Briefings in Bioinformatics, 2024, **25(3)**, bbae228
https://doi.org/10.1093/bib/bbae228
Problem Solving Protocol

scMNMF: a novel method for single-cell multi-omics clustering based on matrix factorization

Yushan Qiu[1], Dong Guo[1], Pu Zhao[2], Quan Zou[3,*]

[1]School of Mathematical Sciences, Shenzhen University, 518000, Guangdong, China
[2]College of Life and Health Sciences, Northeastern University, Shenyang, 110169, China
[3]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 610056, China
[*]Corresponding author. Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610056, China.
E-mail: zouquan@nclab.net

Abstract

Motivation: The technology for analyzing single-cell multi-omics data has advanced rapidly and has provided comprehensive and accurate cellular information by exploring cell heterogeneity in genomics, transcriptomics, epigenomics, metabolomics and proteomics data. However, because of the high-dimensional and sparse characteristics of single-cell multi-omics data, as well as the limitations of various analysis algorithms, the clustering performance is generally poor. Matrix factorization is an unsupervised, dimensionality reduction-based method that can cluster individuals and discover related omics variables from different blocks. Here, we present a novel algorithm that performs joint dimensionality reduction learning and cell clustering analysis on single-cell multi-omics data using non-negative matrix factorization that we named scMNMF. We formulate the objective function of joint learning as a constrained optimization problem and derive the corresponding iterative formulas through alternating iterative algorithms. The major advantage of the scMNMF algorithm remains its capability to explore hidden related features among omics data. Additionally, the feature selection for dimensionality reduction and cell clustering mutually influence each other iteratively, leading to a more effective discovery of cell types. We validated the performance of the scMNMF algorithm using two simulated and five real datasets. The results show that scMNMF outperformed seven other state-of-the-art algorithms in various measurements.
Availability and implementation: scMNMF code can be found at https://github.com/yushanqiu/scMNMF.

Keywords: *single-cell multi-omics; non-negative matrix factorization; joint learning*

INTRODUCTION

Cells, which are the basic components of all living things, perform vital biological functions that keep life running normally. Numerous unique cell types with varying roles make up multicellular organisms. The structure and function of tissues can be understood by classifying the different cell types [1]. However, because an exact and uniform description of cell types is still lacking, identifying cell types remains difficult. Further advancements in bioinformatics have enabled the analysis of transcriptome data of individual cells. Single-cell transcriptomic analysis technology dates back to 2009 [2] and has been used extensively to analyze single-cell RNA sequencing (scRNA-seq) data. The accumulation of scRNA-seq data offers the chance to investigate the characteristics and actions of diverse cell types, and scRNA-seq methods for cell classification have been developed based on transcriptomic similarity. However, the numerous drawbacks of scRNA-seq make cell clustering very challenging [3]. In particular, sequencing technologies for single-cell transcriptome data pose certain dimensionality issues, because they disrupt the distinct expression profiles within an entire genome. Therefore, relying solely on intercellular similarity is one-sided, because the distance differences would have already changed. The analysis tools for single-cell multi-omics data have become increasingly mature and have provided significant impetus for biological research. Single-cell studies differ from traditional bulk methods in that they can handle data heterogeneity. Clustering, as an unsupervised learning method [4], can group similar single cells into the same cluster, thereby detecting differences and similarities among cells [5], and inferring biological information. In most single-cell studies, clustering analysis is a crucial and well-researched phase. Biological processes can be investigated by clustering data, which cannot be achieved by studying large datasets alone [6]. However, clustering analysis for single-cell multi-omics data remains a challenging problem. With the rapid development of various clustering analysis techniques, existing algorithms can be broadly classified as early fusion, late fusion and intermediate fusion algorithms.

Yushan Qiu is an associate professor in Shenzhen University. Her research interest are bioinformatics and machine learning.
Dong Guo is a graduate student in Shenzhen University. His research interests include bioinformatics and machine learning.
Pu Zhao is an associate professor in Northeastern University. Her majors in cell biology and neurobiology.
Quan Zou is a professor in the University of Electronic Science and Technology of China. He is a senior member of IEEE and ACM. He majors in bioinformatics, machine learning and algorithms.

Received: January 31, 2024. **Revised:** April 2, 2024. **Accepted:** April 22, 2024

Early fusion is the simplest of the three methods, where all omics matrices are concatenated, and single-omics clustering algorithms are applied to the merged matrix [7, 8]. Representative methods of early fusion include ICluster [9], which concatenates multi-omics data and then uses probabilistic modeling to reduce the dimensionality of the data, and LRACluster [10], which builds a low-rank approximation model based on multi-omics data to identify their common subspace and subsequently perform clustering operations. However, early fusion methods have several drawbacks: (i) if not correctly standardized, they may give greater weight to omics that have greater characteristics, (ii) they ignore the various data distributions across various omics and (iii) in situations where the data dimensionality is already high, they further increase the data dimensionality, making it more difficult to differentiate between samples.

Late fusion involves partitioning using different clustering methods on different omics datasets. The advantage of this method is that any clustering algorithm can be applied to each omics dataset [11, 12]. However, the disadvantage is that if the clustering results are obtained by integration, the weaker information in each omics dataset may be lost. Late fusion methods include SAME-clustering [11], which is based on graph partitioning for clustering integration, and KLIC [13], which integrates clustering through kernel learning.

Intermediate fusion is distinct from the other two fusion methods and primarily has three categories. Methods in the first category use similarity-based methods, where clustering is performed by constructing a similarity matrix between data points. CiteFuse [14] is one such method that calculates similarity matrices among different omics data, then merges the matrices using a fusion method and finally clusters the merged similarity matrix. Seurat [15] is another method in this category. Methods in the second category are based on dimensionality reduction model algorithms, which assume that there is an invariant low-dimensional structure between the data, and the low-dimensional structures correspond to the number of clusters. Methods in this category include CoupledNMF [16], MOFA+ [17] and TotalVI [18]. The third category of methods is statistical algorithms based on data modeling. The advantage of these models is that they can utilize biological information while determining the distribution functions. Representative methods in this category include BREM-SC [19] and Clonealign [20]. Similarity-based methods cannot explicitly consider dropout events and, for different omics data, they simply concatenate the data features or remember them directly, and therefore cannot fully exploit the information between different omics. Most statistical-based methods require data to have a specific distribution, but this may not necessarily be applicable in practical applications.

Inspired by the problems associated with the methods described above, we developed a non-negative matrix factorization (NMF)-based method that differs from the traditional approach of separate dimensionality reduction and clustering processes. The aim was to better analyze the intrinsic correlations between different omics data and investigate the mutual influences between joint dimensionality reduction and clustering. To extract the intrinsic correlations between different omics data, we used NMF to process the data and obtain their common feature matrix. Then, we employed joint dimensionality reduction to study the clustering results using different omics datasets. By integrating dimensionality reduction and clustering analysis, our algorithm demonstrates the advantage of iterative influence between feature selection in dimensionality reduction and cell clustering, leading to a significant improvement in clustering performance.

MATERIALS AND METHODS

Model description

We take M matrices $\mathbf{X}_1, \ldots \mathbf{X}_M$ as inputs, representing M omics datasets. $\mathbf{X}_k \in R^{n \times J_k}$, where n represents the count of cells and J_k represents the count of features for block k. In our method, we offer a model built using methods for matrix factorization, i.e.

$$X_k \approx WH_k, \tag{1}$$

where $\mathbf{W} \in R^{n \times p}$ is a shared basis matrix that allows simultaneous clustering of cells across M omics blocks. $\boldsymbol{H_k} \in R^{p \times J_k}$ is the coefficient matrix for block k. The quantity of latent variables in this model is denoted by the variable p. Here, we impose non-negativity and sparsity constraints on matrix $\mathbf{W}$ to ensure interpretability of the model. H_k is subject to a sparsity constraint to carry out variable selection concurrently with cell clustering. A deeper understanding of the factors influencing the cell clustering is ensured by sparsity. Then, the following is an extension of Equation (1):

$$\begin{gathered} R_{W,H_1,\ldots,H_M} = \sum_{k=1}^{M} \|X_k - WH_k\|^2 + \lambda_k \|H_k\|_1 + \sum_{i=1}^{n} \mu_i \|w_{i\cdot}\|_1 \\ \text{s.t. } W \geqslant 0, \end{gathered} \tag{2}$$

where $w_{i\cdot}$ is the ith row, $\|\cdot\|$ is the Frobenius parametrization and $\|\cdot\|_1$ is the l_1-norm. Both λ_k and μ_i are sparse constraint parameters. B and F are both non-negative, and their product is used to further estimate the decomposition matrix W to gain the possible characteristics of every single cell.

$$W \approx BF, \quad \text{s.t.} B \geqslant 0, F \geqslant 0, \tag{3}$$

where $\boldsymbol{B} \in R^{n \times k_1}$ and $\boldsymbol{F} \in R^{k_1 \times p}$ are the feature and basis matrix, and k_1 is the number of cell types. Equation (3) can be solved by minimizing the approximation, i.e.

$$Q_{B,F} = \|W - BF\|^2 , \ \text{s.t.} B \geqslant 0, F \geqslant 0. \tag{4}$$

Nevertheless, NMF is unable to unveil the intrinsic geometric arrangement of cells, such as manifold embeddings within high-dimensional spaces. We expect $\boldsymbol{B}$ to preserve the inherent geometric structure of matrix $\mathbf{W}$ [21]. In particular, if two cells, denoted as 'c_i' and 'c_j', are close together in the initial space, their representations in the matching space, i.e. s_i and s_j, should also exhibit proximity, and vice versa. We construct a Laplacian graph $\mathbf{G}$ to represent the close relationships between cells contained in matrix $\mathbf{W}$. The vertices of graph $\mathbf{G}$ describe cells, while the edges represent the similarity between two vertices. Let $\boldsymbol{A_{ij}}$ denote the weight matrix, and given that $\boldsymbol{U_p}(c_i)$ represents the p nearest neighbors of c_i, we have

$$A_{ij} = \begin{cases} e^{-\frac{\|c_i - c_j\|^2}{\sigma}}, & c_i \in U_p(c_j) \text{ or } c_j \in U_p(c_i) \\ 0, & \text{others} . \end{cases} \tag{5}$$

We denote $L = D - A$ as the Laplacian matrix of graph $\mathbf{G}$, where $D_{ii} = \sum_j A_{ij}$ is the diagonal matrix. Cai *et al.* [22] demonstrated that trace optimization can be used to define the local topological structure preservation, i.e.

$$\frac{1}{2}\sum_{i,j} a_{ij}||s_i - s_j||^2 = \mathbf{Tr}\left(BLB^T\right). \tag{6}$$

Thus, binding the representation of Equation (6), the objective function of clustering can be defined as

$$Q_{B,F} = \|W - BF\|^2 + \alpha\, \mathbf{Tr}\left(BLB^T\right), \text{s.t.} B \geqslant 0, F \geqslant 0. \tag{7}$$

and α decides the significance of the regularization. Finally, by combining Equations (2) and (7), we obtain the final objective function of the joint analysis model as

$$\begin{aligned} \min O =& R_{W,H_1,\cdots H_M} + Q_{B,F} \\ =& \sum_{k=1}^{M} \|X_k - WH_k\|^2 + \lambda_k \|H_k\|_1 + \sum_{i=1}^{n} \mu_i \|w_i\|_1 \\ &+ \|W - BF\|^2 + \alpha\, \mathbf{Tr}\left(BLB^\top\right) \\ &\text{s.t.} W \geqslant 0, B \geqslant 0, F \geqslant 0. \end{aligned} \tag{8}$$

The framework of the method we constructed is illustrated in Figure 1.

Optimization of the algorithm

Taking into account the non-convexity of the objective function, we employ an alternating iterative algorithm to optimize it, by fixing all other variables and making the optimization problem of the remaining variable convex. The algorithm continues to iterate until it converges or reaches a stopping condition. The update details of Equation (8) are provided in Supplementary Section 1 (see Supplementary Data available online at http://bib.oxfordjournals.org/). The convergence analysis of the updating rule is provided in Supplementary Section 2 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

Related parameters

scMNMF has a total of five parameters. Parameter p is the count of features after dimensionality reduction, k_1 is the count of cell clusters and λ_k, μ_i, and α determine the importance of the corresponding sparsity constraint and regularization. The method for selecting these five parameters is detailed in Supplementary Section 3 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

PERFORMANCE EVALUATION

Accuracy

ACC is a commonly used performance metric in machine learning and data analysis, which measures the percentage of correctly classified instances in a dataset.

$$ACC = \max_{m} \frac{\sum_{i=1}^{k} 1\left\{l_i = m\left(u_i\right)\right\}}{k}, \tag{9}$$

where k represents the count of cells, l_i represents the true labels and $m(\cdot)$ denotes all possible one-to-one mappings between cluster assignments and true labels.

Adjusted rand index

Adjusted rand index (ARI) is a metric used to measure the consistency between clustering results and true labels, and its definition involves the matching of pairs of samples in the clustering results.

$$ARI = \frac{\begin{pmatrix} k \\ 2 \end{pmatrix}(s+n) - [(s+t)(s+m) + (n+t)(n+m)]}{\begin{pmatrix} k \\ 2 \end{pmatrix} - [(s+t)(s+m) + (n+t)(n+m)]}, \tag{10}$$

where k is the total number of cells, s is the number of cell pairs with the same cell label corresponding to the same cluster, t is the number of cell pairs with different cell labels corresponding to different clusters, m is the number of cell pairs with the cell labels corresponding to different clusters and n is the number of cell pairs with different cell labels corresponding to the same cluster.

Normalized mutual information

Mutual Information (MI) is defined as follows:

$$MI = \sum_{s}\sum_{t} f(s,t) \log \frac{f(s,t)}{f(s)f(t)}, \tag{11}$$

where $f(s,t)$ is the joint distribution of random variables S and T, and $f(s)$ and $f(t)$ are the marginal distributions of S and T. NMI is defined as

$$NMI = \frac{2MI}{h(P) + h(Q)}, \tag{12}$$

where $h(P)$ and $h(Q)$ represent the entropy of the clustering result, denoted as P, and the true labels, denoted as Q, respectively.

Adjusted mutual information

Adjusted mutual information (AMI) measures the similarity between cluster assignments and true labels, defined as

$$AMI = \frac{MI - E[MI]}{\max(h(P), h(Q)) - E[MI]}, \tag{13}$$

where MI is defined as above, where $E[MI]$ is its mathematical expectation, and $h(P)$ and $h(Q)$ are the entropies of the predicted labels and true labels, respectively.

EXPERIMENTAL RESULTS

Datasets

We used two simulated and five real datasets to evaluate the performance of the scMNMF method. We preprocessed all datasets, and specific steps can be found in Supplementary Section 4 (see Supplementary Data available online at http://bib.oxfordjournals.org/). Detailed information about these datasets is provided in Table 1.

- Simulated datasets
  We generated two single-cell multi-omics datasets (Sim1 and Sim2), each comprising gene expression omics and epigenetics omics data [23]. Sim1 contained 530 cells with 3 cell types, and Sim2 contained 250 cells with 5 cell types.
- 10X_10K peripheral blood mononuclear cell (PBMC) dataset
  The PBMC dataset was obtained from the 10x Genomics website. Each cell in this dataset has matched scRNA-seq and antibody-derived tags (ADT) data. The cell types for

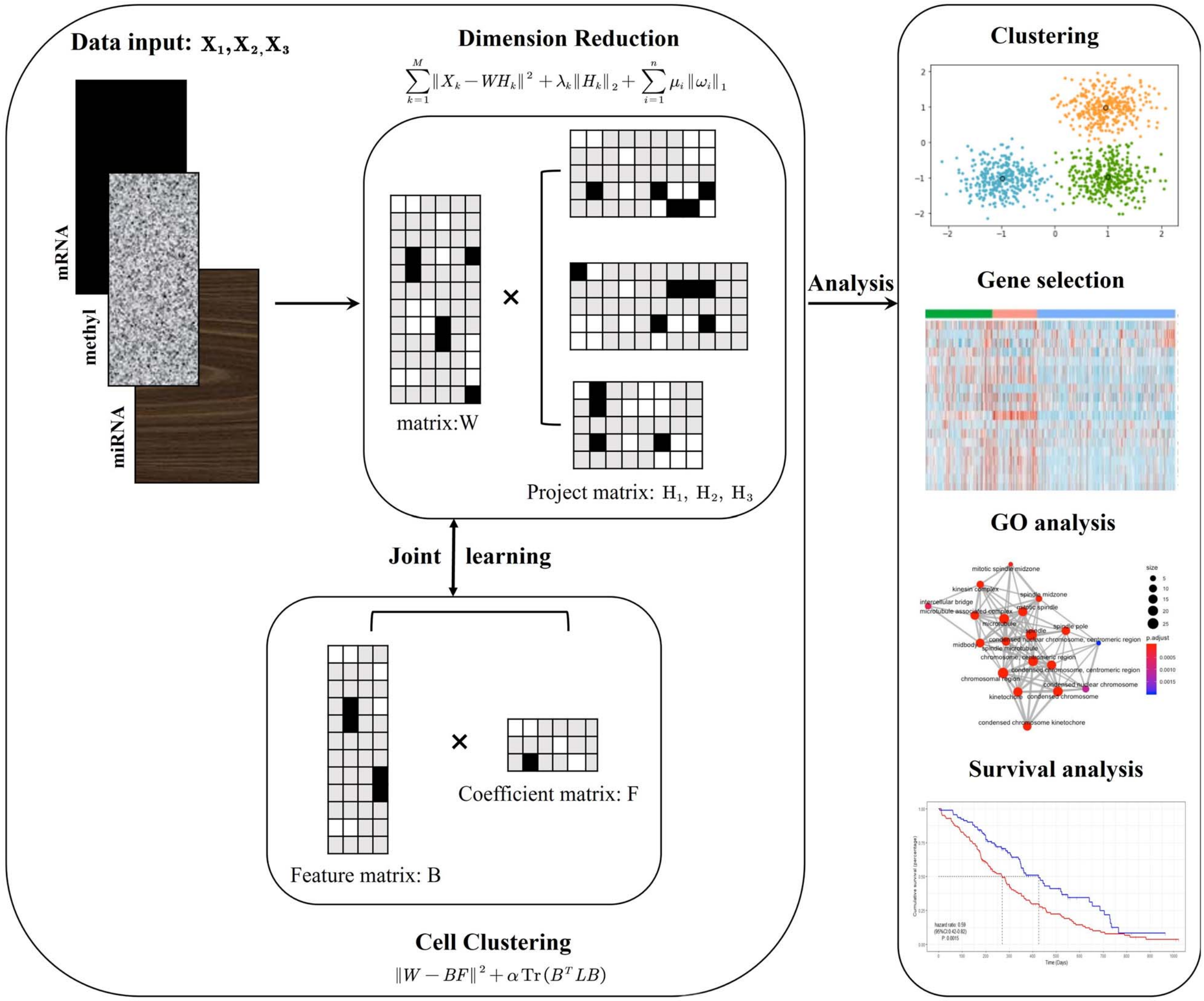

Data input: $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$
mRNA
methyl
miRNA
Dimension Reduction
$\sum_{k=1}^{M} \|X_k - WH_k\|^2 + \lambda_k \|H_k\|_2 + \sum_{i=1}^{n} \mu_i \|\omega_i\|_1$
×
matrix:W
Project matrix: H_1, H_2, H_3
Joint learning
×
Coefficient matrix: F
Feature matrix: B
Cell Clustering
$\|W - BF\|^2 + \alpha \mathrm{Tr}(B^T LB)$
Analysis
Clustering
Gene selection
GO analysis
Survival analysis

Figure 1. Overview of the scMNMF model. This is a joint analysis model that differs from the traditional approach of separately conducting dimensionality reduction and clustering.

each sample were determined from biological information of protein and gene markers. The dataset includes 6661 cells representing 7 different cell clusters.

- PBMC Specter dataset
  The Specter dataset consists of 3762 cells, divided into 16 cell types. The true labels in this dataset are from Specter [24]. The dataset can be obtained from GitHub (https://github.com/canzarlab/Specter).
- Spleen & Lymph nodes (SLN111) dataset
  Specific information about the spleen dataset and cell type labels can be found in the TotalVI method [18].
- SMAGE dataset
  The SMAGE dataset consists of 11 020 cells, divided into 12 cell types. It includes two types of omics data. The scRNA-seq data can be used directly for subsequent analysis, whereas the ATAC gene counts were obtained using tools from Lin *et al.* [25].
- Human bone marrow mononuclear cell (BMNC) dataset
  The BMNC dataset contains of cells from eight donors and was curated by the Human Cell Atlas. The cell type labels for this dataset can be obtained using Seurat software [15].

Table 1: Summary of the two simulated datasets and five real datasets

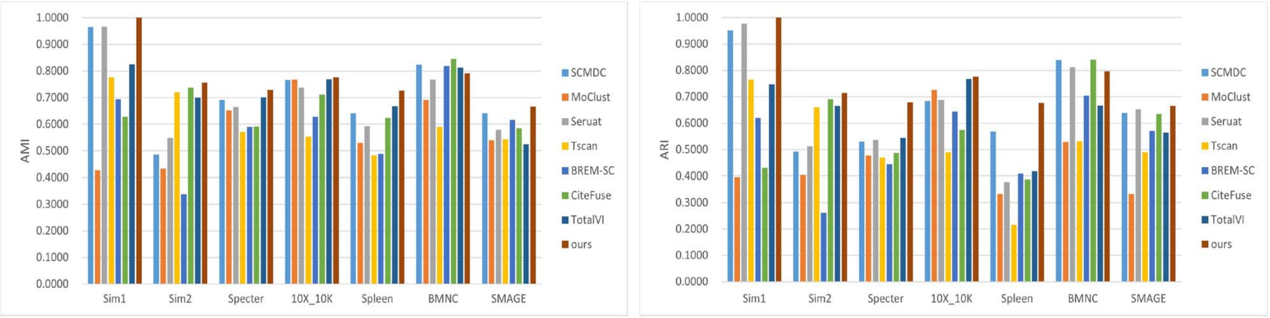
Dataset	Cell	RNA	ADT	ATAC	Type
Sim1	530	2000		5000	3
Sim2	250	2500		5000	5
Specter	3762	33 538	49		16
10X_10K	6661	33 538	17		7
SMAGE	11 020	36 611		20 010	12
Spleen	16 828	13 553	112		35
BMNC	30 672	17 009	25		12

Type, number of cell types.

Methods used for performance comparisons

We compared our scMNMF method with seven advanced methods.

- SCMDC
  SCMDC [25] is a deep learning model that processes different omics datasets in an end-to-end manner and then conducts



(A)

(B)

Figure 2. Clustering performance of different methods on two simulated datasets and five real datasets. (**A**) Performance measured by AMI. (**B**) Performance measured by ARI.

clustering analysis for the next step based on jointly learned latent deep embedding features.

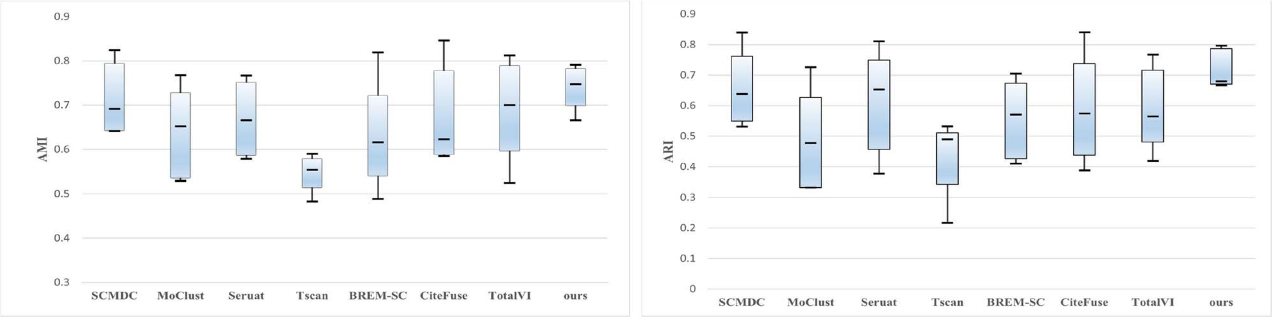
- MoClust
  MoClust [26] is a newly proposed efficient clustering method for handling data. It introduces modality-specific autoencoders to describe multi-omics data and employs distribution alignment with contrastive learning to flexibly merge the omics blocks into omics-invariant representations.
- Seurat
  After reducing the data to a lower-dimensional space, Seurat [15] constructs a weighted nearest neighbor graph and then uses the Leiden algorithm or other related algorithms to iteratively optimize and merge cells together.
- Tscan
  Tscan [27] is a traditional clustering method that processes omics data by reducing dimensionality through principal component analysis and then applies a mixture model to the reduced data for further clustering analysis.
- BREM-SC
  BREM-SC [19] is one of the earliest models proposed for clustering analysis of CITE-seq data, capable of characterizing different omics data. This method conducts clustering analysis by establishing Bayesian models for two blocks of omics data.
- CiteFuse
  CiteFuse [14] independently calculates similarity matrices between two different omics datasets and then applies an efficient fusion algorithm to merge the matrices based on this similarity. Finally, the merged similarity matrix undergoes clustering analysis based on existing graph fusion algorithms.
- TotalVI
  TotalVI [18] uses a variational autoencoder to process two different single-cell multi-omics datasets, then evenly distributes the learned modal encoder data and finally uses algorithms for subsequent clustering analysis.

Clustering performance

We used four popular performance metrics, ACC, ARI, AMI and NMI, to compare the clustering results. The clustering performance of seven algorithms and scMNMF on the two simulated datasets and five real datasets measured by AMI and ARI is shown in Figure 2. The clustering performance measured by ACC and NMI is given in Supplementary Figure 1 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

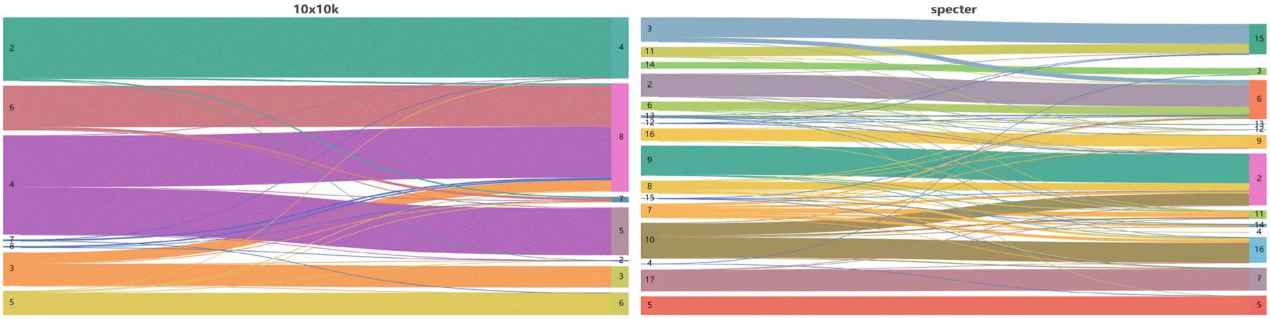
Figure 2 shows that scMNMF outperformed the other seven competing methods on all the datasets, except the BMNC dataset. The alternating influence of feature selection in dimensionality reduction and cell clustering of the scMNMF algorithm significantly improved its clustering performance. Tscan [27] produced the worst performance across the five real datasets, indicating that its traditional single-cell clustering method was unable to fully leverage the advantages of multi-omic data to enhance clustering performance and was not suitable for multi-omics data. However, Tscan performed quite well on the simulated datasets, further indicating that traditional clustering methods, including the one implemented in Seurat [15], are still somewhat unstable when dealing with single-cell multi-omics data. Although BREM-SC [19] is a mature clustering analysis algorithm, the results showed that it performed poorly on the seven datasets we ran. Its poor performance can be attributed to several limitations of the BREM-SC algorithm, particularly its inability to represent dropout events. SCMDC [25] and TotalVI [18] produced better clustering metrics than the other methods on all seven datasets, indicating that deep learning algorithms have an advantage in handling high-dimensional data. Our scMNMF model did not perform well on the BMNC dataset, whereas CiteFuse [14] produced the best performance on this dataset. This difference in performance may be because CiteFuse [14] is sensitive to high computational complexity, whereas scMNMF requires iterative optimization for the optimal solution, resulting in suboptimal performance when dealing with large-scale datasets such as BMNC. For ACC, scMNMF performed better than the other seven algorithms on the two simulated datasets and all the real datasets, except the BMNC dataset (Supplementary Figure 1, see Supplementary Data available online at http://bib.oxfordjournals.org/). For NMI, the scMNMR model ranked in the top two for all the datasets, except the Specter dataset. Overall, scMNMF performed well on all the datasets compared the performances of the other methods tested.

To further validate the feasibility of our model on five real datasets, we generated box plots based on the clustering results, to showcase the overall clustering effect of scMNMF and the other comparative methods. We calculated the lower quartile, maximum value, minimum value, median and upper quartile of the clustering metric scores for each method across different datasets. Then, we visualized the obtained values in a box plot to compare the overall performance of the methods. The height of the box indicates the interquartile range (IQR). A small IQR indicates concentrated and stable clustering results, whereas a large IQR indicates unstable clustering results, implying the method is unstable. The results for AMI and ARI are shown in Figure 3, and the results for ACC and NMI are in



(A) (B)

Figure 3. Boxplot of scMNMF and other related algorithms on five real datasets. (**A**) Performance measured by AMI. (**B**) Performance measured by ARI.



(A) (B)

Figure 4. Sankey plots for two of the real datasets. (**A**) The 10X_10K dataset. (**B**) The Specter dataset. The numbers on the left of the plots indicate the true label categories, and the numbers on the right of the plots indicate the predicted label categories.

Supplementary Figure 2 (see Supplementary Data available online at http://bib.oxfordjournals.org/). The plots show that scMNMF produced relatively concentrated performances across all datasets, indicating that its overall performance was better than those of the other methods tested. We created Sankey diagrams to illustrate the cluster assignments (Figure 4, the diagrams for the other datasets are given in Supplementary Figure 3, see Supplementary Data available online at http://bib.oxfordjournals.org/). When the number of labels was small, the clustering results were more pronounced. However, when the dataset contains a larger number of cell types, including some sparsely represented cells, there may be a certain degree of misalignment, resulting in suboptimal classification results. Nevertheless, overall, our scMNMF model produced a relatively advantageous clustering performance across different datasets.

We further investigated whether Laplacian regularization improved the clustering performance. Because Laplacian regularization affected the iterative formulation of the feature matrix B, we modified the iterative formula and conducted clustering performance comparison experiments on the five real datasets. The results indicate that when the Laplacian regularization term was added, the clustering performance of the scMNMF algorithm significantly improved compared with its performance without the Laplacian regularization term (Figure 5). This result suggests that Laplacian regularization effectively preserved the intrinsic geometric structure of the feature matrix, thereby enhancing clustering performance.

We used UMAP [28] to visualize the clustering performance. Taking the 10X_10K dataset as an example (Figure 6), the results show that cells were intertwined in the original data, and after processing with scMNMF, the cells were effectively separated. Similar effects were observed for the other datasets (Supplementary Figure 4, see Supplementary Data available online at http://bib.oxfordjournals.org/). Thus, we demonstrated that our scMNMF model can indeed help to improve the cell clustering.

Downstream analysis

We performed some downstream analysis of the scMNMF results, including gene ontology (GO) [29] and survival analysis. Gene expression is closely related to the characteristics and morphology of cells, so identifying the relevant informative genes is important. The specific steps that we used to select informative genes are provided in Supplementary Section 5 (see Supplementary Data available online at http://bib.oxfordjournals.org/). We identified 28 informative genes from the 10X_10K dataset that were differentially expressed. Details of the 28 genes are listed in Supplementary Table 1 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

To evaluate the biological significance of the 28 genes, we performed GO and KEGG pathway enrichment analysis as described previously [30]. Briefly, we considered all the genes in the genome as enrichment references. GO terms with *P*-values ¡0.01, minimum count of 3 and enrichment factor ¿1.5 were grouped based on the similarity of their members. The results of GO analysis are shown in Figure 7A. Carbohydrate derivative biosynthetic process, protein N-linked glycosylation via asparagine, mitotic spindle assembly, lymphocyte mediated immunity, *in utero* embryonic development, cell activation, wound healing, regulation of post-translational protein modification and angiogenesis were significantly enriched. Among the enriched GO terms, lymphocyte mediated immunity, wound healing and regulation of angiogenesis play pivotal roles in tumor invasion and metastasis. Protein–protein interaction (PPI) enrichment analyses were carried out

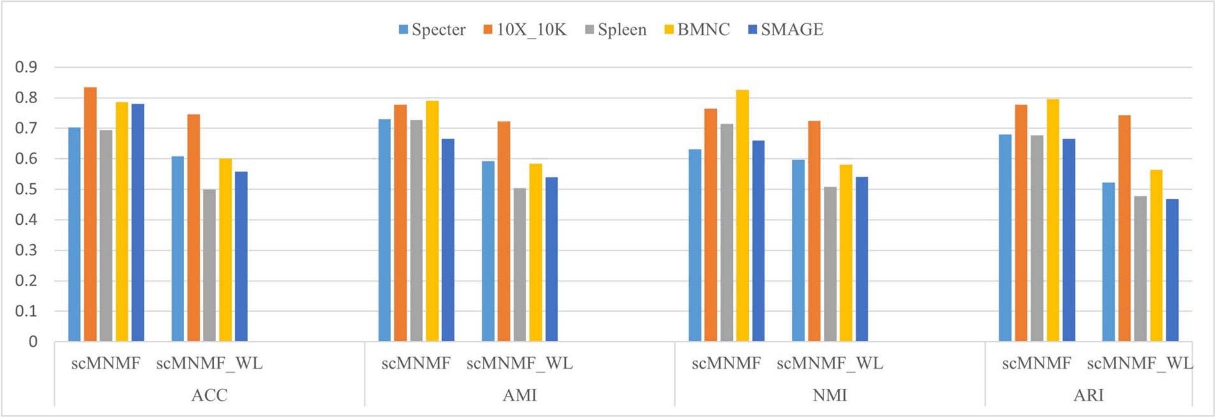


Figure 5. Clustering performance of scMNMF with and without Laplace regularization on five real datasets. scMNMF, with Laplace regularization; scMNMF_WL, without Laplace regularization.

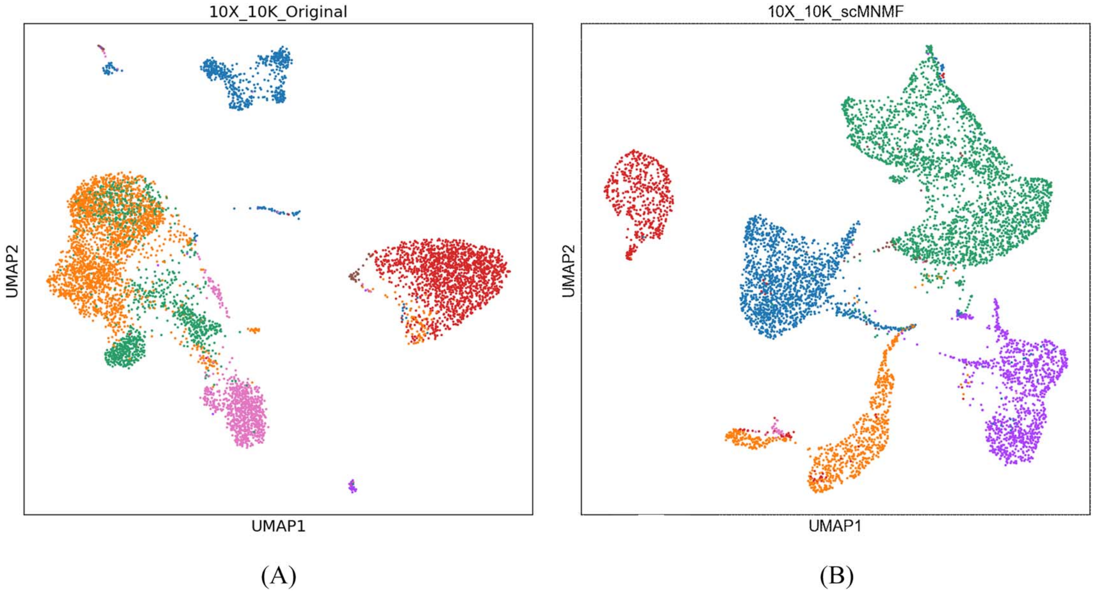


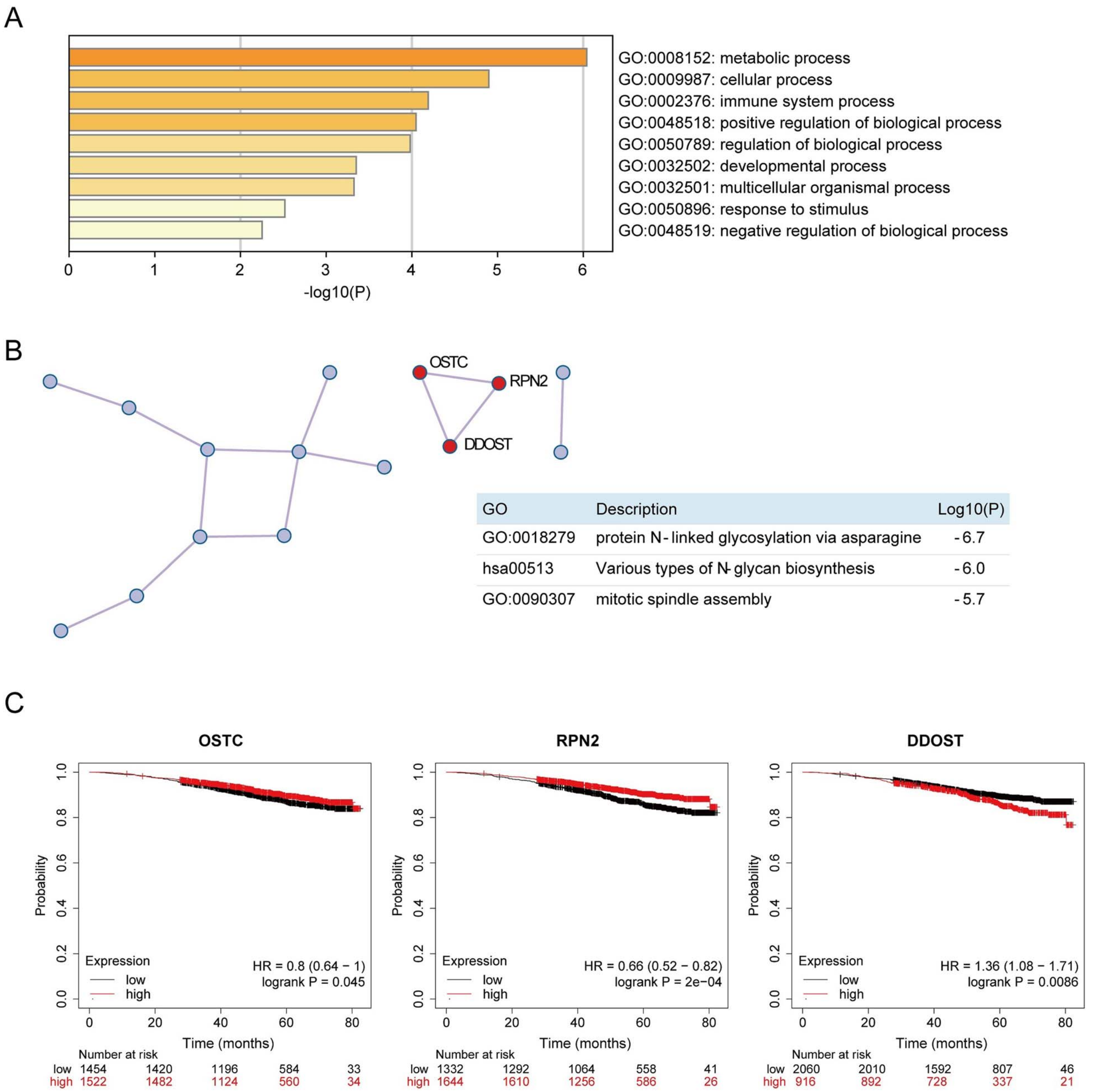
Figure 6. UMAP visualization of the clustering performance of scMNMF with the 10X_10K dataset. (**A**) Original 10X_10K dataset. (**B**) 10X_10K dataset after clustering by scMNMF.

using STRING [31], BioGrid [32], OmniPath [33] and InWeb IM, and the results are shown in Figure 7B. The protein complex screened by the PPI analysis plays an important role in protein N-linked glycosylation, and N-linked glycosylation of PD-L1 at N35, N192, N200 and N219 is required for PD-L1-dependent tumor metastasis, implying the involvement of the three genes in the complex (informative genes) in tumor progression and tumor immune escape. To analyze the correlation between the informative genes and the survival time of patients, we used the Kaplan–Meier tool (https://kmplot.com) to plot survival curves for the three informative genes, then analyzed their impact on survival. We found that out of the 28 selected informative genes, 22 were correlated with patient survival time (Figure 7C and Supplementary Figure 5, see Supplementary Data available online at http://bib.oxfordjournals.org/). The molecular complex composed of OSTC, RPN2 and DDOST is involved in protein N-linked glycosylation and regulates the survival of patients with breast cancer. OSTC and RPN2 promote patient survival, whereas DDOST shortens survival (Figure 7C). In summary, scMNMF can effectively facilitate research on breast cancer progression and is a highly effective tool for biological predictions.

CONCLUSION

Cell clustering has become an important and rapidly developing direction in single-cell research in recent years. Clustering combines different types of single-cell data, such as gene expression, protein expression and chromatin states, to comprehensively classify and cluster individual cells. This integrated analysis provides a comprehensive cell-type identification and functional interpretation, which aids in the understanding of the complex biological characteristics of cells. Many new methods have emerged for clustering single-cell multi-omics data, but typically most of them independently perform dimensionality reduction and then apply the reduced features to the clustering step. This separate two-step approach can lead to suboptimal results.

Here, we propose scMNMF to deal with this problem. scMNMF is an unsupervised method that jointly performs dimensionality reduction and clustering, thereby providing a more accurate and efficient algorithm for cell type discovery. The advantage of this approach is that the feature selection for dimensionality reduction and cell clustering mutually influence each other in an iterative manner, facilitating the discovery of cell



GO	Description	Log10(P)
GO:0018279	protein N- linked glycosylation via asparagine	-6.7
hsa00513	Various types of N-glycan biosynthesis	-6.0
GO:0090307	mitotic spindle assembly	-5.7

Figure 7. Downstream analysis of the scMNMF results. (**A**) Functional enrichment analysis of marker genes based on GO. (**B**) Protein–protein interactions of the marker genes. Molecular Complex Detection was used to identify a molecular complex (i.e., OSTC‘ RPN2 and DDOST). (**C**) Kaplan–Meier survival analysis for marker genes *OSTC*, *RPN2* and *DDOST*.

types. The results also show that the scMNMF algorithm is more accurate and robust than most algorithms on various datasets.

In the future, the fundamental concepts of scMNMF can be further explored and improved. First, in the data preprocessing step, we can consider operations for addressing dropout phenomena, which can impact clustering performance. Integrating preprocessing, dimensionality reduction and clustering together may lead to improvement. Second, because of the heterogeneity between different omics data, the issue of contribution between different omics data can be considered by assigning them different weights. This will be helpful for our learning process and also one of the directions for further study.

Key Points

- The paper provided a novel joint model which introduced dimension reduction, and clustering to the single-cell multi-omics data.
- The paper developed a novel method (scMNMF) and compared the state-of-the-art models of the cell clustering. Aiming at the shortcoming of their models, a novelty model was designed.
- Biological analysis is also conducted to validate the biological significance of our method, including GO, and survival analysis.

- Experimental results have shown that scMNMF has excellent predictive and generalization ability.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

ACKNOWLEDGEMENTS

The authors are deeply grateful to the anonymous reviewers for their helpful and constructive comments.

FUNDING

National Natural Science Foundation of China (62372303, 62002234, 62131004, 81971015).

DATA AVAILABILITY

The data is also accessible from the link.

REFERENCES

1. Zeng H, Sanes JR. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat Rev Neurosci* 2017;**18**(9): 530–46.
2. Tang F, Barbacioru C, Wang Y, *et al.* Mrna-seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**(5): 377–82.
3. Guo M, Hui Wang S, Potter S, *et al.* Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS Comput Biol* 2015;**11**(11):e1004575.
4. Grün D, Lyubimova A, Kester L, *et al.* Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature* 2015;**525**(7568):251–5.
5. Bersanelli M, Mosca E, Remondini D, *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics* 2016;**17**:167–77.
6. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell rna-seq data. *Nat Rev Genet* 2019;**20**(5):273–82.
7. Zitnik M, Nguyen F, Wang B, *et al.* Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion* 2019;**50**:71–91.
8. Vincent P, Larochelle H, Lajoie I, *et al.* Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res* 2010;**11**(12):3371–408.
9. Shen R, Mo Q, Schultz N, *et al.* Integrative subtype discovery in glioblastoma using icluster. *PLoS One* 2012;**7**(4):e35236.
10. Dingming W, Wang D, Zhang MQ, Jin G. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics* 2015;**16**(1):1–10.
11. Huh R, Yang Y, Jiang Y, *et al.* Same-clustering: S ingle-cell a ggregated clustering via m ixture model e nsemble. *Nucleic Acids Res* 2020;**48**(1):86–95.
12. Zhu Y, Zhang D-X, Zhang X-F, *et al.* Ec-pgmgr: ensemble clustering based on probability graphical model with graph regularization for single-cell rna-seq data. *Front Genet* 2020;**11**:572242.
13. Cabassi A, Kirk PDW. Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics* 2020;**36**(18):4789–96.
14. Kim HJ, Lin Y, Geddes TA, *et al.* Citefuse enables multi-modal analysis of cite-seq data. *Bioinformatics* 2020;**36**(14): 4137–43.
15. Butler A, Hoffman P, Smibert P, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**(5):411–20.
16. Duren Z, Chen X, Zamanighomi M, *et al.* Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci* 2018;**115**(30): 7723–8.
17. Argelaguet R, Arnol D, Bredikhin D, *et al.* Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**(1):1–17.
18. Gayoso A, Steier Z, Lopez R, *et al.* Joint probabilistic modeling of paired transcriptome and proteome measurements in single cells. *Biorxiv* 2020;**18**:2020–05.
19. Wang X, Sun Z, Zhang Y, *et al.* Brem-sc: a bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Res* 2020;**48**(11):5814–24.
20. Campbell KR, Steif A, Laks E, *et al.* Clonealign: statistical integration of independent single-cell rna and dna sequencing data from human cancers. *Genome Biol* 2019;**20**:1–12.
21. Li P, Jiajun B, Bin X, *et al.* Locally discriminative spectral clustering with composite manifold. *Neurocomputing* 2013;**119**: 243–52.
22. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intell* 2010;**33**(8):1548–60.
23. Jin S, Zhang L, Nie Q. Scai: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol* 2020;**21**:1–19.
24. Ringeling FR, Canzar S, *et al.* Linear-time cluster ensembles of large-scale single-cell rna-seq and multimodal data. *Genome Res* 2021;**31**(4):677–88.
25. Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nat Commun* 2022;**13**(1):7705.
26. Yuan M, Chen L, Deng M. Clustering single-cell multi-omics data with moclust. *Bioinformatics* 2023;**39**(1):btac736.
27. Ji Z, Ji H. Tscan: pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic Acids Res* 2016;**44**(13): e117–7.
28. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*. 2018;**3**:861.
29. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25–9.
30. Qiu Y, Yan C, Zhao P, Zou Q. Ssnmdi: a novel joint learning model of semi-supervised non-negative matrix factorization and data imputation for clustering of single-cell rna-seq data. *Brief Bioinform* 2023;**24**(3):bbad149.
31. Szklarczyk D, Gable AL, Lyon D, *et al.* String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**(D1):D607–13.
32. Stark C, Breitkreutz B-J, Reguly T, *et al.* Biogrid: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**(suppl_1):D535–9.
33. Li T, Wernersson R, Hansen RB, *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat Methods* 2017;**14**(1):61–4.