osteoblast differentiation | 8.77E-05 | GO:0045596 | negative regulation of cell differentiation | 7.24E-04 |
| GO:0035137 | hindlimb morphogenesis | 9.56E-05 | GO:0002683 | negative regulation of immune system process | 8.04E-04 |
| GO:0035295 | tube development | 9.85E-05 | GO:0042127 | regulation of cell proliferation | 8.36E-04 |
| GO:1902105 | regulation of leukocyte differentiation | 1.08E-04 | GO:0030203 | glycosaminoglycan metabolic process | 8.78E-04 |
| GO:0007155 | cell adhesion | 1.10E-04 | GO:0051592 | response to calcium ion | 8.80E-04 |
| GO:0035584 | calcium-mediated signaling using intracellular calcium source | 1.14E-04 | GO:0035148 | tube formation | 8.80E-04 |

**Figure S14:** Enriched GO terms ($p < 10^{-3}$) in the $244$ DEC up-regulated genes that are only detected in scImpute's imputed data by DESeq.

| GO Term | Description | P-value | GO Term | Description | P-value |
|---|---|---|---|---|---|
| GO:0032502 | developmental process | 7.42E-12 | GO:0051252 | regulation of RNA metabolic process | 1.06E-04 |
| GO:0048856 | anatomical structure development | 7.59E-12 | GO:0042481 | regulation of odontogenesis | 1.07E-04 |
| GO:0048646 | anatomical structure formation involved in morphogenesis | 2.99E-11 | GO:0009887 | animal organ morphogenesis | 1.12E-04 |
| GO:0030154 | cell differentiation | 1.30E-08 | GO:0003002 | regionalization | 1.12E-04 |
| GO:0050793 | regulation of developmental process | 2.21E-08 | GO:0072073 | kidney epithelium development | 1.13E-04 |
| GO:0009653 | anatomical structure morphogenesis | 2.47E-08 | GO:0023051 | regulation of signaling | 1.22E-04 |
| GO:0035116 | embryonic hindlimb morphogenesis | 3.14E-08 | GO:0061312 | BMP signaling pathway involved in heart development | 1.25E-04 |
| GO:0048869 | cellular developmental process | 4.97E-08 | GO:0061180 | mammary gland epithelium development | 1.25E-04 |
| GO:2000026 | regulation of multicellular organismal development | 7.95E-08 | GO:0048762 | mesenchymal cell differentiation | 1.26E-04 |
| GO:0060411 | cardiac septum morphogenesis | 1.81E-07 | GO:0010646 | regulation of cell communication | 1.26E-04 |
| GO:0051094 | positive regulation of developmental process | 2.06E-07 | GO:0001958 | endochondral ossification | 1.29E-04 |
| GO:0035137 | hindlimb morphogenesis | 2.69E-07 | GO:0036075 | replacement ossification | 1.29E-04 |
| GO:0048513 | animal organ development | 3.45E-07 | GO:0060317 | cardiac epithelial to mesenchymal transition | 1.29E-04 |
| GO:0051093 | negative regulation of developmental process | 5.96E-07 | GO:0001568 | blood vessel development | 1.48E-04 |
| GO:0032501 | multicellular organismal process | 8.84E-07 | GO:0048523 | negative regulation of cellular process | 1.53E-04 |
| GO:0022603 | regulation of anatomical structure morphogenesis | 1.47E-06 | GO:0060349 | bone morphogenesis | 1.55E-04 |
| GO:0003198 | epithelial to mesenchymal transition involved in endocardial | 1.60E-06 | GO:0045596 | negative regulation of cell differentiation | 1.99E-04 |
| GO:0098609 | cell-cell adhesion | 1.98E-06 | GO:0045892 | negative regulation of transcription, DNA-templated | 2.01E-04 |
| GO:0048598 | embryonic morphogenesis | 2.32E-06 | GO:1903507 | negative regulation of nucleic acid-templated transcription | 2.12E-04 |
| GO:0006357 | regulation of transcription from RNA polymerase II promoter | 2.65E-06 | GO:0003179 | heart valve morphogenesis | 2.18E-04 |
| GO:0060429 | epithelium development | 2.78E-06 | GO:2000241 | regulation of reproductive process | 2.19E-04 |
| GO:0048729 | tissue morphogenesis | 2.80E-06 | GO:1902679 | negative regulation of RNA biosynthetic process | 2.19E-04 |
| GO:0045893 | positive regulation of transcription, DNA-templated | 3.77E-06 | GO:0030500 | regulation of bone mineralization | 2.30E-04 |
| GO:1903508 | positive regulation of nucleic acid-templated transcription | 3.77E-06 | GO:0030512 | negative regulation of transforming growth factor beta receptor | 2.30E-04 |
| GO:1902680 | positive regulation of RNA biosynthetic process | 3.85E-06 | GO:2000677 | regulation of transcription regulatory region DNA binding | 2.33E-04 |
| GO:0051239 | regulation of multicellular organismal process | 4.30E-06 | GO:0035904 | aorta development | 2.38E-04 |
| GO:0001763 | morphogenesis of a branching structure | 4.75E-06 | GO:0048583 | regulation of response to stimulus | 2.42E-04 |
| GO:0045595 | regulation of cell differentiation | 5.27E-06 | GO:0002009 | morphogenesis of an epithelium | 2.46E-04 |
| GO:0003148 | outflow tract septum morphogenesis | 5.44E-06 | GO:0001667 | ameboidal-type cell migration | 2.47E-04 |
| GO:0001525 | angiogenesis | 6.50E-06 | GO:0051173 | positive regulation of nitrogen compound metabolic process | 2.49E-04 |
| GO:0051254 | positive regulation of RNA metabolic process | 7.01E-06 | GO:0007411 | axon guidance | 2.74E-04 |
| GO:0048522 | positive regulation of cellular process | 7.62E-06 | GO:1903845 | negative regulation of cellular response to transforming growth | 2.78E-04 |
| GO:0009888 | tissue development | 8.20E-06 | GO:0031325 | positive regulation of cellular metabolic process | 3.00E-04 |
| GO:0035239 | tube morphogenesis | 9.92E-06 | GO:0097094 | craniofacial suture morphogenesis | 3.01E-04 |
| GO:0007155 | cell adhesion | 1.02E-05 | GO:0098742 | cell-cell adhesion via plasma-membrane adhesion molecules | 3.12E-04 |
| GO:0090092 | regulation of transmembrane receptor protein serine  signaling | 1.05E-05 | GO:0009966 | regulation of signal transduction | 3.13E-04 |
| GO:0022610 | biological adhesion | 1.15E-05 | GO:0001503 | ossification | 3.16E-04 |
| GO:0061138 | morphogenesis of a branching epithelium | 1.32E-05 | GO:0048589 | developmental growth | 3.21E-04 |
| GO:0090287 | regulation of cellular response to growth factor stimulus | 1.34E-05 | GO:0071657 | positive regulation of granulocyte colony-stimulating production | 3.50E-04 |
| GO:0006355 | regulation of transcription, DNA-templated | 1.46E-05 | GO:2001055 | positive regulation of mesenchymal cell apoptotic process | 3.50E-04 |
| GO:0048754 | branching morphogenesis of an epithelial tube | 1.48E-05 | GO:1901258 | positive regulation of macrophage colony-stimulating factor | 3.50E-04 |
| GO:0017015 | regulation of transforming growth factor beta receptor signaling | 1.67E-05 | GO:0051147 | regulation of muscle cell differentiation | 3.54E-04 |
| GO:0006928 | movement of cell or subcellular component | 1.69E-05 | GO:0042127 | regulation of cell proliferation | 3.55E-04 |
| GO:1903506 | regulation of nucleic acid-templated transcription | 1.92E-05 | GO:0035148 | tube formation | 3.83E-04 |
| GO:1903844 | regulation of cellular response to transforming growth factor | 1.99E-05 | GO:0040007 | growth | 3.84E-04 |
| GO:0030326 | embryonic limb morphogenesis | 2.05E-05 | GO:0001822 | kidney development | 4.07E-04 |
| GO:0035113 | embryonic appendage morphogenesis | 2.05E-05 | GO:0051253 | negative regulation of RNA metabolic process | 4.28E-04 |
| GO:0035115 | embryonic forelimb morphogenesis | 2.06E-05 | GO:0048562 | embryonic organ morphogenesis | 4.33E-04 |
| GO:2001141 | regulation of RNA biosynthetic process | 2.13E-05 | GO:0023019 | signal transduction involved in regulation of gene expression | 4.63E-04 |
| GO:0051240 | positive regulation of multicellular organismal process | 2.56E-05 | GO:0048519 | negative regulation of biological process | 4.76E-04 |
| GO:0010628 | positive regulation of gene expression | 2.71E-05 | GO:0035295 | tube development | 4.76E-04 |
| GO:0048518 | positive regulation of biological process | 2.90E-05 | GO:0009893 | positive regulation of metabolic process | 4.89E-04 |
| GO:0009891 | positive regulation of biosynthetic process | 3.55E-05 | GO:2000243 | positive regulation of reproductive process | 5.02E-04 |
| GO:2000678 | negative regulation of transcription regulatory DNA binding | 3.57E-05 | GO:0003139 | secondary heart field specification | 5.04E-04 |
| GO:0061311 | cell surface receptor signaling pathway involved in heart | 3.57E-05 | GO:0003128 | heart field specification | 5.04E-04 |
| GO:0045597 | positive regulation of cell differentiation | 3.71E-05 | GO:0060363 | cranial suture morphogenesis | 5.04E-04 |
| GO:0043392 | negative regulation of DNA binding | 3.82E-05 | GO:0070167 | regulation of biomineral tissue development | 5.09E-04 |
| GO:0051270 | regulation of cellular component movement | 3.87E-05 | GO:0060560 | developmental growth involved in morphogenesis | 5.10E-04 |
| GO:0045944 | positive regulation of transcription from RNA polymerase II | 8.08E-05 | GO:0008283 | cell proliferation | 6.73E-04 |
| GO:0097485 | neuron projection guidance | 8.08E-05 | GO:0001657 | ureteric bud development | 6.78E-04 |
| GO:0042692 | muscle cell differentiation | 8.62E-05 | GO:0007369 | gastrulation | 6.78E-04 |

**Figure S15:** Enriched GO terms ($p < 10^{-3}$) in the $339$ DEC up-regulated genes that are only detected in scImpute's imputed data by MAST.

| GO Term | Description | P-value |
|---|---|---|
| GO:0018401 | peptidyl-proline hydroxylation to 4-hydroxy-L-proline | 5.97E-05 |
| GO:0030029 | actin filament-based process | 1.58E-04 |
| GO:0065007 | biological regulation | 1.61E-04 |
| GO:0034113 | heterotypic cell-cell adhesion | 1.67E-04 |
| GO:0030036 | actin cytoskeleton organization | 2.30E-04 |
| GO:1903829 | positive regulation of cellular protein localization | 3.29E-04 |
| GO:0034446 | substrate adhesion-dependent cell spreading | 6.18E-04 |
| GO:0065008 | regulation of biological quality | 6.25E-04 |
| GO:0048583 | regulation of response to stimulus | 6.79E-04 |

**Figure S16:** Enriched GO terms ($p < 10^{-3}$) in the $249$ DEC up-regulated genes that are only detected in raw data by DESeq.

| GO Term | Description | P-value | GO Term | Description | P-value |
|---|---|---|---|---|---|
| GO:0071840 | cellular component organization or biogenesis | 2.69E-08 | GO:0071310 | cellular response to organic substance | 2.30E-04 |
| GO:0006260 | DNA replication | 7.56E-08 | GO:0044806 | G-quadruplex DNA unwinding | 2.33E-04 |
| GO:0016043 | cellular component organization | 8.79E-08 | GO:0120035 | regulation of plasma membrane bounded cell projection organization | 2.36E-04 |
| GO:0032508 | DNA duplex unwinding | 1.14E-06 | GO:0033627 | cell adhesion mediated by integrin | 2.63E-04 |
| GO:0048869 | cellular developmental process | 2.57E-06 | GO:0030334 | regulation of cell migration | 2.66E-04 |
| GO:0071897 | DNA biosynthetic process | 2.67E-06 | GO:0048518 | positive regulation of biological process | 2.86E-04 |
| GO:0022604 | regulation of cell morphogenesis | 3.23E-06 | GO:0070252 | actin-mediated cell contraction | 3.01E-04 |
| GO:0032392 | DNA geometric change | 3.82E-06 | GO:0031344 | regulation of cell projection organization | 3.05E-04 |
| GO:0001649 | osteoblast differentiation | 3.87E-06 | GO:0009987 | cellular process | 3.19E-04 |
| GO:0048523 | negative regulation of cellular process | 6.03E-06 | GO:0030048 | actin filament-based movement | 3.25E-04 |
| GO:0048519 | negative regulation of biological process | 6.05E-06 | GO:0022616 | DNA strand elongation | 3.33E-04 |
| GO:0030198 | extracellular matrix organization | 6.38E-06 | GO:0070887 | cellular response to chemical stimulus | 3.38E-04 |
| GO:0032502 | developmental process | 6.68E-06 | GO:0009719 | response to endogenous stimulus | 3.48E-04 |
| GO:0043062 | extracellular structure organization | 6.69E-06 | GO:0034330 | cell junction organization | 3.90E-04 |
| GO:0043504 | mitochondrial DNA repair | 7.06E-06 | GO:0050767 | regulation of neurogenesis | 4.03E-04 |
| GO:0000732 | strand displacement | 8.07E-06 | GO:0008285 | negative regulation of cell proliferation | 4.06E-04 |
| GO:0042127 | regulation of cell proliferation | 1.26E-05 | GO:0007010 | cytoskeleton organization | 4.42E-04 |
| GO:0035987 | endodermal cell differentiation | 1.28E-05 | GO:0000904 | cell morphogenesis involved in differentiation | 4.58E-04 |
| GO:0030516 | regulation of axon extension | 1.49E-05 | GO:0008284 | positive regulation of cell proliferation | 4.77E-04 |
| GO:0007049 | cell cycle | 1.93E-05 | GO:0051093 | negative regulation of developmental process | 4.77E-04 |
| GO:0022610 | biological adhesion | 2.01E-05 | GO:0022607 | cellular component assembly | 4.90E-04 |
| GO:0010975 | regulation of neuron projection development | 2.08E-05 | GO:0010769 | regulation of cell morphogenesis involved in differentiation | 4.99E-04 |
| GO:0071103 | DNA conformation change | 2.78E-05 | GO:0051276 | chromosome organization | 5.04E-04 |
| GO:0051716 | cellular response to stimulus | 2.99E-05 | GO:0006928 | movement of cell or subcellular component | 5.06E-04 |
| GO:1901796 | regulation of signal transduction by p53 class | 3.53E-05 | GO:0050793 | regulation of developmental process | 5.19E-04 |
| GO:0060284 | regulation of cell development | 3.66E-05 | GO:1901566 | organonitrogen compound biosynthetic process | 5.68E-04 |
| GO:0010811 | positive regulation of cell-substrate adhesion | 4.11E-05 | GO:0045664 | regulation of neuron differentiation | 5.87E-04 |
| GO:0007155 | cell adhesion | 4.23E-05 | GO:1903047 | mitotic cell cycle process | 6.44E-04 |
| GO:0048513 | animal organ development | 4.37E-05 | GO:0090100 | positive regulation of transmembrane receptor protein serine/threonine | 6.48E-04 |
| GO:0051052 | regulation of DNA metabolic process | 5.68E-05 | GO:0048583 | regulation of response to stimulus | 6.52E-04 |
| GO:0061387 | regulation of extent of cell growth | 5.84E-05 | GO:0051128 | regulation of cellular component organization | 6.61E-04 |
| GO:0010810 | regulation of cell-substrate adhesion | 5.94E-05 | GO:0051270 | regulation of cellular component movement | 6.92E-04 |
| GO:0006259 | DNA metabolic process | 8.21E-05 | GO:0040012 | regulation of locomotion | 7.05E-04 |
| GO:0000731 | DNA synthesis involved in DNA repair | 8.64E-05 | GO:0010033 | response to organic substance | 7.25E-04 |
| GO:0030029 | actin filament-based process | 8.93E-05 | GO:0050794 | regulation of cellular process | 7.28E-04 |
| GO:0042493 | response to drug | 1.08E-04 | GO:0009611 | response to wounding | 7.48E-04 |
| GO:0010721 | negative regulation of cell development | 1.16E-04 | GO:0034446 | substrate adhesion-dependent cell spreading | 7.63E-04 |
| GO:0097435 | supramolecular fiber organization | 1.25E-04 | GO:0030049 | muscle filament sliding | 7.63E-04 |
| GO:0042221 | response to chemical | 1.50E-04 | GO:0033275 | actin-myosin filament sliding | 7.63E-04 |
| GO:2000145 | regulation of cell motility | 1.73E-04 | GO:0010977 | negative regulation of neuron projection development | 7.88E-04 |
| GO:0050678 | regulation of epithelial cell proliferation | 1.74E-04 | GO:0023051 | regulation of signaling | 7.95E-04 |
| GO:0048856 | anatomical structure development | 1.77E-04 | GO:0010646 | regulation of cell communication | 8.19E-04 |
| GO:0022603 | regulation of anatomical structure | 1.81E-04 | GO:0043687 | post-translational protein modification | 8.58E-04 |
| GO:0050770 | regulation of axonogenesis | 1.98E-04 | GO:0045773 | positive regulation of axon extension | 9.69E-04 |
| GO:0030154 | cell differentiation | 2.03E-04 | GO:0007044 | cell-substrate junction assembly | 9.69E-04 |
| GO:0051782 | negative regulation of cell division | 2.04E-04 | GO:0007420 | brain development | 9.79E-04 |
| GO:0043200 | response to amino acid | 2.08E-04 | | | |

**Figure S17:** Enriched GO terms ($p < 10^{-3}$) in the $339$ DEC up-regulated genes that are only detected in scImpute's imputed data by MAST.

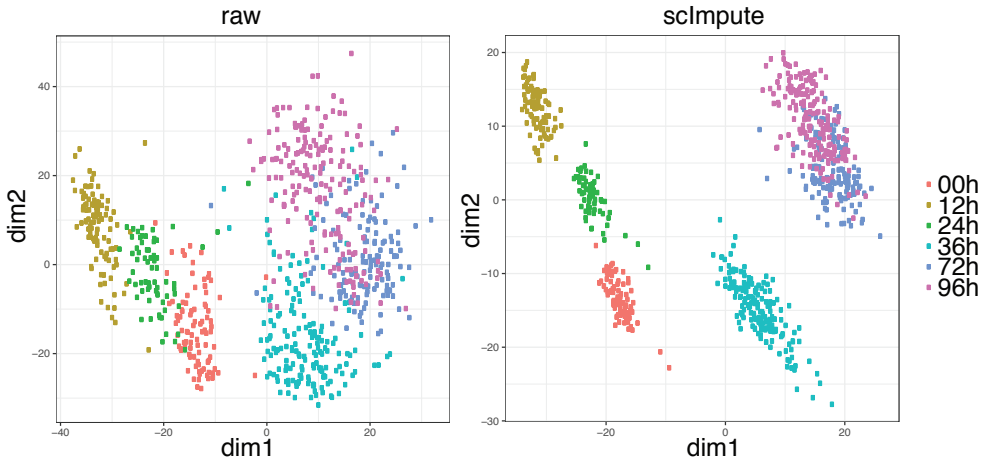


**Figure S18:** The first two dimensions of PCA results calculate from raw and imputed time-course ESC data.

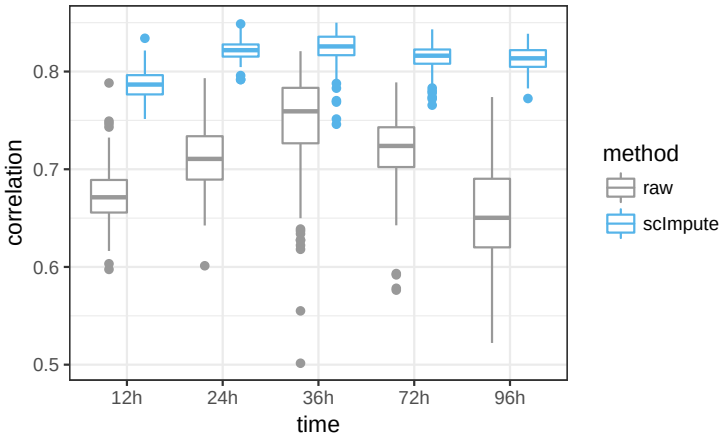


**Figure S19:** Correlation between gene expression in single-cell and bulk data. The Pearson correlation coefficients are calculated between each individual cell and averaged bulk data, at each time point. The correlations based on the imputed data are significantly increased compared with the raw data.

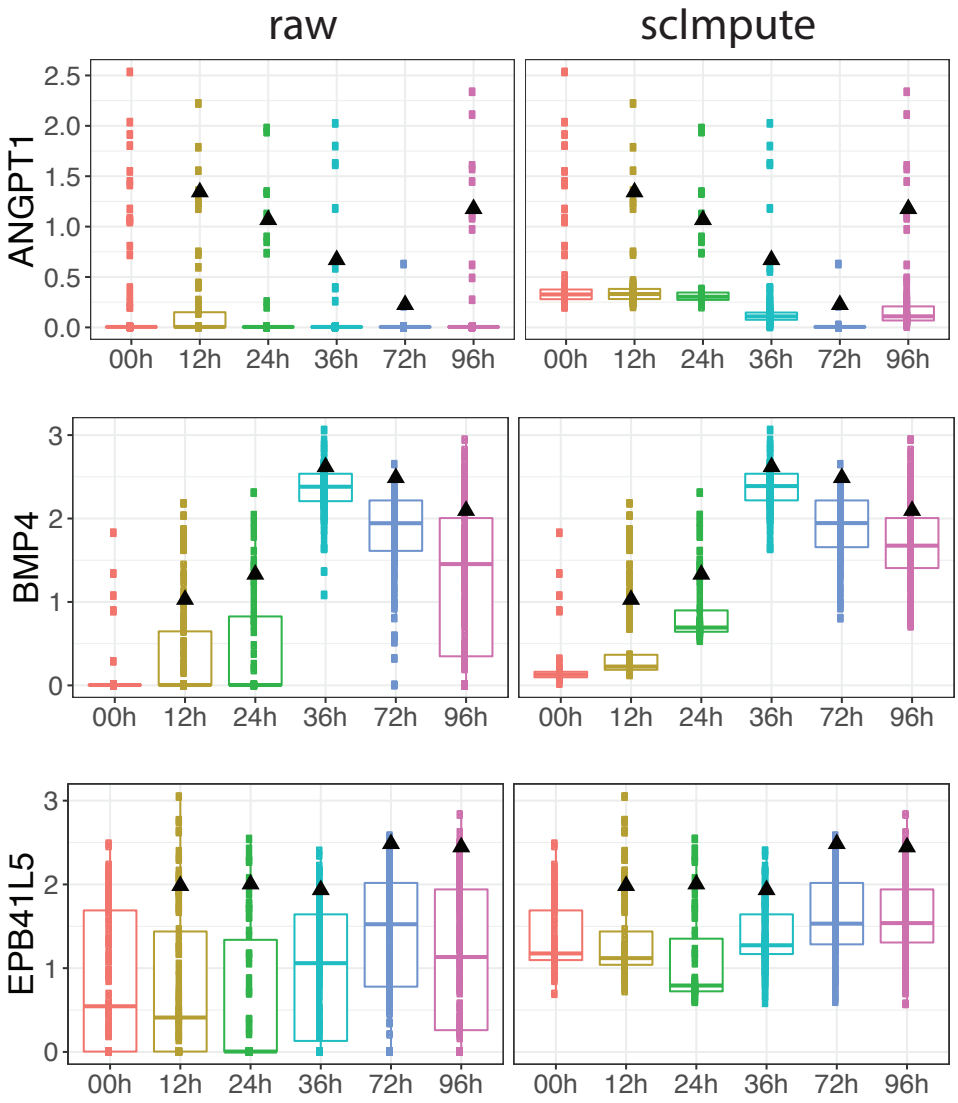


**Figure S20:** Time-course expression patterns of four genes that are annotated with GO term endoderm development.

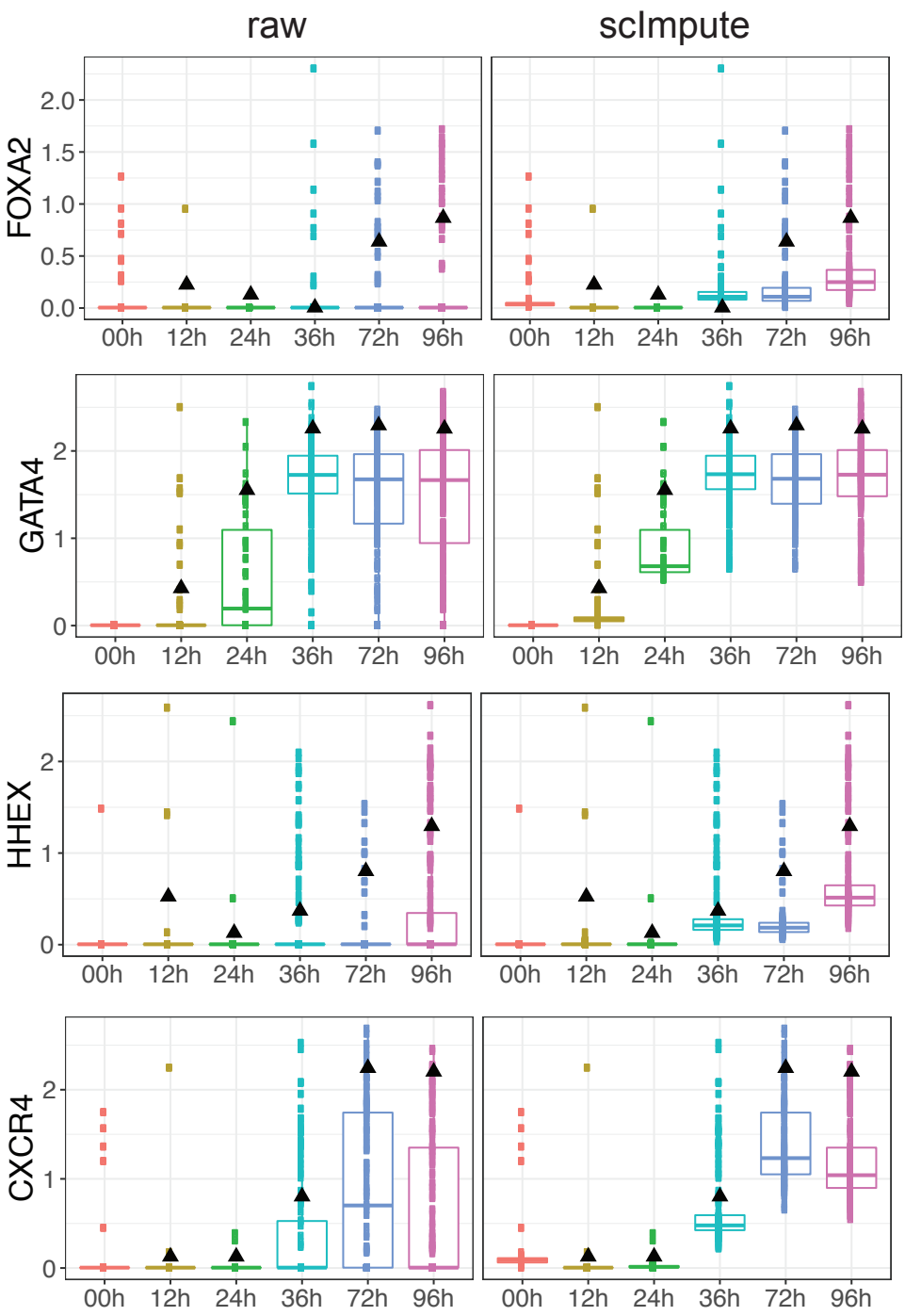


**Figure S21:** Time-course expression patterns of four marker genes of DECs.

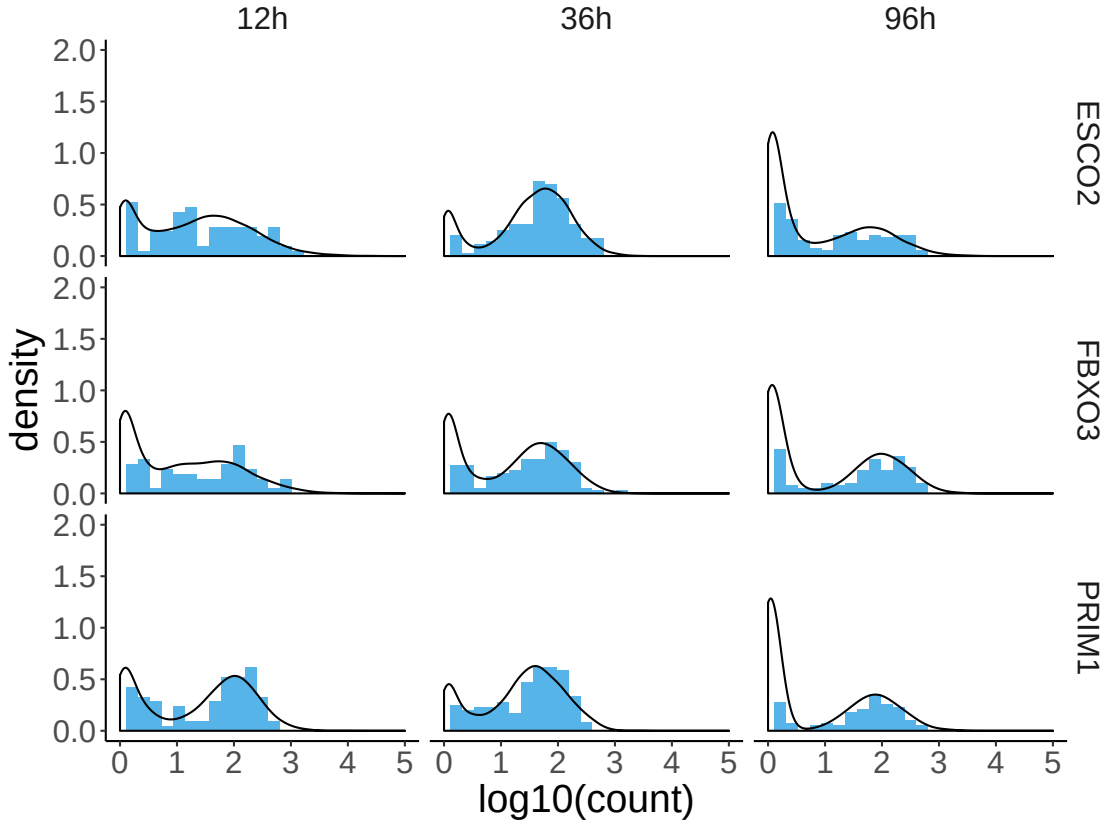


**Figure S22:** Three example genes (*ESCO2, FBXO3*, and *PRIM1*) for comparison of observed and fitted expression distribution in three different cell types. The results are based on the human ESC data. Blue histogram represents observed distribution and black line represents fitted distribution by the Gamma-Gaussian mixture model (4.1).

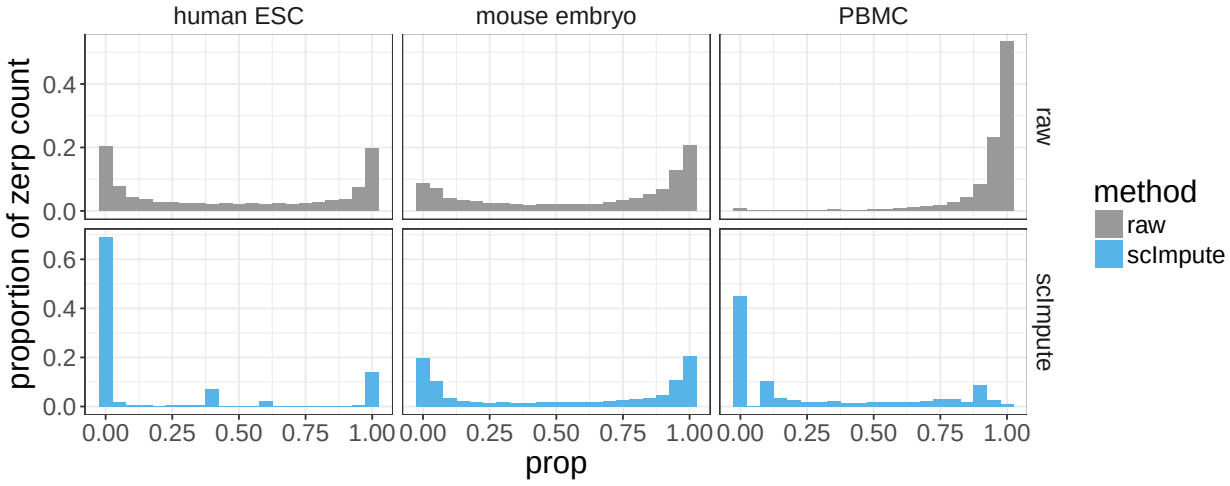


**Figure S23:** Three examples in real data showing the number of genes with various proportions of zero expression values in the raw and imputed data. The proportion of zero values in the imputed data is significantly reduced compared with the raw data.

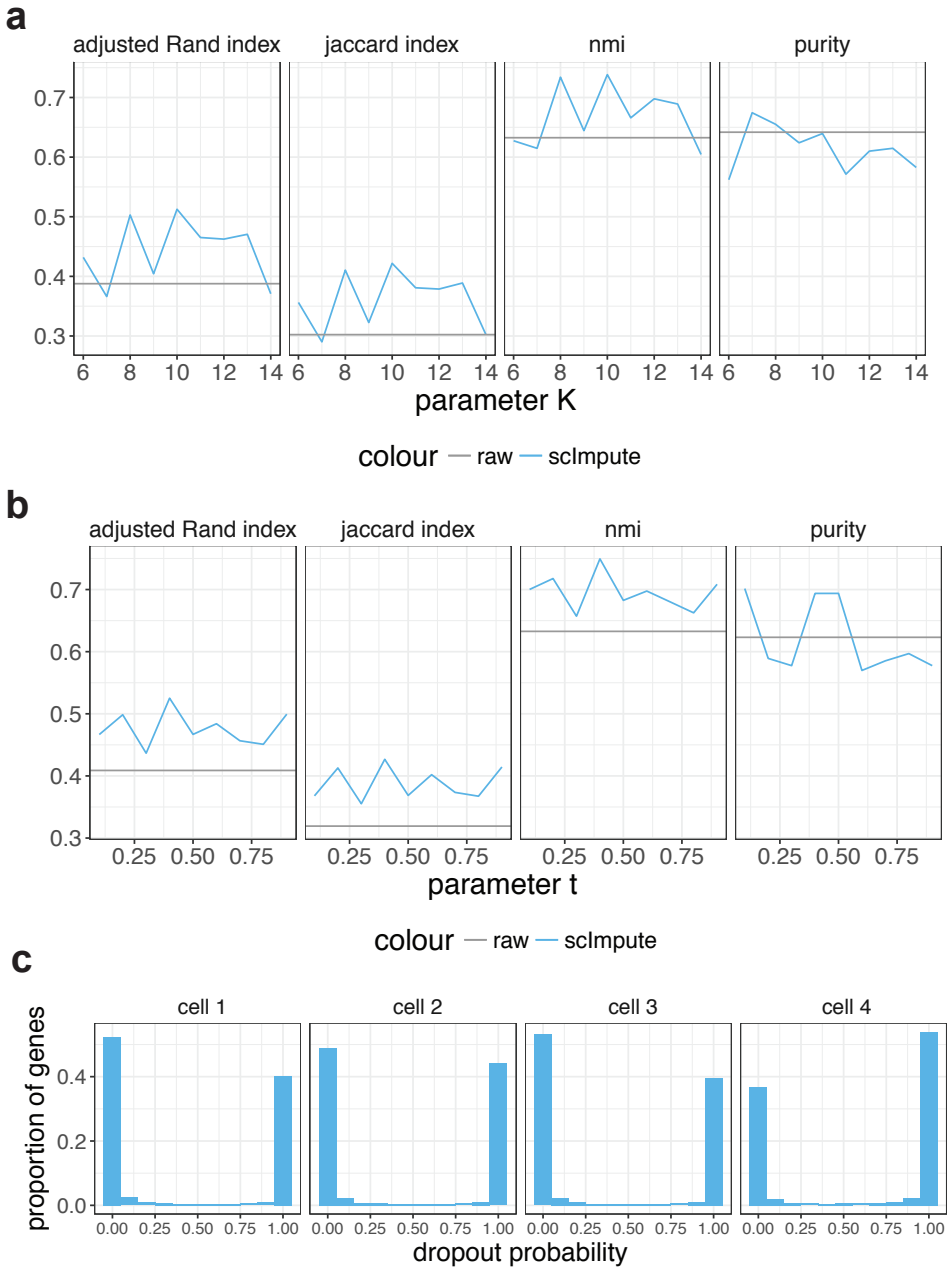


**Figure S24:** Sensitivity analysis based on the mouse embryo data. **a**: Clustering results of imputed data when different values of parameter $K$ are used in scImpute. **b**: Clustering results of imputed data when different values of parameter $t$ are used in scImpute. **c**: The distribution of dropout probabilities in four randomly selected cells from the mouse embryo data. Most genes have dropout probabilities that very close to either $0$ or $1$.