Ma *et al. BMC Genomics* (2025) 26:132
https://doi.org/10.1186/s12864-025-11319-0

BMC Genomics

RESEARCH Open Access

scMFG: a single-cell multi-omics integration method based on feature grouping

Litian Ma[1], Jingtao Liu[1], Wei Sun[2], Chenguang Zhao[2*] and Liang Yu[1*]

Abstract

Background Recent advancements in methodologies and technologies have enabled the simultaneous measurement of multiple omics data, which provides a comprehensive understanding of cellular heterogeneity. However, existing methods have limitations in accurately identifying cell types while maintaining model interpretability, especially in the presence of noise.

Methods We propose a novel method called scMFG, which leverages feature grouping and group integration techniques for the integration of single-cell multi-omics data. By organizing features with similar characteristics within each omics layer through feature grouping. Furthermore, scMFG ensures a consistent feature grouping approach across different omics layers, promoting comparability of diverse data types. Additionally, scMFG incorporates a matrix factorization-based approach to enable the integrated results remain interpretable.

Results We comprehensively evaluated scMFG's performance on four complex real-world datasets generated using diverse sequencing technologies, highlighting its robustness in accurately identifying cell types. Notably, scMFG exhibited superior performance in deciphering cellular heterogeneity at a finer resolution compared to existing methods when applied to simulated datasets. Furthermore, our method proved highly effective in identifying rare cell types, showcasing its robust performance and suitability for detecting low-abundance cellular populations. The interpretability of scMFG was successfully validated through its specific association of outputs with specific cell types or states observed in the neonatal mouse cerebral cortices dataset. Moreover, we demonstrated that scMFG is capable of identifying cell developmental trajectories even in datasets with batch effects.

Conclusions Our work presents a robust framework for the analysis of single-cell multi-omics data, advancing our understanding of cellular heterogeneity in a comprehensive and interpretable manner.

Keywords Single-cell, Multi-omics, Feature grouping, Integration

*Correspondence:
Chenguang Zhao
zhao_chenguang@outlook.com
Liang Yu
lyu@xidian.edu.cn
[1]School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China
[2]Department of Rehabilitation Medicine, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China

Background

Cellular heterogeneity plays a pivotal role in various biological processes [1–3]. The advent of advanced single-cell sequencing technologies has empowered us to quantitatively dissect this heterogeneity [4–7]. In recent years, significant advancements in single-cell multi-omics sequencing technologies have enabled the simultaneous exploration of multiple molecular layers. For instance, techniques such as sci-CAR [8], SNARE-seq [9], SHARE-seq [10], and 10x multiome, which combines transcription and chromatin accessibility sequencing, as well as scNMT-seq [11], which profiles chromatin accessibility, DNA methylation, and gene expression jointly, have emerged. These technologies offer invaluable opportunities to investigate cellular heterogeneity and enhance the refined identification of cell types [12, 13]. Consequently, there is a growing demand for developing effective integration methods that facilitate integrated analyses to uncover the complexities of cellular heterogeneity.

Researchers have developed various approaches to integrate single-cell multi-omics data [12, 14–28]. These methods can be categorized into three main categories based on their underlying methodologies: matrix factorization, neural networks, and network analysis [29]. Matrix factorization methods decompose the omics data matrix into the product of a weight matrix and a factor matrix. These methods are straightforward and offer clear interpretations of the factors. However, the presence of noise in single-cell data poses challenges to their analysis [30]. Experimental protocols, library preparation, amplification, and sequencing can introduce noise. Treating each omics layer as a whole can introduce additional noise that hinders accurate cell type identification. This additional noise can arise from irrelevant features, which are those that do not significantly contribute to the distinguishing characteristics of different cell types. For example, when trying to identify a particular cell type, certain genes may play a crucial role in distinguishing and characterizing that specific cell type. However, including all genes without discrimination, including those that are not relevant or informative for that cell type, can introduce noise and confound the identification process. Deep generation models based on neural networks have emerged as a powerful framework for modeling high-dimensional data [31–34]. These models leverage multiple nonlinear layers to capture complex relationships and learn the underlying structure of high-dimensional data, even in the presence of noise. However, neural network methods can lack interpretability, making it challenging to understand the intricate details and decision-making processes of the model [35]. Another approach is the network-based method, which utilizes weighted graphs to represent relationships between cells [29]. However, this approach overlooks the similarity between features. Features can possess distinct biological meanings and regulatory mechanisms, and treating them as a whole may obscure subtle variations and correlations. Therefore, there is a need for further development of integration methods that balance interpretability, noise robustness, and feature-level analysis to effectively uncover the complexity of single-cell multi-omics data.

To bridge these gaps, our research proposes a novel method called scMFG, which leverages feature grouping and group integration techniques for the integration of single-cell multi-omics data. Our approach utilizes the Latent Dirichlet Allocation (LDA) model [36] to group related features, effectively mitigating the impact of noise and reducing the dimensionality of the data. This dimensionality reduction is particularly crucial when dealing with large-scale single-cell omics data that possess complexity and high dimensionality. By employing the same feature grouping method across different omics layers, we establish a consistent reference framework, facilitating effective comparison and correlation of results between different data types. The integration of multiple omics feature groups in scMFG is achieved through incorporating the MOFA + component [24]. By incorporating MOFA+, we capture the shared variability among different omics feature groups, thereby enhancing our understanding of cellular heterogeneity. Compared to other single-cell multi-omics integration methods, scMFG not only identifies cell types but also provides enhanced interpretability by linking cell state with the joint embedding. We have also provided a user-friendly tool, following the standard best practices as outlined in recent guidelines [37].

Methods

Data sources

Data collection

We collected six datasets from two publicly available databases, namely the Gene Expression Omnibus (GEO) and 10x Genomics databases (www.10xgenomics.com/resources/datasets). The datasets include kidney with the GEO number GSE117089, snare_p0 with the GEO number GSE126074, 10x_lymph_node, 10x_pbmc, share_skin with the GEO number GSE140203, and neuips with the GEO number GSE194122.

The kidney dataset, comprising 11,296 cells from mouse kidneys, was sequenced using sci-CAR [8]. Cells with less than 200 gene or peak expressions were filtered out post-quality control. For clustering evaluation, reference cell annotations of kidney was provided by the authors [8]. The snare_p0 dataset, with 5,081 cells from neonatal mouse cortex, was sequenced using SNARE-seq [9], applying the same quality control criteria. Reference cell annotations of snare_p0 dataset was provided by the authors [9]. Similarly, the 10x_lymph_node (fresh

frozen lymph node with B cell lymphoma dataset) and 10x_pbmc datasets (human PBMCs), both sequenced using 10x Genomics technology, underwent comparable quality control. Reference cell annotations of them were downloaded from 10X Genomics website (www.10xgenomics.com/resources/ datasets). The share_skin dataset, consisting of 34,774 skin cells, was sequenced using SHARE-seq [10]. In each dataset, cells below the 200 expressions threshold were excluded. The annotation of SHARE-seq skin dataset was provided by paper [10]. The neuips dataset features human bone marrow mononuclear cells, captured in 13 batches using 10X Genomics. Reference cell annotations of neuips was provided by the authors [38]. Refer to Supplemental Tables 1 and 2 for further details.

Data preprocessing

For the preprocessing of single-cell RNA sequencing (scRNA-seq) data, we utilized established pipelines available in the scanpy [39] package. These pipelines encompassed normalization, logarithmic transformation, and feature selection steps. In the case of the share_skin dataset, we specifically chose to select 5,000 highly variable genes for analysis, while for the other datasets, we followed the standard approach and selected 3,000 highly variable genes. Regarding the analysis of single-cell ATAC sequencing data, we first performed binarization, followed by the same preprocessing steps as scRNA-seq data. This included normalization, logarithmic transformation, and feature selection. In this case, we identified and selected the top 10,000 highly variable peaks for subsequent experimentation.

To benchmark scMFG against other available methods, including MOFA+ [24], Cobolt [26], scMVP [25], Seurat v4 [12], GLUE [28] and scJoint [27], we conducted integrations using all algorithms on the same dataset. For each method, we followed their respective tutorials and utilized default settings unless stated otherwise.

Methods

The scMFG model was specifically developed to address the integration of simultaneous profiling of multiple omics data in single cells. As illustrated in Fig. 1, scMFG initially performs feature grouping based on expression

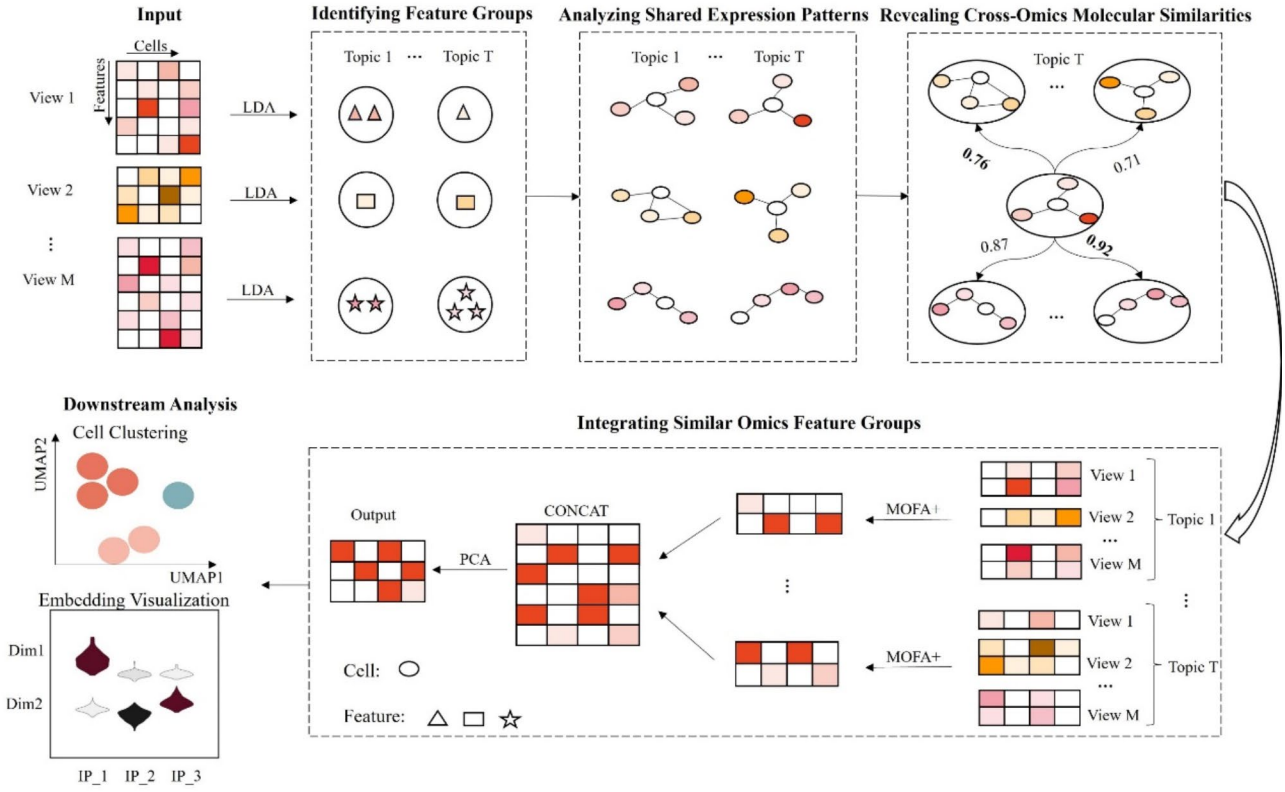

Fig. 1 Overview of scMFG integration method for single-cell multi-omics data. In scMFG, the initial step involves grouping the features within each omics layer. These feature groups are then subjected to analysis using K-nearest neighbor graphs to uncover shared expression patterns, establishing a foundational layer of intercellular relationships. In addition to this, similarity metrics are employed to further elucidate molecular expression patterns across different omics layers, allowing for the detection of subtle biological processes. To integrate these molecular expression patterns, the MOFA + framework is applied, synergistically combining insights across the omics feature groups. Finally, the results from each integrated omics group are concatenated, and dimensionality reduction is performed to reduce complexity and facilitate further analysis

patterns within each omics dataset. Subsequently, we identify and integrate the most similar feature groups across different omics modalities. This iterative process repeats until integration is performed for all feature groups across the various omics layers. Consequently, our model consists of four main steps: (1) Identification of feature groups with similar expression patterns within each omics; (2) Analysis of shared expression patterns within each feature group; (3) Uncovering similar molecular expression patterns across different omics and (4) Integration of the identified similar omics groups. Further comprehensive details regarding this approach can be found in the Methods section provided below.

Identifying feature groups with similar expression patterns within omics

To accurately identify different cell types and address challenges such as noise and similar molecular expressions, our method incorporates the LDA model, a widely accepted Bayesian probabilistic model extensively used in various domains, including single-cell analysis [39, 40]. By leveraging the capabilities of the LDA model, we can effectively distinguish and isolate features with similar expressions from the overall analysis. In our approach, we consider the expression matrix for the m-th omic, denoted as Y_m. We categorize the features of each omics layer into T distinct groups, each representing a unique biological pattern. Generally, when the number of cells is small (typically less than ten thousand), the value of T is around 15–20; when the number of cells is large (the number of cells exceeds ten thousand), the value is around 20–30. Initially, we generate a topic distribution θ for each the mth omic by sampling from a Dirichlet distribution, guided by the hyperparameter α:

$$\theta_m \sim Dirichlet(\alpha) \tag{1}$$

Here, the hyper-parameter α is a T-dimensional parameter that represents the prior weights of the T groups, typically set to $1/T$. For the t-th group, we define the prior distribution of features as a Dirichlet distribution:

$$\beta_t^m \sim Dirichlet(\phi) \tag{2}$$

Where ϕ is a hyperparameter vector that defines the distribution of features within the t-th group of omic m. For the t-th feature in the m-th omic, we determine the group index $\mathrm{Z_{m,n}}$ by sampling from a multinomial distribution:

$$Z_{m,n} \sim Multinomial(\theta_m) \tag{3}$$

Here, $\mathrm{Z_{m,n}}$ is a categorical variable representing the group assignment for the n-th feature in omic m, with possible values from 1 to T. For each assigned group $\mathrm{Z_{m,n}}$, the probability distribution of the observed features is given by:

$$W_{m,n} \sim Multinomial\left(\beta_{Z_{m,n}}^m\right) \tag{4}$$

For estimating the LDA model parameters, we utilize online variational inference [41], which allows for the iterative updating and optimization of parameters. This process is implemented using the scikit-learn package in Python [42]. Following the methodology described, we partition each omics dataset into T distinct groups, with varying feature lengths in each group. For example, in the mth modality Y_m, the features are divided into T groups. let G_t represent the number of features in the t -th group. The features are organized such that:

$$F_{m,i} \cap F_{m,j} = \varnothing \tag{5}$$

Where $F_{m,i}$ represents the feature within the i-th group. Note that the sum of features across all groups, $G_1 + G_2 + , \ldots , G_T$ equals N_m, the total number of features in the mth omic. It is important to note that there is no overlap between these feature groups.

This partitioning strategy allows us to effectively focus on specific subsets of features within each omic, enabling a more targeted exploration of cellular heterogeneity and functional aspects.

Analyzing shared expression patterns within each feature group

Building upon the feature groups identified via the LDA model, we next focus on analysing shared expression patterns among cells. This analysis is crucial for understanding cellular relationships and biological processes. We employ K-nearest neighbors (KNN) graphs, constructed from post-grouping feature expressions of the mth modality Y_m, to identify these patterns. The KNN graph edges represent shared expressions between cells, highlighting similar expression behaviors and potentially related cellular states. We typically set the KNN parameter K to 15, based on its minimal impact on model performance. After constructing KNN graphs for each group, we collate the K neighbors for all cells in the t -th group of the mth modality, storing them in a set S_m^t:

$$S_m^t = \bigcup_{i=1}^{V} \{knn_{i,1}{}^t, knn_{i,2}{}^t, \ldots , knn_{i,K}{}^t\} \tag{6}$$

Where $knn_{i,K}{}^t$ refers specifically to the kth neighbour of the ith cell in the tth group, V represents the total number of cells in the mth modality.

Uncovering similar molecular expression patterns across omics

After obtaining the k-nearest neighbours of cells in the t-th group of mth modality S_m^t, we employ the Jaccard similarity metric [43] to compare the cellular relationships between different feature groups across omics datasets. we employ the Jaccard similarity metric to compare the cellular relationships between different feature groups across omics datasets. The Jaccard similarity between two sets S_m^t and S_j^k is defined as:

$$J\left(S_m^t, S_j^p\right) = \frac{|S_m^t \cap S_j^p|}{|S_m^t \cup S_j^p|} \tag{7}$$

In this context, m and j are indices representing distinct omics datasets, ensuring that $m \neq j$. Additionally, t and p are used to denote different groups within these datasets. Note that the values of t and p range from 1 to T . The $|S_m^t \cap S_j^p|$ represents the number of common elements (cells) shared between the two sets S_m^t and S_j^p. It measures the overlap of k-nearest neighbors between the two groups from different omics datasets. And the $|S_m^t \cup S_j^p|$ represents the total number of unique elements (cells) present in either of the two sets S_m^t and S_j^p , it accounts for all the cells that are part of either group's k-nearest neighbors. The ratio $J\left(S_m^t, S_j^p\right)$ gives the Jaccard similarity, a value between 0 and 1. A Jaccard similarity of 1 indicates that the sets are identical, while a value of 0 indicates that they have no elements in common. This metric helps quantify the similarity in molecular expression patterns across different omics datasets by comparing the neighborhoods of cells. Comparing these scores across feature groups from different omics dataset allows us to pinpoint groups with maximal similarity:

$$Corre\_index_m^t = \bigcup_{j=1 \text{ and } j \neq m}^{M} \{argmax_p\, J(S_m^t, S_j^p)\} \tag{8}$$

Where $corre\_index_m^t$ represents the set of indices identifying the most similar groups across different omics datasets to Y_m^t, the term $argmax_p\, J(S_m^t, S_j^p)$ in this context specifies finding the group index p within each modality j (excluding m) where the similarity is maximized. This approach effectively reveals shared molecular characteristics across different datasets, providing crucial insights into the intricate relationships and functions within the biological systems.

Integrating similar omics feature groups

Following the identification of similar groups through Jaccard similarity, we proceed to integrate these groups to gain a multi-layered understanding of biological processes. In smog analysis, the integration of multiple omics datasets can be approached using MOFA+ [24] methodology, which provides a robust and interpretable framework for data integration.

In initiating the integration process, we first select an expression matrix Y_m^t from the tth group of the mth omics dataset. Using Jaccard similarity calculations, we identify the $M-1$ groups most similar to Y_m^t:

$$Group_m^t = Y^{corre\_index_m^t} \bigcup Y_m^t \tag{9}$$

Where $group_m^t$ refers to the selected feature expressions of other omics groups that are most similar to Y_m^t expression, including Y_m^t itself. This crucial step allows us to establish meaningful connections between related groups across different omics datasets, setting the stage for a comprehensive integration of the data.

Subsequently, we employed the muon.tl.mofa function [44] to implement the integration of the $group_m^t$.

$$Res_m{}^t = MOFA + \left(Group_m^t\right) \tag{10}$$

In this term, $Res_m{}^t$ refers to the output of MOFA+. The integrated results of the T groups are then concatenated to form a unified dataset, ensuring that information from all omics datasets is included.

$$Res = CONCAT\left(Res_m{}^1, \ldots, Res_m{}^T\right) \tag{11}$$

However, the integrated result obtained from concatenation may still be high-dimensional, containing a large number of features. To facilitate easier visualization, interpretation, and analysis, dimensionality reduction is performed using principal component analysis(PCA) [45], which allowing us to capture the most informative features while reducing the complexity of the integrated dataset. Therefore, the resulting low-dimensional representation, known as joint embedding, is used in downstream analyses to provide a unified view of the data across different omics.

$$Output = PCA\,(Res) \tag{12}$$

Results

scMFG effectively identified cell types

In this study, we conducted experiments to evaluate the effectiveness of feature grouping in improving the accuracy of cell type identification in single-cell multi-omics integration. To assess the performance of our method, we compared it with other paired dataset integration tools. In our evaluation, we applied the scMFG method to four complex datasets: 10x_lymph node, share_skin, kidney, and 10x_pbmc. For the clustering analysis, we employed

the Leiden algorithm [40] and evaluated the performance of the model using established clustering accuracy metrics such as the Adjusted Rand Index (ARI) [41] and Normalized Mutual Information (NMI) [42]. Higher ARI and NMI scores indicate improved clustering accuracy and alignment with the reference standards. These metrics have been widely used in previous literature [43] to assess clustering performance in single-cell analysis. Our method achieved the highest ARI scores on the kidney and 10x_pbmc datasets. In terms of the NMI metric, our method also attained the highest scores, except for the 10x_lymph dataset, where scJoint outperformed others with the highest ARI and NMI scores. (Fig. 2, A and B). This outcome provides evidence for the effectiveness of feature grouping in enhancing the accuracy of cell type identification.

To investigate the impact of feature grouping on cell subtype identification, we designed a simulation study using five synthetic datasets derived from the kidney dataset. In the first dataset, we selected four well-defined subtypes of tubule cells, resulting in a total of 5,183 cells. From this subset, we randomly selected 5,000 features from the gene expression data and 20,000 features from the chromatin accessibility data to construct our initial simulated dataset. This feature selection process was carefully controlled to ensure that the datasets retained sufficient biological variability for accurate subtype identification. For the remaining datasets, we retained the same feature set but varied the number of cells by random selection to create different levels of sampling sparsity. Specifically, we created the second, third, fourth, and fifth datasets by randomly selecting 4,000, 3,000, 2,000, and 1,000 cells, respectively, as shown in Table S3. The result showed that our approach excelled in performance with a reduced number of cells (fig. S1). We used UMAP to visualize the integrated results of the five methods on the fifth dataset. As depicted in Fig. 2C, our method achieved optimal clustering results, outperforming these methods in accurately separating these four cell subtypes.

As research increasingly focuses on identifying rare cell types, we applied our method to this task. Typically, a cell type is considered rare if it makes up less than 3% of the total cell population. However, some rare cell types can be even less prevalent. To test the efficacy of each tool in identifying these extremely rare cell types, we set the proportion of rare cells to 1% and evaluated the models accordingly. Using the neuips dataset, we selected data from batches s1d1, s2d1, and s4d1 to create four simulated datasets, each replicated five times. The first dataset includes only two cell types, while the remaining three datasets feature four main cell types and one rare type, with the number of cells increasing. Details about these datasets are provided in table S4. Our evaluation used NMI and Purity [44] as metrics and compared our method with three clustering-like tools: GiniClust [45], RaceID [46], and SCMER [47]. For visualization, we examined the results on the second simulated dataset. The scatter plot showed that our method achieved the highest metrics, with GiniClust as the closest competitor (Fig. 2, D and E). Results for the other datasets are available in the supplementary material (Supplemental Tables 5 to 8).

Overall, these findings highlight the crucial role of feature grouping in improving the identification of cell types, enabling a more comprehensive and nuanced understanding of cellular heterogeneity.

scMFG unraveled functional diversity in transit amplifying cells

Single-cell data often contains a significant amount of noise, which poses challenges for analysis. To demonstrate the importance of feature grouping, we applied the scMFG method to the share_skin dataset, the cell type labels for this dataset are known. We employed the UMAP [48] to visualize the cellular distribution of highly variable genes. Interestingly, we observed a noticeable overlap between two transit amplifying cells (TACs) subtypes (Fig. 3A). By employing the LDA-based feature grouping method, we grouped the highly variable genes into multiple groups. In particular, we focused on the 14th group due to its inclusion of the marker genes specific to the TAC_2 subtype. Subsequently, we utilized the grouped features for clustering analysis, which resulting in a more accurate classification of the TACs.

To further validate the performance of scMFG, we also visualized the results of other methods in identifying these two TAC subtypes on the same dataset (Fig. 3A). Methods like scMVP [25], Cobolt [26], and MOFA+ [24] showed a significant overlap between the two TAC subtypes, indicating a less effective separation. Seurat [12] and GLUE [28] performed better than the aforementioned methods, with some overlap but an overall improved separation. scJoint [27] showed partial overlap for the two subtypes, with some regions demonstrating clear separation and others less so. This comparative analysis emphasizes the importance of feature grouping in accurately identifying and distinguishing cell types in the presence of noise.

Additionally, if we have prior knowledge of certain marker genes, we can further refine our feature grouping. For instance, in this dataset, we selected three known marker genes for the TAC_2 cell subtype: Shh, Fbp1, and Krt73. We used Spearman correlation [49] to calculate the correlation between each gene and these three marker genes, selecting the top 10% of genes with the highest correlation for each marker gene. We then identified the intersection of these three sets of genes, resulting in a total of 86 genes. Using these 86 genes as

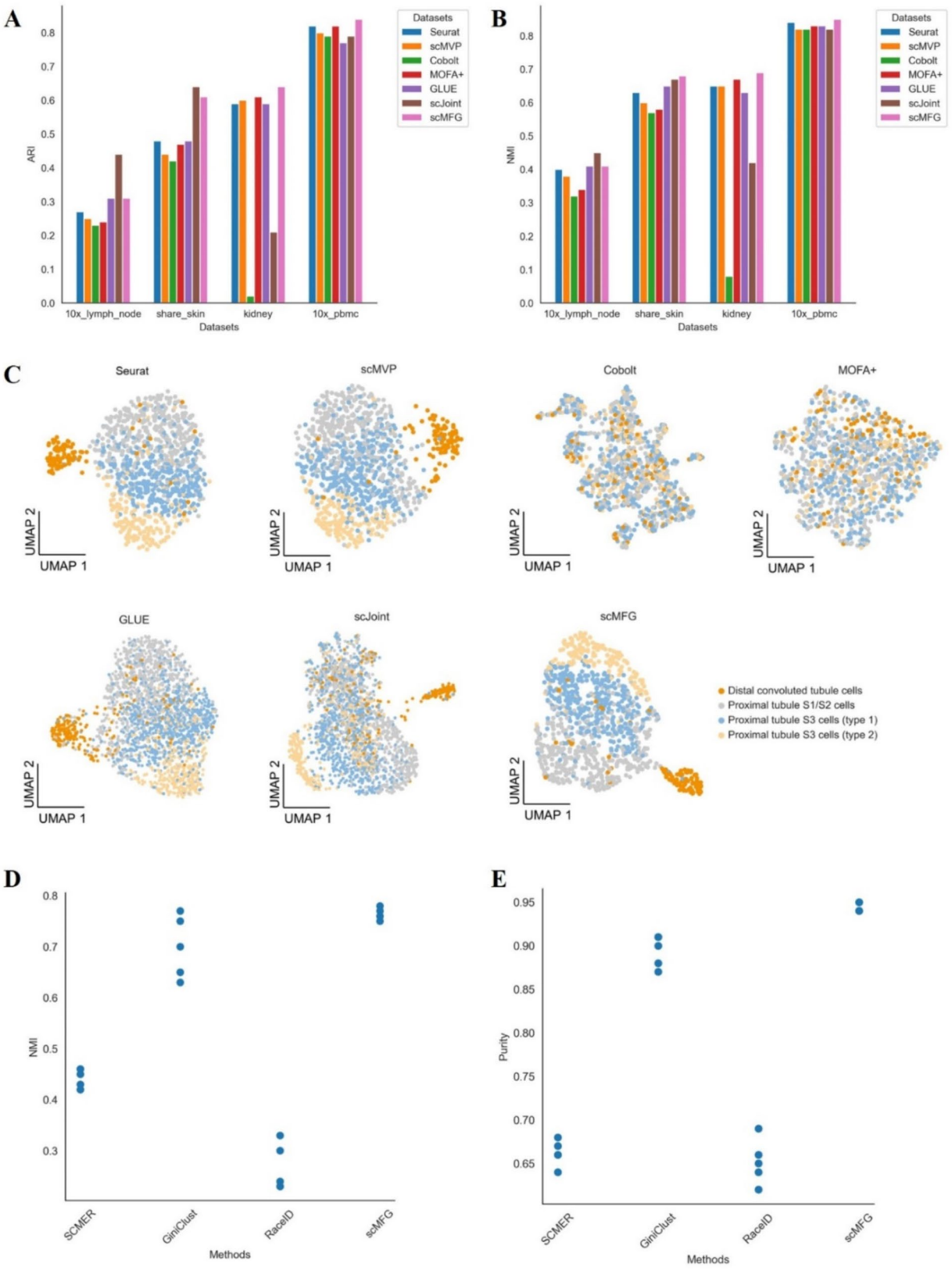

Fig. 2 scMFG effectively identified cell types. (**A**) ARI metric with the varying of the datasets. (**B**) NMI metric with the varying of the datasets. (**C**) The UMAP visualizations of the reduced dimensionality space generated by Seurat, scMVP, Cobolt, MOFA+, GLUE, scJoint and scMFG. In these visualizations, cells are color-coded according to their corresponding cell types. (**D**) NMI metric with the varying of the methods. (**E**) Purity metric with the varying of the methods

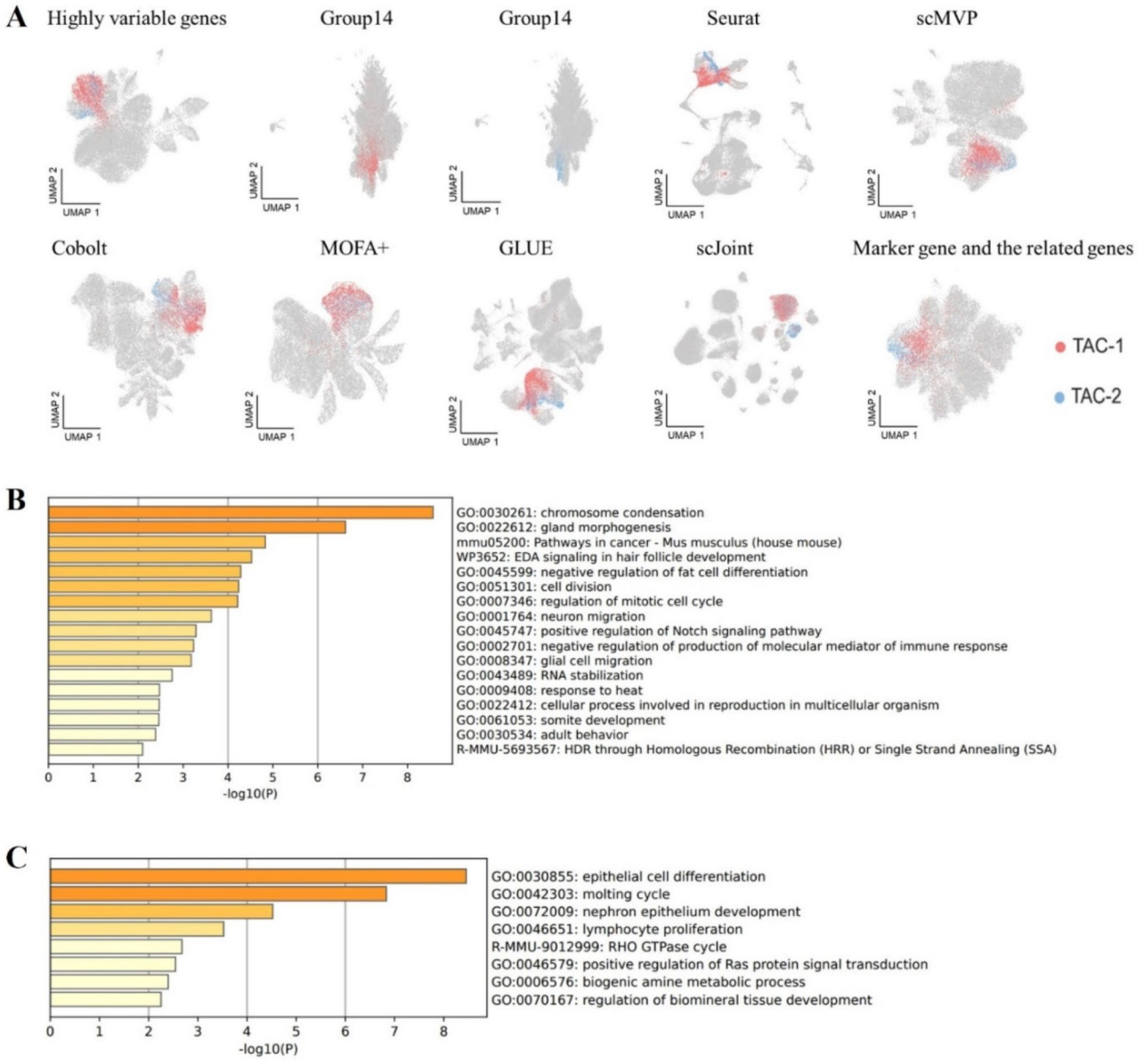

Fig. 3 scMFG unraveled functional diversity in transit amplifying cells. (**A**) Visualization results of TAC cell subtypes TAC-1 and TAC-2 using all highly variable genes, genes from the 14th group obtained through feature grouping, and results from Seurat, scMVP, Cobolt, MOFA+, GLUE, scJoint, and genes selected based on their high correlation with marker genes. (**B**) Enrichment analysis results of the group 13, indicating functions primarily related to cell division, such as chromatin condensation. (**C**) Enrichment analysis results of the group 14, showing functions mainly associated with cell differentiation

a feature group, we performed clustering and visualization (Fig. 3A). The results showed a better separation of the TAC subtypes, demonstrating that if we have prior knowledge, we can leverage it for more precise grouping, thereby improving identification accuracy. The reason behind the effectiveness of feature grouping lies in its ability to bring together functionally relevant genes or features that contribute to specific biological processes. By grouping these features, scMFG focuses on the informative aspects of the data, effectively reducing the influence of noise and non-discriminative genes that can lead to mixed or ambiguous clusters.

To gain further insights into the functionality of the grouped features, we conducted enrichment analysis using the Metascape tool [50]. This analysis revealed distinct functions associated with the grouped features within the TACs, as depicted in Fig. 2, B and C. For example, the enrichment analysis of the 13th group (Fig. 3B) showed that the functions of this group are primarily related to cell division, including chromatin condensation and other cell division processes. On the other hand, the enrichment analysis of the 14th group (Fig. 3C) indicated that the functions of this group are mainly associated with cell differentiation. These results suggest

that our method can reveal the dual role of TACs in both cell division and differentiation. The enrichment of these features in cell differentiation suggests their contribution to cell maturation and the differentiation pathway. Additionally, their enrichment in cell division indicates their involvement in promoting rapid cell turnover. The clear differentiation observed, facilitated by scMFG emphasizes the dual role of TACs in maintaining tissue integrity and regeneration processes [51]. In the context of transit amplifying cells, the feature grouping approach effectively captured and highlighted their unique gene expression patterns related to cell differentiation and division. This not only improves the separation of cell types but also uncovers the functional relevance and biological processes associated with specific cell types.

Overall, our study provides valuable insights into the significance of feature grouping for robust cell type identification and functional characterization.

scMFG enhanced interpretability by linking cell state with joint embedding

To investigate the biological implications embedded in the joint representation, we employed the scMFG method on the snare P0 dataset, which consists of 19 distinct cell types [9]. By utilizing stacked violin plots, we visualized the associations between the six dimensions of the joint embedding and the various cell types (Fig. 4A). The results revealed that specific dimensions of the latent vectors, particularly latent0, effectively captured the unique cell states of Ex6_Tle4 cells. Furthermore, latent1 demonstrated the ability to distinguish IP_Hmgn2 cells, while latent2 successfully differentiated In_Nxph1(In_1) cells. These findings underscore the effectiveness of scMFG in establishing meaningful connections between cell states and the joint embedding.

In the subsequent analysis, we performed Spearman correlation calculations [49] to examine the relationship between all genes and the latent vectors. Our findings revealed an intriguing pattern: for each dimension of the latent vectors, there was a specific gene exhibiting the highest correlation, serving as a marker gene for a particular cell type. For example, the gene Hs3st4 displayed the highest correlation with latent0, serving as a marker gene specifically associated with the Ex6_Tle4 cell [9] (Fig. 4, B and C). This observation was further supported by its unique expression pattern. Similar results were observed for the Trps1 and Galntl6 genes (Fig. 4, D and E). Furthermore, as previously mentioned, latent0 effectively captured the Ex6_Tle4 cell. Therefore, this observation implies that even in the absence of prior knowledge about cell types, our integrated dataset and subsequent generation of latent representations enable the identification of

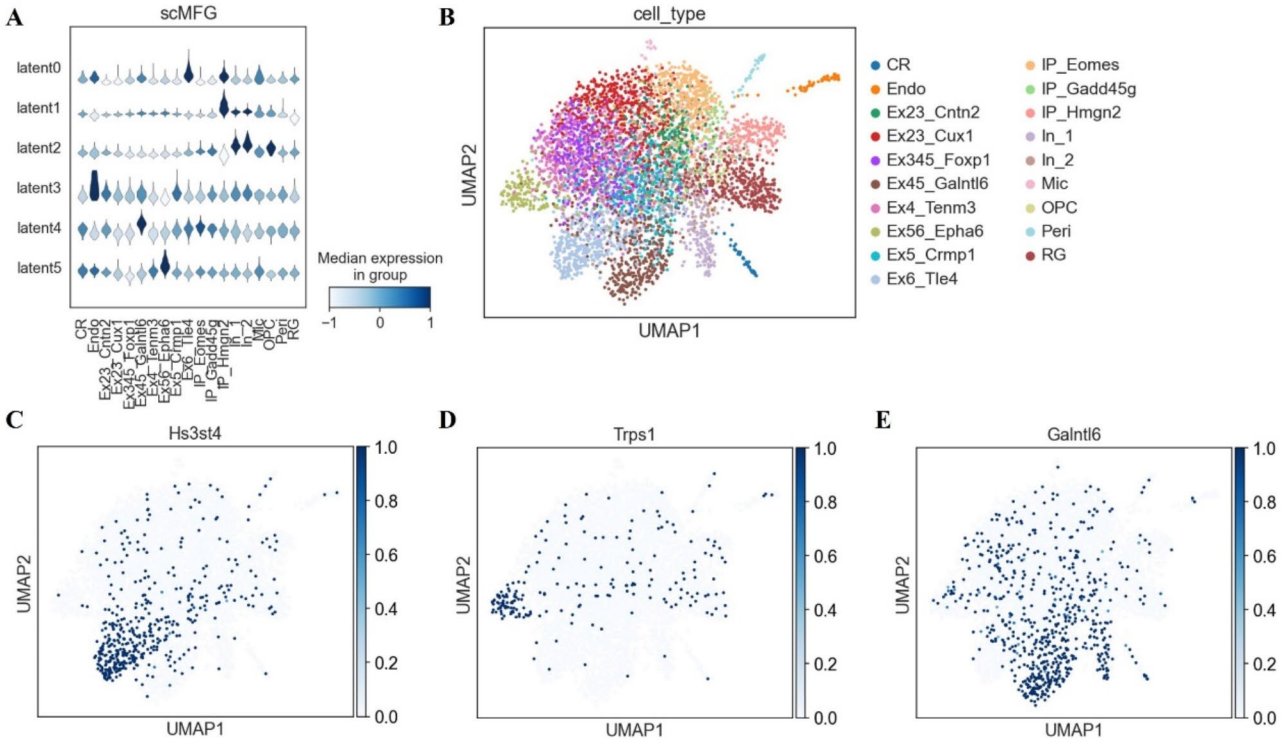

Fig. 4 scMFG enhanced interpretability by linking cell state with joint embedding. (**A**) Visualizations of the joint embedding of five dimensions. (**B**) Visualizations the distribution of every cell and cells are color-coded according to their corresponding cell types. (**C**) Expression pattern visualization of the Hs3st4 gene. (**D**) Expression pattern visualization of the Trps1 gene. (**E**) Expression pattern visualization of the Galntl6 gene

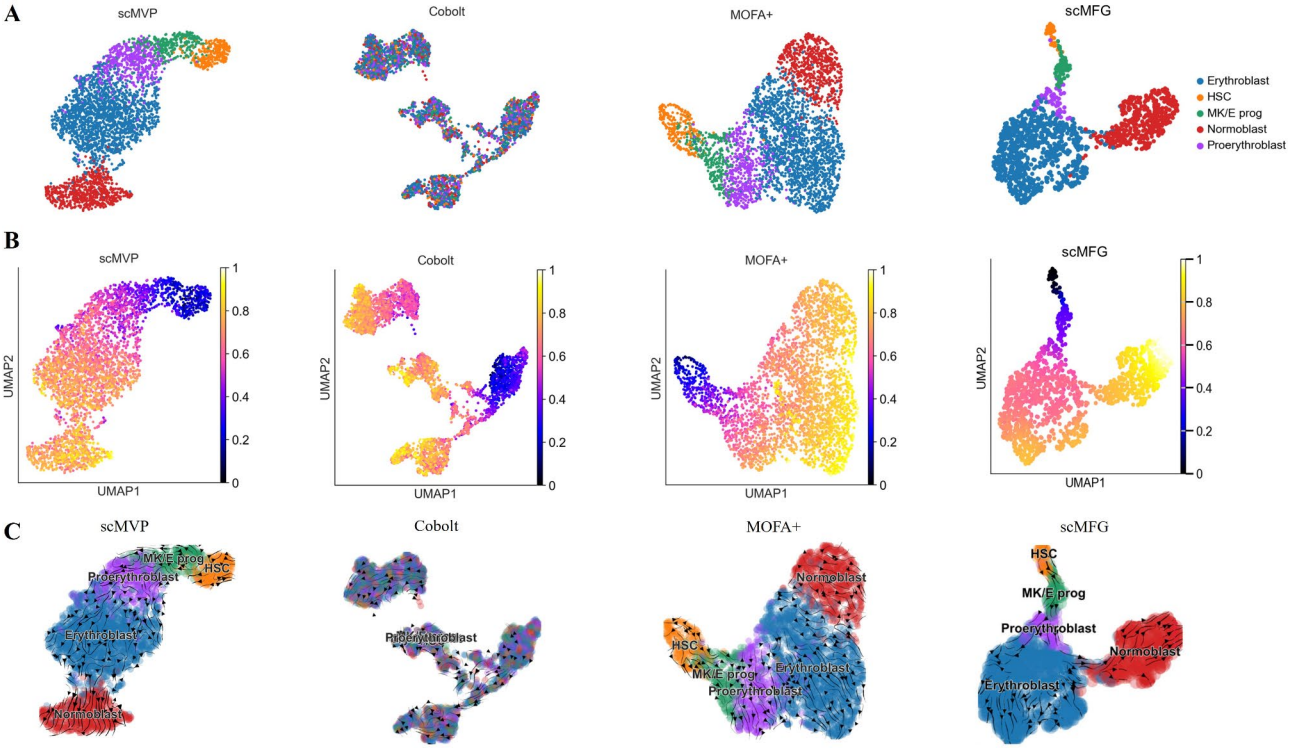

Fig. 5 scMFG facilitated the understanding of lineage relationships by leveraging joint embedding. (**A**) UMAP visualization of the distribution and relationships inferred between the five cell types using the integrated embeddings. (**B**) The UMAP plot for pseudo time-series analysis. (**C**) Visualize k-NN graph-based transition matrices

potential marker genes. It underscores the robustness of our approach in uncovering novel insights into cellular heterogeneity and facilitating marker gene discovery.

scMFG facilitated the understanding of lineage relationships

Building on the success of scMFG in cell type identification and enhanced interpretability, we extended its application to pseudotime analysis. We investigated the dynamics of cellular differentiation on neuips dataset. Our validation process focused on three distinct scenarios to comprehensively evaluate the effectiveness of scMFG in trajectory inference. The first scenario addressed technical variation. We selected batches s1d1, s2d1, and s4d1 from the neuips dataset, all derived from the same donor but measured at different sites, thus capturing variations primarily introduced by technical factors. The second scenario involved biological variation. For this, we selected batches s4d1, s4d8, and s4d9 from the neuips dataset. These batches are from different donors but measured at the same site, thus containing primarily biological variations. The third scenario encompassed both biological and technical variations. We selected batches s1d1, s2d1, and s3d3 from the neuips dataset. The first two batches (s1d1 and s2d1) are from the same donor measured at two different sites, while the third batch (s3d3) is from a different donor measured at a third site. This dataset includes both biological and technical variations.

To provide a comprehensive evaluation of our model, we focused on two critical aspects: batch effect correction and trajectory inference quality. For batch effect correction, we employed the Modified Average Silhouette Width of batch (batch ASW) [52]. This metric ranges from 0 to 1, with higher values indicating better performance in maintaining batch integrity. While most methods require batch labels as input, both scMFG and scJoint operate without them. Notably, although Cobolt achieved the highest batch ASW score, it demonstrated poor separation of cell types, as illustrated in Fig. 5A. Furthermore, scJoint relies on cell annotation information, whereas our method does not. Despite these differences, our results show that scMFG performs comparably to other leading methods on this metric (fig. S2). Seurat was excluded from this comparison as it lacks the feature data necessary for calculating inter-cell distances, which limits its applicability for this specific metric.

For trajectory inference, we focused on the development of Hematopoietic stem cells (HSC) to Normoblast cells, evaluating performance using the Trajectory Conservation Score [53]. This score, based on Spearman's rank correlation coefficient between pseudotime values

before and after integration, is scaled between 0 and 1, with higher values indicating better trajectory conservation. GLUE and scJoint were excluded from this comparison because they concatenate results from multiple modalities, leading to each cell appearing twice in the analysis. This duplication makes it unclear which representation should be used for pseudotime analysis, complicating a fair comparison. Thus, we compared scMFG with three other approaches, our method achieved the highest scores across all three datasets (fig. S3), demonstrating scMFG's superiority in preserving accurate pseudotime trajectories.

To further illustrate these results, we visualized the cell type identification, pseudotime analysis, and cell transition matrices based on pseudotime for scenario 1 (Fig. 5A). Using Palantir [54], we ordered the cells along a pseudotime trajectory with joint embedding (Fig. 5B), and employed CellRank [55] to compute a directed cell-cell transition matrix according to the Palantir pseudotime, visualizing KNN graph-based transition matrices via streamlines (Fig. 5C).

The results revealed that although Cobolt achieved the highest batch ASW value, it mixed all cell types together, indicating poor separation. While scMVP and MOFA + maintained trajectory conservation, scMVP's pseudotime analysis exhibited less distinct temporal ordering among the five cell types compared to scMFG and MOFA+. MOFA + showed trajectory changes but less clarity in the cell transition matrix, particularly in the transition from Proerythroblast to Erythroblast. In contrast, scMFG demonstrated a more pronounced and clearer transition between these cell types, highlighting its superior performance in delineating cell differentiation processes.

In a word, through the analysis of joint embedding latent features, scMFG offered a comprehensive view of the hierarchical relationships and developmental trajectories within complex biological systems.

Discussion

Integrating single-cell multi-omics data is crucial for achieving a comprehensive understanding of cellular heterogeneity. However, single-cell data inherently contains noise, and treating each omics layer as a whole can introduce additional noise that hinders accurate cell type identification. This additional noise can arise from irrelevant features, which are those that do not significantly contribute to the distinguishing characteristics of different cell types. To overcome these limitations, we propose scMFG, a novel method that employs feature grouping techniques for the effective integration of single-cell multi-omics data. In scMFG, we achieve feature grouping by assigning each feature within every omics layer to a specific topic. This innovative approach in scMFG significantly reduces the influence of irrelevant features, thereby enhancing the accuracy and reliability of the analysis. Moreover, different omics layers, such as gene expression, DNA methylation, and chromatin accessibility, provide unique data types and feature information. To ensure comparability across these diverse data types, scMFG applies the same feature grouping approach. In order to maintain the interpretability of the integrated feature groups, scMFG introduces matrix factorization-based methods. Overall, our work presents a robust framework for the analysis of single-cell multi-omics data.

However, it is essential to acknowledge the limitations of our approach. The choice of variables can significantly influence the quality of the integrated results [56, 57]. While we employed a generic method to select highly variable features, we recognize that variable selection methods can be susceptible to outliers, heavy-tailed errors, and model misspecifications. These concerns highlight the necessity of incorporating robustness checks in future work to evaluate how different variable selection methods may impact the outcomes of multi-omics integration. Additionally, the effectiveness of variable selection can vary based on the biological context and the specific characteristics of the datasets analyzed, underscoring the need for a tailored approach in future studies. And the fixed number of groups for each omics modality presents a limitation in capturing the unique patterns and structures inherent in the data. Future research could focus on developing adaptive methods for feature grouping based on data characteristics, enhancing the precision and depth of multi-omics integration.

Conclusions

In this study, we introduced scMFG, a novel method that leverages the grouping of features and integrates them using a group-based approach. This design enabled scMFG to comprehensively handle sequencing datasets that measure multiple omics within the same cell. We've shown its effectiveness in accurately identifying cell types, enhancing interpretability of joint embeddings, and facilitating the understanding of lineage relationships. These experiments underscore scMFG's utility in revealing complex biological patterns and developmental trajectories, positioning it as a significant advancement in single-cell analysis and a valuable tool for deciphering cellular heterogeneity in multi-omics data.

Abbreviations

LDA	Latent Dirichlet Allocation
KNN	Nearest neighbors
PCA	Principal component analysis
GEO	Gene Expression Omnibus
10x_lymph_node	Fresh frozen lymph node with B cell lymphoma dataset
human PBMCs	10x_pbmc datasets

TACs	Transit amplifying cells
ARI	Adjusted Rand Index
NMI	Normalized Mutual Information
In_1	In_Nxph1
batch ASW	Modified Average Silhouette Width of batch
HSC	Hematopoietic stem cells

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11319-0.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
Prof. L.Y. thanks to all those who maintain excellent databases and to all experimentalists who enabled this work by making their data publicly available. And our work is supported by High-Performance Computing Platform of Xidian University.

Author contributions
All authors contributed to the article. L.T.M, J.T.L. and L.Y. formulated the model. L.T.M. was responsible for implementing the algorithm and performing simulation studies in this research. C.G.Z. and W.S. performed a baseline comparison on real data. All authors participated in the real data analysis presented in this paper. L.T.M. and L.Y. were responsible for writing the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China [grant nos. 62072353 and 62272065], the Shaanxi Science and Technology Foundation [grant nos. 2024JC-YBMS-620].

Data availability
The datasets analyzed in this study are available from the GEO repository under the following accession numbers: GSE117089 [8], GSE126074 [9], GSE140203 [10], GSE194122 [37]. 10x_lymph_node and 10x_pbmc datasets are obtained from www.10xgenomics.com/resources/datasets.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 20 August 2024 / Accepted: 3 February 2025
Published online: 11 February 2025

References

1. Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. Nat Rev Nephrol. 2018;14(8):479–92.
2. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34(11):1145–60.
3. Altschuler SJ. LF Wu 2010 Cellular heterogeneity: do differences make a difference? Cell 141 4 559–63.
4. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33(2):155–60.
5. Dai C, Jiang Y, Yin C, Su R, Zeng X, Zou Q, Nakai K, Wei L. scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods. Nucleic Acids Res. 2022;50(9):4877–99.
6. Wang J, Chen Y, Zou Q. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. PLoS Genet. 2023;19(9):e1010942.
7. Zhao M, He W, Tang J, Zou Q, Guo F. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. Brief Bioinform. 2022;23(2):bbab568.
8. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5.
9. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019;37(12):1452–7.
10. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, Ding J, Brack A, Kartha VK, Tay T. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell. 2020;183(4):1103–16. e1120.
11. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018;9(1):781.
12. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87. e3529.
13. Zhang ZL, Cui FF, Su W, Dou LJ, Xu AQ, Cao C, Zou Q. webSCST: an interactive web application for single-cell RNA-sequencing data and spatial transcriptomic data integration. Bioinformatics. 2022;38(13):3488–9.
14. 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.
15. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. Brief Bioinform. 2021;22(4):bbaa287.
16. Zuo C, Dai H, Chen L. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. Bioinformatics. 2021;37(22):4091–9.
17. Gayoso A, Lopez R, Steier Z, Regier J, Streets A, Yosef N. A joint model of RNA expression and surface protein abundance in single cells. Biorxiv 2019:791947.
18. Martinez-de-Morentin X, Khan SA, Lehmann R, Qu S, Maillo A, Kiani NA, Prosper F, Tegner J, Gomez-Cabrero D. Adaptative machine translation between paired single-cell Multi-omics Data. BioRxiv. 2021;2021(2001):2027–428400.
19. Ma A, Wang X, Li J, Wang C, Xiao T, Liu Y, Cheng H, Wang J, Li Y, Chang Y. Single-cell biological network inference using a heterogeneous graph transformer. Nat Commun. 2023;14(1):964.
20. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics. 2020;36(14):4137–43.
21. Wang X, Sun Z, Zhang Y, Xu Z, Xin H, Huang H, Duerr RH, Chen K, Ding Y, Chen W. BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. 2020;48(11):5814–24.
22. Singh R, Hie BL, Narayan A, Berger B. Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. Genome Biol. 2021;22(1):1–24.
23. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124.
24. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):1–17.
25. Li G, Fu S, Wang S, Zhu C, Duan B, Tang C, Chen X, Chuai G, Wang P, Liu Q. A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data. Genome Biol. 2022;23(1):20.
26. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 2021;22(1):1–21.
27. Lin Y, Wu T-Y, Wan S, Yang JY, Wong WH, Wang YR. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40(5):703–10.
28. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022;40(10):1458–66.

29. Stanojevic S, Li Y, Ristivojevic A, Garmire LX. Computational methods for single-cell multi-omics integration and alignment. Genom Proteom Bioinform. 2022;20(5):836–49.
30. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133–45.
31. Zeng X, Wang F, Luo Y, Kang S-g, Tang J, Lightstone FC, Fang EF, Cornell W, Nussinov R, Cheng FJCRM. Deep generative molecular design reshapes drug discovery. Cell Rep Med. 2022;4:100794.
32. Wang R, Jiang Y, Jin J, Yin C, Yu H, Wang F, Feng J, Su R, Nakai K, Zou Q. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Res. 2023;51(7):3017–29.
33. Yang Y, Gao D, Xie X, Qin J, Li J, Lin H, Yan D, Deng K. DeepIDC: a Prediction Framework of Injectable Drug Combination based on heterogeneous information and deep learning. Clin Pharmacokinet. 2022;61(12):1749–59.
34. Xu J, Xu J, Meng Y, Lu C, Cai L, Zeng X, Nussinov R, Cheng FJCRM. Graph embedding and gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. Cell Rep Methods 2023:100382.
35. Azad A, Vafaee F. Single cell data explosion: deep learning to the rescue. arXiv Preprint arXiv:190106105 2019.
36. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
37. Alessandri S, Ratto ML, Rabellino S, Piacenti G, Contaldo SG, Pernice S, Beccuti M, Calogero RA, Alessandri L. CREDO: a friendly customizable, REproducible, DOcker file generator for bioinformatics applications. BMC Bioinformatics. 2024;25(1):110.
38. Luecken MD, Burkhardt DB, Cannoodt R, Lance C, Agrawal A, Aliee H, Chen AT, Deconinck L, Detweiler AM, Granados AA. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*: 2021; 2021.
39. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
40. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233.
41. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
43. Lee MY, Kaestner KH, Li M. Benchmarking algorithms for joint integration of unpaired and paired single-cell RNA-seq and ATAC-seq data. Genome Biol. 2023;24(1):244.
44. Jain H, Grover R, LIET A. Clustering analysis with purity calculation of text and sql data using k-means clustering algorithm. IJAPRR. 2017;4(44557):47–58.
45. Jiang L, Chen H, Pinello L, Yuan G-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17:1–13.
46. Grün D. Revealing dynamics of gene expression variability in cell state space. Nat Methods. 2020;17(1):45–9.
47. Liang S, Mohanty V, Dou J, Miao Q, Huang Y, Müftüoğlu M, Ding L, Peng W, Chen K. Single-cell manifold-preserving feature selection for detecting rare cell populations. Nat Comput Sci. 2021;1(5):374–84.
48. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv Preprint arXiv:180203426 2018.
49. Spearman C. The proof and measurement of association between two things. 1961.
50. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.
51. Rangel-Huerta E, Maldonado E. Transit-amplifying cells in the fast lane from stem cells towards differentiation. Stem Cells Int 2017, 2017.
52. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9.
53. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41–50.
54. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'Er D. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol. 2019;37(4):451–60.
55. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, Ansari M, Schniering J, Schiller HB. CellRank for directed single-cell fate mapping. Nat Methods. 2022;19(2):159–70.
56. Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. Brief Bioinform. 2015;16(5):873–83.
57. 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.

Publisher's note