Received: 9 November 2024 | Revised: 27 December 2024 | Accepted: 16 January 2025

IET Systems Biology

DOI: 10.1049/syb2.70003

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

SeqBMC: Single-cell data processing using iterative block matrix completion algorithm based on matrix factorisation

Gong Lejun[1] | Yu Like[1] | Wei Xinyi[1] | Zhou Shehai[1] | Xu Shuhua[2]

[1]Jiangsu Key Lab of Big Data Security & Intelligent Processing, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

[2]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, China

Correspondence

Xu Shuhua.
Email: 2277621025@qq.com

Funding information

Natural Science Foundation of Nanjing University of Posts and Teclecommunications, Grant/Award Number: NY223093; Natural Science Foundation of Zhejiang Province under, Grant/Award Number: LGG22F020040; Open Research Fund of State Key Laboratory of Digital Medical Engineering, Grant/Award Number: 2024-M10; Wenzhou Scientific Research Project, Grant/Award Number: ZG2024013

Abstract

With the development of high-throughput sequencing technology, the analysis of single-cell RNA sequencing data has become the focus of current research. Matrix analysis and processing of downstream gene expression after preprocessing is a hot topic for researchers. This paper proposed an iterative block matrix completion algorithm, called SeqBMC, based on matrix factorisation. The algorithm is used to complete the missing value of the gene expression matrix caused by the defect of sequencing technology. The gene frequency of the matrix is used to block the matrix, and then the matrix factorisation algorithm is used to complete the small matrix after the block, and then the biological zeros that may exist in the block matrix are retained. Experimental results show that the matrix completion algorithm can significantly improve the classification performance of the gene expression matrix after completion with 86.81% F1 score, which is conducive to the recognition of cell types in sequencing data. Moreover, this completion method can be completed only by the machine learning method without too much prior knowledge related to biology and has good effects. Compared with ALRA, SeqBMC increased 5.47% accuracy and 5.03% F1 score. It indicates that SeqBMC has significant advantages in the matrix completion of single-cell RNA sequencing data.

KEYWORDS

big data, biocomputers, biocomputing, biology computing, data analysis, data mining, decision making

1 | INTRODUCTION

At present, the mainstream cell identification process in the field of single-cell RNA sequencing in biology is to use manual labelling to infer the cell type according to the highly expressed genes in the gene expression matrix, and this manual labelling method has many disadvantages. With the continuous development of sequencing technology, the number of sequenced cells is also increasing [1, 2], which makes single-cell transcriptome sequencing widely used [3, 4]. If the traditional method of manually labelling cell types is considered for each cell, the workload will be very large [5, 6]. Therefore, if the method of machine learning is used to extract and learn the features of the gene expression matrix and directly predict and classify cell types and cell subtypes, the time cost and technical cost consumed by manual cell type annotation can be greatly reduced [7, 8]. Therefore, finding a suitable automatic classification algorithm of cell types has the possibility of realisation and very important practical significance for single-cell sequencing technology [9].

Because of the limitations of single-cell RNA sequencing technology, measuring gene expression in a single cell requires the amplification of really small amounts of mRNA, which results in a phenomenon known as "dropout" [10]. In this phenomenon, some transcripts of genes with low expression levels are not detected, so the expression value is zero. Under this "missing" phenomenon, the gene expression matrix shows an excessive number of zeros. Some of these zeros are biological "biological zeros" genes that are not expressed in cells in the first place [11]. Some are due to "technical zeros" caused by sequencing technology in which expressed genes are not detected because of "loss". The technical zeros should be

completed as far as possible, while the biological zeros should be retained when the single-cell gene expression matrix is completed. At present, most of the completion of the single-cell RNA gene expression matrix is to complete the entire gene expression matrix without differentiating between technical zero and biological zero. This results in a large number of biological zeros being incorrectly completed, so a matrix completion algorithm specifically designed for single-cell RNA sequencing is needed to compensate for this deficiency [12].

The gene expression matrix completion algorithm of single-cell RNA can solve the gene loss phenomenon caused by sequencing technology or the quality of the sample to be sequenced [13]. Given the unique characteristics of single-cell RNA sequencing data, the completion of gene expression matrices should address the issue of zeros and should not merely fill in the biological zeros. Therefore, a specialized approach should be designed for the gene expression matrix completion algorithm. After the completion algorithm of the matrix, the matrix is more close to the actual cell gene expression level. It lays a foundation for the accuracy of downstream analysis of single-cell RNA sequencing [14, 15]. George et al. proposed zero-retention estimation (ALRA) for single-cell RNA sequencing data using the low-rank approximation [16]. This method categorises the "zeros" in the gene expression matrix into "biological zeros" and "technical zeros". Based on the assumption that the underlying true expression matrix is non-negative, low-rank, and contains many zeros, low-rank approximation is performed through singular value decomposition (SVD), and it is observed that elements corresponding to biological zeros for each gene are symmetrically distributed around zero. A specific threshold is set, and zeros outside this threshold are considered technical zeros, which are subject to completion, while zeros within the threshold are considered biological zeros and are not completed. This method is overly conservative in its completion operations, preserving nearly all biological zeros but also incorrectly preserving nearly half of the technical zeros, leading to an insufficient completion of technical zeros. Huang et al. proposed a transgene-based single-cell sequencing data processing method (SAVER) [17]. Accurate expression estimates of all genes are obtained by cross-gene and cell information. This is a method of using gene-to-gene relationships to restore the true expression level of each gene in each cell, eliminating technical differences while preserving biological differences across cells. The method is also too conservative for the completion operation, saving most of the biological zeros, and the completion of the technical zeros is not sufficient. At the same time, the calculation efficiency of the method is low, and the performance of the computer is high. At present, there are also some other completion algorithms for single-cell RNA sequencing data processing, but they all have some shortcomings [18–21]. Thus, the performance of the completed genes in cell clustering is not perfect, and the inference of cell trajectory needs to be improved.

In this paper, an iterative block matrix completion method based on matrix factorisation is proposed, called SeqBMC. This method can be used to complete the missing value of the gene expression matrix caused by the defect of sequencing technology. In the matrix completion, biological zeros and technical zeros in the matrix are distinguished, and only technical zeros caused by technical defects are completed. Compared with existing ALRA and SAVER methods, SeqBMC theoretically has some advantages. Firstly, SeqBMC uses a gene frequency-driven partitioning method to more accurately distinguish between the technical zero and biological zero, thereby preserving more biological information during the completion process. Secondly, SeqBMC adopts an iterative matrix factorisation algorithm, which is more efficient in processing small matrices and can better capture the local features of single-cell data. In addition, SeqBMC automatically completes matrix completion through machine learning methods without requiring a large amount of prior biological knowledge, surpassing the ALRA and SAVER methods that rely on a large amount of biological knowledge in terms of automation and computational efficiency. The following sections describe this approach in detail.

2 | METHODS AND MATERIALS

2.1 | Pipeline of SeqBMC algorithm

SeqBMC is an iterative block matrix completion algorithm based on matrix decomposition to process single-cell RNA data. It consists of the following steps as shown in Figure 1.

The pipeline includes the following steps:

(1) Getting and preprocessing the single-cell data to obtain gene expressing matrix M.
(2) Calculating and sorting the frequency of each gene in the gene expression matrix M to obtain the sorted matrix P.
(3) The matrix P was partitioned into columns and segmented into k block matrices V of the same size to obtain the columns of the small matrix, where n is the number of genes, the behaviour M of the small matrix, and m is the number of cells.
(4) Each segmented matrix V is decomposed into the product of two matrices W and H by the low-rank matrix factorisation algorithm.
(5) Update the matrices W and H, and calculate the error E between the product matrix of W and H and the block matrix V.
(6) When the error E is not less than the set threshold, return to Step 5; when the error E is less than the set threshold, the product matrix of W and H is taken as the iterative matrix obtained by decomposition, and the next step is taken.
(7) The iterative matrix obtained by the decomposition was compared with the block matrix V. The non-zeros in the block matrix V were kept unchanged, the biological zeros were retained without completion, and the technical zeros were replaced and completed to obtain the block gene expression matrix except the biological zeros.

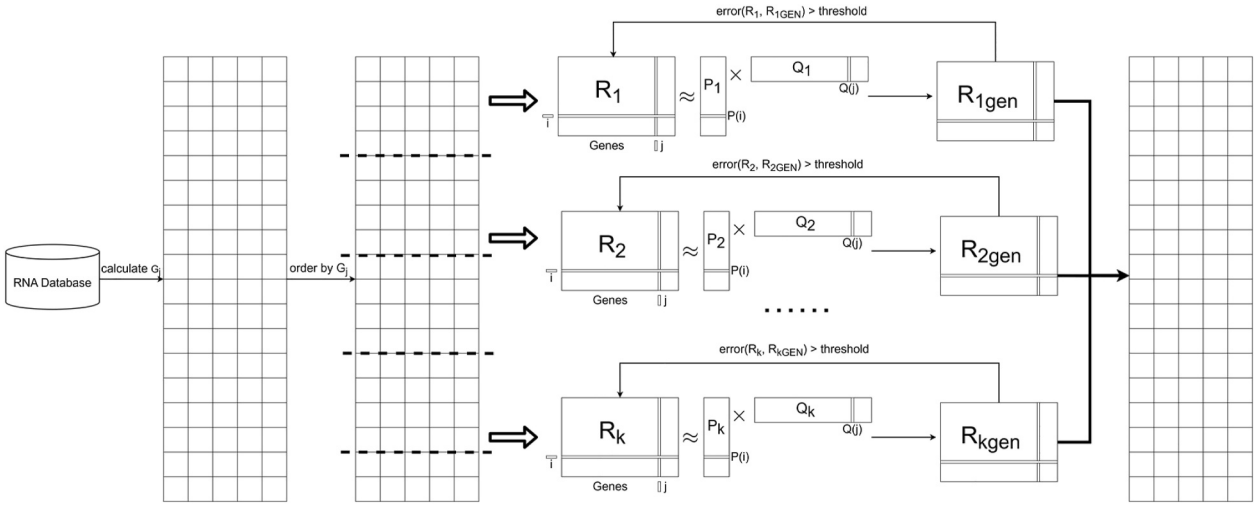

FIGURE 1 Pipeline of SeqBMC algorthim.

(8) The segmented gene expression matrix obtained in step 7 except the biological zeros after completion was spliced to form the complete matrix, which was the complete gene expression matrix after completion. The supervised learning classification algorithm was used to construct the classification model, and the predicted cell types were obtained by the input of the complete gene expression matrix data after the completion.

2.2 | Dataset

This paper performs the matrix completion for the phenomenon of gene "loss" in single-cell RNA sequencing data due to the defect of sequencing technology. In this study, single-cell sequencing data of human chronic myelogenous leukaemia (CML) was collected from the gene expression database of the National Centre for Biotechnology Information (NCBI) (data key number GSE76312) [22]. So far, NCBI has collected more than 51,500 single-cell RNA sequencing data. The sequencing data of human chronic myelogenous leukaemia single cells selected in this paper contained five different cell types: acute phase chronic myelogenous leukaemia cells, chronic phase chronic myelogenous leukaemia cells, erythroid leukaemia cells, normal haematopoietic stem cells, and pre-acute phase chronic myelogenous leukaemia cells. This dataset contains 1102 single-cell sequencing data, and it is a labelled dataset that has been manually labelled by biological researchers with cell types. This labelled dataset is more suitable for the subsequent evaluation of the classification performance of the completed matrix. The distribution of the experimental dataset is shown in Figure 2.

In the Figure 2, *a* type represents acute phase chronic myelogenous leukaemia cells, *b* type represents chronic phase chronic myelogenous leukaemia cells, *c* type represents erythroid leukaemia cells, *d* type represents normal haematopoietic stem cells, and *e* type represents pre-acute phase chronic myelogenous leukaemia cells. The distribution of the

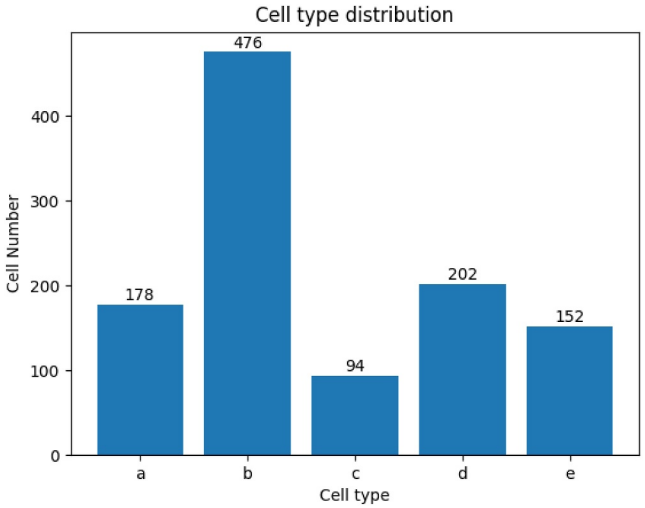

FIGURE 2 Cell type distribution of experimental single-cell sequencing data.

sequencing data was not uniform among different cells. The number of chronic myeloid leukaemia cells in the chronic phase was high, while the number of erythroid leukaemia cells was low. The number of other cell types was basically similar, and there was no extreme cell type with too many or too few cells.

2.3 | Data preprocessing

Because the distribution of human CML single-cell sequencing data is not uniform among cells, in order to ensure that the subsequent classification algorithm will not have a performance gap due to too many or too few cells of a certain type, it is necessary to use the oversampling algorithm to oversample the original five types of cells. The SMOTE [23] oversampling method is not merely a simple replication of the minority class samples. Instead, it is a synthetic sampling technique. It starts

from the minority class samples, identifies neighbouring samples, and generates new synthetic samples. If the original random oversampling technique is used to simply replicate the samples of a few classes, it will only improve the number of samples without improving the quality of samples, and the subsequent data classification performance will not be greatly improved. In the preprocessing, single-cell transcriptome data analysis and processing tool CellRanger were used for transformation, and the SMOTE oversampling method was used for oversampling the uneven data. The number of cells of each type after treatment was 476, and the total number of cells in the dataset was 2380.

2.4 | Gene frequency calculation and sequencing

In order to facilitate subsequent block processing of the gene expression matrix and to distinguish between biological zeros that do not need to be completed and technical zeros that do, the first step is to calculate the gene frequency for all genes in the sequencing data and then sort the gene expression matrix in descending order based on gene frequency. To ensure the accuracy of gene frequency calculation, we conducted statistical tests to evaluate whether there is any bias in gene frequency distribution. The test results indicate that the distribution of gene frequencies is uniform and there is no significant statistical bias, which provides a reliable basis for subsequent matrix completion. The essence of gene expression in the matrix is determined by the occurrence times of different genes in each cell in sequencing. Therefore, the calculation of gene frequency requires the sum of the corresponding gene expression levels of all cells, as shown in Formula 1.

$$G_j = \sum_{i=1}^{m} M_{ij} \tag{1}$$

where M_{ij} represents the gene expression of the jth gene in the ith cell of the original gene expression matrix M, and M is the original gene expression matrix, which is a matrix with m rows and n columns.

G_j represents the sum of the jth gene expressed in all cells. The value can be used to approximate the frequency of the gene, that is, the larger the value of G_j, the larger the proportion of the gene in the sequencing and the higher the importance of the gene in the sequenced cells. The deletion of genes in this column in a cell may be the technical zero point caused by inadequate sequencing technology and should be completed at this time. However, a small value of G_j indicates that the expression of this gene in the original gene expression matrix is very low. When the corresponding gene in this column is missing in a certain cell, it is possible to be a biological zero, that is, matrix completion is not necessary.

After calculating the G_j, the original gene expression matrix needs to be sorted by gene columns, that is, the gene expression matrix M should be reordered in descending order based on G_j. This will place genes with significant expression levels at the top positions in the matrix.

Aiming at the dataset, G_j is calculated and sorted as shown in Figure 3.

As shown in Figure 3, from 50 to 1102, the distribution of genes corresponding to the number of cells expressing them is relatively even, with genes such as MTRNR2L10, H3F3B, and EEF1A1 being expressed in all cells. However, there is a significant increase in the number of genes that are expressed in fewer than 10 cells, with a large number of genes, such as LOC101928107, MIR3663HG, and MIR3649, not being

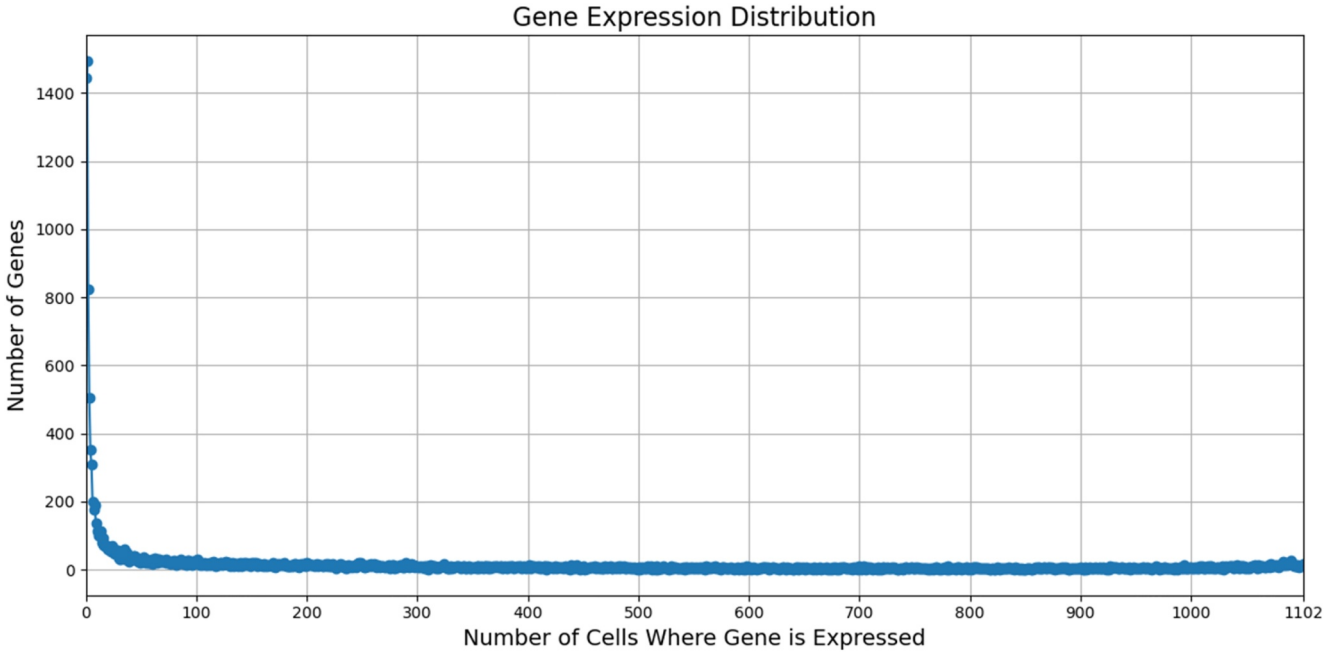

FIGURE 3 Gene expression distribution.

detected in any of the five cell types. Completing these genes could significantly reduce the quality of the completion, making the method of preserving the biological zero point based on gene frequency particularly effective.

2.5 | Block-gene expression matrix

In the previous section, the gene expression matrix is sorted by gene frequency from largest to smallest. In this section, the gene expression matrix is sorted by gene frequency from largest to smallest for block operation. The matrix of original size $m \times n$ is partitioned into k small matrices of the same size by columns, where each small matrix has size $m \times \frac{n}{k}$. Then, each small matrix is completed by an iterative matrix factorisation algorithm. A diagram of matrix block completion is shown in Figure 4.

A critical challenge in matrix blocking lies in the selection of the parameter k. An excessively large value of k will result in submatrices that are too small after blocking, thereby compromising the effectiveness of the matrix completion algorithm. Conversely, an overly small value of k will lead to overly coarse granularity of the blocked matrices, thus diminishing the rationale for blocking based on differences in gene frequencies. Therefore, choosing the appropriate parameter k will have a great influence on the performance of the matrix completion model.

2.6 | Iterative matrix factorisation algorithm for matrix completion

Matrix factorisation algorithm is a common method in a matrix completion field [24]. This algorithm applies matrix completion techniques to small matrices that have already been partitioned, filling in the zeros while preserving and not altering the values of the non-zero elements. The original matrix V is decomposed into the product of two small matrix both W and H. The matrix both W and H are updated based on the difference matrix E. This process continues until the value of E falls below a specified threshold. W and H of the iterative updating formula is as follows:

$$W_{ij} \leftarrow W_{ij} \frac{(EH^T)_{ij}}{(WHH^T)_{ij}} \tag{2}$$

$$H_{ij} \leftarrow H_{ij} \frac{(W^T E)_{ij}}{(W^T WH)_{ij}} \tag{3}$$

The product matrix of W and H of the final iteration can be used to complete the elements of the corresponding position in the matrix. If the corresponding position in the original matrix V is zero, the product of W and H can be used for completion. If the elements of the corresponding position are not zero, the value of the original matrix can be retained. The formula is as follows:

$$V'(i,j) = WH(i,j)(V(i,j) = 0) \tag{4}$$

$$V'(i,j) = V(i,j)(V(i,j) \neq 0) \tag{5}$$

2.7 | Retention of biological zeros in the matrix

There are many biological zeros in the gene expression matrix. These biological zeros indicate that the gene is not expressed in

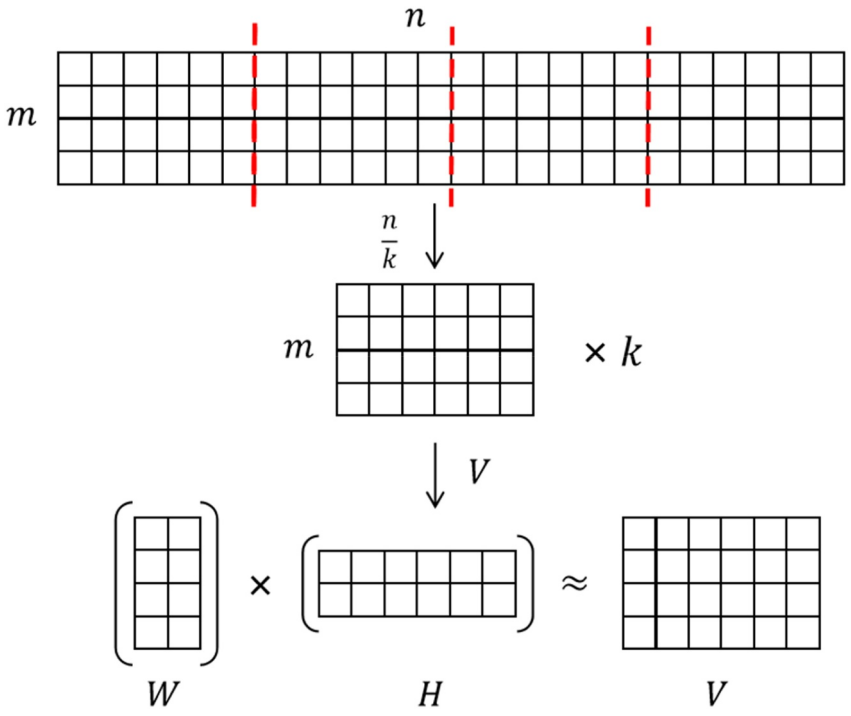

FIGURE 4 Matrix block completion diagram.

TABLE 1 Classification performance for different k values in the block matrix.

k	Accuracy	Precision	Recall	F1-score
1	0.8534	0.8536	0.8534	0.8535
$\frac{\sqrt{n}}{4}$	0.8570	0.8576	0.8588	0.8582
$\frac{\sqrt{n}}{2}$	0.8601	0.8621	0.8596	0.8608
$\sqrt{n}$	**0.8651**	**0.8663**	**0.8631**	**0.8647**
$2\sqrt{n}$	0.8534	0.8536	0.8534	0.8535
$4\sqrt{n}$	0.8325	0.8345	0.8311	0.8328
n	0.8080	0.8277	0.8080	0.8177

Note: The bold values represent the best performances.

TABLE 2 Classification performance for different t values.

t	Accuracy	Precision	Recall	F1-score
$m\sqrt{n}$	0.8080	0.8277	0.8080	0.8177
$\frac{1}{2}m\sqrt{n}$	0.8239	0.8235	0.8241	0.8238
$\frac{1}{4}m\sqrt{n}$	0.8315	0.8365	0.8250	0.8307
$\frac{1}{8}m\sqrt{n}$	0.8512	0.8531	0.8410	0.8470
$\frac{1}{16}m\sqrt{n}$	**0.8672**	**0.8833**	0.8535	**0.8681**
$\frac{1}{32}m\sqrt{n}$	0.8632	0.8675	0.8510	0.8591
0	0.8651	0.8663	**0.8631**	0.8647

Note: The bold values represent the best performances.

the cell. In this case, the zeros in the matrix are biological zeros, which are not caused by the insufficiency of sequencing technology but should be zero in the first place. These biological zeros should not be completed but should be retained in the final matrix.

In the gene expression matrix, the expression level of a gene is actually determined by the frequency of the gene, and the cells to be sequenced are generally the same tissue or organ. Therefore, when a gene is missing in a cell and the gene is highly expressed in other cells, it is highly likely to be the technical zero point caused by technical reasons and needs to be completed. On the contrary, if a gene is missing in a cell and its expression is low or not expressed at all in other cells, the zero point at this time is very likely to be a biological zero, and no completion is needed. The zero point can be retained in the completed matrix. Since the gene expression matrix has been sorted and partitioned previously, it only needs to consider the gene frequencies of each small matrix after being partitioned when the zeros are reserved. The zeros of genes with lower frequencies in the small matrix after partitioning were reserved, and the matrix decomposition algorithm was not used for completion. Set the threshold as t. When the number of genes expressed in the whole matrix is less than t, the zeros in the small matrix after the block are considered to be biological zeros with high probability and need not be completed. Since the size of the whole small matrix is $m\sqrt{n}$, this paper tests the performance of the matrix classification algorithm when t is $m\sqrt{n}$, $\frac{1}{2}m\sqrt{n}$, $\frac{1}{4}m\sqrt{n}$, $\frac{1}{8}m\sqrt{n}$, $\frac{1}{16}m\sqrt{n}$, $\frac{1}{32}m\sqrt{n}$, 0, respectively. The experimental results are shown in Table 2.

3 | RESULTS AND DISCUSSIONS

3.1 | Importance of gene features

An intuitive judgement of the matrix completion effect is the change of the importance of features in the matrix. For the gene expression matrix, it is the change of gene importance in the matrix. The top 20 gene importance scores of the human chronic myelogenous leukaemia single-cell sequencing dataset before and after completion are shown in Figure 5.

Figure 5 shows the changes in gene ranking, which may be caused by multiple factors. For example, changes in gene expression levels, specific expression of cell types, and unclear differentiation between technical and biological zeros can all lead to changes in gene rankings. We further analysed the biological functions of these genes and their expression patterns in different cell types and found that genes with higher rankings are often closely related to cell differentiation and disease status, while genes with lower rankings may be universally expressed in multiple cell types, lacking specificity [25].

The gene score in Figure 4 above is determined by the expression levels of corresponding genes in the matrix. It can be seen from Figure 5 that the top three highest-scoring genes in the matrix before and after completion are the same, all of which are LYZ, S100A9 and TCL1A. The scores of these three genes are all greater than 40, far exceeding other genes. At the same time, because the completion algorithm made many technical zeros in these three genes completely, the scores of all three genes became higher. The gene with the fourth characteristic score before and after completion was different. For the matrix before completion, the gene with the fourth characteristic score was S100A4, while for the matrix after completion, it was S100A6. After completion, the S100A4 gene ranking dropped one place to fifth. The top 20 genes showed changes in rank and score, and the top genes also improved their score. This indicates that the iterative block matrix completion algorithm based on matrix factorisation has an obvious completion effect on genes with high expression (often the marker genes).

3.2 | SMOTE analysis

SMOTE has certain advantages in addressing data imbalance problems. By generating new samples through interpolation, it effectively expands the minority class data, thereby improving the model's ability to learn from imbalanced data and enhancing overall prediction performance [26]. However, in the context of gene expression matrices, SMOTE also has some limitations. Firstly, SMOTE may compromise the biological authenticity of the data, as the interpolated samples

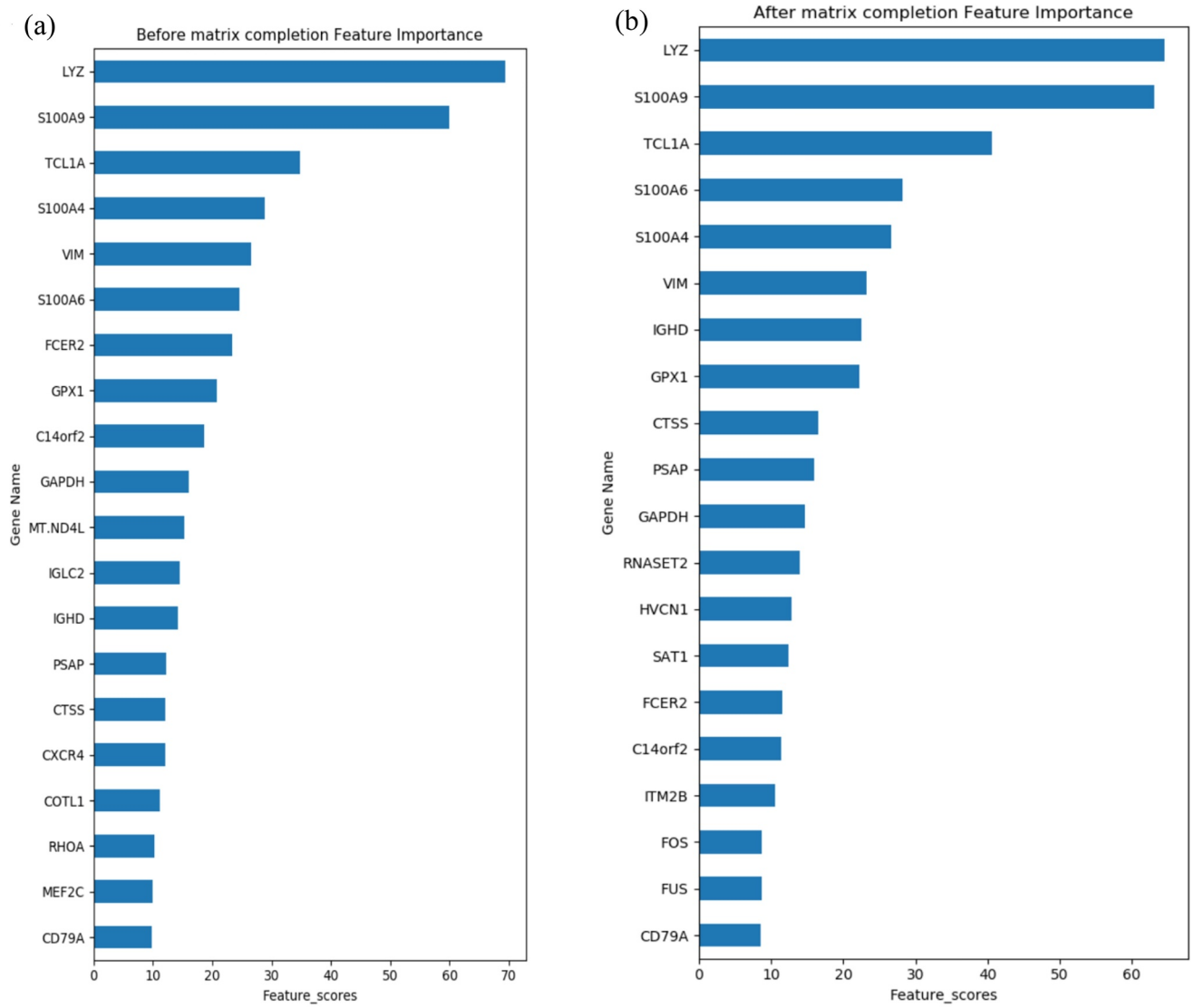

FIGURE 5 (a) Importance score of top 20 genes before matrix completion. (b) Importance score of top 20 genes after matrix completion.

might lack actual biological significance, potentially introducing noise or bias that affects the model's learning outcomes. Additionally, this method does not account for the unique gene correlations or structural characteristics inherent in gene expression matrices, which could result in synthetic samples deviating from the true data distribution, failing to fully capture the characteristics of real data.

3.3 | Performance comparisons

In this paper, standard machine learning classification performance evaluation indicators are used to evaluate the experimental results, including accuracy, precision, recall, and comprehensive evaluation value F1 − score. These metrics are used to evaluate the performance of different classification models, and the formulas of these metrics are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} - \text{score} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FN} + \text{FP}} \tag{9}$$

TP is true positive, which refers to the samples that are correctly predicted as positive classes, where the true value is also the number of positive classes; FN is false negative, which refers to the number of positive classes in the samples incorrectly predicted to be negative. FP is a false positive, which refers to the number of negative classes in the samples incorrectly predicted as positive classes. TN is true negative,

which refers to the number of negative classes among the samples correctly predicted as negative classes.

In order to optimise the parameters k and t, we conducted a series of experiments. The parameter k controls the size of the matrix block, directly affecting the efficiency and effectiveness of the completion algorithm. We tested different values of k, ranging from 1 to n, and recorded the classification performance in each case. By comparing the F1 score at different k values, we found that the model achieved the best classification performance when k is equal to $\sqrt{n}$. The parameter t is used to determine the threshold for retaining biological zeros. We tested different values of t from 0 to $m\sqrt{n}$ and analysed their impact on the completion matrix. The experimental results show that the model can effectively complete the technical zeros while retaining the biological zeros with the $\frac{1}{16}m\sqrt{n}$ of t value, thereby improving the accuracy of classification. The details and data of these parameter selections are as follows:

Choosing the appropriate parameter k will have a great influence on the performance of the matrix completion model. Since the total number of genes in the gene expression matrix is n, this paper tests the cell classification performance of the gene expression matrix after block completion when k is 1, $\frac{\sqrt{n}}{4}$, $\frac{\sqrt{n}}{2}$, $\sqrt{n}$, $2\sqrt{n}$, 4 $\sqrt{n}$, n (n is the number of columns of the gene expression matrix, that is, the number of all genes). In this paper, the random forest classification algorithm is used to evaluate the performance of the above test k values and select the optimal solution of the parameter k values. Table 1 shows the classification performance for different k values in the block matrix. The bold values represent the best performances in Tables.

From the above experimental results in Table 1, it can be seen that when the value of k is $\sqrt{n}$, the performance of the matrix partitioning algorithm reaches the optimum. At the same time, it is worth noting that when $k = 1$, it is equivalent to the iterative matrix completion algorithm of matrix factorisation without the matrix partitioning algorithm. When $k = n$, each column of the original matrix is segmented once, and the segmented matrix is a long vertical bar matrix with m rows and 1 column. The completion of this matrix by the iterative matrix factorisation algorithm will be invalid. Therefore, at this time, the original matrix is directly classified without any completion operation.

The random forest algorithm is used to measure the classification performance of different t values. The experimental results are shown in Table 2.

According to the experimental results, when t is $\frac{1}{16}m\sqrt{n}$, the accuracy, precision and F1 score of the algorithm are the highest, and the recall rate is also high. Therefore, the threshold t is chosen as $\frac{1}{16}m\sqrt{n}$, that is, when the number of non-zero values in the small matrix after the block is less than $\frac{1}{16}m\sqrt{n}$, it is considered that there is a high possibility of biological zeros in the matrix and zeros are reserved. It is worth noting that when t is $m\sqrt{n}$, it means that all the block matrices retain zeros, which is equivalent to the random forest classification of the original matrix without the completion operation. When t is 0, it means that the zero preservation operation is not required for all matrices. In this case, it is equivalent to only completing the matrix after the block without a biological zero preservation operation. Finally, the pieced small matrix after completion is concatenated, and the large matrix obtained by concatenation is the final result of matrix completion.

In this study, t represents the number of gene expressions within each subdivided small matrix. When the matrix completion operation is applied to small matrices with low gene expression levels, it may lead to the misjudgement of biological zeros, thereby affecting the overall completion performance. To improve predicted accuracy, this study proposes setting a threshold for the t-value to determine whether completion should be performed. For regions with low gene expression levels, a no-completion strategy is adopted, which significantly enhances predicted accuracy. However, even within small matrices with low gene expression levels, a certain number of non-biological zeros may still exist. As shown in the analysis of Table 2, when $t = \frac{1}{16}m\sqrt{n}$, the best balance between completing non-biological zeros and avoiding the completion of biological zeros can be achieved. This result indicates that appropriately setting the t-value threshold not only optimises the completion performance but also meets the practical needs of biological data analysis.

High-dimensional datasets typically require dimensionality reduction techniques to map them onto a lower dimensional subspace, thereby enabling effective visualisation. In this study, we employ the T-SNE method to compare the completion results of SeqBMC, aiming to assess their impact on the visualisation of the GSE76312 dataset. The visualisation results are shown in Figure 6.

Figure 6 illustrates the two-dimensional distribution of data before and after imputation using T-SNE dimensionality reduction. From the visualisation of the original data, it can be observed that the overall structure in the low-dimensional space appears relatively scattered, with indistinct differences between categories and poor clustering performance. In contrast, after imputation, the clustering performance improves significantly, with data points from different categories forming clearer clusters and more distinct boundaries. This demonstrates that the SeqBMC method has significant advantages in completing gene expression matrices, effectively improving the data structure and enhancing clustering performance.

Using the single-cell sequencing dataset of human chronic myelogenous leukaemia, the number of cells in the original dataset was 1102, which was divided into 5 types of cells. The SMOTE oversampling technique was used to make the number of different types of cells consistent, and the final dataset was 2380 cells. The number of genes in the gene expression matrix is 11,235. To test the performance of the matrix completion algorithm, the dimensionality reduction of features will reduce the performance gap of different completion algorithms, so the principal components analysis (PCA) is no

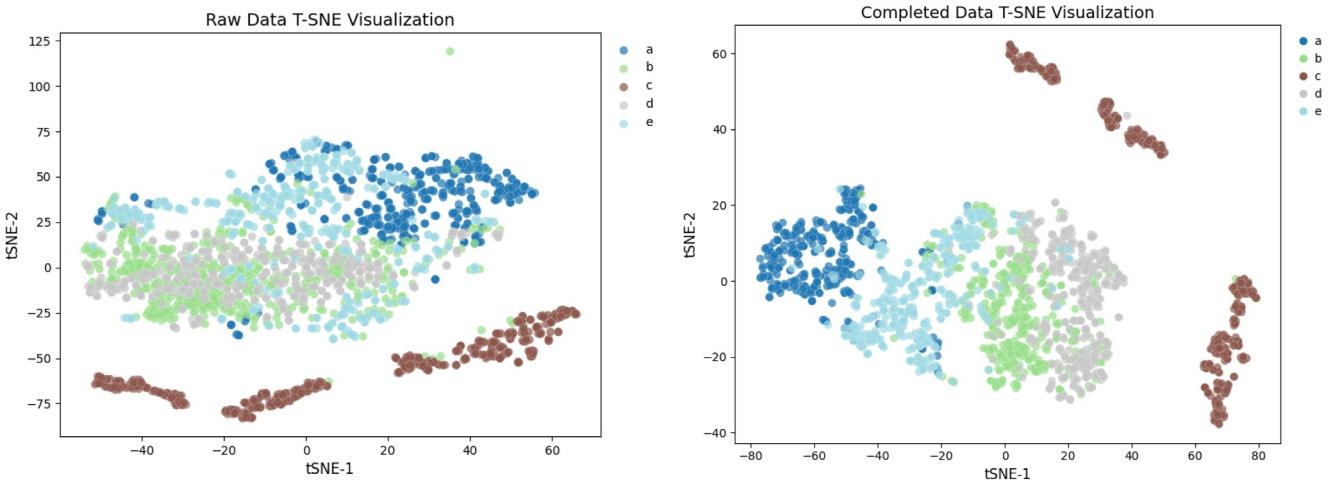

FIGURE 6 T-SNE visualisation in raw data and completed data.

longer used for the dimensionality reduction of the gene expression matrix.

The experiment uses a random forest classifier to compare the classification performance of several matrix completion algorithms. Several completion algorithms include ALRA [16], SAVER [17], DrImpute [27], scWMC [28]and SeqBMC. ALRA algorithm uses the characteristics of the gene expression matrix to complete a large number of "zero-preserving" matrix, which is by far the most widely used single-cell RNA gene expression matrix completion algorithm. SAVER algorithm is by far the highest performance of the single-cell RNA gene expression matrix completion algorithm, and this algorithm by getting different relations between genes and genes, to estimate the exact expression of all genes in the matrix and restore the true expression level of each gene in each cell, eliminated the technical differences, while preserving the biological differences across the cell. A strong collection of biological information is required. DrImpute interpolates dropout events in single-cell RNA sequencing data by considering the correlation between cells. It first identifies similar cells based on clustering results and then interpolates through the average expression values of similar cells. scWMC is a method based on weighted matrix completion, which utilises prior subspace information to complete single-cell RNA sequencing data. By using a weighting strategy, it may better handle noise and outliers in the data and improve the accuracy of completion. SeqBMC is an iterative block matrix completion algorithm based on matrix factorisation. The algorithm combines matrix partitioning, matrix factorisation and biological zeros and distinguishes biological zeros from technical zeros in the completed gene expression matrix. At the same time, the frequency of genes in the gene expression matrix was used to complete the matrix in blocks and preserve the biological zeros of genes. Finally, the machine learning algorithm was used to complete the matrix. Performance comparisons between SeqBMC and different methods are shown in Table 3.

The baseline is the result of using the random forest [29] directly on the original data. It can be seen from the experimental data that ALRA algorithm has a poor completion effect

TABLE 3 Performance comparisons between SeqBMC and different methods.

Methods	Accuracy	Precision	Recall	F1-score
Baseline	0.8080	0.8277	0.8080	0.8177
ALRA	0.8125	0.8125	0.8232	0.8178
SAVER	0.8671	0.8723	**0.8614**	0.8668
DrImpute	0.8533	0.8536	0.8533	0.8525
scWMC	0.8587	0.8621	0.8587	0.8540
SeqBMC	**0.8672**	**0.8833**	0.8535	**0.8681**

Note: The bold values represent the best performances.

because it retains too many biological zeros. The proposed SeqBMC algorithm in cell classification was 86.72%, 88.33%, 85.35% and 86.81% with the accuracy, precision, recall and F1-score, respectively. After completion, compared with the baseline, the accuracy of cell classification was increased by 5.92%, the accuracy was increased by 5.56%, the recall rate was increased by 4.55%, and the F1-score was increased by 5.04%. The proposed algorithm has the highest performance among all completion algorithms in terms of accuracy and F1-score, and the performance of the proposed algorithm is second only to the SAVER algorithm in terms of recall. Compared with the SAVER algorithm, the biggest advantage of the proposed algorithm is that it can be completed only by computational means without a lot of prior biological knowledge. The advantage of SeqBMC over DrImpute lies in its use of a gene frequency-based blocking strategy, which enables more accurate differentiation between the technical zero and biological zero during the completion process. This method helps to preserve more biological information as it only completes data points lost due to technical limitations rather than all zero values, thereby improving the biological relevance of the data and the accuracy of analysis. The advantage of SeqBMC over scWMC lies in its iterative matrix factorisation algorithm, which makes it more efficient in processing small block matrices and capturing local features of single-cell data. In

addition, SeqBMC does not require a large amount of prior biological knowledge, which gives the advantages in automation and computational efficiency, especially when processing large-scale single-cell RNA sequencing datasets. Therefore, the experiment shows that the SeqBMC algorithm proposed in this paper has better performance in the completion of the single-cell RNA gene expression matrix and can effectively produce better classification results for cell classification algorithms, which is helpful for the downstream data analysis of single-cell RNA sequencing data.

4 | CONCLUSIONS

Based on the matrix factorisation algorithm, an iterative block matrix completion algorithm SeqBMC is proposed in this paper. The approach uses the gene frequency of the matrix to block the matrix uses the matrix factorisation algorithm to complete the small matrix after the block, and then retains the biological zeros that may exist in the block matrix. Experimental results show that the matrix completion algorithm can significantly improve the classification performance of the gene expression matrix after completion, which is conducive to the identification of cell types in sequencing data. Moreover, this proposed completion method can be completed only by computational means without too much prior knowledge related to biology and has a good effect. Compared with ALRA, SeqBMC increased 5.47% accuracy and 5.03% F1-score, while compared with SAVER, it increased 0.01% accuracy and 0.13% F1-score. These improvements highlight the significant advantages of SeqBMC in matrix completion for single-cell RNA sequencing data. Thus, SeqBMC is promising for the identification of single-cell subtype and provides a way for classification of single-cell subtype, which is helpful for the downstream data analysis of single-cell RNA sequencing data.

AUTHOR CONTRIBUTIONS

Gong Lejun: Conceptualisation; writing original draft. **Yu Like**: Data analysis; revising the work. **Wei Xinyi**: Data analysis; reviewing the work. **Zhou Shehai**: Writing original draft. **Xu Shuhua**: Funding acquisition; writing—review and editing.

ACKNOWLEDGEMENTS

This research is supported by the Open Research Fund of State Key Laboratory of Digital Medical Engineering (Grant Nos. 2024-M10), Natural Science Foundation of Zhejiang Province under Grant No. LGG22F020040, the Wenzhou Scientific Research Project (Grant No. ZG2024013), and Natural Science Foundation of Nanjing University of Posts and Teleccommunications (Grant No. NY223093).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The experimental data related to single-cell sequencing data of human chronic myelogenous leukaemia is from the gene expression database of National Centre for Biotechnology Information (NCBI) (Data key number GSE76312), which is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76312.

ORCID

Gong Lejun https://orcid.org/0000-0001-7971-3000

REFERENCES

1. Papalexi, E., Satija, R.: Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. 18(1), 35–45 (2018). https://doi.org/10.1038/nri.2017.76
2. Rantalainen, M.: Application of single-cell sequencing in human cancer. Briefings in Functional Genomics 17(4), 273–282 (2018). https://doi.org/10.1093/bfgp/elx036
3. Potter, S.S.: Single-cell RNA sequencing for the study of development, physiology and disease. Nat. Rev. Nephrol. 14(8), 479–492 (2018). https://doi.org/10.1038/s41581-018-0021-7
4. Svensson, V., et al.: Exponential scaling of single-cell RNA-seq in the past decade. Nat. Protoc. 13(4), 599–604 (2018). https://doi.org/10.1038/nprot.2017.149
5. Hu, C., et al.: CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 51(D1), D870–D876 (2023). https://doi.org/10.1093/nar/gkac947
6. Lu, J., et al.: scRNA-seq data analysis method to improve analysis performance. IET Nanobiotechnol. 17(3), 246–256 (2023). https://doi.org/10.1049/nbt2.12115
7. Vieth, B., et al.: A systematic evaluation of single cell RNA-seq analysis pipelines. Nat. Commun. 10(1), 4667 (2019). https://doi.org/10.1038/s41467-019-12266-7
8. Chen, G., Ning, B., Shi, T.: Single-cell RNA-seq technologies and related computational data analysis. Front. Genet. 10, 317 (2019). https://doi.org/10.3389/fgene.2019.00317
9. Wang, B., et al.: Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. Sci. Rep. 11(1), 1141 (2021). https://doi.org/10.1038/s41598-020-80881-2
10. Malec, M., Kurban, H., Dalkilic, M.: ccImpute: an accurate and scalable consensus clustering based algorithm to impute dropout events in the single-cell RNA-seq data. BMC Bioinf. 23(1), 291 (2022). https://doi.org/10.1186/s12859-022-04814-8
11. Azim, R., Wang, S., Dipu, S.A.: CDSImpute: an ensemble similarity imputation method for single-cell RNA sequence dropouts. Comput. Biol. Med. 146, 105658 (2022). https://doi.org/10.1016/j.compbiomed.2022.105658
12. Ting, D.T., et al.: Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep. 8(6), 1905–1918 (2014). https://doi.org/10.1016/j.celrep.2014.08.029
13. Wang, C.Y., et al.: Unsupervised cluster analysis and gene marker extraction of scRNA-seq data based on non-negative matrix factorization. IEEE J. Biomed. Health Inf. 26(1), 458–467 (2021). https://doi.org/10.1109/jbhi.2021.3091506
14. Senra, D., Guisoni, N., Diambra, L.: ORIGINS: a protein network-based approach to quantify cell pluripotency from scRNA-seq data. MethodsX 9, 101778 (2022). https://doi.org/10.1016/j.mex.2022.101778
15. Soemartojo, S.M., et al.: Iterative bicluster-based Bayesian principal component analysis and least squares for missing-value imputation in microarray and RNA-sequencing data. Math. Biosci. Eng. 19(9), 8741–8759 (2022). https://doi.org/10.3934/mbe.2022405
16. Linderman, G.C., Zhao, J., Kluger, Y.: Zero-preserving imputation of scRNA-seq data using low-rank approximation. bioRxiv, 397588 (2018)
17. Huang, M., et al.: SAVER: gene expression recovery for single-cell RNA sequencing. Nat. Methods 15(7), 539–542 (2018). https://doi.org/10.1038/s41592-018-0033-z
18. Jin, K., et al.: scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation. Bioinformatics

36(10), 3131–3138 (2020). https://doi.org/10.1093/bioinformatics/btaa108
19. Chen, C., et al.: scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. Bioinformatics 36(10), 3156–3161 (2020). https://doi.org/10.1093/bioinformatics/btaa139
20. Tang, W., et al.: bayNorm: bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. Bioinformatics 36(4), 1174–1181 (2020). https://doi.org/10.1093/bioinformatics/btz726
21. Talwar, D., et al.: AutoImpute: autoencoder based imputation of single-cell RNA-seq data. Sci. Rep. 8(1), 16329 (2018). https://doi.org/10.1038/s41598-018-34688-x
22. Giustacchini, A., et al.: Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat. Med. 23(6), 692–702 (2017). https://doi.org/10.1038/nm.4336
23. Chawla, N.V., et al.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357 (2002). https://doi.org/10.1613/jair.953
24. Zuo, Z.L., et al.: Double matrix completion for circRNA-disease association prediction. BMC Bioinf. 22(1), 307 (2021). https://doi.org/10.1186/s12859-021-04231-3
25. Zhao, S., et al.: A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. Nat. Genet. 55(2), 346–354 (2023). https://doi.org/10.1038/s41588-022-01278-7
26. Elreedy, D., Atiya, A.F., Kamalov, F.: A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach. Learn. 113(7), 4903–4923 (2024). https://doi.org/10.1007/s10994-022-06296-4
27. Gong, W., et al.: DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinf. 19, 1–10 (2018). https://doi.org/10.1186/s12859-018-2226-y
28. Su, Y., et al.: scWMC: weighted matrix completion-based imputation of scRNA-seq data via prior subspace information. Bioinformatics 38(19), 4537–4545 (2022). https://doi.org/10.1093/bioinformatics/btac570
29. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20(8), 832–844 (1998). https://doi.org/10.1109/34.709601

How to cite this article: Lejun, G., et al.: SeqBMC: Single-cell data processing using iterative block matrix completion algorithm based on matrix factorisation. IET Syst. Biol. e70003 (2025). https://doi.org/10.1049/syb2.70003