SCC: An accurate modication method for scRNA-seq dropouts based on mixture model

yan zheng (yan.zheng@nwpu-bioinformatics.com)
Northwestern Polytechnical University https://orcid.org/0000-0002-7348-3264
Jialu Hu
Northwestern Polytechnical University
Xuequn Shang
Northwestern Polytechnical University

Research article

Keywords: scRNA-seq, noise, mixture model, dropouts identification, gene expression estimation

Posted Date: May 13th, 2020

DOI: https://doi.org/10.21203/rs.2.22114/v2

Version of Record: A version of this preprint was published on January 6th, 2021. See the published version at https://doi.org/10.1186/s12859-020-03878-8.

RESEARCH

SCC: An accurate modification method for scRNA-seq dropouts based on mixture model

Yan Zheng*, Jialu Hu and Xuequn Shang

*Correspondence: yan.zheng@nwpu-bioinformatics.com
School of Computer Science, Northwestern Polytechnical University, West Youyi Road 127, Xi'an, China
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Single-cell RNA sequencing (scRNA-seq) technology has advanced in past decade, which enables gene expression analysis to be carried out at higher resolution. This technology is of great significance for exploring the dynamic development process of life, studying the regulation mechanism of genes and discovering new cell types. However, there are still some limitations in scRNA-seq technology. scRNA-seq technology only detect 5-15 percent of the mRNA molecules. Therefore, low-expression genes are difficult to detect in scRNA-seq data.

Method: The scRNA-seq data tend to be bimodal expression distribution because the expression is either strongly zero or high expression. In this paper, we proposal a method scRNA-seq complementation (SCC) to solve the dropouts in scRNA-seq data. Firstly, we find the nearest neighbor cells of every cell, and then use a mixture model to impute the dropouts of scRNA-seq data. The model can identifies the possibility of dropouts and estimates the reasonable gene expression value.

Results: We use SCC and two existing algorithms to test performance on three scRNA-seq datasets (Kolod, Pollen and Usoskin). The result shows that SCC outperforms existing tools. SCC significantly reduce the intra-class distance of cells and enhance the clustering of cell subpopulation, which is significant for future research on gene expression.

Conclusions: SCC is an effective tool to resolve the noise in scRNA-seq data. The code is freely accessible at the website:https://github.com/nwpuzhengyan/SCC.

Keywords: scRNA-seq; noise; mixture model; dropouts identification; gene expression estimation

Background

With the development of gene sequencing technology, genome research has become more and more popular in the past decades [1]. However, the studies of gene expression have been limited to analysis of pooled populations of cells. Although the methods based on bulk RNA-seq can obtain the genome-wide RNA sequence expression information, the resulting gene expression profiles are only the average values of the different cell types. Therefore, the heterogeneity in cells is neglected and mutations present only in a few cells are substantially hidden (such as early cancer cells) [2]. The analysis of cell clusters does not show cell heterogeneity, which is also an important feature of organ development [3]. In the process of organ development, the progenitor cells undergo diverse differentiation decisions to become specific cell

types. In order to better understand the differentiation and heterogeneity of cells, we need a new technology to help us define the gene expression of individual cells.

A powerful technology has advanced in the past decade, which enables gene expression analysis to be carried out at higher resolution. Every gene expression can now be defined in a single cell. This technology, known as single-cell RNA sequencing (scRNA-seq), enables rapid determination of the precise gene expression patterns of tens of thousands of individual cells [1]. This technology is vital for exploring the dynamic development process of life and studying the regulation mechanism of genes, which also can be used to discover new cell types. However, there are still some limitations in scRNA-seq technology. The main limitation is the noise of the scRNA-seq data, which is mainly caused by the poor sensitivity of scRNA-seq technology [4]. Current scRNA-seq technology can detect only about 10 percent of the mRNA molecules that are actually present. Therefore, low-expression genes are difficult to detect in scRNA-seq data [3]. In addition, the expression of genes are not in a steady state manner in different periods. Batch effect can also cause noise. It is estimated that there are eighty percent noise is caused by technical limitations and the remaining twenty percent is estimated to be of biological origin. The primary challenge in the scRNA-seq data analysis is how to resolve the noise of data.

As shown in Fig.1, the scRNA-seq data tend to be bimodal expression distribution [5]. Although many zero counts in scRNA-seq data are true absence of expression, a big part is caused by technical factors. There are many approaches for solving the noise of scRNA-seq data. scImpute estimates the true expression of gene through clustering similar cells and SAVER recovers the true expression levels of gene by a method that takes advantage of gene-to-gene relationships [6, 7]. We propose a method to get the true expression of gene by integrating nearest neighbor cells. Because of the poor sensitivity of scRNA-seq technology, the scRNA-seq data we obtained is incomplete. The detection rate of mRNA is only 5-15 percent, so genes with low expression levels are difficult to detect [1]. Because of the detection rate of mRNA is only about 5-15 percent, the low-expression genes are difficult to detect. But if we detect mRNA in multiple cells with the same type, low expression genes are likely to be detected in a small fraction of cells. Once the cells with same cell types are clustered, we can combine all gene expression data from same type cells to impute a complete gene expression data [7]. Therefore, the main idea of our method is that obtaining more complete gene expression data by integrating gene expression data of similar cells, which is similar with scImpute. However, scImpute may remove the cell-to-cell heterogeneity because scImpute impute scRNA-seq data by clustering cells with same types. And clustering cells into true types is very difficult. Cell-to-cell heterogeneity is also of great significance for exploring cell heterogeneity.

We have said that scImpute may remove cell-to-cell heterogeneity because correct clustering is difficult. In fact, in scRNA-seq, even in the same cell types, cells with different volumes still have very different mRNA transcript number [5]. Therefore, even if scImpute can clustering cells into true types, the heterogeneity in same cell types will be neglected. However, the cell-to-cell heterogeneity is a very important part of scRNA-seq. In a word, scImpute is a method of clustering similar cells first and then imputing gene expression data. For the methods that obtain complete data by clustering, the gene expression data is over smoothing and the

cell-to-cell heterogeneity is likely to be neglect. Therefore, an accurate expression recovery method that can preserve heterogeneity is essential. In addition, volume is an important factor affecting gene expression in scRNA-seq. Therefore, the mRNA transcript numbe in different cell volumes is different, which leads to the different mRNA capture number. The methods like scImpute and SAVER ignore this factor. In this paper, we propose a method named scRNA-seq complementation (SCC), which can modify the data of scRNA-seq and reduce the intra-class distance of cells. In SCC, we replace clustering similar cells with find nearest neighbour cells of each cells. This method can not only obtain the complete gene expression data but also preserve cell-to-cell heterogeneity.

The main idea of SCC is shown in Fig 2. In Fig 2, there are three different cell types represented by different colors and shapes. The sizes of quadrilaterals represent different cell volumes and the holes in quadrilaterals represent the dropouts in scRNA-seq data. In order to solve dropouts of scRNA-seq data, we bring together similar cells and obtain the complete cell data by complementing similar cells. In SCC, for every cell, we find the nearest neighbour cells with similar volume and modify the dropouts of the cell by the complement of neighbour cells. Compared with scImpute, we retain the cell-to-cell Heterogeneity. The result shows that our method can reduce the intra-class distance of cells and enhance the clustering of cell subpopulation.

Methods

In order to solve the noise of scRNA-seq data, we have developed a new method, SCC, which can be used to recover the gene expression of genes with dropouts. The basic process is shown in Fig 3. As shown in Fig 3, the core algorithm of SCC consists of three steps. The first step is the filtration of outliers. Because we recover the gene expression by nearest cells, the outliers have a great impact on the modified result. The second step is the detection of nearest neighbour cells. In the second step, we find nearest neighbour cells by another method scmap. The third step is the modification of gene expression. We propose a mixture model to describe the distribution of gene expression and estimate the real value of genes. The estimation of parameters is the process of estimating parameters by EM algorithm. The detailed description of each step is introduced in the following sections.

The filtration of outliers

The input of our method is the matrix $X_{g\times c}$ of scRNA-seq data and the g (row) represent genes and c (column) represent cell names. Outliers are the result of technical limitations, which have a great influence on the cells clustering [8]. At first, we use principal component analysis (PCA) to reduce dimension on matrix X and calculate the distance matrix on the data with dimension reduction. The matrix of scRNA-seq data is a sparse matrix and PCA can resolve the sparse problem and accelerate the speed of calculation [9]. We select first two principal components to calculate the distance matrix $D_{c\times c}$.

The distance matrix $D_{c\times c}$ can be calculated based on the PCA-transformed data. For each cell c, we select nearest cell and calculate the nearest distance dis_c. For the set of dis_c, we find the first quartile dis_{q1} and third quartile dis_{q3}. For the cells

that satisfy the follow formula, $dis_c > dis_{q1} + 1.5(dis_{q3} - dis_{q1})$, we regard them as outlier cells [6]. For the outlier cells, we think them as the result of technical errors and accidents. These cells have a great influence on the results of imputation. We delete the cells and implement our experiment in the rest cells matrix $X_{g \times c1}$ ($c1$ represent the rest cells).

The detection of nearest neighbour cells

In the step, we find top twenty nearest neighbour cells of each cells by calculate the Euclidean distance first, but this step spend a lot of time in the calculation of distance. In fact, projecting individual cells onto most similar neighbour cells is an important method to reduce the batch effect of scRNA-seq data. We detect nearest neighbour cells by another method SCmap rather than calculation of the Euclidean distance. SCmap is a very convenient method to project a cell to the nearest neighbors, which use the cosine similarity, Pearson and Spearman correlations to calculate similarities [10]. An important feature of SCmap is very fast and it takes only about one minute to select features (important genes) and calculate nearest neighbour cells for 40000 cells. We choose SCmap to select 20 nearest neighbour cells $C = \{c_1, c_2, c_3, \ldots, c_{20}\}$ for each cell c. And then we filter lower similarity cells (lower similarity means that the similarity difference is greater than 0.1). The rest cells set $C = \{c_1, c_2, c_3, \ldots, c_n\}$ will be retained for recovering the gene expression profiles.

The modification of gene expression

After we obtain the nearest neighbour cells set $C = \{c_1, c_2, c_3, \ldots, c_n\}$ of each cell c, we need to predict the real value of gene expression of each cell [11]. For the genes of each cell, we classify genes into three categories: high-expression, low-expression and zero-expression. The high-expression genes have a large amount of mRNA, so they are likely to be detected in scRNA-seq technology. The low-expression genes have less mRNA number, which in poor sensitivity leading to most dropouts in scRNA-seq data. Moreover, the zero-expression means that the real value of gene expression is zero. For the different expression genes, we construct a mixture model to determine whether a zero value is a real value or a dropout. Because the low-expression genes are difficult to detect, the most genes scRNA-seq data tend to be bimodal expression distribution. Therefore, we describe pattern by a mixture model with three components. The first component is a normal distribution used to represent the high-expression genes, and the gene expression of high-expression genes in same cell types are different because of biological factors. The second component is a binomial distribution used to represent the distribution of low-expression genes and the probability p is the mRNA detection rate. The third component is a zero distribution to represent genes that the real value is zero. As shown in Fig 4, the distribution of mixture model has similar bimodality with gene expression distribution. Therefore, the mixture model can describe the distribution of scRNA-seq data well. Most methods log the input matrix with a pseudo count 1 for dealing with scRNA-seq data. In our method, the cells matrix $X_{g \times c1}$ is not transformed by log with pseudo count, so it is reasonable to describe the true distribution of scRNA-seq genes by the binomial distribution.

For each cell c, we construct a different mixture models to modify the cell c gene expression. The different models have different proportions and parameters. For each gene g, its expression is modeled in cells set $C = \{c_1, c_2, c_3, \ldots, c_n\}$. The model formula is as follows:

$$f(x) = \lambda_1 Normal(x, \mu, \sigma) + \lambda_2 Binomial(x, p, N) + \lambda_3 0$$

$$subject \quad to \quad \lambda_1 + \lambda_2 + \lambda_3 = 1$$

$$x \in \{X_{g,c_1}, X_{g,c_2}, \ldots, X_{g,c_n}\}$$

The $\lambda_1, \lambda_2, \lambda_3$ are the probabilities of three distribution and the sum of λ is one, the μ, σ are the mean and standard deviation of Normal distribution and the p is the probability of success and the N is the total number of observations. If the λ_1 of a gene is high and λ_3 is low, we think that the zero in the gene is likely a dropouts. On the other hand, if the λ_3 of a gene is high, the real value of the gene is likely zero. X_{g,c_n} is the value of gene g in cell c_n. We obtain every value of λ and identify which λ is greater than one-third. If λ_1 is greater than one-third, the value of $X_{g,c}$ is uncharged to retain the cell-to-cell heterogeneity. If the λ_2 is greater than one-third, the modified value of $X_{g,c}$ is $N * p$, and if the λ_3 is greater than one-third, the real value of $X_{g,c}$ is likely zero.We put the modified value into the matrix and get the modified data.

The estimation of parameters

The advantage of this model is that it assumes that the value of dropouts relate to nearest neighbour cells. We obtain the modified gene expression by complement of neighbour cells. The parameters in the model are estimated by Expectation Maximization (EM) algorithm [12]. The expectation-maximization algorithm is to find the maximum likelihood estimation or the maximum posteriori estimation of parameters in the probabilistic model, in which the probabilistic model depends on the hidden variables that cannot be observed. We first set an initial value for every parameter and then calculate every value's probability in three component. The number of iteration is set to 100 and the iteration will stop when the difference of parameters is small (The threshold is set to 0.01).

The E step: The initial value of μ is the mean of nearest neighbour cells and the initial value of σ is the standard deviation. The initial of p is 0.1 because the detection rate of mRNA is about 5-15 percent. The N is an integer number less than 10. All the λ are set to one-third. For the gene g in nearest cells set $C = \{c_1, c_2, c_3, \ldots, c_n\}$, we calculate the probabilities of every gene expression in three component for modifying the parameters in M step.

The M step: We have obtained the probabilities $Pro_{n\times 3}$ of gene g in nearest cells set $C = \{c_1, c_2, c_3, \ldots, c_n\}$. Then we calculate the new probabilities $Pro_{n\times 3} = Pro_{n\times 3}/rowsum(Pro)$ that the values belong to three components. We set $P1 = sum(Pro[, j])$, $P2 = sum(Pro[, j] * matrix[g,])$. The mean of every component j will be calculated by $Mean_j = P2/P1$ and the value of $\lambda_j = P1/n$ (n is the number

of nearest cells). We set $P3 = sum(Pro[,j] * (matrix[g,] - Mean_j)^2)$ and the standard deviation of every component j will be calculated by $Dev_j = sprtP3/P1$. After obtaining the values, we can calculate the parameters: μ,σ,N,p,λ.

Results

The output of SCC is a modified matrix consisting g row and $c1$ columns. SCC does not need the type of cells and the scRNA-seq matrix is the only input. Besides, SCC is memory-efficient because it only modify one cell at a time. First, we use simulation data to verify the method. And then we performed SCC in three scRNA-seq datasets (Kolod, Pollen and Usoskin). The results shows that the SCC can significantly reduce the intra-class distance of cells and enhance the clustering of cell subpopulation.

The simulate data

We use the scSimulator function to create the simulate data. The simulate data contains 3 cell types, 150 cells and 8180 genes. For the simulate data, we get the modified data by SCC and visualize the raw data and modified data by PCA. The visualization is shown in the Fig 5. The left part is the raw simulate data and right part is the modified data. As shown in the Fig 5, the modified data is more intensive. The Adjusted Rand Index (ARI) can be used to calculate similarity between the clustering and real types [8]. The range of ARI values is between -1 and 1. The negative value means that the clustering result is bad, which indicating that the labels are independently distributed. The values of good clustering results are positive (1 is the best result), which indicating that the distribution of the two labels is identical. We use K-means to cluster the cell and calculate the ARI values in raw data and modified data. The ARI values are 0.4394281 and 0.5233112 respectively. The modified data result is better than raw data. So we can conclude that SCC is effective method to recover the gene expression.

The real data

We alse run SCC in the real datasets (Kolod, Pollen and Usoskin). Kolod (704 cells, 13473 genes) is a dataset about pluripotent cells under different environment conditions and Pollen (249 cells, 6982 genes) is a dataset contains Eleven cell populations including neural cells and blood cells. Usoskin (622 cells, 17772 genes) is a neuronal cells dataset with sensory subtypes.

SCC can resolve the dropouts in scRNA-seq

We counted the number of zero values in the GSE76381 (Adult), which is downloaded from NCBI website. The num of genes is 18219 and the num of cells is 243. 3429466 of 4427217 gene expression values are zeros. Zero values accounts for 77.46 percent of the total gene expression values. After the modification of gene expression, the ratio of zero values dropped to 68 percent. Therefore, we can draw a conclusion that SCC can assign value to dropouts and retain a part of real zero values.

SCC can reduce the the intra-class distance of cells

We use intra-class distance and inter-class distance to evaluate the performance of SCC modification. The intra-class distance is the mean square distance between sample cells of the same cell types and the inter-class distance is the mean square distance of the sample cells of the different cell types. The smaller the intra-class distance is, the better the modification result is. The smaller the inter-class distance is, the worse the modification result is [13]. We hope that the intra-class distance of cells is smaller and the inter-class distance of cells is larger. We assume that the K_i is the K_i class, N_i is the number of cell in the K_i class and X_k^i is the kth cell value in the K_i class. The two formulas are as follows:

$$Dis_{intra} = \sum_{K_i} \frac{1}{N_i * N_i} \sum_{k=1}^{N_i} \sum_{l=1}^{N_i} Dis(X_k^i, X_l^i)$$

$$Dis_{inter} = \sum_{K_i} \sum_{K_j} \frac{1}{N_i * N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} Dis(X_k^i, X_l^j)$$

After we obtain the matrix of modification, we calculate the intra-class distance and inter-class distance by the formulas above. However, the two distances are used to measure the performance of result is not convenient. We use another value $Dis = Dis_{intra}/Dis_{inter}$ to describe the performance of result. For the new distance Dis, the smaller the value is, the better the performance of result is. We run SCC, scImpute and SAVER in three different public scRNA-seq datasets and calculate the Dis (The three datasets are Kolod, Pollen and Usoskin). Compared with other methods, SCC can significantly reduce the intra-class distance of cells. The result as shown in Table 1.

As shown in the Table 1, SCC reduce the distance of same type cells compared with raw data. Compared with other methods, SCC has best performance in most scRNA-seq datasets except for Kolod dataset. The result shows that SCC can significantly make the cells with same type closer. We add one to the raw matrix and modified matrix and transform them by log. Finally, we perform principal component analysis on new matrix. The first two principal components are used for visualization [14]. The result is shown in follow Fig 6. As shown in the Fig 6, we can clearly observe that SCC can shorten the distance of same type cells.

SCC can enhance the clustering of cell subpopulation

After the recovery of gene expression, we further cluster the modified cells data. In the existing clustering algorithm, K-means clustering algorithm is a popular iterative solution-clustering algorithm [15]. We run SCC, scImpute and SAVER in three datasets and perform k-means clustering in the modified results. Finally, we calculate the ARI values in different methods. The result is shown in Fig 7. The detailed values are shown in Table 2.

As shown in the Table 2, the ARI values of SCC are obviously higher than other methods, especially in Kolod dataset, the ARI value of clustering is 1. The value of ARI in Kolod is 1, which indicating that the two labels of clustering and real cell types are the same. From this table, we can draw a conclusion that SCC can significantly enhance the clustering of cell subpopulation in most datasets.

Discussion

The primary limitation is the high noise in scRNA-seq technology. There are many reasons causing noise, the most important one of which is the poor sensitivity of scRNA-seq technology. The poor sensitivity makes low-expression genes hard to detect. scImpute and SAVER are existing tools for solving the noise in scRNA-seq data. They estimate the true expression of gene by clustering similar cells and taking advantage of gene-to-gene relationships. However, they ignore cell-to-cell heterogeneity in same types, which is important for cell heterogeneity. Compared with existing tools, SCC can retain the cell heterogeneity by modifying the expression of each cell by the nearest neighbour cells. In the intra-class distance of cells and the clustering of cell subpopulation, SCC has better performance compared with other tools.

Conclusions

In conclusion, we propose a method SCC (single-cell complementation) to resolve the noise (especially dropouts) in scRNA-seq data. SCC focuses on the gene expression that are largely affected by poor sensitivity of mRNA, while retaining the expression of genes with high expression level. The main idea of method is the complementation of similar cells. For each cell, we find the nearest neighbour cells by scmap and recover the value of dropouts by a mixture model. Compared with other methods, we can retain the cell heterogeneity by replacing clustering with detecting the nearest neighbour cells. We perform SCC and other two methods in three different scRNA-seq datasets. The result shows that SCC can significantly reduce the intra-class distance of cells and enhance the clustering of cell subpopulation. Another advantage of SCC is memory-efficient and it can run in tens of thousands cells on a laptop.

In future, we will continue our research based on the previous scRNA-seq work. After the modification of gene expression, we will cluster the cells in the modified cells data. In the existing clustering algorithm, some methods (such as SC3) have a good performance in clustering of scRNA-seq data [16]. We will verify the defects of the existing clustering methods and decide how to propose a better clustering methods for scRNA-seq data.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The code is freely accessible at the website:https://github.com/nwpuzhengyan/SCC.

Competing interests
The authors declare that they have no competing interests.

Funding
Publication costs were funded by the National Natural Science Foundation of China (Grant No. 61702420); This project has been funded by the National Natural Science Foundation of China (Grant No. 61332014, 61702420 and 61772426); the China Postdoctoral Science Foundation (Grant No.2017M613203); the Natural Science Foundation of Shaanxi Province (Grant No.2017JQ6037); the Fundamental Research Funds for the Central Universities (Grant No. 3102018zy032).

Author's contributions
YZ designed the computational framework and implemented the algorithm. JH modify the program and the manuscript. XS is the major coordinator, who contributed a lot of time and efforts in the discussion of this project. All authors read and approved the final manuscript.

Acknowledgements
Many thanks go to Yiqun Gao and Yang Qi for their help in the discussion of this methods.

References

1. Potter, Steven, S.: Single-cell rna sequencing for the study of development, physiology and disease. Nature Reviews Nephrology
2. Baslan, T., Hicks, J.: Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nature Reviews Cancer **17**(9), 557–569 (2017)
3. Papalexi, E., Satija, R.: Single-cell rna sequencing to explore immune cell heterogeneity. Nature Reviews Immunology
4. Park, J., Shrestha: Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease
5. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome Biology **16**(1), 278
6. Li, W., Li, J.: An accurate and robust imputation method scimpute for single-cell rna-seq data **9**(1), 997 (2018)
7. van Dijk, D.e.a.: a diffusion-based imputation method reveals genegene interactions in single-cell rna-sequencing data
8. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Hemberg, M.: Sc3: Consensus clustering of single-cell rna-seq data. Nature Methods **14**(5), 483–486 (2017)
9. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., Batzoglou, S.: Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. Nature Methods **14**(4), 414–416
10. Kiselev, V.Y., Yiu, A., Hemberg, M.: scmap: projection of single-cell rna-seq data across data sets. Nature Methods
11. Lin, P., Troup, M., Ho, J.W.K.: Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. Genome Biology **18**(1), 59
12. Lakhal-Chaieb, L., Greenwood: A smoothed em-algorithm for dna methylation profiles from sequencing-based methods in cell lines or for a single cell type
13. Zhang, J.: Intra- and inter-class spectral variability of tropical tree species at la selva, costa rica: Implications for species identification using hydice imagery. Remote Sensing of Environment **105**(2), 129–141
14. Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., Wang, W.: Normalization and noise reduction for single cell rna-seq experiments. Bioinformatics **31**(13), 2225–2227
15. Chlis, N.K., Alexander Wolf, F., Theis, F.J.: Model-based branching point detection in single-cell data by k-branches clustering. Bioinformatics
16. Nikolenko, S.I., Korobeynikov, A.I., Alekseyev, M.A.: Bayeshammer: Bayesian clustering for error correction in single-cell sequencing. Bmc Genomics **14**(S1), 7 (2012)

Figures

Figure 1 The bimodal expression distribution. The typical bimodal expression distribution of scRNA-seq data (The eleventh gene in mECS, which is a dataset about mammary epithelial cells). The detection rate of mRNA is 5-15 percent. The genes with low expression levels are difficult to detect. Therefore, the expression of scRNA-seq data is either strongly zero or high expression. We call this distribution a bimodal distribution.

Figure 2 The complementation of scRNA-seq data. A quadrilateral represents a cell and the different Colors represent different cell types and different sizes represent different cell volumes. The data of scRNA-seq is incomplete and the holes in quadrilaterals represent the missing information. Although the data of each cells is incomplete, we can still bring together similar cells. We obtain the complete cell data by complementing similar cells.

Figure 3 The flow chart of the core algorithm of SCC. The input of SCC is the matrix of scRNA-seq data and there are three main steps in SCC. The first step is the filtration of outliers.In second step, we obtain nearest neighbour cells set of every cell c by SCmap. In third step, we solve the dropouts of the cell c according to the nearest neighbour cells.

Figure 4 The distribution of mixture model. We use mixture model to describe the distribution of scRNA-seq. It can be seen that the distribution of mixture has similar bimodality with the distribution of scRNA-seq. Therefore, the distribution of scRNA-seq can be described by the mixture model.

Figure 5 The visualization of simulate data. We run SCC in the simulate dataset. The left chart is the raw simulate data and the right data is the simulate data after modification.

Figure 6 The PCA visualization of raw data and modified data. We perform SCC in three different datasets. The three datasets are Kolod, Pollen and Usoskin. Next, we perform principal component analysis on raw and new matrixs and use the first two principal components to draw the scatter plot. The left plots are the scatter plot of raw data and right plots are the scatter plot of modified data. As shown in the Fig, we can draw a conclusion that SCC can make cells with same types aggregate more closely.

Figure 7 The barchart of ARI values. We run SCC, scImpute and SAVER in three different datasets(Kolod, Pollen and Usoskin) and perform K-means clustering in modified data. Finally, we calculate the ARI value and draw the barchart of different methods.

Tables

Table 1 Table of cell distance

	raw data	SCC	scImpute	SAVER
Kolod	27272.63	21138.75	27193.77	15774.53
Pollen	157.074	118.77	561.21	167.564
Usoskin	21467.77	19498.95	23502.52	20341.59

Table 2 Table of ARI value

	SCC	scImpute	SAVER
Kolod	1	0.957535	0.995985
Pollen	0.70043	0.694498	0.654509
Usoskin	0.24264	0.07003	0.131585

Figures

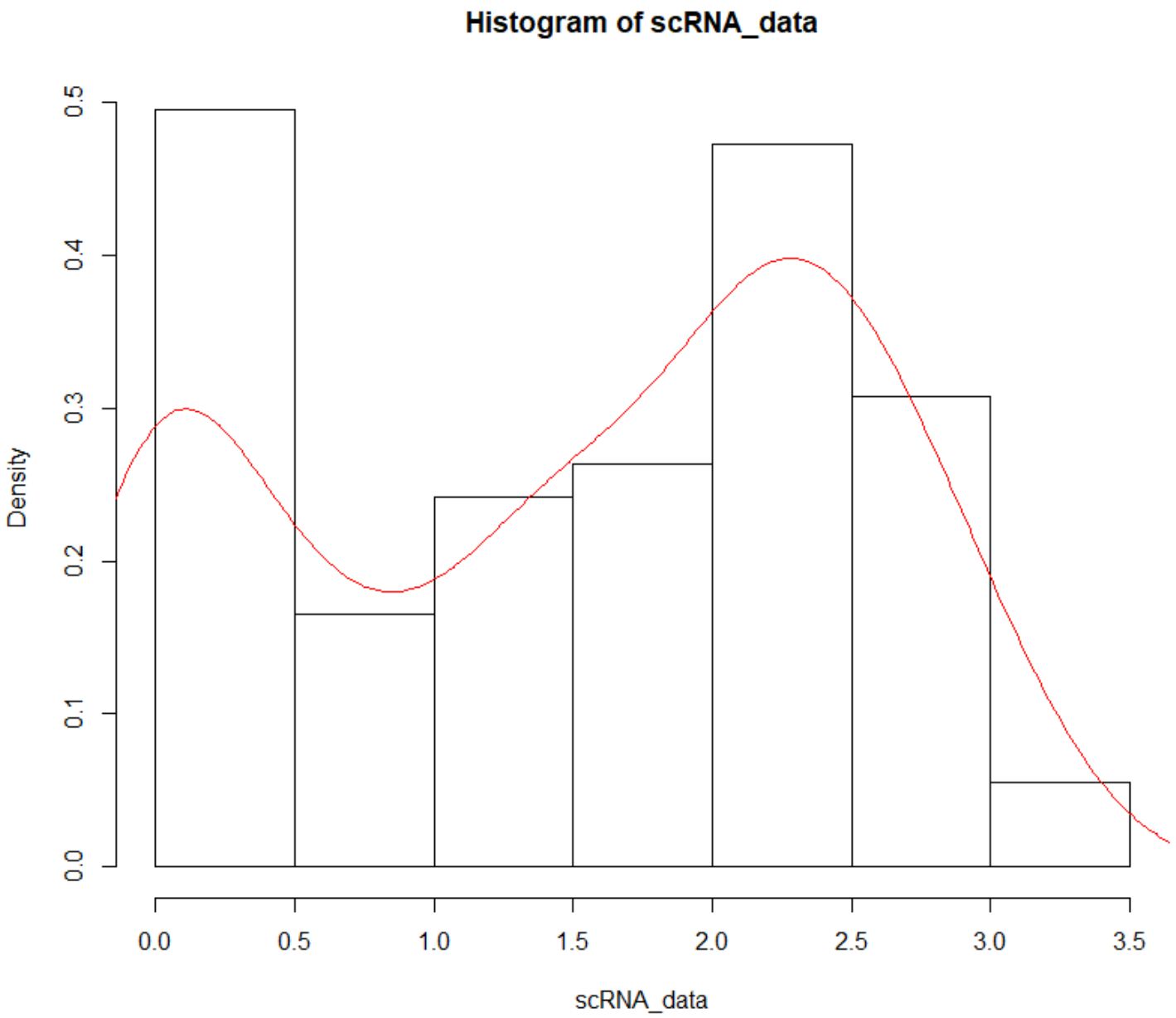

Figure 1

The bimodal expression distribution. The typical bimodal expression distribution of scRNA-seq data (The eleventh gene in mECS, which is a dataset about mammary epithelial cells). The detection rate of mRNA is 5-15 percent. The genes with low expression levels are di cult to detect. Therefore, the expression of scRNA-seq data is either strongly zero or high expression. We call this distribution a bimodal distribution.

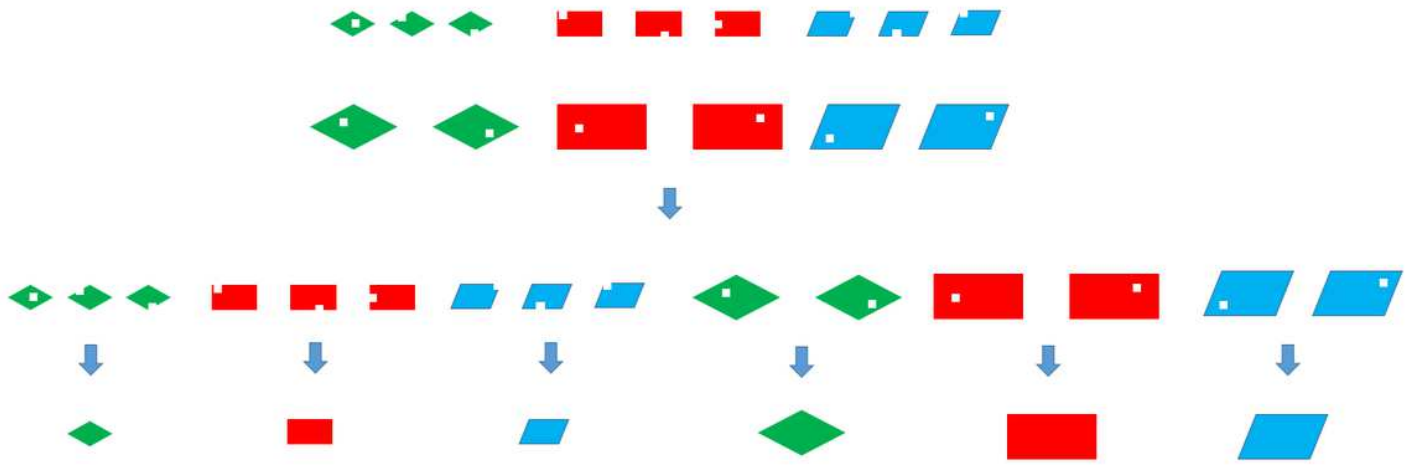

Figure 2

The complementation of scRNA-seq data. A quadrilateral represents a cell and the di erent Colors represent di erent cell types and di erent sizes represent di erent cell volumes. The data of scRNA-seq is incomplete and the holes in quadrilaterals represent the missing information. Although the data of each cells is incomplete, we can still bring together similar cells. We obtain the complete cell data by complementing similar cells.

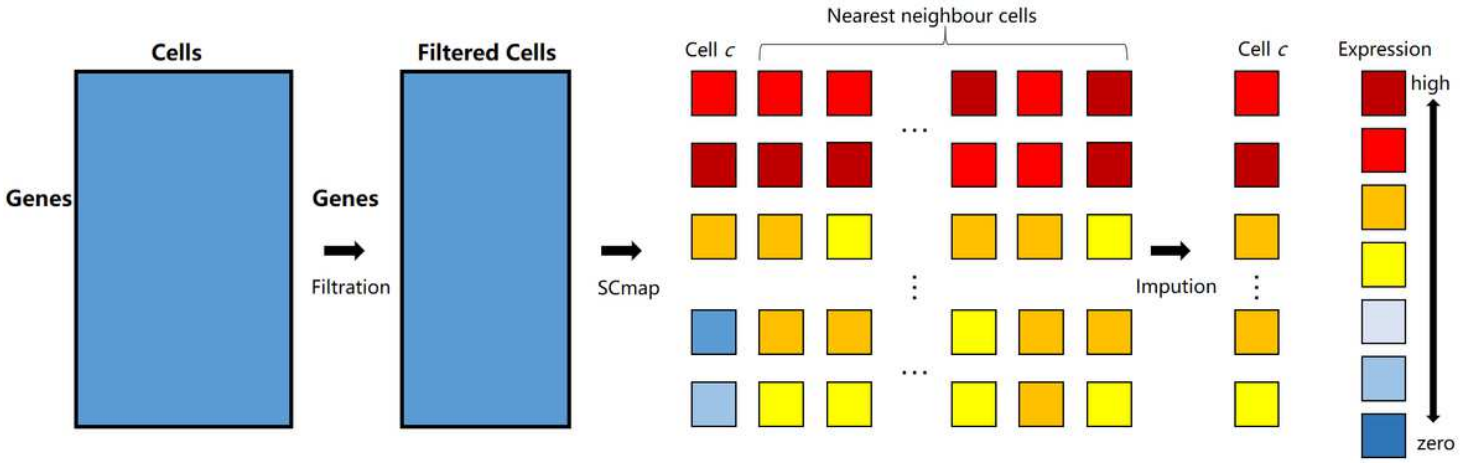

Figure 3

The ow chart of the core algorithm of SCC. The input of SCC is the matrix of scRNA-seq data and there are three main steps in SCC. The rst step is the ltration of outliers.In second step, we obtain nearest neighbour cells set of every cell c by SCmap. In third step, we solve the dropouts of the cell c according to the nearest neighbour cells.

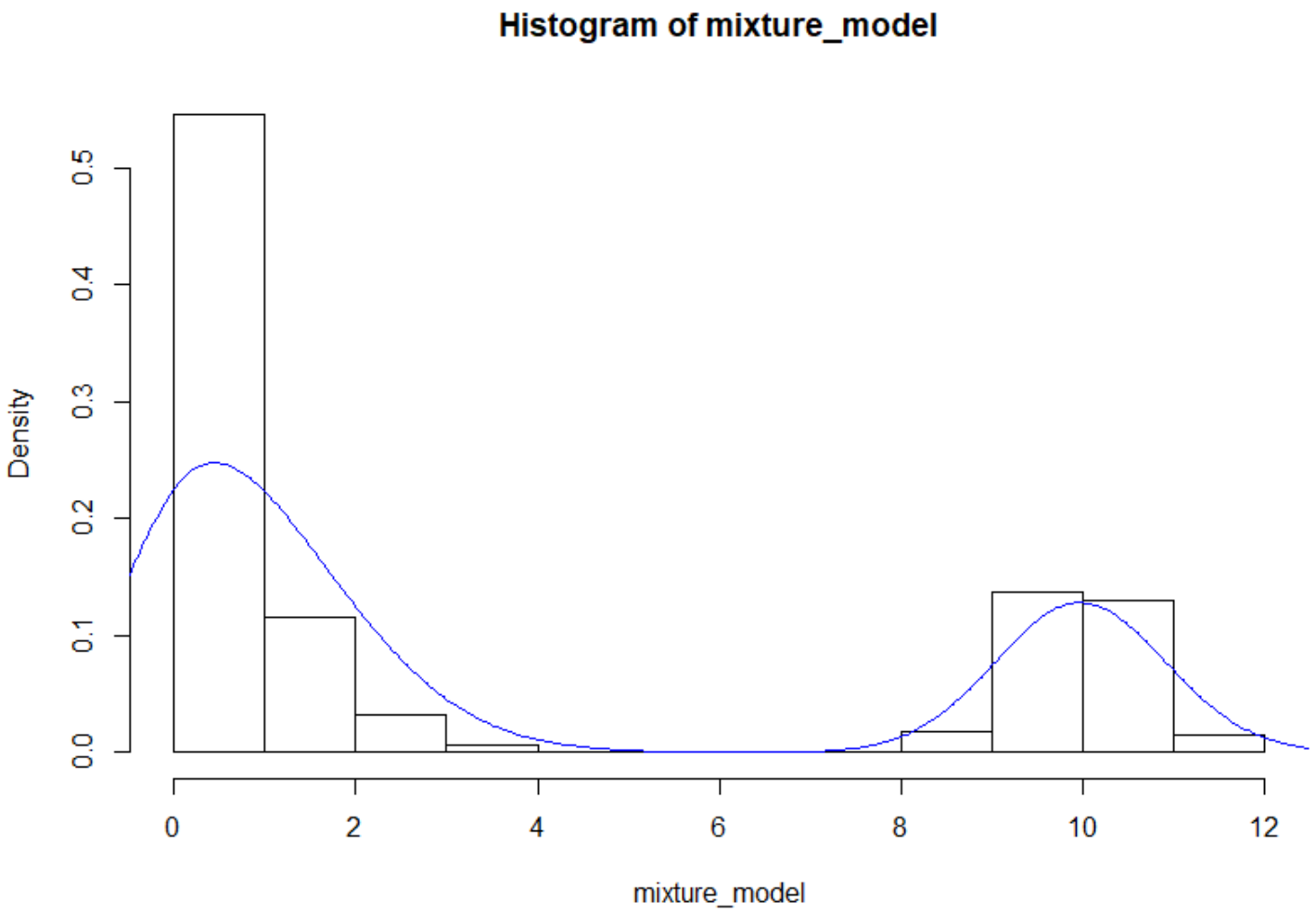

Figure 4

The distribution of mixture model. We use mixture model to describe the distribution of scRNA-seq. It can be seen that the distribution of mixture has similar bimodality with the distribution of scRNA-seq. Therefore, the distribution of scRNA-seq can be described by the mixture model.

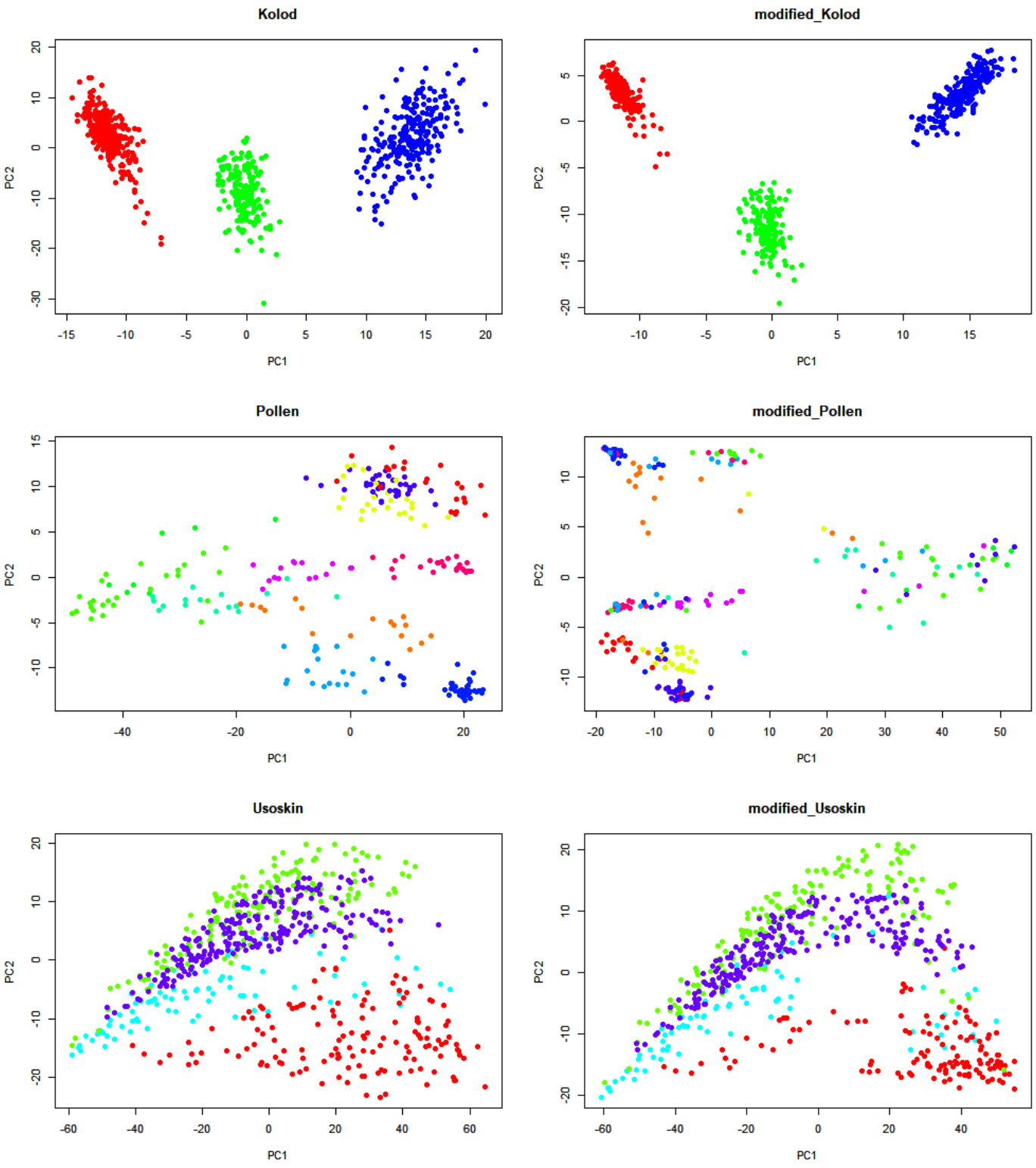

Figure 5

The visualization of simulate data. We run SCC in the simulate dataset. The left chart is the raw simulate data and the right data is the simulate data after modication.

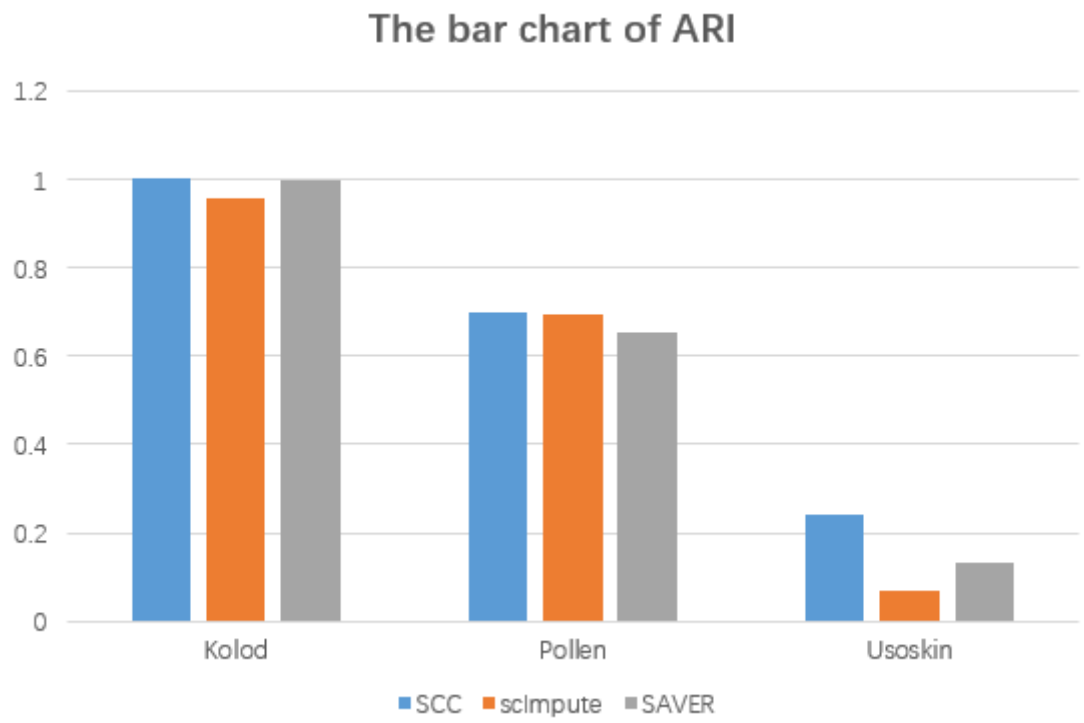

Figure 6

The PCA visualization of raw data and modied data. We perform SCC in three di erent datasets. The three datasets are Kolod, Pollen and Usoskin. Next, we perform principal component analysis on raw and new matrixs and use the rst two principal components to draw the scatter plot. The left plots are the scatter plot of raw data and right plots are the scatter plot of modied data. As shown in the Fig, we can draw a conclusion that SCC can make cells with same types aggregate more closely.