BMC Genomics

RESEARCH

Open Access

# JSNMFuP: a unsupervised method for the integrative analysis of single-cell multi-omics data based on non-negative matrix factorization

Bai Zhang[1], Mengdi Nan[1], Liugen Wang[2], Hanwen Wu[1], Xiang Chen[1], Yongle Shi[1], Yibing Ma[1] and Jie Gao[1*]

## Abstract

With the rapid advancement of sequencing technology, the increasing availability of single-cell multi-omics data from the same cells has provided us with unprecedented opportunities to understand the cellular phenotypes. Integrating multi-omics data has the potential to enhance the ability to reveal cellular heterogeneity. However, data integration analysis is extremely challenging due to the different characteristics and noise levels of different molecular modalities in single-cell data. In this paper, an unsupervised integration method (JSNMFuP) based on non-negative matrix factorization is proposed. This method integrates the information extracted from the latent variables of each omic through a consensus graph. High-dimensional geometrical structure is captured in the original data and biologically-related feature links across modalities are incorporated into the model using regularization terms. JSNMFuP can be utilized for data visualization and clustering, facilitating marker characterization and gene ontology enrichment analysis, providing rich biological insights for downstream analysis. The application on real datasets shows that JSNMFuP has superior performance in cell clustering. The factors are interpretable, making it an effective method for analyzing cell heterogeneity using single-cell multi-omics data.

**Keywords** Single-cell multi-omics data, Non-negative matrix factorization, Data integration

## Introduction

Single-cell sequencing technologies allow us to probe multiple biological layers. Single-cell level data, such as gene copy number, gene expression, chromatin accessibility, and protein abundance, enable us to comprehensively analyze cell heterogeneity. Single-cell multi-omics data analysis can provide unprecedented insights into cellular state and biological processes. However, integrating various omics data is a challenging task. More and more computational tools are being developed to integrate single-cell multi-omics data. The Integration problems in single-cell biology can be categorized as the integration problems of matched and unmatched data [1]. In recent years, several algorithms have been developed to integrate unmatched data. Seurat V3 [2] constructs a gene activity matrix from scATAC-seq data and integrates it with scRNA-seq data by matching shared genes. Both MATCHER [3] and UnionCom [4] implement data integration through manifold learning. MATCHER [3] assumes that all cells are distributed along a one-dimensional structure. UnionCom [4] embeds each modality into a distance matrix that captures the intrinsic low-dimensional structure of each single-cell dataset. GLUE

*Correspondence:
Jie Gao
gaojie@jiangnan.edu.cn
[1] School of Science, Jiangnan University, Wuxi, Jiangsu, China
[2] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu, China

[5] is a deep learning method for integrating unmatched single-cell multi-omics data and inferring regulatory interactions based on the entire input dataset. Most integration methods for unmatched data require identifying similar cells in a sample and accurately aligning cross-experimental datasets to discover new insights.

With the advancement of sequencing technolog [6–10], more and more single-cell multi-omics data from the same cells can be used for integrative analysis, but methods for unmatched data are not applicable to the setting that features are measured from the same cell. single-cell multi-omics data analysis greatly improves our ability to resolve cell states, which requires computational methods that can define cell states based on matched data. BREM-SC [11], based on a probability generation model, assumes that each gene has multiple distributions in each cell type, which used to obtain RNA and protein count matrices using CITE-seq. Seurat V4 [12] is a late integration method that uses weighted nearest neighbor (WNN) to synthesize intercellular affinity. Both scAI [13] based on non-negative matrix factorization (NMF) and MOFA+ [14] based on factor analysis, integrate data through a latent space. JSNMF [15] is also based on NMF to integrate single-cell multi-omics data, but it assumes different latent variables for two molecular morphologies and uses a consensus graph to combine the information carried by different molecular modalities. Meanwhile, several other ensemble algorithms for mismatched single-cell omics, for example, iNMF [16] is an online, continuously iterative, on-line single-cell data integration algorithm that scales to an arbitrarily large number of cells using fixed memory and iteratively merges new datasets as they are generated. CCNMF [17] is to link multi-omics single cells by linking copy number and general concordance of gene expression profiles. Couple NMF [18] is for the generation of different types of functional genomic data on single cells from different cell samples from the same heterogeneous population should be coupled to the clustering behaviour of cells in different samples.

The features in multi-omics data do not exist in isolation, and there are complex interactions between them. These interactions often reflect complex regulatory networks in organisms, and introducing such a priori information can improve the accuracy of data analysis. The information provided by different histological data is often complementary, and by integrating these data and introducing feature interactions, a more comprehensive view can be formed and new information and patterns that cannot be revealed by single histological data can be discovered.The Correlation-based Local Approximation of Membership (CLAM) algorithmic framework is one of the methods to integrate multi-omics data and introduce known molecular interactions during gene module identification. In this paper, we propose a novel multi-view algorithm based on non-negative matrix factorization (NMF) for the integrative analysis of single-cell multi-omics data obtained from the same cell, referred to as Jointly Semi-Orthogonal Non-negative Matrix Factorization using Prior knowledge (JSNMFuP). JSNMFuP not only captures the high-dimensional geometrical structure of each omics in the original data, but also considers the related features across modalities. Compared with JSNMF [15], it effectively improves clustering performance, as validated on mouse brain and kidney datasets. We utilize JSNMFuP for the analysis of three modalities dataset. This method not only accurately distinguishes HepG2 cells (a human hepatoblastoma-derived cell line) from human hepatocellular carcinoma (HCC) cells, but also provides biological insights into the classification of HCC subpopulations in the dataset.

## Method

### NMF

Given a non-negative original matrix $X \in R^{f \times s}$, where $f$ is the number of shared genes and $s$ is the number of cells. NMF [19] aims to approximate it by the product of two non-negative low-rank matrices $W \in R^{f \times K}$ and $H \in R^{K \times s}$, i.e., $X \approx WH$. $K$ is the number of factors and its value is less than $f$ and $s$. Solving NMF to obtain the base matrix $W$ and coefficient matrix $H$ can be considered as solving a constrained optimization problem, whose objective function is:

$$\min \|X - WH\|_F^2 \text{ s.t. } W \geq 0, H \geq 0. \tag{1}$$

where $\|\|_F$ denotes the Frobenius paradigm number of a matrix.

### JSNMF

JSNMF [15] is suitable for the joint analysis of transcriptomic and epigenomic profiles. For feature matrices $X_1 \in R^{f_1 \times s}$ and $X_2 \in R^{f_2 \times s}$, JSNMF requires that they have the same samples, but the features can be different. JSNMF enables the construction of a consensus graph, which integrates various molecular patterns within the same cells to analyze cellular heterogeneity. The objective function is as follows:

$$\begin{aligned} \min_{W^{(i)},H^{(i)},S,\lambda^{(i)}} J &= \sum_{i=1}^{2} \left\|X^{(i)} - W^{(i)}H^{(i)T}\right\|_F^2 \\ &+ \frac{\alpha}{2} \sum_{i=1}^{2} \left\|S - H^{(i)}H^{(i)T}\right\|_F^2 \\ &+ \sum_{i=1}^{2} \frac{\varphi(i)}{2} \left\|H^{(i)T}H^{(i)} - I\right\|_F^2 \\ &+ \eta \left\|\mathbf{1}^T S - \mathbf{1}^T\right\|_F^2 \\ &+ \gamma \sum_{i=1}^{2} \left(\lambda^{(i)}\right)^2 tr\left(H^{(i)T}L^{(i)}H^{(i)}\right) \\ \text{s.t. } W^{(i)}, H^{(i)}, S, \lambda^{(i)} &\geq 0, \text{ for } i \in \{1,2\}; \sum_{i=1}^{2} \lambda^{(i)} = 1 \end{aligned} \tag{2}$$

where $\alpha$ , $\eta$ and $\gamma$ are hyperparameter weights, and $\lambda^{(i)}$ can define the weight of the $i$ th data modality in the term with a Laplacian graph by adaptive learning. The first term represents the standard NMF loss function, which quantifies the difference between the original matrix and the reconstructed matrix. The second term refers to the cell-cell similarity matrix $S \in R^{s\times s}$ that integrates multiple molecular modalities. The third term is a semi-orthogonal constraint on $H^{(i)}$ and the fourth term is a normalized term of $S$. The fifth term refers to graph Laplacian regularization, which aims to preserve the high-dimensional geometrical structure of each modality in the original data space.

#### JSNMFuP

JSNMF points out that incorporating the regulatory relationships between genes and regulatory regions into the framework is one of the future development directions. The approach of incorporating prior information about interactions into the objective function in the form of regularization terms has been successfully utilized in various fields [18, 20]. To improve the performance of JSNMF algorithm, we introduce an adjacency matrix to connect the regulatory relationships between different modality features. We use the adjacency interaction matrix $R_{i,j}$ to connect the features of one modality to the features of another modality. If there is a regulatory relationship between the two features of distinct omics layers, the corresponding element in $R_{i,j}$ is 1; otherwise, it is 0. In the guide graph constructed by GLUE[5], nodes represent omics features (such as genes and ATAC peaks), and edges represent prior regulatory relationships between these features. By default, if ATAC peaks are located near the promoter of a gene, they will be connected. Inspired by GLUE, these regions and features are considered connected when constructing adjacency matrices. Based on JSNMF, we further define the objective function of JSNMFuP as follows:

$$\begin{aligned}\min_{W_i,H_i,S,\lambda_i} F &= \sum_{i=1}^{M}\left\|X_i - W_i{H_i}^T\right\|_F^2 + \frac{\alpha}{2}\sum_{i=1}^{M}\left\|S - H_i{H_i}^T\right\|_F^2 \\ &+ \sum_{i=1}^{M}\frac{\varphi_i}{2}\left\|{H_i}^T H_i - I\right\|_F^2 + \eta\left\|\mathbf{1}^T S - \mathbf{1}^T\right\|_F^2 \\ &+\gamma_1\sum_{i=1}^{M}(\lambda_i)^2 tr({H_i}^T L_i H_i) + \gamma_2\sum_{1\le i<j\le M} tr({W_i}^T R_{i,j} W_i) \\ s.t. W_i, H_i, S, \lambda_i \ge 0, \textit{ for } i \in \{1,2,\cdots,M\};\ &\sum_{i=1}^{M}\lambda_i = 1\end{aligned} \tag{3}$$

where $\gamma_2 \sum_{1\le i<j\le M} tr({W_i}^T R_{i,j} W_i)$ is a new term that represents the relationship between modality features through network regularization. Parameter $\gamma_2$ represents the weight of the network regularization constraint. The JSNMFuP framework is shown in Fig. 1.

Since the objective function of JSNMFuP is non-convex, it is relatively difficult to solve it directly. For this reason, we use the method of optimizing one matrix at a time while keeping the other variables constant. JSNMFuP first uses the Non Negative Double Singular Value Decomposition (NNDSVD) algorithm [21] to compute the optimization problem $\min_{W_i,H_i\ge 0}\left\|X_i - W_i{H_i}^T\right\|_F^2$ to initialize $W_i$ and $H_i$, and uses Similar Network Fusion (SNF) algorithm [22] to initialize $S$. Then the objective function is optimized using multiplicative updates (MU). The optimization problem is divided into four sub-problems for iterative solution.

#### Selection of hyperparameters

In JSNMFuP, We refer to the JSNMF [15] to initialize the adaptive weights and set the default values of each parameter to: $\varphi_i = \frac{1}{M}, \eta = 0.5$. The graph regularization parameters $\alpha$ and $\gamma_1$, and the feature links regularization parameter $\gamma_2$ are set as follows:

$$\alpha = \frac{M\sum_{i=1}^{M}\left\|X_i - \hat{W}_i\hat{H}_i^T\right\|_F^2}{10\sum_{i=1}^{M}\left\|S - \hat{H}_i\hat{H}_i^T\right\|_F^2} \tag{4}$$

$$\gamma_1 = \frac{M\sum_{i=1}^{M}\left\|X_i - \hat{W}_i\hat{H}_i^T\right\|_F^2}{10\sum_{i=1}^{M} tr(\hat{H}_i^T L_i\hat{H}_i)} \tag{5}$$

$$\gamma_2 = \frac{M(M-1)\sum_{i=1}^{M}\left\|X_i - \hat{W}_i\hat{H}_i^T\right\|_F^2}{20\sum_{1\le i<j\le M} tr({W_i}^T R_{i,j} W_i)} \tag{6}$$

In the results, we perform robust analysis of the hyperparameters $\alpha$ , $\gamma_1$ and $\gamma_2$. The results indicate that within a certain range of parameter value, the overall performance of the model is relatively robust. $K$ is the number of factors. In our experiments, we set the number of factors in JSNMFuP to equal the number of cell types. At the same time, we test the robustness of JSNMFuP by varying the number of factors from 10 to 50 to assess its sensitivity to the number of $K$.

#### Constructing Laplacian graph

If two cells are close to each other in the original data space, they should also be close to each other in the low-dimensional latent space. We capture the high-dimensional geometrical structure of each dataset using Laplace graphs. The Laplace matrix $L_i$ can be defined as

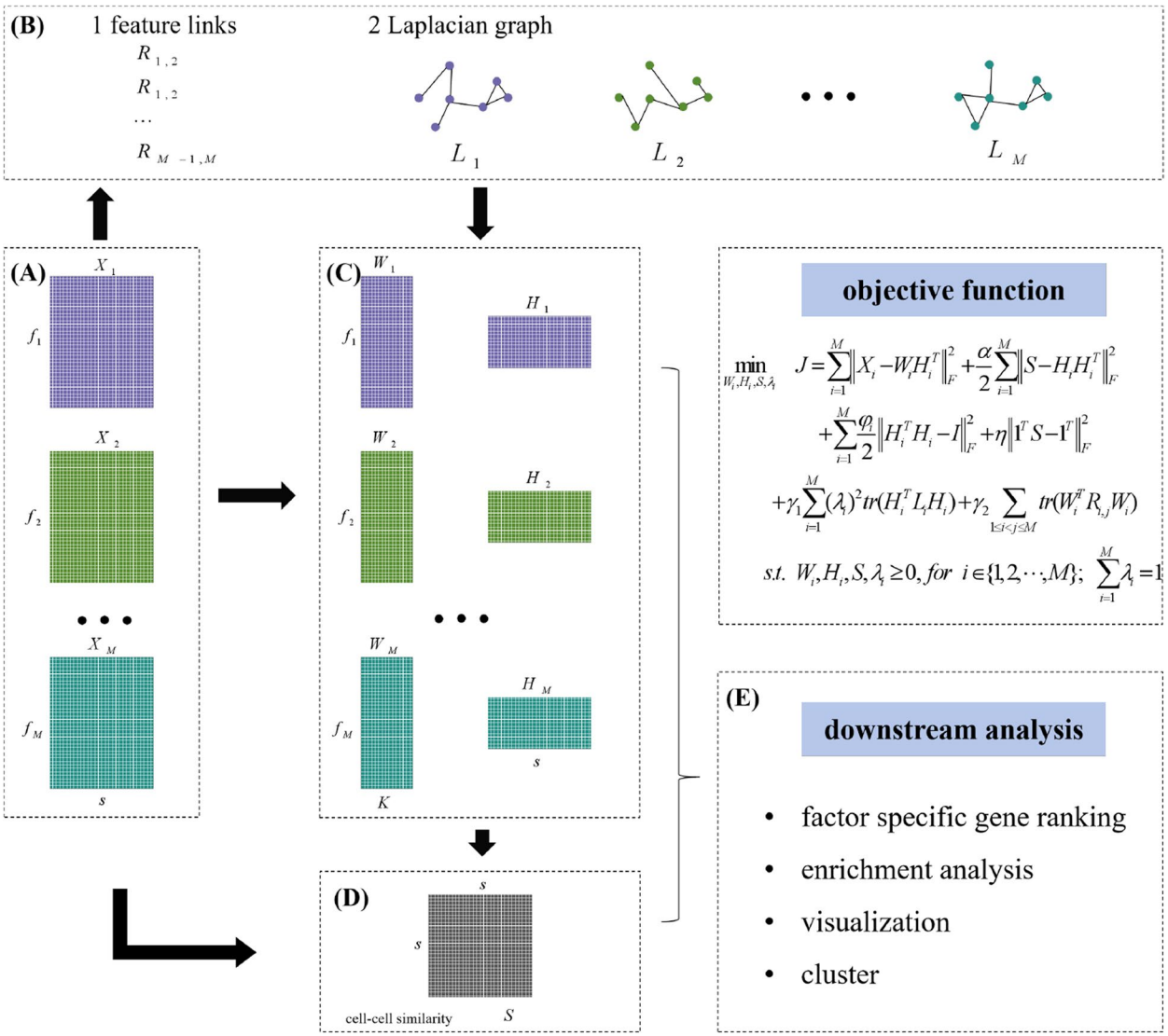


**Fig. 1** Overview of JSNMFuP. **A** Integration analysis of single-cell multi-omics data from the same cells (**B**) Construction of feature interaction matrix and Laplacian graph (**C**) Learning of feature loading matrix and factor loading matrix (latent variables) (**D**) Learning of cell-cell similarity matrix (**E**) Data integration for downstream analysis

$L_i = D_i - A_i$, where $A_i$ is the adjacency matrix and $D_i$ is the diagonal matrix. We use an exponential similarity kernel function to calculate the similarity.

$$(A_i)_{m,n} = exp\left(-\frac{\left\|(x_i)_m - (x_i)_n\right\|_F^2}{\beta \times (\mu_i)_{m,n}}\right) \tag{7}$$

$$(\mu_i)_{m,n} = \frac{mean(d(m,N_m)) + mean(d(m,N_m)) + d_{mn}}{3} \tag{8}$$

where $\beta$ is empirically set to 0.5 and $(\mu_i)_{m,n}$ is used to eliminate the scaling problem. $d_{mn}$ denotes the squared Euclidean distance between cell $m$ and $n$, $mean(d(m,N_m))$ denotes the average of the squared Euclidean distance between cell $m$ and its 20 nearest neighbors, and $mean(d(n,N_n))$ denotes the average of the squared Euclidean distance between cell $n$ and its 20 nearest neighbors.

### Evaluation metrics

We construct a k-nearest neighbor (KNN) graph (k=50) using the cell-cell similarity matrix $S$, and then cluster the cells using Louvain [23] based on the KNN graph. The adjusted rand index [24] (ARI), normalized mutual information (NMI), and residual average Gini index [25] (RAGI) are calculated to evaluate algorithm

performance. ARI and NMI require the true labels of the data to compare the similarity and consistency between the cluster prediction labels and the true labels. RAGI does not require true labels of the data, as it measures the difference between the variability of marker gene expression and the variability of housekeeping gene expression across cell clusters. Marker genes are highly expressed in specific cell types, while housekeeping genes need to be stably expressed in all cells. The higher the values of the three metrics, the better the clustering effect of the algorithm. The true label is determined by the cell label provided in the original publication, and the selection of marker genes and housekeeping genes is consistent with JSNMF in the same datasets. For the three modalities data, we calculate the Calinski-Harabasz (CH) index [26] and the Silhouette Coefficient [27] to determine the number of clusters for analysis. The higher values indicate that the cluster is more compact and farther away from other clusters.

### Competitive methods

The three commonly used vertical integration methods are scAI [13], MOFA+ [14], and MNN [12]. scAI aggregates sparse epigenetic signals in cell like structures learned in an unsupervised manner through iterative learning, allowing for coherent fusion with transcriptome measurements. MOFA+ reconstructs low dimensional representations of data using computationally efficient variational inference and supports flexible sparsity constraints, allowing for joint modeling of variations between multiple sample groups and data patterns. MNN can learn the relative utility of each data type in each cell, thereby achieving integrated analysis of multiple patterns. And JSNMF [15] is our main method based on, which assumes two molecular patterns of different latent variables and integrates transcriptome and epigenome data with consensus graph fusion to better address different features and noise levels in different molecular patterns in single-cell multi omics data. Therefore, we choose the above four methods to compare with our method. MNN and MOFA+ are composed of corresponding R packages, while other methods are used by Python packages. For each method, we use the same cells according to the required data preprocessing steps, and each method uses default hyperparameters.

### Datasets

1. Mouse brain dataset. Histone modifications and gene expression profiles of 7465 adult mouse frontal cortex and hippocampus from Paired-Tag are downloaded from the Gene Expression Omnibus (GEO) (GSE152020).
2. Mouse kidney dataset. Chromatin accessibility and gene expression profiles of 8837 adult mouse kidney cells obtained from sci-CAR are downloaded from GEO (GSM3271044 and GSM3271045).
3. Hou dataset. 25 HCC cells and 6 single HepG2 cells are sequenced by scTrio-seq, and the sequencing data are downloaded from GEO (GSE65364).

For the mouse brain dataset and mouse kidney dataset, we filter out cells with read counts for the genes less than 500 in the expression data or read counts for the regions less than 200 in the epigenomic data. Then use Seurat [2] to normalize the data of the two modalities, perform logarithmic transformation, select features, and extract the first 5000 highly variable genes and the first 10000 highly variable regions for analysis. If the highly variable region overlaps with the gene body or promoter region of a highly variable gene, they will be connected. For the three modalities dataset (Hou dataset), we convert the downloaded gene expression data from FPKM to TPM standardization. For DNA methylation data, we first download the GTF annotation file from the Gencode database (https://www.gencodegenes.org/). Then, extract the CpG sites located within the range of 2000bp upstream to 2000bp downstream of transcription start sites, calculate the average methylation level and normalize the methylation values. We utilize the R package CopyKAT [28] to acquire gene copy number variation (CNV) data, which is inferred from single-cell gene expression data. Then, normalize the CNV data. Select the first 2000 highly variable features from each modality and create an adjacency matrix of feature links based on features associated with the same gene.

## Result

### Application to the mouse brain dataset

In the first case, we investigate jointly profiles of histone modifications and gene expression in mouse brain cells. We compare JSNMFuP with scAI, WNN, MOFA+, and JSNMF. In order to quantify the clustering effect, we optimize the clustering resolution to align with the number of clusters in the publication. JSNMFuP enables the data visualization using uniform manifold approximation and projection (UMAP) [29], and annotate cells with the original labels (Fig. 2A). It can be seen that cells of the same cell type can be closely clustered together, and JSNMFuP has the highest values of the three indicators in comparison, representing the best clustering performance (Fig. 2B). Our analysis focuses on factors 17 and 14, with a detailed examination of both. The violin plots demonstrate that

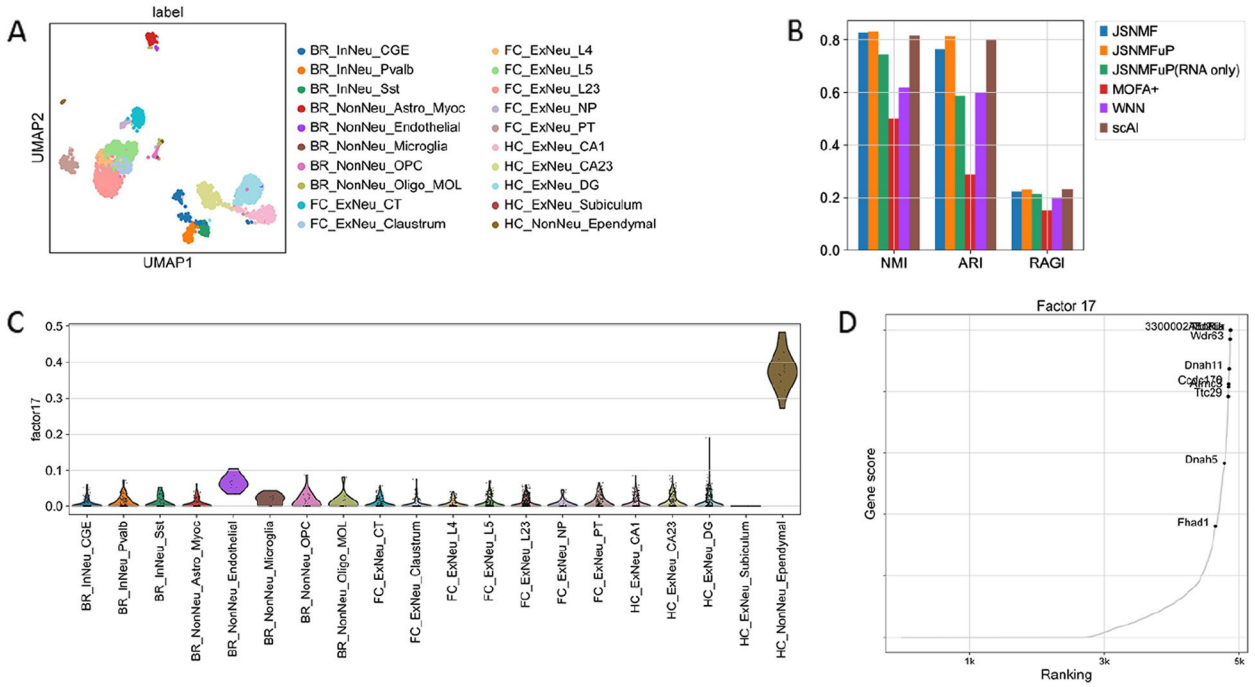


**Fig. 2** Analysis of the mouse brain dataset (**A**) UMAP visualization of real cell types in the mouse brain dataset (**B**) Comparison of clustering results, evaluated by NMI, ARI, and RAGI (**C**) The violin plot shows the expression of different cell types in factor 17 (**D**) Gene ranking plot for the factor 17

mouse brain ependymal cells (HC NonNeu Ependymal) exhibit elevated $H_1$ in Factor 17 (Fig. 2C), while brain inhibitory neurons expressing growth inhibitors (BR_InNeu_Sst) display heightened $H_1$ values in Factor 14 (Fig. S3B). Gene rankings are plotted using scAI to define gene scores, which is the expression value of each gene divided by the sum of the expression values of all genes in that factor (Fig. 2D) (Fig. S3C). The figure highlights 10 genes that have been identified as marker genes of mouse brain epithelial cells in CellMarker 2.0 [30]: Wdr63, Enkur, 3300002A11Rik, Ccdc170, Armc3, Fhad1, Ttc29, Dnah11, Ttc21a. Furthermore, 10 marker genes are identified in growth inhibitor-expressing brain inhibitory neurons (Ccnb1, Cep170b, Igf1, Igfbpl1, Pdyn, Ptprm, Reln, Rnaseh2b, Ubash3b, Unc13c). These marker genes generally have high gene scores. Furthermore, we rank genes based on their expression values in each factor of the gene loading matrix $W_1$, and the top 100 genes are considered as factor-specific genes. Use the R package ClusterProfiler to conduct Gene Ontology (GO) Biological Process (BP) enrichment analysis on these 100 genes. In the GO BP enrichment analysis, the corrected $p$-value (Fisher's accurate test) is obtained using the Benjamini-Hochberg method, where Log10(p.adjust) represents the logarithm of the corrected $p$-value to the base 10. We set the corrected $p$-value threshold to 0.01, meaning that we screen with a threshold of Log10(p.adjust) $= -2$. A total of 20 biological processes are screened. As shown in Supplementary Table S1, the results reveal biological processes closely associated with cilia and microtubules. This is consistent with the characteristics of ependymal cells. Ependymal cells, a type of glial cells located in the central nervous system, produce cerebrospinal fluid and contribute to the blood-brain barrier. In the brain ventricles, the cilia on the ependymal cells fluctuate back and forth to circulate cerebrospinal fluid [31]. The prominent feature of the apical surface of the ependymal cells is the presence of motile cilia, and the main skeleton of the motile cilia is a "9+2" microtubule structure. Online GO BP enrichment analysis is conducted using GREAT (http://bejerano.stanford.edu/great/public/cgi-bin/greatWeb.php) for the top 1000 loci in each factor ranking. The Binom Raw $P$-Value is set to be less than 0.01 and the Binom FDR Q-Value is set to be less than 0.05, resulting in hexose metabolic process ($p$-value=$6.80 \times 10^{-5}$), monosaccharide metabolic process ($p$-value=$6.35 \times 10^{-5}$), response to insulin($p$-value=$8.06 \times 10^{-5}$). And previous histological and in vitro studies have shown that the posterior ependymal cells play a role as glucose sensors [32]. This indicates that enrichment analysis provides

consistent and rich functional insights into the cell types identified.

### Application to the mouse kidney dataset

In the second case, we conduct an integrative analysis of single-cell gene expression and chromatin accessibility in mouse kidney cells. We also compare JSNMFuP with four other methods and implement UMAP visualization (Fig. 3A). In this dataset, JSNMFuP still shows the best clustering performance in all indicators (Fig. 3B). We analysed Factor 9 and Factor 3 and found that distal convoluted tubule cells in mouse kidneys had higher levels of Factor 9 (Fig. 3C) and proximal tubule S3 cells (type 2 cells) in kidneys had higher levels of Factor 3 (Fig. S4C). In CellMarker 2.0, 10 marker genes (Wnk1, Slc12a3, Lhx1, Sgms2, Slc16a7, Gm15848, Abca13, Trpm6, Pvalb, Hoxb5os) are identified. As can be seen from the gene sequencing plot, the scores and rankings tend to be higher for factor 9. Whereas (Eci3, Ghr, Guca2b, Ldhd, Mep1a, Nudt19, Slc6a18, Slc7a13, Slco1a6, Snhg11) tend to have higher scores and rankings in factor 3. We also perform GO BP enrichment analysis on the top 100 factor-specific genes using the R package ClusterProfiler. Identify 5 biological processes with corrected *P*-values less than 0.01, i.e., Log10 (p.adjust) less than −2. The distal tubules and the collecting duct in the posterior half are the primary locations for secreting potassium ions. They can also reabsorb sodium chloride and water in appropriate proportions to regulate sodium and potassium homeostasis [33]. In the GO BP enrichment analysis results, we also observe processes related to potassium ion homeostasis and sodium ion transport (Supplementary Table S2). Through the GO BP enrichment analysis on $W_2$, we obtain the following biological processes: positive regulation of stress-activated MAPK cascade ($p$-value=$6.52 \times 10^{-5}$), positive regulation of stress-activated protein kinase signaling cascade ($p$-value=$7.25 \times 10^{-5}$), positive regulation of JNK cascade ($p$-value=$2.46 \times 10^{-4}$). This may be due to the activation of the MAPK signaling pathway by factors such as osmotic pressure and shear stress induced by fluid flow over the cell surface in curved tubular cells at the distal end of the mouse kidney.

### Application to three modalities dataset

In this case, we integrate and analyze 31 liver cells data from the Hou dataset. We calculate the CH index and Silhouette coefficient for various numbers of clusters across different factors to identify the optimal number of clusters. From Fig. 4A, it is clear that when the number

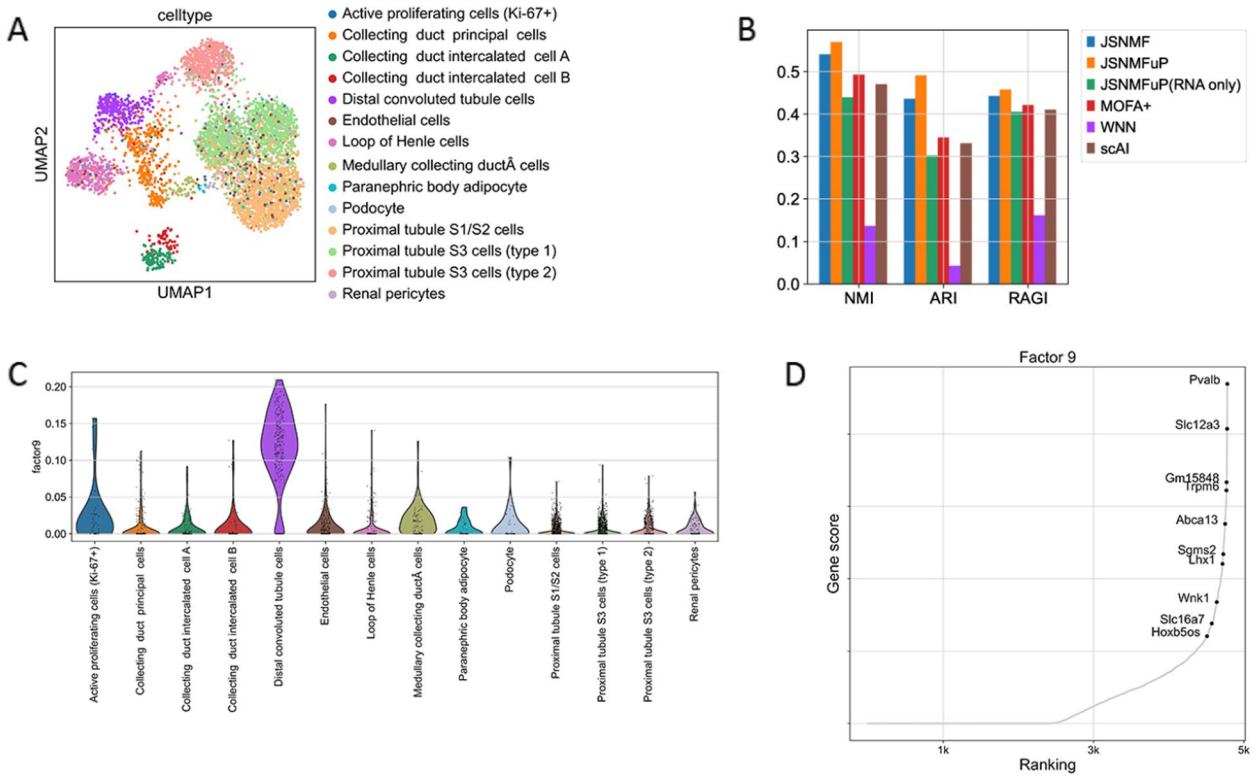


**Fig. 3** Analysis of the mouse kidney dataset (**A**) UMAP visualization of real cell types in the mouse kidney dataset (**B**) Comparison of clustering results, evaluated by NMI, ARI, and RAGI (**C**) The violin plot shows the expression of different cell types in factor 9 (**D**) Gene ranking plot for the factor 9

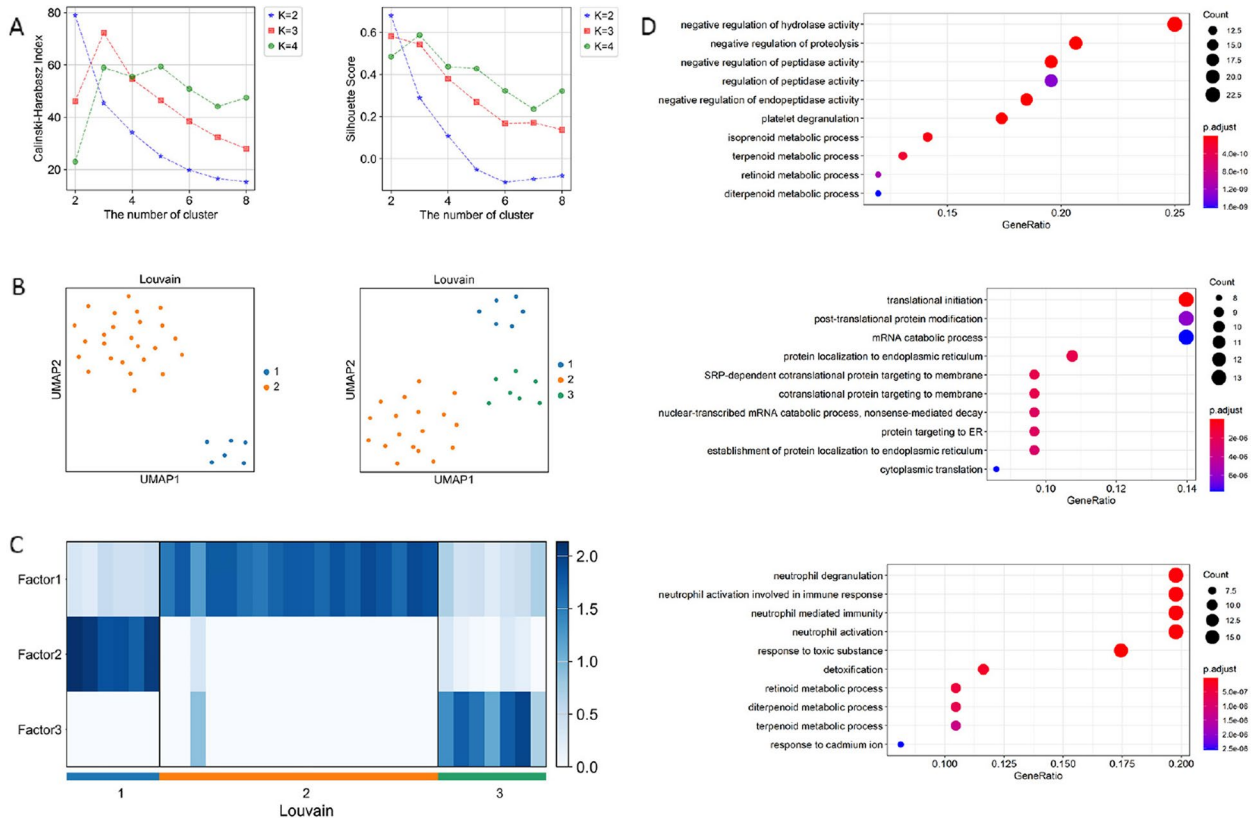


**Fig. 4** Analysis of Hou dataset (**A**) CH index and Silhouette coefficient corresponding to different cluster numbers in different factors (**B**) UMAP visualization of Louvain clustering (Left: K=2, Right: K=3) (**C**) Heatmap of factor loading matrix obtained by JSNMFuP (**D**) Dotplots for GO BP analysis (upper: Cluster 1, middle: Cluster 2, bottom: Cluster 3)

of factors is 2 and the number of clusters is 2, both the CH index and silhouette coefficient have the highest values. Therefore, we set $K$=2 in JSNMFuP and optimize the clustering resolution to 2 clusters. Figure 4B displays the UMAP visualization of the clustering results. 25 HCC cells and 6 HepG2 cells can be correctly clustered. Overall, when the number of factors is 3 and the number of clusters is 3, a higher CH index and Silhouette coefficient are also obtained. Therefore, we also analyze this situation. By comparing the left and right subgraphs of Fig. 4B, It can be seen that when $K$ changes from 2 to 3, Cluster 2 divides into Cluster 2 and Cluster 3. That is, 25 HCC cells are divided into two subpopulations, which is consistent with previous research findings [33]. These two subpopulations are associated with Factor 1 and Factor 3, respectively (Fig. 4C). We utilize the top 100 genes with the highest gene expression values in each factor for GO BP analysis (Fig. 4D). Factor 2 at $K$=3 exhibits correlation with HepG2 cells, and the analysis results focus on protein transcription and translation, as well as endoplasmic reticulum-related biological processes. HepG2 cell line closely resembles human liver tissue in both morphology and function, and is frequently utilized in the research on drug metabolism and liver toxicity. The liver is highly active in protein synthesis and metabolism, with the endoplasmic reticulum playing a key role in the synthesis of proteins and lipids. For these reasons, The biological explanations we obtain in the factors are consistent with existing knowledge. HCC subpopulation 1, corresponding to Cluster 2, is at the forefront in blood coagulation, regulation of various enzymes, and metabolism in the biological process of gene enrichment. HCC subpopulation 2, corresponding to Cluster 3, focuses on neutrophil activation and detoxification. This indicates that HCC subpopulation 1 exhibits a stronger response to immune recognition, while HCC subpopulation 2 tends to evade immune recognition and is more aggressive.

### Hyperparameters and factor stability

To assess the robustness of JSNMFuP to hyperparameters and the number of factors, we create line plots to observe its performance. The experimental results are depicted in

the figures. As can be seen in the results, for the hyperparameters $\alpha$ , $\gamma_1$ , $\gamma_2$ and the factor $K$, the clustering performance evaluated by NMI, ARI, and RAGI does not show significant differences with their changes (Supplementary Fig. S1), indicating that the performance of JSNMFuP is robust.

### Convergence analysis

As a non-convex problem, JSNMFuP is solved by an iterative optimization algorithm. Supplementary Fig. S2 shows the objective function changes in each iteration. The horizontal axis represents the number of iterations, and the vertical axis represents the value of the objective function. It can be clearly seen from the figure that the convergence trend of the objective function curve is obvious, with the target value steadily decreasing in each iteration. The algorithm JSNMFuP guarantees convergence.

## Conclusion

The rapid development of single-cell multi-omics sequencing technologies has led to an increasing availability of single-cell multi-omics data from the same cell. This has driven research on cell heterogeneity. In this study, we propose a method called JSNMFuP for integrating and analyzing multi-omics data from the same cells.

Firstly, JSNMFuP demonstrates good clustering performance. In comparing the clustering performance of the mouse brain dataset and the mouse kidney dataset, we used metrics such as NMI, which utilizes raw label information, and RAGI, which uses the expression of marker genes and steward genes. On one hand, JSNMFuP ranks first in clustering comparison with the other four methods (scAI, WNN, MOFA+, JSNMF). On the other hand, we construct a model that only uses gene expression data for clustering. The results show that JSNMFuP outperforms this model in clustering performance by utilizing two omics information. This indicates that JSNMFuP can integrate complementary and compatible information from each modality, and incorporating multi-omics data into integrative analysis is beneficial.

Secondly, the factors in JSNMFuP are interpretable. In the detailed analysis of factor 17 in the mouse brain dataset, it is evident that factor 17 shows a strong correlation with ependymal cells in the mouse brain. GO enrichment analysis is performed on the genes ranked at the top of factor 17, revealing biological processes closely related to the unique cilia found in ependymal cells. In the analysis of factor 9 in the mouse kidney dataset, we also obtain biological explanations related to the distal convoluted tubule cells of the mouse kidney. Using data from three modalities, JSNMFuP accurately distinguishes HepG2 cells from HCC cells and divides HCC cells into two subpopulations. This allows for the study of regulatory mechanisms of biological pathways related to HCC. Finally, we investigate the robustness of hyperparameters $\alpha$ , $\gamma_1$ , $\gamma_2$ and the factor $K$. The test results confirm that the clustering performance of the JSNMFuP for the examples is stable, as evidenced by the variation in their values. The JSNMFuP algorithm ensures convergence and tends to converge within a hundred iterations.

The matrix decomposition methods such as scAI, MOFA+ and JSNMF may not be suitable for CITE-seq datasets due to the small second modal dimension of surface proteins. Our proposed JSNMFuP, as well as other matrix decomposition methods, are more suitable for handling single-cell multi-omics datasets with larger modal dimensions, especially for analyses involving transcriptomic and epigenomic data. Based on JSNMF, JSNMFuP incorporates the regulatory relationships between different omics features into the non-negative matrix factorization model. At present, JSNMFuP simply links the features associated with the same gene as an interaction.In the future,we might be able to mine the database to find the corresponding high-confidence feature relationships that indicate interaction. The adjacency matrix of JSNMFuP features can be further optimized for linking terms.

In our future work, we will further explore the interaction between NMF-based multi-omics data integration methods [33] and various variable selection methods, as well as their differences in performance across different datasets. In addition, we will explore methods to design simulation studies to fully evaluate the performance of our approach in different scenarios. In particular, we will explore how to accurately model complex single-cell regulatory relationships and evaluate the performance of methods. For the problem of data contamination, we will refer to the study of Wu et al. [36] and design simulation studies to evaluate the robustness of our method under different degrees of multi-omics data contamination, with a view to providing more reliable and robust analytical tools.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11462-8.

Supplementary Material 1.

**Acknowledgements**

This work has been supported by the National Natural Science Foundation of China (Grant No. 12271216, 11831015).

Thank you to all colleagues in the laboratory for their help.

**Authors' contributions**

GJ and ZB conceived and designed the project. ZB and NM managed the entire trial, conducted software code development and wrote the manuscript.

WL, WH, CX, SY, and MY helped with data collection and analysis. All authors read and approved the final manuscript.

**Funding**
This work has been supported by the National Natural Science Foundation of China (Grant No. 12271216, 11831015).

**Data availability**
The datasets in this study are available from the Gene Expression Omnibus (GEO) repository with the following accession numbers: GSE152020, GSE3271045, and GSE65364. The JSNMFuP source code, demo script, and demo data are freely available on the GitHub website (https://github.com/ZB-JN/JSNMFuP).

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 22 April 2024 Accepted: 10 March 2025
Published online: 20 March 2025

**References**

1. Miao Z, Humphreys BD, McMahon AP, Kim J. Multi-omics integration in the age of million single-cell data. Nat Rev Nephrol. 2021;17(11):710–24.
2. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
3. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 2017;18(1):1–19.
4. Cao K, Bai X, Hong Y, Wan L. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics. 2020;36(Supplement_1):i48–i56.
5. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022;40:1458–66.
6. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5.
7. Zhu C, Zhang Y, Li YE, Lucero J, Behrens MM, Ren B. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nat Methods. 2021;18(3):283–92.
8. Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science. 2018;6418(362):1060–3.
9. Guo F, Li L, Li J, Wu X, Hu B, Zhu P, et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. Cell Res. 2017;27(8):967–88.
10. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018;9(1):781.
11. Wang X, Sun Z, Zhang Y, Xu Z, Xin H, Huang H, 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.
12. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87.
13. 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):25.
14. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111.
15. Ma Y, Sun Z, Zeng P, Zhang W, Lin Z. JSNMF enables effective and accurate integrative analysis of single-cell multiomics data. Brief Bioinforma. 2021;23(3):bbac105.
16. Gao C, Liu J, Kriebel AR, Preissl S, Luo C, Castanon R, et al. Iterative single-cell multi-omic integration using online learning. Nat Biotechnol. 2021;39:1000–7.
17. Bai X, Duren Z, Wan L, Xia LC. Joint inference of clonal structure using single-cell genome and transcriptome sequencing data. NAR Genomics Bioinforma. 2024;6(3):bbac105.
18. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci U S A. 2018;115(30):7723–8.
19. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
20. Wang Y, Zhou G, Guan T, Wang Y, Xuan C, Ding T. A network-based matrix factorization framework for ceRNA co-modules recognition of cancer genomic data. Brief Bioinforma. 2022;23(5):bbac154.
21. Boutsidis C, Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recog. 2008;41(4):1350–62.
22. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.
23. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):P10008.
24. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50.
25. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. 2019;20(1):241.
26. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3(1):1–27.
27. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
28. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39:599–608.
29. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for Dimension Reduction. Genome Biol. 2017;18(1):1–19.
30. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, 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. 2023;51(1):870–6.
31. MacDonald A, Lu B, Caron M, Caporicci-Dinucci N, Hatrock D, Petrecca K, et al. Single cell transcriptomics of ependymal cells across age, region and species reveals cilia-related and metal ion regulatory roles as major conserved ependymal cell functions. Front Cell Neurosci. 2021;15:703951.
32. Sato M, Minabe S, Sakono T, Magata F, Nakamura S, Watanabe Y, et al. Morphological analysis of the hindbrain glucose sensor-hypothalamic neural pathway activated by hindbrain glucoprivation. Endocrinology. 2021;162(9):bqab125.
33. Poirion O, Zhu X, Ching T, Garmire LX. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. Nat Commun. 2018;9(1):1–194892.
34. Gallafassi EA, Bezerra MB, Rebouças NA. Control of sodium and potassium homeostasis by renal distal convoluted tubules. Braz J Med Biol Res. 2023;56:e12392.
35. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A selective review of multi-level omics data integration using variable selection. High Throughput. 2019;8(1):4.
36. Wu C, Zhang Q, Jiang Y, Ma S. Robust network-based analysis of the associations between (epi) genetic measurements. J Multivar Anal. 2018;168:119–30.

## Publisher's Note