Published online 28 August 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 3 **1**
doi: 10.1093/nargab/lqaa064

Dimensionality reduction for single cell RNA sequencing data using constrained robust non-negative matrix factorization

Shuqin Zhang[1,*], Liu Yang[2], Jinwen Yang[1], Zhixiang Lin[3] and Michael K. Ng[4,*]

[1]School of Mathematical Sciences, Fudan University, Shanghai 200433, China, [2]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, [3]Department of Statistics, Chinese University of Hong Kong, Shatin Hong Kong, China and [4]Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong, China

Received January 20, 2020; Revised August 10, 2020; Editorial Decision August 17, 2020; Accepted August 19, 2020

ABSTRACT

Single cell RNA-sequencing (scRNA-seq) technology, a powerful tool for analyzing the entire transcriptome at single cell level, is receiving increasing research attention. The presence of dropouts is an important characteristic of scRNA-seq data that may affect the performance of downstream analyses, such as dimensionality reduction and clustering. Cells sequenced to lower depths tend to have more dropouts than those sequenced to greater depths. In this study, we aimed to develop a dimensionality reduction method to address both dropouts and the non-negativity constraints in scRNA-seq data. The developed method simultaneously performs dimensionality reduction and dropout imputation under the non-negative matrix factorization (NMF) framework. The dropouts were modeled as a non-negative sparse matrix. Summation of the observed data matrix and dropout matrix was approximated by NMF. To ensure the sparsity pattern was maintained, a weighted ℓ_1 penalty that took into account the dependency of dropouts on the sequencing depth in each cell was imposed. An efficient algorithm was developed to solve the proposed optimization problem. Experiments using both synthetic data and real data showed that dimensionality reduction via the proposed method afforded more robust clustering results compared with those obtained from the existing methods, and that dropout imputation improved the differential expression analysis.

INTRODUCTION

Since single-cell RNA sequencing (scRNA-seq) technology was first reported by Tang *et al.* in 2009 (1), it has received increasing attention. Unlike bulk RNA-seq, scRNA-seq is a powerful tool that can capture the transcriptome-wide cell-to-cell variations (2–4), and thus can be used in a range of areas such as investigation of the cellular heterogeneity within cell populations and complex tissues (5–7), characterization of the organ development of early embryonic cells (8) and exploration of the transcriptomic diversity of human brains (9). Several efficient sequencing protocols, such as Smart-seq, Drop-seq, CEL-seq, SCRB-seq and the commercial device 10× chromium3′ have been developed.

An important characteristic of scRNA-seq data is its high proportion of zero entries, which primarily derive from two different sources. First, a proportion of genes that are not expressed in a single cell give true zero expression entries. Second, mRNA transcripts that are present in low concentrations in a single cell may be lost during the reverse transcription and amplification steps, meaning that these transcripts are undetectable in the following sequencing steps. This is denoted as 'dropout' phenomenon. Such dropout entries cannot be distinguished from true zero expression entries, which makes the analysis of scRNA-seq data more challenging.

Although the traditional methods developed for analysis of bulk RNA-seq data can still be applied to analysis of scRNA-seq data, the performance of these methods is greatly affected by the presence of dropouts. Recently, several statistical models and computational algorithms have been developed for analyzing scRNA-seq data from different perspectives including imputation of the dropouts, differential expression analysis, dimensionality reduction and clustering (10–25). These methods can be divided into two groups, based on whether the dropouts are explicitly modeled. Traditional methods developed for bulk

*To whom correspondence should be addressed. Tel: +86 21 65647484; Fax: +86 21 65646073; Email: zhangs@fudan.edu.cn
Correspondence may also be addressed to Michael K. Ng. Tel: +852 28592260; Fax: +852 25592225; Email: mng@maths.hku.hk

RNA-seq data analysis do not consider the dropout phenomenon. For example, the commonly used dimensionality reduction methods such as Principal Component Analysis (PCA) and *t*-Distributed Stochastic Neighbor Embedding (tSNE) enable easy visualization of the data, but do not consider the existence of missing values (26,27). A recently published dimensionality reduction and visualization method that uses Uniform Manifold Approximation and Projection (UMAP) to consider the nonlinear relations between the cells also does not consider the dropouts (28). Some clustering methods proposed for scRNA-seq data use multiple kernels to capture the relations between the cells on different scales. Although the dropouts are not explicitly considered in the models, good clustering results can be obtained (16,21,24). Several recent methods take the dropouts into account, modeling them as random variables generated from a Bernoulli distribution or as missing values in the data matrix (12–14,17,23,25,29–30). Recently, a dropout imputation method that uses deep learning has been proposed (DeepImpute) (11). Using such methods, the dropouts are typically imputed first, and then the available methods (such as those developed for bulk RNA-seq data) are used for further analysis. A comparison of the current algorithms for imputation is presented in a previous report (25). The existing methods for differential expression analysis primarily consider the dropouts as random variables and are developed on a model-based framework (15,31–32). However, compared with imputation methods, few methods explicitly model the dropouts in dimensionality reduction and clustering (18–19,22). Zero-Inflated Factor Analysis (ZIFA), the first dimensionality reduction method that considers dropouts, is a probabilistic generative model in which the real data are generated from a low-dimension latent space via linear combinations and the zero entries are generated from a Bernoulli distribution. An iterative EM algorithm is then applied for statistical inference. The ZIFA algorithm does not consider the true domain of the counts, because the counts cannot be less than zero (22). Clustering through Imputation and Dimensionality Reduction (CIDR) is an scRNA-seq data dimensionality reduction and clustering method that also considers dropout imputation before the dimensionality reduction step via Principal Coordinate Analysis (PCoA). However, its imputation value is dependent on the pairwise cells, and is not fixed, which means that further analysis using the imputed data is difficult (18). On other work, Lin *et al*. (19) developed a probabilistic clustering model for the joint analysis of scRNA-seq and single cell ATAC-Seq data with consideration of the randomness from dropout and data integration.

In this study, we develop a dimensionality reduction method termed Constrained Robust Non-negative Matrix Factorization (CRNMF) for scRNA-seq data analysis. Non-negative Matrix Factorization (NMF) has been successfully applied in many fields for dimensionality reduction, feature selection and clustering. As scRNA-seq data are non-negative, and have an approximately low-rank structure because of the gene expression similarities between cells of the same type, it is reasonable to develop the model under the NMF framework. Shao and Hofer previously applied NMF to classify the cells and obtained good results, but did not consider dropouts (33). In this study, the dropouts in the data were modeled as a sparse matrix, with the nonzero entries corresponding to imputed values for the dropout events. Due to the variations in sequencing depth across the cells, the proportion of dropouts also varied across different cells. Thus, we used a weighted ℓ_1 penalty to take account for the sequencing variations. We formulated the model as an optimization problem with constraints and developed an efficient algorithm to solve it. Using this model, dropout imputation was implemented simultaneously with dimensionality reduction. The performance of our proposed method was demonstrated by both simulation studies and real data analysis.

MATERIALS AND METHODS

Given the count matrix for n cells and p genes, we normalized the count matrix by the library size of each cell and took $\log(1 + x)$ to obtain the matrix $X_{p\times n}$, where each X_{ij} denoted the expression level of gene i in the jth cell. We then developed our method based on X, as follows.

Given the fact that scRNA-seq data are non-negative and the data matrix is likely of low rank because of the similar gene-expression patterns in the same cell types, NMF is a natural choice to do dimensionality reduction. In the standard NMF method, X is approximated by the product of two non-negative matrices, i.e. $X_{p\times n} \approx W_{p\times r}H_{r\times n}$, $W_{ij} \geq 0$, $H_{ij} \geq 0$ and $r < p$, n is the number of reduced dimensions. W and H can be obtained by multiplicative iterations (34). When a dropout occurs, the corresponding entry becomes zero in the observed X. To recover the value of the dropouts, we define a matrix $S_{p\times n}$, where $S_{ij} > 0$ when the ijth entry is a dropout and $S_{ij} = 0$ otherwise. The data matrix with dropout recovery can then be written as $X + S$, which can be approximated by WH, i.e. $X + S \approx WH$. As the zero entries include both dropouts and true zero expressions, and the exact position of all the dropouts is not known, we can reasonably assume that S is a sparse matrix. We let $\Omega = \{(i, j), X_{ij} = 0, 1 \leq i \leq p, 1 \leq j \leq n\}$, and $\Omega^c = \{(i, j), X_{ij} > 0, 1 \leq i \leq p, 1 \leq j \leq n\}$. The dropout set is then a subset of the index set Ω. Thus, we formulate the optimization model as follows:

$$\begin{aligned} \min_{W,H,S} \ & \frac{1}{2}\|X + S - WH\|_F^2 + \lambda\|S\|_1 \\ s.t. \ & W_{ij} \geq 0, H_{ij} \geq 0, \\ & S_{ij} \begin{cases} \geq 0, \text{if } (i, j) \in \Omega, \\ = 0, \text{ if } (i, j) \in \Omega^c, \end{cases} \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ are the Frobenious norm and elementwise ℓ_1 norm of a matrix, respectively. The ℓ_1 norm penalty is imposed to encourage the sparsity of S, and λ is the parameter to control the sparsity level of S. Figure 1 gives an illustrative example of this method. After applying this method, we can obtain cell expression in a lower dimension (as shown in H), and obtain the estimated dropouts as S with the recovered data matrix being $X + S$.

In the scRNA-seq data, the sequencing depth varies between cells. The cells undergoing deeper sequencing are expected to have fewer dropouts than those undergoing shallower sequencing. Ideally, we should impute more for the

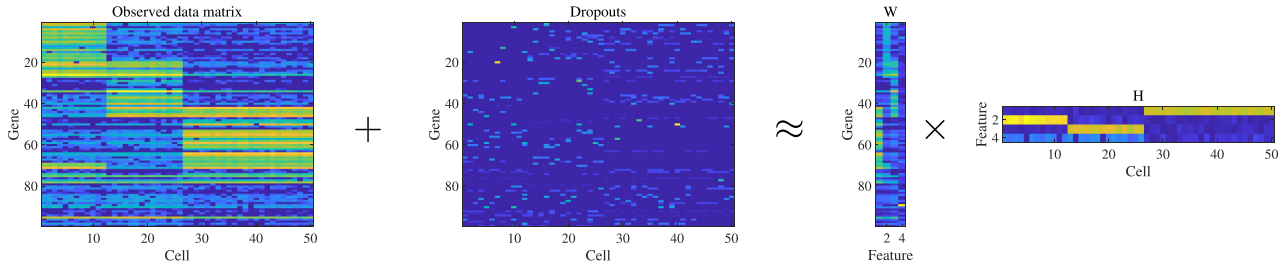


Figure 1. Illustration of the CRNMF model.

cells with low sequencing depths and less for those with greater sequencing depths. To account for the variations in sequencing depth, we propose to add a weight μ_j to each single cell j, which is a parameter related to the sequencing depth, the magnitude of which increases with sequencing depth increasing. We denote $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)$. The model is now presented as:

$$\begin{aligned} \min_{W,H,S} \ & \frac{1}{2}\|X+S-WH\|_F^2 + \lambda \sum_{j=1}^{n} \mu_j \|S_j\|_1 \\ s.t. \ & W_{ij} \geq 0,\ H_{ij} \geq 0, \\ & S_{ij} \begin{cases} \geq 0, \text{ if } (i,j) \in \Omega, \\ = 0, \text{ if } (i,j) \in \Omega^c, \end{cases} \end{aligned} \tag{2}$$

where S_j is the jth column of S. The model (1) is called unweighted Constrained Robust NMF (CRNMF), and the model (2) is called weighted CRNMF.

To solve the optimization problems (1) and (2), we alternatively update W, H and S. When S is fixed, the optimization problem has the traditional NMF framework, and we can iteratively use multiplicative update to obtain W and H. When W and H are fixed, the problem becomes ℓ_1 regularized optimization problem, and S can be solved elementwisely. The algorithm for (2) is summarized in Algorithm 1.

Algorithm 1: Weighted CRNMF

Input: scRNA-seq data matrix: X, parameter: λ, μ, the rank of matrix W, H: r;

Initialization: $S^{(0)} = \mathbf{0}$;

Repeat until convergence

(1) Fix S, given $W^{(0)}$, $H^{(0)}$, solve W, H;
$H_{ij} = H_{ij}(W^T(X+S))_{ij}/((W^TWH)_{ij} + 10^{-9})$;
$W_{ij} = W_{ij}((X+S)H^T)_{ij}/((WHH^T)_{ij} + 10^{-9})$;
Repeat until convergence

(2) Fix W, H, solve S;
$S_{ij} = (Soft_{\lambda\mu_j}((WH-X)_{ij}))_+, (i,j) \in \Omega$,
$S_{ij} = 0, (i,j) \in \Omega^c$,
where
$Soft_y(x) := sgn(x)(|x| - y)_+, (x)_+ = \max(x, 0)$.

Output: W, H, S.

Having obtained W, H and S, we can consider H as the r-dimensional representation of the cells and then use k-means or other traditional methods to perform clustering. We can also perform dropout imputation by adding S to X, and considering $X + S$ as the observed data for further analysis, including differential expression analysis.

The parameters $\boldsymbol{\mu}$, r and λ in the model (2) must be prespecified. Each μ_j should depend on the sequencing depth of the jth cell. We let $\mu_j = \frac{\sum_{i=1}^{p} X_{ij}}{\text{median}\{\sum_{i=1}^{p} X_{ij}, \forall j\}}$. If the value $\sum_{i=1}^{p} X_{ij}$ for cell j is the median of $\{\sum_{i=1}^{p} X_{ij}, \forall j\}$, $\mu_j = 1$ and with the increase of sequencing depth, the value of μ_j increases. This is in consistent with the fact that the more deeply sequenced cells are expected to have fewer dropouts. For unweighted CRNMF, each μ_j is 1. For the parameter r, as the rank of WH should be r and WH approximates X, a good choice of r is the rank of X. Although a cross-validation-based method can be applied to choose r, for simplicity, we estimate r from the singular values of X (35). The small singular values should correspond to the noise in X, thus, we choose r as the number of singular values of proper scale. Suppose the singular values of X are: $d_1 \geq d_2 \geq \cdots \geq d_n$, then we set the default value of r as: $r = \min_i\{|\frac{d_i - d_{i-1}}{d_{i-1} - d_{i-2}}| < \delta, |\frac{d_{i-1} - d_{i-2}}{d_{i-2} - d_{i-3}}| < \delta\}$. Here δ is set to be chosen by the users, and should be close to 1. This ensures that the singular values less than r should be little-changed, and thus can be considered as noise. In our analysis, we set δ to be 1.05. As we have used μ_j to scale the depth differences between different cells, a stable region was obtained for λ, which gave similar results. Our experiments showed that a good choice for the value of λ was $\sim$1.5. A more accurate λ can be chosen based on the clustering results using cross validation (36). If $(WH)_{ij} - X_{ij} > \lambda\mu_j$, we impute the entry (i,j) with $(WH)_{ij} - X_{ij} - \lambda\mu_j$; otherwise, we simply consider it to be the true zero expression. By choosing λ, we provide a threshold that differentiates between the dropouts and the true zero expressions. To reduce the effect of initialization in the proposed method, we consider the initial value of W, H based on singular value decomposition (SVD), as described in a previous study (37).

RESULTS

To evaluate the performance of our proposed method, we first compared the performance of the weighted CRNMF with that of three existing dimensionality reduction methods and then with the performance of the unweighted CRNMF and original NMF.

For the first comparison, the following three dimensionality reduction methods were selected: PCA, CIDR (18)

and Single-cell Interpretation via Multi-kernel LearRning (SIMLR) (24). PCA is a commonly used dimensionality reduction method that does not consider dropouts. CIDR performs dimensionality reduction with dropout imputation, wherein the imputation of dropouts depends on the pairwise distances between each cell pair, and they are not fixed. Thus it is difficult to use the imputed dropouts in the downstream analysis. CIDR has shown better dimensionality reduction performance than ZIFA (22), which was not considered in our study. SIMLR integrates different kernel-based similarities to visualize the cells without considering the dropouts (24). It also performs dimensionality reduction before visualization and clustering. We applied all of these four methods to both simulated datasets and real biological datasets for dimensionality reduction. To illustrate their performance, we performed k-mean clustering. The results were judged according to three criteria: Adjusted Rand Index (ARI), ACCuracy (ACC) and Normalized Mutual Information (NMI). For CIDR and SIMLR, we ran the R packages `cidr` and `SIMLR`, to obtain a lower-dimensional representation of the cells. To see the performance of dropout imputation, we compared CRNMF with DeepImpute, a recently developed dropout imputation method that uses deep learning. DeepImpute has shown more stable results than the existing methods (11). We used the package `DeepImpute` from Github for this comparison and for simplicity used 'CRNMF' to denote the weighted CRNMF.

Simulation study

We first evaluated the performance of CRNMF using synthetic datasets. To generate the scRNA-seq data, we directly used the R package: `splatter`, which simulates scRNA-seq data and demonstrates good consistency with the real datasets (38). We simulated three settings, depending on the dropout types. The parameters were set as follows: nGenes=10 000, nCells=500, group.prob=(0.05, 0.10, 0.25, 0.60), mean.shape=2, mean.rate=0.3, de.prob=(0.1, 0.1, 0.05, 0.05), de.facLoc=(0.5, 1, 0, 1.5) and de.downProb=(0.3, 0.3, 0.5, 0.5). The dropout types include 'experiment', 'cell' and 'batch', wherein 'experiment' is the global dropout and uses the same parameters for every cell; 'cell' uses a different set of parameters for each cell; and 'batch' uses the same parameters for every cell in each batch. For each dropout type, we varied the parameters describing the distribution of dropouts and applied three different parameter settings. Table 1 shows the dropout parameters. In the dropout type 'cell', '1.00 : 2.00' indicated that the dropout.mid was from 1.00 to 2.00 with an equal step size for 500 cells, and '(1.00, 2.00)' in the dropout type 'batch' indicated that the dropout.mid was 1.00 and 2.00 in the two batches. Other parameters were set as default. We also added the ratio of zeros in each dataset in the table.

We first show the results for dimensionality reduction. The number of cell types was set to be the true number of clusters. For CRNMF, CIDR and SIMLR, the parameters were set as default. The number of reduced dimensionality value of PCA was set to be the same as that of CRNMF, which was obtained from SVD. Figure 2 shows the clustering results. CRNMF performed quite well when the ratio of zeros was not very high. In the setting 'experiment', CRNMF performed the best among all the three considered settings. In the setting 'cell', CRNMF performed the best in the first two cases, whereas SIMLR performed the best in the third case. We checked the detailed results of CRNMF. The selected parameter is $r = 5$ using the default setting. When we set the parameter as 6, the values of ARI, ACC, and NMI were close to 1. Figure 3A shows the dimensionality reduction results of CRNMF when $r = 6$. The pattern of the sixth feature in Figure 3A disappeared when the parameter r was 5 because of the small size of the

Table 1. Summary of the dropout parameters in the simulated datasets

dropout.type	Setting	dropout.mid	dropout.shape	Rate of zeros
Experiment	1	1.20	−1.00	76.24%
	2	1.80		82.11%
	3	2.40		86.97%
Cell	1	1.00 : 2.00	−0.60: −1.20	77.84%
	2	1.25 : 2.25		78.80%
	3	1.50 : 2.50		79.47%
Batch	1	(1.20, 1.50)	(− 0.60, −1.20)	78.88%
	2	(1.00, 2.00)	(− 0.60, −1.00)	81.04%
	3	(1.00, 2.00)	(− 0.60, −1.20)	83.01%

In the dropout type 'cell', '1.00 : 2.00' means the dropout.mid is from 1.00 to 2.00 with equal step size for 500 cells, and '(1.00, 2.00)' in the dropout type 'batch' means the dropout.mid is 1.00 and 2.00 in the two batches.

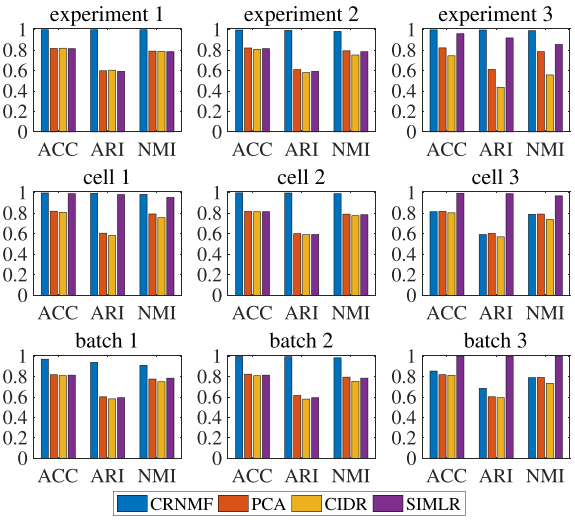
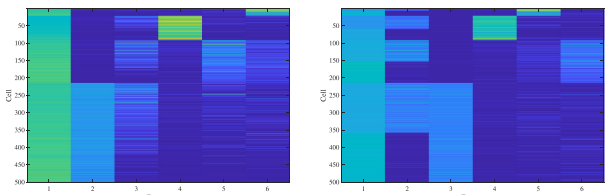


Figure 2. Clustering results for the cells after dimension reduction in the simulated datasets.



(a) Cell 3 (b) Batch 3

Figure 3. Dimension reduction for the simulated datasets 'cell 3' and 'batch 3'.

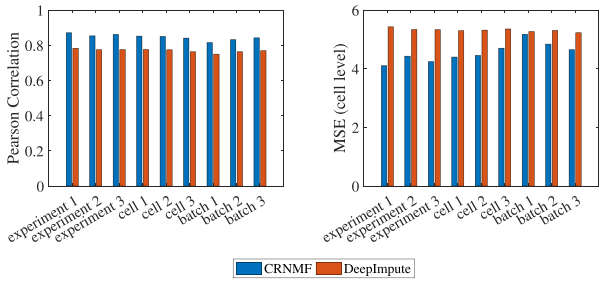


Figure 4. Correlation and MSE between the imputed data matrix $X + S$ and the true data matrix in the simulation study.

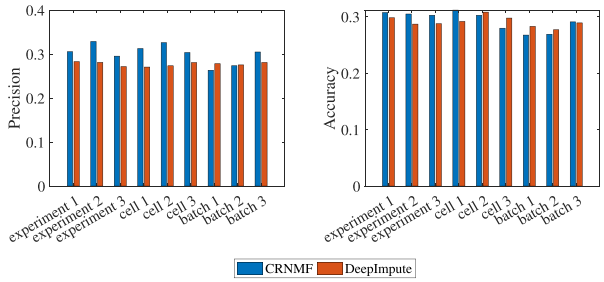


Figure 5. Precision and accuracy for DE gene detection in the simulated datasets after dropout imputation.

first group, and thus the clustering results were not good enough. In this case, it performs similar to PCA and CIDR. In the dropout type 'batch', CRNMF performed similar to that in the dropout type 'cell', i.e. it performed the best in the first two settings, and SIMLR performed the best in the third setting. We further checked the dimensionality reduction results for CRNMF. Figure 3B shows the transposition of H after dimension reduction. The features from 3 to 6 clearly show the cluster pattern, whereas the second feature explains the batch effect of the dataset. If this column is removed in the clustering step, ARI, ACC and NMI will have values very close to 1. In this case, if we can know the cell numbers of each batch in advance, we may first remove the batch effect and then perform the clustering.

To assess the performance of CRNMF in dropout imputation, we first computed the correlation and mean squared error (MSE) between the imputed data matrix $X + S$ and the true data matrix, which was recorded in the simulation. For comparison, we performed dropout imputation using DeepImpute and computed these values as well. The results in Figure 4 show that CRNMF provided higher correlations and lower MSEs. We then performed differential gene expression analysis for the data with imputation of the possible dropouts, i.e. $X + S$, using `limma` (39,40). Parameters of all algorithms in `limma` were set to the default values. We calculated both precision and accuracy to measure the results as we knew the true differentially expressed (DE) genes in the simulations. To compute the precision, the genes with adjusted P-values less than 10^{-3} were considered to be DE genes. Precision was defined as the ratio between the detected true DE genes and the total number of detected DE genes. In addition, we selected the genes with the smallest adjusted P-values and the number of selected DE genes was the same as that of the true DE genes. The detection accuracy was defined as the ratio between the detected true DE genes and the total number of true DE genes. The results in Figure 5 show that CRNMF generated similar results to those generated by DeepImpute. The false positive rate of DE gene detection is also calculated and shown in the Supplementary Materials.

Real data analysis

We applied weighted CRNMF and other three methods including PCA, CIDR and SIMLR to analysis of eight real biological datasets, the cell types of which have been described in their respective original publications. The key information of these datasets is summarized in Table 2. After dimensionality reduction using the above mentioned four methods, we applied k-means to perform clustering. For the evaluation and fair comparison, we tuned the parameter corresponding to the number of reduced feature dimensionality in all four methods to obtain the maximum ARI. Other parameters were set as default. The number of cell types was known beforehand. In the following experiments, each data matrix was normalized by the library size of each cell and was taken $\log(1 + x)$ to obtain the matrix $X_{p \times n}$. We also conducted the experiments with data matrix normalization using deconvolution and sctransform (41,42). The results are put in the Supplementary Materials.

Human brain scRNA-seq dataset ('Darmanis'). We used the same dataset as used by Lin *et al.* (18), which comprises 420 cells with eight cell types (9,18). The methods CRNMF, PCA, CIDR and SIMLR were applied for dimensionality reduction, and all the cells were visualized in the lower-dimensional space using UMAP. The visualization results are shown in Figure 6A, and the clustering results are shown in Figure 6B. CIDR and CRNMF performed similarly well in both visualization and clustering, with the margins between different cell types being large, aside from a few wrongly assigned cells. When PCA was applied, the resulting separation of different cell types was poor. In the case of SIMLR, although the margins between different cell types were large, some cell types were mixed.

In the real datasets, the underlying number of true dropouts were unknown, thus it was not easy to compare the performance of dropout imputation. Therefore, we performed DE analysis of the dataset after dropout imputation and clustered the cells based only on the detected DE genes. Better detection of DE genes was expected to provide better clusters. We directly applied the R package `limma` to the imputed data $X + S$ using CRNMF and DeepImpute, and set the adjusted P-value 10^{-2} as the cut-off for DE-gene selection. Figure 7A shows the visualization of the cells with all DE genes using UMAP. The DE genes detected from the data imputed using CRNMF had much better separated cell clusters, indicating that dropout imputation using CRNMF improved DE gene detection.

Table 2. Summary of the eight real datasets

Reference	Protocol	Data size	Rate of zeros	Cell types
Darmanis *et al.* (9)	SMARTer	22 085 × 420	81.40%	8
Deng *et al.* (43)	Smart-Seq(2)	22 431 × 268	60.29%	6
Segerstolpe *et al.* (44)	Smart-Seq	25 525 × 1099	73.35%	9
Klein *et al.* (45)	inDrop	24 175 × 2717	65.76%	4
Baron *et al.* (46)	inDrop	20 125 × 1937	90.44%	14
Tabula Muris-Tongue (47)	10×	23 433 × 3101	87.53%	11
Tabula Muris-Limb (47)	10×	23 433 × 4536	93.42%	15
Butler *et al.* (48)	10×	32 738 × 2638	97.40%	8

Mouse preimplantation embryos scRNA-seq dataset ('Deng'). This dataset comprises 268 cells of six cell types (43) corresponding to different stages of mouse preimplantation development. After applying CRNMF, we obtained the lower-dimensional representation of cells H and the possible dropouts S. Figure 6B shows the clustering results after dimensionality reduction using the four methods. CRNMF demonstrated better clustering results than SIMLR in terms of ARI and NMI and similar results in terms of ACC, all of which were much better than those obtained by the other two methods. Visualization of the cells with lower-dimensional representation is shown in Figure 6A. There were overlaps between the cell types '8cell' and '16cell' in all methods. In PCA and SIMLR, there were more overlaps between the cell types 'blast' and '16cell'.

We performed DE analysis of this dataset after dropout imputation using CRNMF and DeepImpute, similar to the analysis performed on the previous dataset. A visualization of the cells composed of all DE genes is shown in Figure 7B. The figure shows that the DE genes detected in the dataset after dropout imputation using CRNMF provide better separation of the cells. This result also indicates that dropout imputation using CRNMF helps DE gene detection.

Human pancreatic islets in type 2 diabetes ('Segerstolpe'). Next, we evaluated human pancreatic islet cells exhibiting type 2 diabetes using our method (44). Before the analysis, we filtered the low-quality cells and cell clusters of fewer than 10 cells, yielding 1099 cells of 9 cell types. The clustering results using the four methods are shown in Figure 6B. CRNMF gave the best clustering results.

DE analysis was also conducted for this dataset after imputation using CRNMF and DeepImpute. Figure 7C shows the visualization of the cells with all of the selected DE genes. Some cell types were divided into different types when imputing the dropouts using DeepImpute. The DE genes selected from the dataset imputed using CRNMF showed much better separation of cells. This result also indicates that dropout imputation using CRNMF provided better DE gene detection.

Mouse embryonic stem cell scRNA-seq dataset ('Klein'). This dataset comprises 2717 mouse embryonic stem cells in four cell types (45) corresponding to the days after leukemia inhibitory factor withdrawal. Figure 6A shows the visualization of the cells after dimensionality reduction, and Figure 6B shows the clustering results, where CRNMF performed the best, followed by PCA, consistent with the visualization results in Figure 6A.

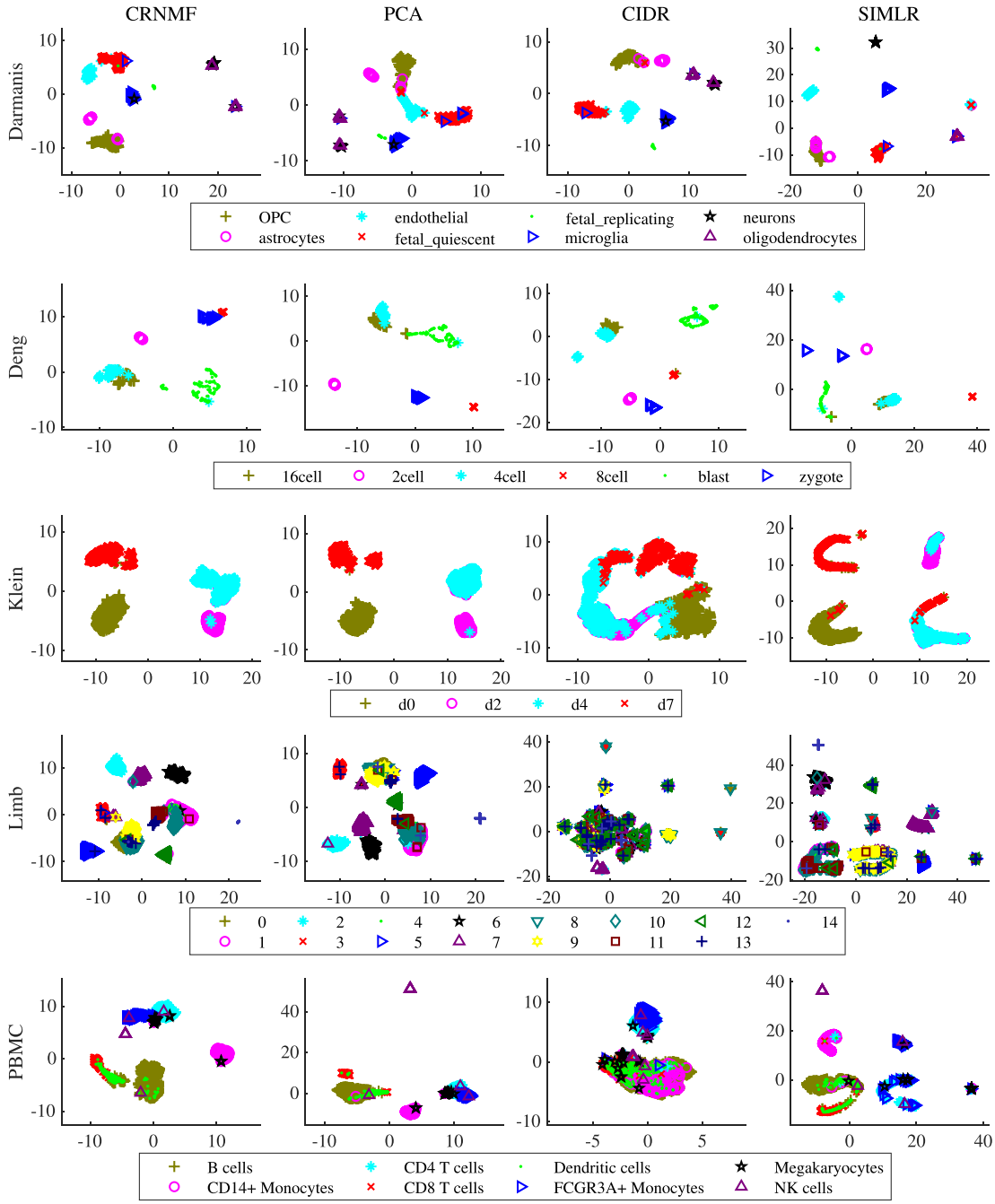
To evaluate the effect of dropout imputation on this dataset, we performed DE analysis after dropout imputation. Figure 7D shows the visualization of the cells with all of the selected DE genes. The DE genes selected after dropout imputation using CRNMF provided a very good cluster structure. The DE genes with dropout imputation using DeepImpute provided some mixed clusters.

Human pancreatic islets scRNA-seq dataset ('Baron'). Of the four donors included in this study, the scRNA-seq data of donor 1 was selected, which comprises 1937 cells in 14 cell types (46). The cells were sequenced using inDrop, which revealed that ∼90.00% entries in the data matrix were equal to zero. Figure 6B shows the comparison of clustering results after dimensionality reduction using the four methods. The results show that CRNMF performed similarly to PCA and had the highest NMI of the four methods. SIMLR showed the highest ARI and ACC in this dataset.

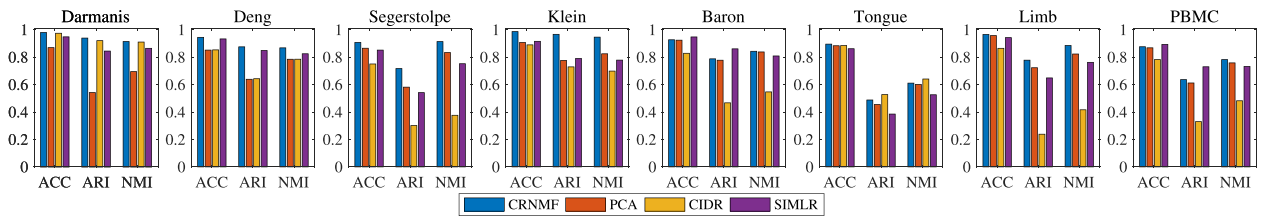
In the DE analysis, the DE genes detected from the dataset after imputation using CRNMF and DeepImpute showed similar clustering results. The visualization is shown in Figure 7E, wherein CRNMF gave a little better separation between the clusters.

Mouse organs' scRNA-seq from Tabula Muris ('Tongue' and 'Limb'). Two scRNA-seq datasets for tongue and limb generated using 10× from Tabula Muris (47) were downloaded. For tongue, we chose the batch 'Tongue-10X_P4_0', which comprises 3101 cells in 11 cell types, and for limb, 4536 cells from two batches were combined together for analysis. Both datasets have very high proportion of zero entries. Figure 6B shows the clustering results after dimensionality reduction. CRNMF performed similarly to PCA and CIDR in 'Tongue', while it performed much better than the other three methods in 'Limb'. Visualization of the cells from 'Limb' in the lower-dimensional space also revealed that CRNMF performed much better than the other three methods.

In the DE analysis, the DE genes detected from the 'Tongue' after imputation using CRNMF and DeepImpute showed similar clustering results. The visualization is shown in Figure 7F, wherein no clearly visible differences can be observed between the clusters. The DE genes detected from the 'Limb' after imputation using CRNMF gave much better cell type separation than those using DeepImpute. The visualization is shown in Figure 7G.



(a)



(b)

Figure 6. Comparison of the dimensionality reduction results: (**A**) UMAP visualization of the cells after dimensionality reduction using the corresponding methods; (**B**) clustering results using *k*-means for cells represented in a lower dimension.

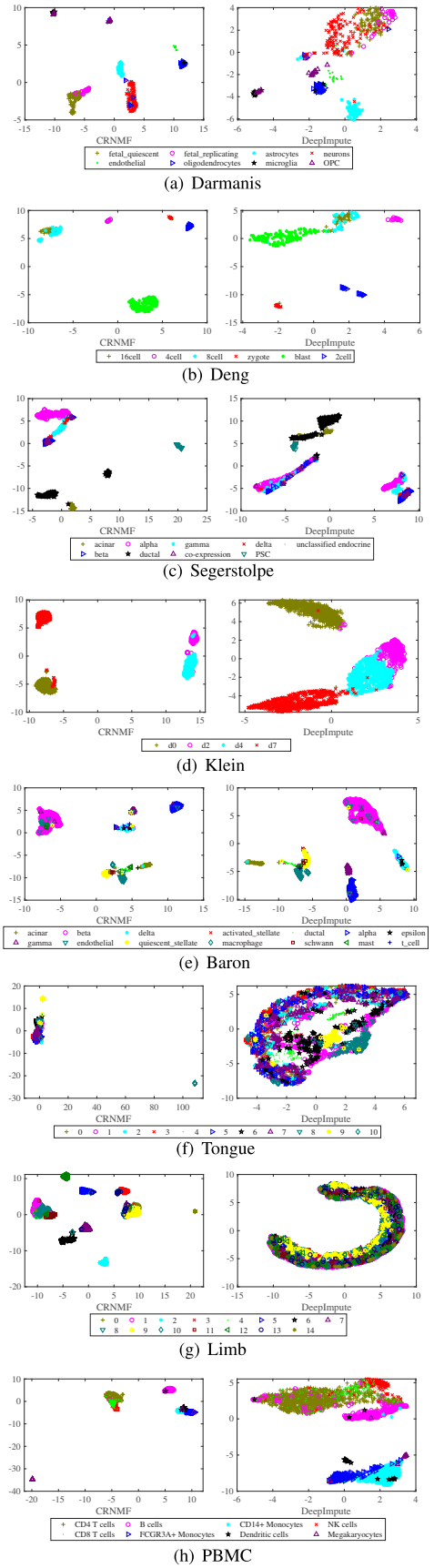


Figure 7. Visualization of the cells with all the DE genes detected in the dataset after dropout imputation using CRNMF and DeepImpute.

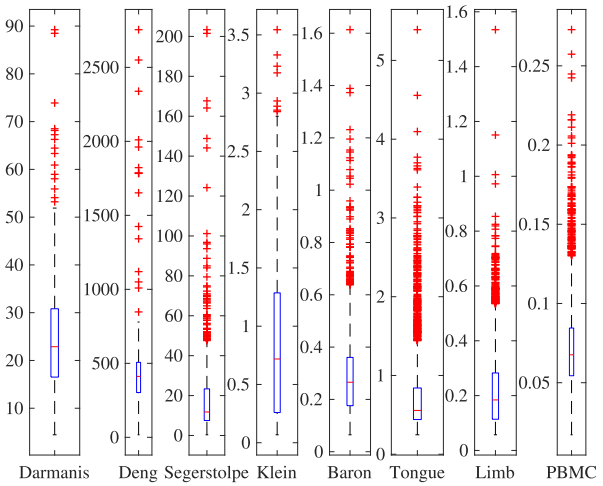


Figure 8. Sequencing depth of the eight real datasets.

Human peripheral blood mononuclear cells from 10X Genomics ('PBMC'). We downloaded a dataset of Peripheral Blood Mononuclear Cells (PBMC) studied in (48). This dataset includes 2638 cells from 8 cell types. After filtering the genes that have zero expression in all the cells, there are a total of 16 579 genes. Figure 6B shows the clustering results after dimensionality reduction. CRNMF achieved the highest NMI, and SIMLR achieved the highest ARI and ACC. The visualization of the cells in the 2D space shows that CRNMF gave better distributed clusters.

A visualization of the cells composed of all the selected DE genes in the two-dimensional space is shown in Figure 7H. Though some clusters were mixed together in both figures, it is clear that dropout imputation using CRNMF gave much better separation of the cell types.

Evaluation of NMF-based methods

Next, we compared the performance of weighted CRNMF with that of unweighted CRNMF and original NMF using the abovementioned eight real datasets to identify the effects of imputing the dropouts and adding the weights into the model.

To determine how the sequencing depth affects dimensionality reduction and clustering for NMF-based methods, we first plotted the average sequencing depth for each dataset in Figure 8. For each dataset, the sum of the frequency of all genes was divided by the total number of genes to obtain an approximate measure of the sequencing depth. Among all of the datasets, the sequencing depth of 'Deng' was the greatest, having the largest variation, followed by that of 'Segerstolpe'.

We further applied weighted CRNMF, unweighted CRNMF and original NMF to the eight real datasets to perform dimensionality reduction and compared their k-means clustering results. We set the same stopping criterion and the same rank of W and H, which was chosen using our proposed method. After following the same procedure, we obtained the clustering results as shown in Figure 9. These results indicate that when the sequencing depth had

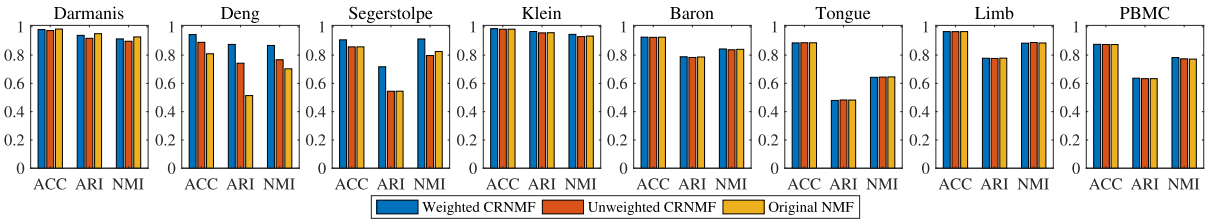


Figure 9. Comparison of clustering results comparison of NMF-based methods in the eight real datasets.

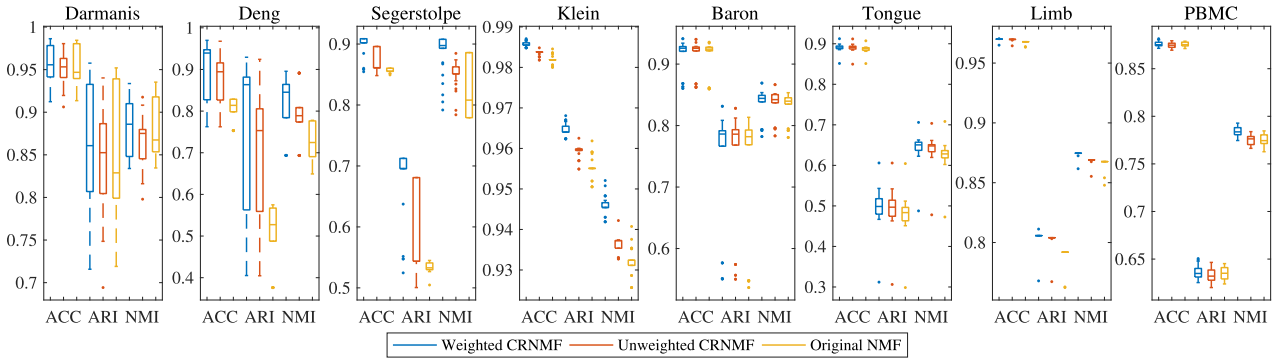


Figure 10. Clustering results of NMF-based methods by downsampling 80% of the cells 30 times in the eight real datasets.

large variation, weighted CRNMF performed much better than original NMF and unweighted NMF. In both 'Deng' and 'Segerstolpe', weighted CRNMF was superior in handling deep sequenced data. In the case of shallow sequenced cells, these three methods performed similarly in terms of clustering, while CRNMF also simultaneously imputed the dropouts.

To further evaluate the stability of these three methods, we downsampled the cells at 80% 30 times, and performed the same procedure as mentioned above to compare the clustering results. The results in Figure 10 show that weighted CRNMF provided the best results for all datasets and that it exhibited more advantages than the other two methods with the increased variation in the sequencing depth. For example, using weighted CRNMF, unweighted CRNMF and original NMF, the median of ARI was 0.864, 0.753 and 0.527 in 'Deng', whereas it was 0.964, 0.960 and 0.955 in 'Klein', respectively. Although the ARIs were closer in 'Baron' using all of the three methods, the results using weighted CRNMF were a little better than those of the other two methods.

The computational complexity of CRNMF depends on the number of cell n, genes p and the reduced dimensionality r. Usually the number of genes varies little, and r is not large, thus the computational time variation across different datasets mainly depends on n. We implemented the program in a MacBook Pro with 3.1GHz double-core Intel Core i5 processor and 8GB 2133 MHz LPDDR3 memory to evaluate the computational time of CRNMF. Given r in the model, computation of dataset 'Deng' takes 11 seconds only, while it takes 578 s for the dataset 'Limb'. The correlation between the computational time and the cell number n is 0.92. This shows that CRNMF can be applicable in real data analysis.

CONCLUSION AND DISCUSSION

The dropout phenomenon creates more challenges for dimensionality reduction in scRNA-seq data analysis. Several methods that consider the dropouts in dimensionality reduction have been proposed. However, as our experiments on different datasets indicated, these methods are not robust, and their performance varies across datasets. Although many reasons may exist for this finding, we considered two possible reasons. First, the dropout imputation usually does not consider the cell-to-cell variation in sequencing depth. Deeper sequencing could lead to fewer dropouts, whereas shallower sequencing could result in more dropouts. Second, the non-negativity of the count data is not considered.

In this study, we developed a dimensionality reduction model based on NMF. We modeled the dropouts using a sparse matrix, and the sparsity was ensured using a weighted ℓ_1 norm, which was related to the sequencing depth. By adding weights, the tuning parameter controlling the sparsity of the dropout matrix became very stable. The rank of the matrices in NMF could be chosen according to the singular values of the observed data matrix, which was also very stable. Extensive data analysis showed that more robust clustering results were achieved after dimensionality reduction using CRNMF. In particular, in the real data clustering, CRNMF achieved the best results for five of the eight datasets based on three criteria and also achieved the highest NMI in the last three datasets. In addition to performing dimensionality reduction, CRNMF simultaneously imputes the dropouts. Comparison with the updated dropout imputation method showed that CRNMF enables better DE gene analysis.

In the scRNA-seq data analysis, one important yet difficult problem is rare cell type detection. When the features of

the rare cell types are not highly distinguishable, the current CRNMF may not be able to detect them. This is because adding such information in the low rank approximation may not reduce much of the objective function. Besides the dimensionality reduction step, detection of rare cell types also depends on the clustering methods. Due to the small sample size of rare cell types, the commonly used clustering methods, such as *k*-means, may not be able to group these cells together. To improve rare cell type detection, we will add the known marker gene information of each cell type in the model to increase the differentiation between different cell types. This is left as one of our future works.

As shown in the data analysis, one limitation of the CRNMF method is that it may consider the strong batch effects as the features in dimensionality reduction. How to model or learn the batch effect before implementing our algorithm is a challenge for future work.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

S. Zhang's research is supported in part by National Natural Science Foundation of China grant No. [11471082], Science and Technology Commission of Shanghai Municipality grant No. [20ZR1407700]. L. Yang's research is supported in part by National Natural Science Foundation of China grant No. [61702358], and in part by the Tianjin Science and Technology Plan Project grant No. [19ZXZNGX00050, 19ZXZNGX00050]. L. Lin is supported by the Chinese University of Hong Kong direct grants No. [4053360] and No. [4053423], the Chinese University of Hong Kong startup grant No. [4930181], and Hong Kong Research Grant Council Grant ECS No. [CUHK 24301419]. M. Ng's research supported in part by the Hong Kong Research Grant Council Grant GRF [12306616, 12200317, 12300218 and 12300519].

Conflict of interest statement. None declared.

REFERENCES

1. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C.C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A. *et al.* , (2009) mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.
2. Kolodziejczyk,A.A., Kim,J.K., Svensson,V., Marioni,J.C. and Teichmann,S.A. (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.
3. Saliba,A., Westermann,A.J., Gorski,S.A. and Vogel,J. (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, **42**, 8845–8860.
4. Vallejos,C.A., Marioni,J.C. and Richardson,S. (2015) BASiCS: Bayesian analysis of Single-Cell sequencing data. *PLOS Comput. Biol.*, **11**, e1004333.
5. Kelsey,G., Stegle,O. and Reik,W. (2017) Single-cell epigenomics: Recording the past and predicting the future. *Science*, **358**, 69–75.
6. Liu,S. and Trapnell,C. (2016) Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res.*, **5**, 182.
7. Stubbington,M.J.T., Rozenblattrosen,O., Regev,A. and Teichmann,S.A. (2017) Single-cell transcriptomics to explore the immune system in health and disease. *Science*, **358**, 58–63.
8. Moignard,V., Woodhouse,S., Haghverdi,L., Lilly,A.J., Tanaka,Y., Wilkinson,A.C., Buettner,F., Macaulay,I.C., Jawaid,W., Diamanti,E. *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, **33**, 269–276.
9. Darmanis,S., Sloan,S.A., Zhang,Y., Enge,M., Caneda,C., Shuer,L.M., Gephart,M.H., Barres,B.A. and Quake,S.R. (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7285–7290.
10. Andrews,T.S. and Hemberg,M. (2018) False signals induced by single-cell imputation. *F1000Res.*, **7**, 1740.
11. Arisdakessian,C., Poirion,O., Yunits,B., Zhu,X. and Garmire,L. X. (2019) DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.*, **20**, 211.
12. Chen,M. and Zhou,X. (2018) VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.*, **19**, 196.
13. Chen,C., Wu,C., Wu,L., Wang,X., Deng,M. and Xi,R. (2020) scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition. *Bioinformatics*, **36**, 3156–3161.
14. Huang,M., Wang,J., Torre,E.A., Dueck,H., Shaffer,S.M., Bonasio,R., Murray,J.I., Raj,A., Li,M. and Zhang,N.R. (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.
15. Jia,C., Hu,Y., Kelly,D., Kim,J., Li,M. and Zhang,N.R. (2017) Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.*, **45**, 10978–10988.
16. Kiselev,V.Y., Kirschner,K., Schaub,M.T., Andrews,T.S., Yiu,A., Chandra,T., Natarajan,K.N., Reik,W., Barahona,M., Green,A.R. *et al.* , (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.
17. Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
18. Lin,P., Troup,M. and Ho,J. W.K. (2017) CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.*, **18**, 59.
19. Lin,Z., Zamanighomi,M., Daley,T., Ma,S. and Wong,W.H. (2020) Model-based approach to the joint analysis of Single-Cell data on chromatin accessibility and gene expression. *Stat. Sci.*, **35**, 2–13.
20. Ntranos,V., Kamath,G.M., Zhang,J., Pachter,L. and Tse,D. (2016) Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.*, **17**, 112.
21. Park,S. and Zhao,H. (2018) Spectral clustering based on learning similarity matrix. *Bioinformatics*, **34**, 2069–2076.
22. Pierson,E. and Yau,C. (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, **16**, 241–241.
23. Prabhakaran,S., Azizi,E., Carr,A. and Peer,D. (2016) Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *Int. Conf. Mach. Learn.*, **45**, 1070–1079.
24. Wang,B., Zhu,J., Pierson,E., Ramazzotti,D. and Batzoglou,S. (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, **14**, 414–416.
25. Zhang,L. and Zhang,S. (2018) Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17**, 376–389.
26. Der Maaten,L.V. and Hinton,G.E. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
27. Jolliffe,I.T. (1986) Principal component analysis and factor analysis. In: *Principal Component Analysis, Springer Series in Statistics*. Springer, NY, pp. 115–128.
28. Becht,E., Mcinnes,L., Healy,J., Dutertre,C., Kwok,I. W.H., Ng,L.G., Ginhoux,F. and Newell,E.W. (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, **37**, 38–44.
29. Van Dijk,D., Sharma,R., Nainys,J., Yim,K., Kathail,P., Carr,A., Burdziak,C., Moon,K.R., Chaffer,C.L., Pattabiraman,D.R. *et al.* (2018) Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, **174**, 716–729.
30. Zhu,L., Lei,J., Devlin,B. and Roeder,K. (2018) A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.*, **12**, 609–632.
31. Kharchenko,P.V., Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

32. Vallejos,C.A., Richardson,S. and Marioni,J.C. (2016) Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.*, **17**, 70.
33. Shao,C. and Hofer,T. (2017) Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics*, **33**, 235–242.
34. Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
35. Owen,A.B. and Perry,P.O. (2009) Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.*, **3**, 564–594.
36. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.
37. Boutsidis,C. and Gallopoulos,E. (2008) SVD based initialization: A head start for nonnegative matrix factorization.. *Pattern Recognition*, **41**, 1350–1362
38. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
39. Phipson,B., Lee,S., Majewski,I.J., Alexander,W.S. and Smyth,G.K. (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.*, **10**, 946–963.
40. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
41. Lun,A. T.L., Mccarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.*, **5**, 2122.
42. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 296–296.
43. Deng,Q., Ramskold,D., Reinius,B. and Sandberg,R. (2014) Single-Cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
44. Segerstolpe,A., Palasantza,A., Eliasson,P., Andersson,E., Andreasson,A., Sun,X., Picelli,S., Sabirsh,A., Clausen,M., Bjursell,M.K. *et al.* , (2016) Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.*, **24**, 593–607.
45. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V.H., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for Single-Cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
46. Baron,M., Veres,A., Wolock,S.L., Faust,A.L., Gaujoux,R., Vetere,A., Ryu,J.H., Wagner,B.K., Shenorr,S.S., Klein,A.M. *et al.* , (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.*, **3**, 346–360.
47. Coordination,O., Coordination,L., Preparation,L., Annotation,C.T. and Investigators,P. (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.
48. Butler,A., Hoffman,P.J., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.