BMC Biology

RESEARCH

Open Access

scDEBGCL: a deep embedding approach based on bipartite graph contrastive learning for single-cell RNA-seq data

Jing Wang[1], Delei Ke[1], Junfeng Xia[2], Ao Liu[3], Yansen Su[4*] and Chun-Hou Zheng[4*]

Abstract

Background Single-cell RNA sequencing (scRNA-seq) allows for the measurement of gene expression at the transcriptomic level with single-cell precision, thereby deepening our comprehension of cellular heterogeneity. However, the high dimensionality and sparsity of scRNA-seq data impede downstream analyses (such as cell clustering and trajectory inference), and learning effective embedded representations of the data has become a key aspect in scRNA-seq data analysis.

Result We present scDEBGCL, a novel deep embedding algorithm based on bipartite graph contrastive learning. scDEBGCL leverages Singular Value Decomposition (SVD) for bipartite graph enhancement and integrates graph contrastive learning, graph reconstruction, and data reconstruction to jointly learn low-dimensional embedded representations of cells, which are used for downstream tasks such as cell clustering, trajectory inference, and marker gene identification. Specifically, scDEBGCL first converts the gene expression matrix into a cell-gene bipartite graph and applies SVD to this bipartite graph for graph enhancement. This strategy effectively preserves the global cell-gene interactions and facilitate the learning of global synergistic signals within the data. To further capture the discriminative cellular representations, scDEBGCL performs contrastive learning between the enhanced graph and the original bipartite graph. Then, scDEBGCL integrates the contrastive learning loss, bipartite graph reconstruction loss, and ZINB distribution-based reconstruction loss to jointly optimize and learn the low-dimensional representations of cells for downstream analyses such as cell clustering, cell trajectory inference, and marker gene identification.

Conclusions Experiments results demonstrate that scDEBGCL is a useful GCL framework for deep embedding in scRNA-seq data, providing a reliable foundation for various downstream analyses.

Keywords Single-cell RNA sequencing, Deep embedding, Bipartite graph, Singular Value Decomposition, Graph contrastive learning

*Correspondence:
Yansen Su
suyansen@ahu.edu.cn
Chun-Hou Zheng
zhengch99@126.com
Full list of author information is available at the end of the article

Background

Single-cell RNA sequencing (scRNA-seq) technology enables the measurement of transcriptome-wide gene expression at the single-cell level, which is crucial for identifying cell types [1, 2], inferring cell trajectories [3–5], investigating cellular heterogeneity [6, 7], and exploring the mechanisms of complex diseases [8–10]. However, scRNA-seq data presents significant analytical challenges due to its complex and uncertain data distribution, as well as the presence of numerous dropout events [11]. Learning the deep embedding representations of cells and genes has become a critical step in scRNA-seq data analysis, as it can provide effective feature information for downstream analyses.

Recently, a series of representation learning methods have been applied by researchers to learn effective embedding representations from scRNA-seq data. In addition to methods specifically designed for analyzing single-cell RNA sequencing (scRNA-seq) data such as SC3 [12], deep learning-based methods have been widely developed for feature learning. Autoencoder-based models accomplish downstream tasks such as cell type identification by independently learning the feature representations of individual cells or genes. For example, scDeepcluster [13] learns feature representation and clustering via explicit modeling of scRNA-seq data generation. DESC [14] gradually removes batch effects through iterative self-learning. Graph neural network (GNN)-based methods aim to capture high-order relationships between nodes through iterative information propagation, making them highly suitable for modeling complex scRNA-seq data. For instance, scGAE [15] construct cellular graphs and employ graph convolutional networks (GCNs) or attention mechanisms to learn the complex structures and feature information of data. scBiG [16] models cell-gene relationships by transforming the gene expression matrix into a cell-gene bipartite graph, and employs a model-driven graph reconstruction strategy to perform self-supervised training of the GNN on this sparse graph, thereby effectively capturing high-order topological features. Contrastive learning-based methods exhibit excellent performance on sparse data by effectively identifying discriminative feature representations. For example, Contrastive-sc [17] constructs positive/negative sample pairs through data augmentation and leverages contrastive loss to learn cell embeddings, thereby mining the latent structure of single-cell data in unsupervised scenarios. ScAGCL [18] generates meaningful clustering representations for single-cell RNA sequencing (scRNA-seq) data via adversarial contrastive learning. This process involves first constructing a cell–cell graph, with the contrastive learning supported by adversarial attacks on both graph structures and node features.

Although these methods have achieved certain success in learning low-dimensional representations of scRNA-seq data, they still have some limitations. For example, auto-encoder-based deep learning models [19–21] ignore the structural information between cells. Most GNN-based algorithms [22–24] are limited by the over-smoothing problem, which leads to indistinguishable feature representations. Furthermore, while models based on contrastive learning or graph contrastive learning such as GraphCL [25] and GCA [26] are capable of learning discriminative feature representations, they typically adopt random augmentation techniques such as node or edge perturbation, or heuristic augmentation strategies to generate contrastive views. These augmentation approaches tend to lose critical data signals, which is not conducive to preserving the global structure. Therefore, there is an urgent need for a method that can accurately capture the critical interaction information between genes and cells, and effectively learn the embedding representations of cells and genes to support reliable downstream analyses.

To address these issues, we propose the scDEBGCL (deep embedding approach based on bipartite graph contrastive learning) algorithm, which integrates SVD-based graph augmentation with GCL to learn cell embedding and intercellular topological structures. Specifically, scDEBGCL first converts the gene expression matrix into a cell-gene bipartite graph, and then retains the top k largest singular values and their corresponding eigenvectors via Singular Value Decomposition (SVD), thereby preserving critical gene-cell interactions. Unlike local graph augmentation methods (e.g., node/edge perturbation, k-nearest neighbor graph construction), which rely solely on the local neighborhood information of individual nodes, this augmentation approach is a global matrix decomposition method that can capture synergistic signals spanning the entire graph. Subsequently, scDEBGCL adopts the LightGCL [27] paradigm to contrast the primary view and the augmented graph, thereby effectively and efficiently learning informative embedding representations of cells. By injecting global collaborative relationships, LightGCL mitigates the problems caused by inaccurate contrast signals. Furthermore, scDEBGCL integrates the contrastive learning loss, bipartite graph reconstruction loss, and ZINB distribution-based reconstruction loss, which enables it to effectively capture the high-order topological patterns and data characteristics of scRNA-seq data. Experimental results demonstrate that scDEBGCL facilitates the extraction of cellular embedding representations and the reconstruction of scRNA-seq data. The learned cellular representations

exhibit excellent performance in various analyses, including cell clustering, trajectory inference, and marker gene identification. The main contributions of this study are as follows:

- We propose a deep embedding approach called scDEBGCL based on bipartite graph contrastive learning for single-cell RNA-seq data. scDEBGCL converts the gene expression matrix into a cell-gene bipartite graph and leverages SVD for graph enhancement, thereby capturing the global key information underlying cell-gene interactions.
- scDEBGCL adopts a highly effective and robust graph contrastive learning (GCL) paradigm for representation learning. It simplifies the contrastive learning (CL) framework by directly contrasting the cell embeddings derived from the SVD-augmented view and the primary view in the InfoNCE loss function.
- We combine the contrastive learning loss, bipartite graph reconstruction loss, and ZINB distribution-based reconstruction loss to jointly optimize and learn the low-dimensional embedded representations of cells.
- Experimental results demonstrate that scDEBGCL surpasses seven existing representation learning methods in single-cell clustering tasks. Furthermore, cell trajectory inference and marker gene identification analyses validate that the learned representations offer a robust foundation for downstream analyses.

Methods

The framework of scDEBGCL

scDEBGCL consists of three main components: data pre-processing & graph construction; graph augmentation & contrastive learning; and downstream analysis (Fig. 1). First, scDEBGCL pre-process the raw single-cell gene expression data and constructs a cell-gene bipartite graph based on the gene expression matrix. Subsequently, a graph convolutional network (GCN) is applied to the bipartite graph for feature extraction, learning embedding of cells and genes (upper half of the figure). Meanwhile, inspired by LightGCL [27], to enhance graph contrastive learning with global-structure-aware representation, we use an SVD-based scheme [28, 29] to efficiently distill key collaborative signals from a global perspective. And the SVD-enhanced bipartite graph likewise uses GCN to extract features of cells and genes (lower half of the figure). Finally, effective cell and gene embedding representations are learned through graph contrastive learning combined with ZINB-distribution-guided data modeling; the resulting cell embedding representations are then employed for downstream tasks such as cell clustering, trajectory inference, and marker

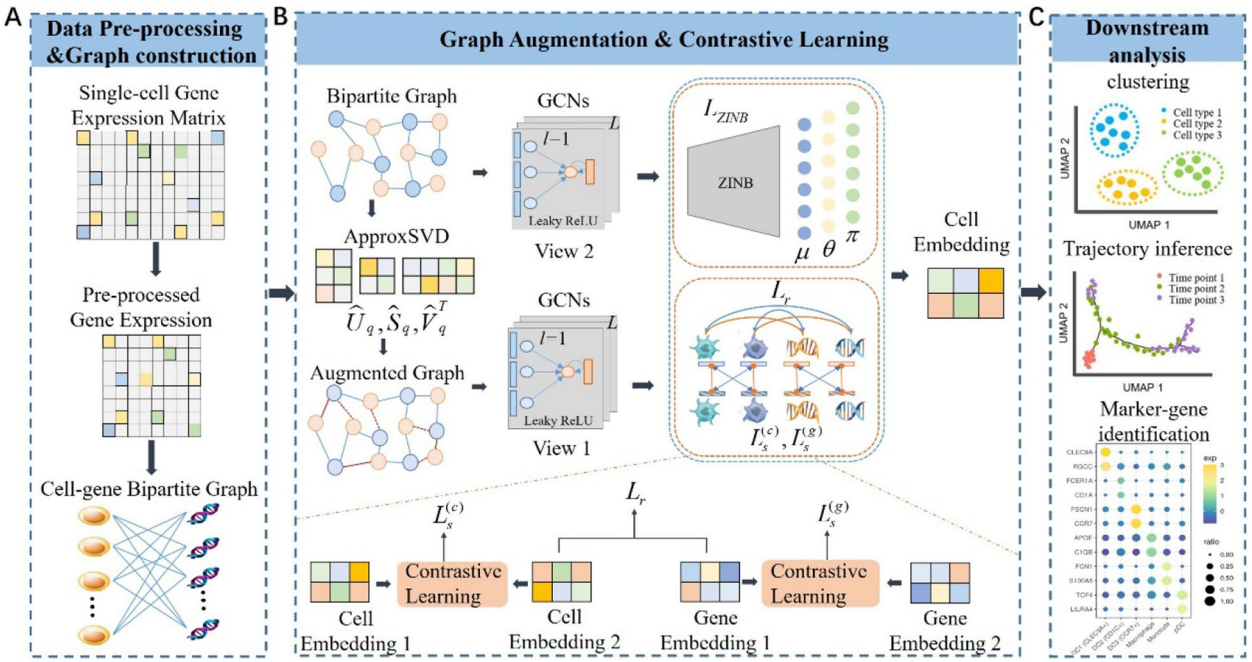

Fig. 1 Overview of the scDEBGCL workflow. **A** Data pre-processing & graph construction; **B** graph augmentation & contrastive learning; **C** downstream analysis

gene identification. The details are illustrated in the following sections.

Data pre-processing and bipartite graph construction

For clustering analysis, we collect nine real scRNA-seq datasets with cell type annotations from different platforms (Additional file 1: Table S1) [30–38]. These datasets are derived from different sequencing platforms (e.g., 10×Genomics, Smart-seq2), covering a variety of human and mouse tissues, with sample sizes ranging from 576 to 14,653 cells. For cell trajectory analysis, we download a human embryo dataset named Yan (90 cells from 6 developmental stages) [39] and a simulated dataset named DPT (500 cells and 10,000 genes) [16] with ground truth and pseudotime orders. We assume that scRNA-seq data are summarized as count matrix, and we preprocess the original count matrix through the following steps. First, we filter out genes expressed in < 3 cells and cells with < 200 expressed genes and denote the remaining matrix as $X = \left[x_{ij}\right] \in \mathbb{R}^{m\times n}$, where x_{ij} represents the expression of gene j in cell i, m represents the number of cells, and n represents the number of genes. Then, for each cell i, we calculate the library size factor l_i as:

$$l_i = \frac{\sum_{j=1}^{n} x_{ij}}{median_{i'}\left\{\sum_{j=1}^{n} x_{i'j}\right\}} \tag{1}$$

According to the size factor, the counts are normalized and log-transformed as:

$$\widetilde{x}_{ij} = \log\left(\frac{x_{ij}}{l_i} + 1\right) \tag{2}$$

In addition, we define the gene factor s_{ij} as the maximum expression value of gene j across all cells:

$$s_j = max_i\left\{\widetilde{x}_{ij}\right\} \tag{3}$$

To represent the relationship between cells and genes, we convert the gene expression matrix into a cell-gene bipartite graph denoted as $A = \left[a_{ij}\right] \in \mathbb{R}^{m\times n}$:

$$a_{ij} = \begin{cases} 1, \ \textit{if gene } j \textit{ is expressed in cell } i, \\ 0, \ \textit{otherwise}. \end{cases} \tag{4}$$

The edge between gene j and cell i is defined as a positive edge if $a_{ij} = 1$ and a negative edge if $a_{ij} = 0$.

Local graph structure learning

To capture the local graph structure information, we assign each cell c_i and gene g_j with an embedding vector $e_i^{(c)}, e_j^{(g)} \in \mathbb{R}^d$, where d is the embedding size. The collections of all cell and gene embeddings are defined as $E^{(c)} \in \mathbb{R}^{m\times d}$ and $E^{(g)} \in \mathbb{R}^{n\times d}$, respectively. Then, we adopt a two-layer GCN to aggregate the neighboring information for each cell and gene. In layer l, the aggregation process is expressed as follows:

$$z_{i,l}^{(c)} = \sigma\left(\widetilde{A}i,: \cdot E_{l=1}^{(g)}\right), \ z_{j,l}^{(g)} = \sigma\left(\widetilde{A}:,j \ \cdot E_{l-1}^{(c)}\right) \tag{5}$$

where $z_{i,l}^{(c)}$ and $z_{j,l}^{(g)}$ denote the lth layer aggregated embedding for cell c_i and gene g_j. $\sigma(\cdot)$ represents the activation function, which is the identity function here. $\widetilde{A}$ is the normalized adjacency matrix which is defined as follows:

$$\tilde{A} = \frac{a_{ij}}{\sqrt{\deg_c(i)}\sqrt{\deg_g(j)}} \tag{6}$$

where $\deg_c(i) = \sum_{j=1}^{n}\prod\left(a_{ij} \neq 0\right)$ and $\deg_g\left(j\right) = \sum_{i=1}^{n}\prod\left(\mathrm{a_{ij}} \neq 0\right)$ represent the row degree and col degree of the adjacency matrix A, respectively. The final embedding for a cell or gene is the sum of its embedding representations across all layers, and the inner product between the final embedding of a cell c_i and a gene g_j predicts the reconstruction of the local graph.

$$e_i^{(c)} = \sum_{l=0}^{L} z_{i,l}^{(c)}, \ e_j^{(g)} = \sum_{l=0}^{L} z_{j,l}^{(g)}, \ \widehat{y}_{ij} = e_i^{(c)T} e_j^{(g)} \tag{7}$$

Global collaborative learning

To effectively capture the global relationships between cells and genes, we introduce Singular Value Decomposition (SVD) in scDEBGCL to emphasize the principal components of the graph by identifying high-expression values in the cell-gene matrix. Specifically, we perform SVD on the normalized adjacency matrix $\widetilde{A}$ as $\widetilde{A} = USV^T$, where U/V is an $I \times I \times J \times J$ orthonormal matrix with columns being the eigenvectors of $\widetilde{A}$'s row-row/column-column correlation matrix. S is an $I \times J$ diagonal matrix storing the singular values of $\widetilde{A}$. We truncate the list of singular values to keep the largest q values, and reconstruct the normalized adjacency matrix with the truncated matrices as $\widehat{A} = U_q \, S_q \, V_q^T$, where $U_q \in \mathbb{R}^{m\times q}$ and $V_q \in \mathbb{R}^{n\times q}$ contain the first q columns of U and V, respectively. $S_q \in \mathbb{R}^{q\times q}$ is the diagonal matrix of the q largest singular values. The reconstructed matrix $\widehat{A}$ is a low-rank approximation of the normalized adjacency matrix $\widehat{A}$, for it holds that $rank\left(\widehat{A}\right) = q$. Based on the $\widehat{A}$, the message is propagated on the reconstructed graph in each layer:

$$g_{i,l}^{(c)} = \sigma\left(\widehat{A}_{i,:} \cdot E_{l-1}^{(g)}\right), \ g_{j,l}^{(g)} = \sigma\left(\widetilde{A}_{:,j} \cdot E_{l-1}^{(c)}\right) \tag{8}$$

To avoid the high cost of SVD, we adopt the randomized SVD algorithm [40] as a replacement.

$$\widehat{U}_q,\ \widehat{S}_q,\ \widehat{V}_q^T\ = ApproxSVD\left(\widehat{A}, q\right),\ \widehat{A}^{(\mathrm{SVD})} = \widehat{U}_q\ \widehat{S}_q\ \widehat{V}_q^T \tag{9}$$

where q is the required rank for the decomposed matrices, and $\widehat{U}_q \in \mathbb{R}^{I\times q}$, $\widehat{S}_q \in \mathbb{R}^{q\times q}$, $\widehat{V}_q \in \mathbb{R}^{J\times q}$ are the approximated versions of U_q, S_q, V_q. Therefore, the message propagation process with the approximated matrices can be rewritten as follows:

$$G_l^{(c)} = \sigma\left(\widehat{A}svd \cdot E_{l-1}^{(g)}\right) = \sigma\left(\widehat{U}_q\ \widehat{S}_q\ \widehat{V}_q^T\ E_{l-1}^{(g)}\right),\ G_l^{(g)} = \sigma\left(\widehat{A}svd \cdot E_{l-1}^{(c)}\right) = \sigma\left(\widehat{U}_q\ \widehat{S}_q\ \widehat{V}_q^T\ E_{l-1}^{(c)}\right) \tag{10}$$

where $G_l^{(c)}$ and $G_l^{(g)}$ are the collections of cell embedding encoded from the augmented graph and $\sigma(\cdot)$ represents the activation function.

Gene expression reconstruction

To better capture data features, we introduce the ZINB distribution into the GCN decoder to model the count of gene j in cell i:

$$ZINB\left(x_{ij}|\pi_{ij},\ \mu_{ij},\ \theta_{ij}\right)\ = \left(1-\pi_{ij}\right)\cdot NB\left(x_{ij}|\mu_{ij},\theta_{ij}\right) + \pi_{ij}\cdot\delta_0\left(x_{ij}\right) \tag{11}$$

where x_{ij} is the count value of gene j in cell i, π_{ij} is the probability of zero-inflation, NB represents the negative binomial distribution with mean μ_{ij} and dispersion θ_{ij}, and $\delta_0(\cdot)$ is the Dirac delta function. By transforming the decoding process into a learning process for the parameters of the ZINB distribution, the expression data of genes and cells can be reconstructed. Inspired by scBig, we also integrate a generalized matrix factorization (GMF) model [41] into the GCN decoder to learn the ZINB parameters. The propagation process can be described as follows:

$$\begin{cases} e = e_i^{(c)}\ \odot e_j^{(g)}, \\ \mu_{ij} = \exp\left(sigmoid\left(W_\mu e\right)\times s_j\right)\times l_i, \\ \theta_{ij} = softplus\left(W_\theta e\times s_j\right), \\ \pi_{ij} = sigmoid\left(W_\pi e\right), \end{cases} \tag{12}$$

where $e_i^{(c)}$ and $e_j^{(g)}$ are the embedding representations of cell i and gene j, $\odot$ denotes the Hadamard product, $W_\mu \in \mathbb{R}^{1\times d}$, $W_\theta \in \mathbb{R}^{1\times d}$, $W_\pi \in \mathbb{R}^{1\times d}$ are weight parameters, and l_i and s_j are the library size factor and gene factor. Sigmoid and softplus are activation functions.

Objective function optimization

To better learn the embedding representations of cells and genes, we first directly contrast the SVD-augmented view embedding representations $g_{i,l}^{(c)}$, $g_{j,l}^{(g)}$ with the main-view embedding representations $z_{i,l}^{(c)}$, $z_{j,l}^{(g)}$ in the InfoNCE loss as follows:

$$L_S^{(c)} = \sum_{i=0}^{I}\sum_{l=0}^{L} -\log\frac{\exp\left(s\left(z_{i,l}^{(c)}, g_{i,l}^{(c)}/\tau\right)\right)}{\sum_{i'=0}^{I}\exp\left(s\left(z_{i,l}^{(c)},\ g_{i',l}^{(c)}/\tau\right)\right)},$$
$$L_S^{(c)} = \sum_{i=0}^{I}\sum_{l=0}^{L} -\log\frac{\exp\left(s\left(z_{i,l}^{(c)}, g_{i,l}^{(c)}/\tau\right)\right)}{\sum_{i'=0}^{I}\exp\left(s\left(z_{i,l}^{(c)}, g_{i',l}^{(c)}/\tau\right)\right)} \tag{13}$$

where $s(\cdot)$ is the cosine similarity and τ is the temperature, respectively.

Additionally, given that the inner product between the final embedding of a cell and a gene predicts the reconstruction of the bipartite graph (Eq. (7)), we designed a pairwise loss function in the model to optimize the low-dimensional representations of cell expression profiles:

$$L_r = \sum_{i=0}^{I}\sum_{s=1}^{S}\ \max\ \left(0, 1 = \hat{y}_{i,p_s}\ + \hat{y}_{i,p_n}\right), \tag{14}$$

where $\hat{y}_{i,p_s}$ and $\hat{y}_{i,p_n}$ denote the predicted scores for a pair of positive and negative items of cell i. This optimization approach of the loss function can reduce the distance between cells with similar expression characteristics in the low-dimensional space, while increasing the distance between cells with significantly different expression characteristics.

To better reconstruct the data, we define the decoder module loss as a negative log-likelihood estimation based on the Zero-Inflated Negative Binomial (ZINB) distribution.

$$L_{ZINB} = -\log\left(ZINB(X|\pi,\ \mu,\ \theta)\right) \tag{15}$$

Finally, the loss function of the entire model is defined as:

$$Loss = L_r + \alpha\left(L_s^{(c)} + L_s^{(g)}\right) + \beta L_{ZINB} \tag{16}$$

where α and β are weight parameters.

Clustering metrics

Two commonly used clustering metrics are used to evaluate the clustering performance of different methods, including normalized mutual information (NMI) [42] and adjusted rand index (ARI) [43]. NMI produce scores from 0 to 1, and the score of ARI ranges from −1 to 1. Higher values represent better clustering performances.

In addition, Q_{local} and Q_{global} [44] metrics are used to evaluate the preservation of local and global structure in the cell embedding representations.

Results and discussion

Cell embedding obtained by scDEBGCL enhance clustering performance

To evaluate the ability of scDEBGCL to learn cell representations, we apply it to nine real scRNA-seq datasets and compared it with seven state-of-the-art methods, namely Contrastive-sc, scAGCL, scGAE, scDeepcluster, DESC, SC3, and scBiG. Among them, Contrastive-sc and scAGCL are contrastive learning-based methods. scGAE, scDeepCluster, and DESC learn cell embedding features using autoencoders or graph autoencoders. SC3 is a single-cell-specific clustering algorithm. scBiG converts the gene expression matrix into a bipartite graph and captures high-order features of the data via GNN. All comparative methods are executed using their default clustering algorithms, while scDEBGCL adopt the Louvain method to ensure their optimal performance. NMI and ARI metrics are employed to evaluate clustering performance, while Q_{local} and Q_{global} metrics are used to evaluate the preservation of local and global structure in the cell embeddings (Additional file 1: Tables S2–S5). To ensure fairness, all methods except SC3 are run ten times to obtain the average results as the final cluster performance. Figure 2A and B shows the heat maps of two metrics values of these nine algorithms in nine real scRNA-seq data sets. The results indicate that scDEBGCL achieves the best clustering performance. Specifically, scDEBGCL achieves higher ARI values than the other seven clustering algorithms on six datasets (excluding Human PBMC, Adam, and Mouse ES), with particularly high ARI values of 0.9788 on the Diaphragm dataset and 0.9259 on the Human pancreas dataset. Despite not achieving the optimal ARI value on the Mouse ES dataset, scDEBGCL still ranks second. Similarly, according to NMI, scDEBGCL ranks first on seven data sets except for Mouse ES and Diaphragm. For Mouse ES, the ARI value of scDEBGCL is only lower than that of the optimal algorithm by 1.74%. For Diaphragm, the NMI value of scDEBGCL is only lower than that of the optimal algorithm by 1.60%. Tables S4 and S5 show Q_{local} and Q_{global} [44] on nine scRNA-seq data sets. The results indicate that the low-dimensional cell embeddings learned by scDEBGCL effectively maintain the global structural features of the data, thereby facilitating the accuracy of downstream analytical tasks.

To more intuitively demonstrate the clustering performance of these methods, we draw UMAP visualizations of the low-dimensional embedding representations learned by different methods. Figure 3A and B shows the UMAP visualizations of the Human pancreas dataset based on eight methods. In Fig. 3A, cells are visualized by the inferred cluster labels, while in Fig. 3B, cells are visualized by the annotated cell types. Each point denotes a cell, and different colors mean different cell clusters or cell types. As can be seen from the figure, the cell visualization of Contrastive-sc, scAGCL, DESC, SC3, and scGAE mix different cell subtypes, leading to errors, while scDEBGCL is able to clearly delineate the distribution boundaries of different pancreatic cell types. Despite achieving high clustering accuracy (ARI = 0.91), scBiG still faces the same challenges as other methods, such as the mixing of other cell types within certain cell types. In contrast, the boundaries of various cell clusters predicted by scDEBGCL are clearer than those of the other

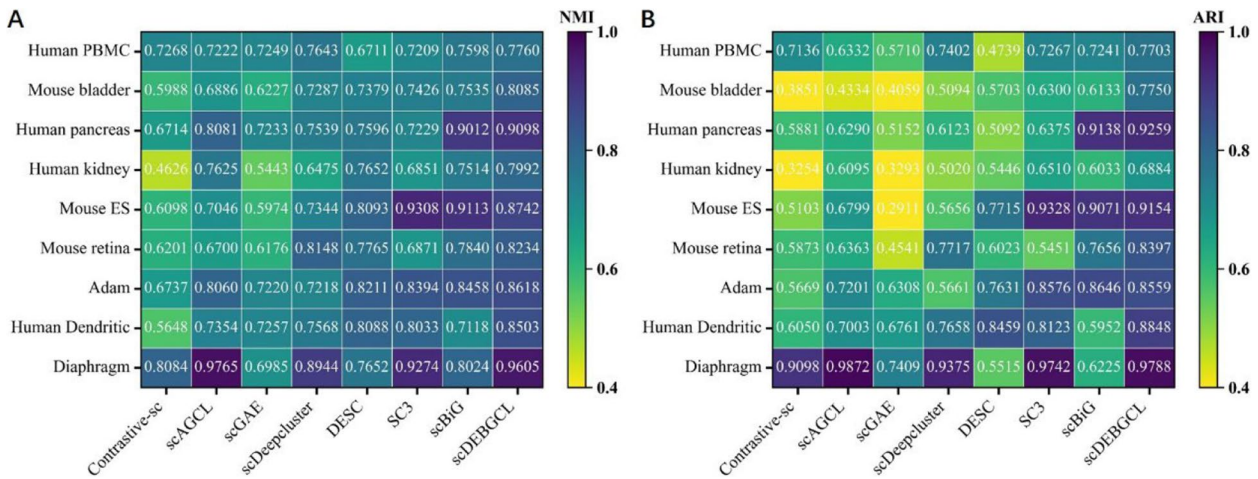

Fig. 2 Comparison of different methods in 9 scRNA-seq data sets in clustering performance. **A** Average ARI value across 10 experiments; **B** average NMI values across 10 experiments

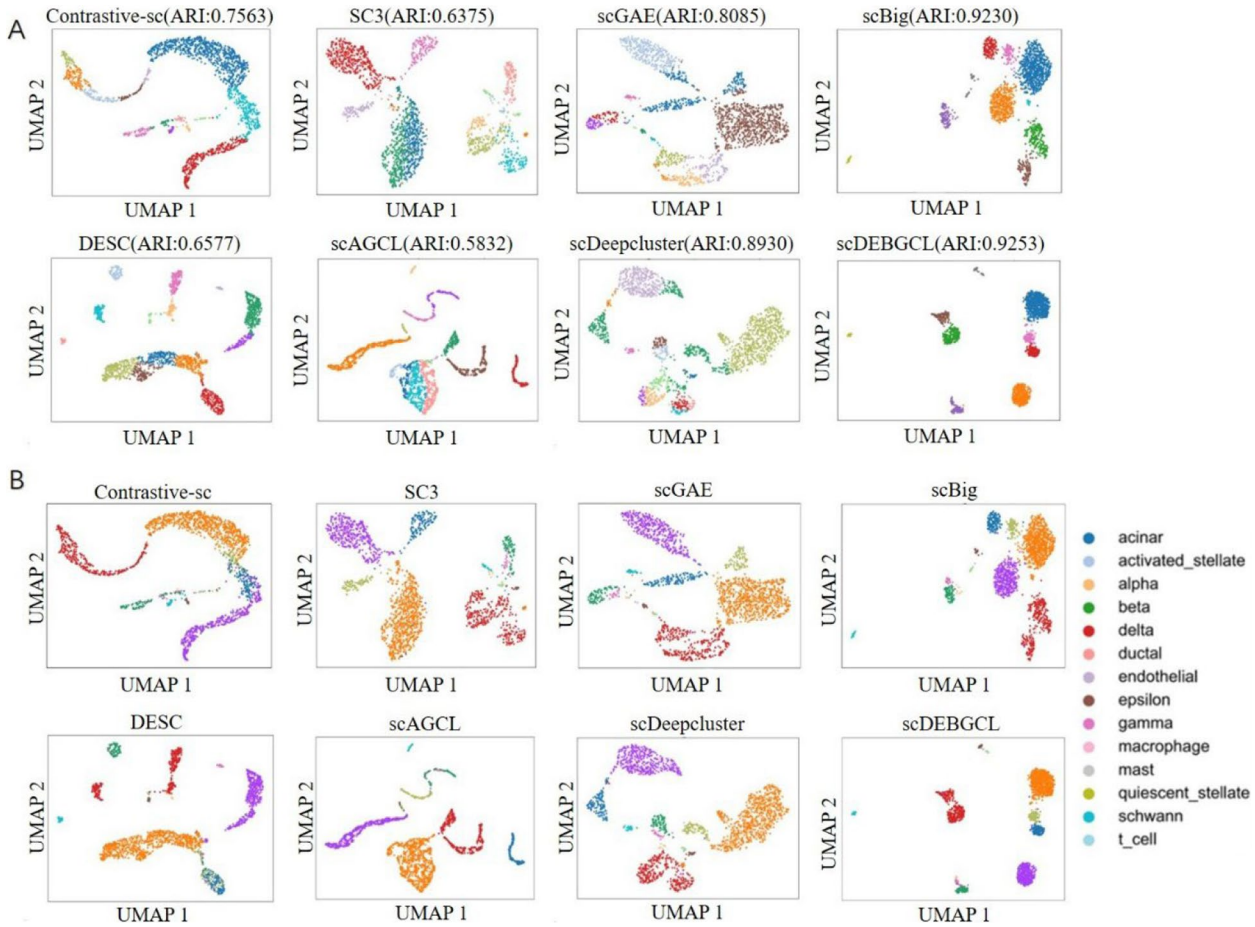

Fig. 3 UMAP visualizations of the Human pancreas dataset based on eight methods. **A** Cells are visualized by the predicted cluster labels; **B** cells are visualized by the annotated cell types

methods, and cell clusters predicted by scDEBGCL are closer to ground truth, implying that scDEBGCL is superior to other competing methods.

Parameter analysis

In scDEBGCL algorithm, the graph augmentation is guided by SVD. In SVD, the rank (q) denotes the number of largest singular values retained, directly controlling how much global collaborative signal is preserved. Applying singular-value decomposition to the cell-gene association matrix yields a low-rank approximation. A larger q retains finer details but risks introducing noise, whereas a smaller q keeps only the principal components, emphasizing dominant patterns and suppressing noise. Thus, q is a critical parameter in our model. To evaluate the impact of parameter q on model performance, we perform experiments with different q values (3, 5, 10, 20, 30) across nine datasets. As shown in Additional file 1: Table S6 and Fig. 4A and B, across the nine datasets, $q=5$ offers an optimal balance—retaining sufficient global information without introducing excessive noise, and yields high clustering performance. In comparison, setting $q=3$ yields lower overall performance across the nine datasets than $q=5$, indicating that the smaller rank cannot adequately capture the data's complexity or provide sufficient contrastive information, thereby impairing clustering performance. When q is set to 10 or a larger value, although this setting can improve the clustering performance on certain datasets (such as Adam), the performance on other datasets is not stable, and the computational cost is relatively high. Therefore, in our model, q is set to 5 to preserve the global structural information of the data and improve computational efficiency.

In addition, highly variable genes can capture more biological information than lowly variable genes that have little influence over determining cell types. To investigate the influence of highly variable genes' number for clustering results, we vary them from 1000 to 5000 and apply scDEBGCL on nine real datasets. We use Additional file 1: Table S7 and bar plot (Fig. 4C) to show the

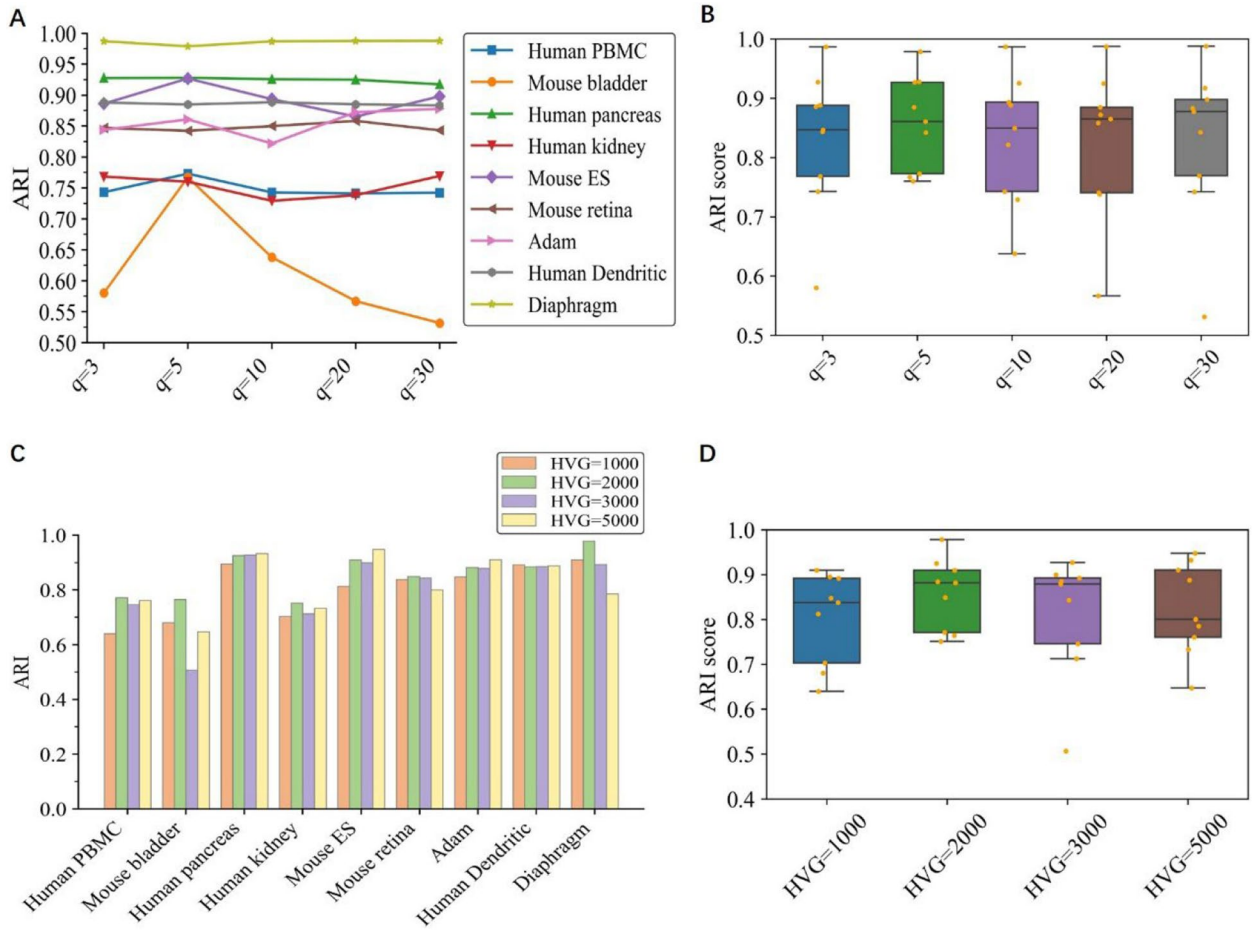

Fig. 4 ARI values of scDEBGCL on nine data sets with different parameter. **A** ARI values of scDEBGCL on nine data sets with different parameter *q*; **B** box plots of ARI scores of scDEBGCL with different *q* values on nine datasets; **C** ARI values of scDEBGCL on nine data sets with highly variable genes; **D** box plots of ARI scores of scDEBGCL with different highly variable genes on nine datasets

ARI values of nine datasets by setting 1000, 2000, 3000, and 5000 highly variable genes, respectively. Figure 4D presents boxplots of the average ARI values across nine datasets under different numbers of highly variable genes. The results indicate that the clustering performance of the model reaches the optimal value when the number of highly variable genes is set to 2000. Furthermore, although setting the number of highly variable genes to 5000 can improve clustering performance on certain datasets, it leads to a significant increase in computational resource requirements, resulting in a marked slowdown in training speed and low computational efficiency. Considering both clustering performance and computational efficiency, we select the top 2000 highly variable genes by default for clustering analysis in our method.

On the other hand, the weights of the contrastive learning loss (α) and ZINB loss (β) significantly impact model performance. To determine optimal weight parameters, we performed experiments on two datasets of varying sizes. Specifically, we assign six values (0.01, 0.05, 0.1, 0.3, 0.5, 0.8) to each weight, evaluating all 36 combinations. Figure 5 and Table S8 show the ARI scores obtained from cell clustering using cell embedding representations learned by scDEBGCL on the Human pancreas and Mouse ES datasets under different weight combinations. The results indicate that when α is greater than 0.1, the clustering performance of the model on the two datasets decreased significantly as the value of α increased. When β is fixed, the smaller the value of α, the better the model performance. This phenomenon indicates that the value of α has a direct impact on the model's clustering performance. Increasing α to make it greater than 0.1 may impair the model's ability to learn discriminative embedded representations, leading to a decline in performance. Additionally, when α is fixed at a value less than 0.1, changes in β have little impact on model performance, which indicates that the contrastive learning

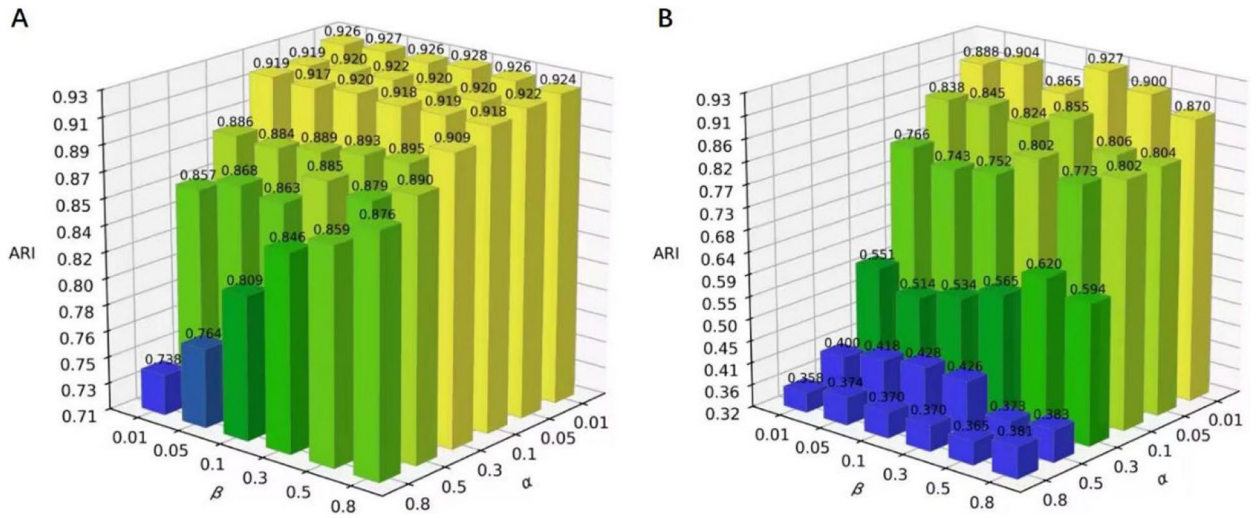

Fig. 5 ARI scores of scDEBGCL on the Human pancreas and Mouse ES datasets under different weight combinations. **A** Human pancreas dataset; **B** Mouse ES dataset

module plays a key role at this point. However, when α is greater than 0.1, the model's clustering performance declines rapidly as the value of β decreases, which suggests that data reconstruction exerts a certain effect on the model. Overall, when $\beta = 0.3$ and $\alpha = 0.01$, the model achieved the highest ARI scores on these datasets. We performed corresponding experiments on additional datasets and found that the combination of $\alpha = 0.01$ and $\beta = 0.3$ achieved the optimal or near-optimal clustering performance across most datasets. Therefore, $\alpha = 0.01$ and $\beta = 0.3$ are ultimately selected as the optimal configuration of loss weights.

We further investigate the impact of the temperature parameter τ in the InfoNCE loss (Eq. (14)). As shown in Additional file 1: Table S9, ARI values across nine datasets indicate that a setting of $\tau = 0.2$ yields the best performance on most datasets. In addition, we investigated the impact of key parameters such as the number of GCN layers and embedding dimension. Additional file 1: Tables S10 and S11 show the ARI values of scDEBGCL on nine datasets with different layer and dimension settings. The results indicate that the overall performance of the scDEBGCL model is optimal with two GCN layers and an embedding dimension of 64.

Ablation study

The scDEBGCL model consists of three components, including contrastive learning, bipartite graph reconstruction, and gene expression simulation based on the ZINB distribution. To evaluate the contribution of each component to the clustering performance of the scDEBGCL algorithm, we remove each module to generate three variants of scDEBGCL: scDEBGCL-CL (with the contrastive learning module removed), scDEBGCL-REC (with the bipartite graph reconstruction module removed), and scDEBGCL-ZINB (with the ZINB module removed). We apply four datasets from different species and sample sizes (PBMC, Mouse ES, Mouse bladder, and Adam) to perform the ablation experiments. For fairness of the comparison, the parameters of the three variants are the same as those of scDEBGCL, and two clustering metrics (NMI and ARI) are employed to evaluate the clustering performance of scDEBGCL and its variants. The results (Fig. 6 and Additional file 1: Table S12) indicate that the clustering performance of scDEBGCL is superior to its three variants on four datasets. For example, on the Mouse_bladder data set, scDEBGCL significantly increased by 10.64%, 5.74%, and 4.45% compared with its three variants (scDEBGCL-CL, scDEBGCL-REC, and scDEBGCL-ZINB) as per the ARI, respectively. The results indicate that each component of the scDEBGCL algorithm plays a crucial role.

Robustness analysis

To verify the robustness of scDEBGCL, we conduct the down-sampling experiment on nine real scRNA-seq data sets. Each data set is randomly down-sampled to obtain partial data sets containing 60%, 70%, 80%, and 90% cells, respectively. We execute pre-processing and representation learning according to the method described in the section Materials and Methods on these down-sampled data. For the testing of scDEBGCL, the same parameters are used for both the down-sampled data and the complete data to ensure consistency. Two clustering metrics (NMI and ARI) are adopted to evaluate the clustering performance of scDEBGCL on these data

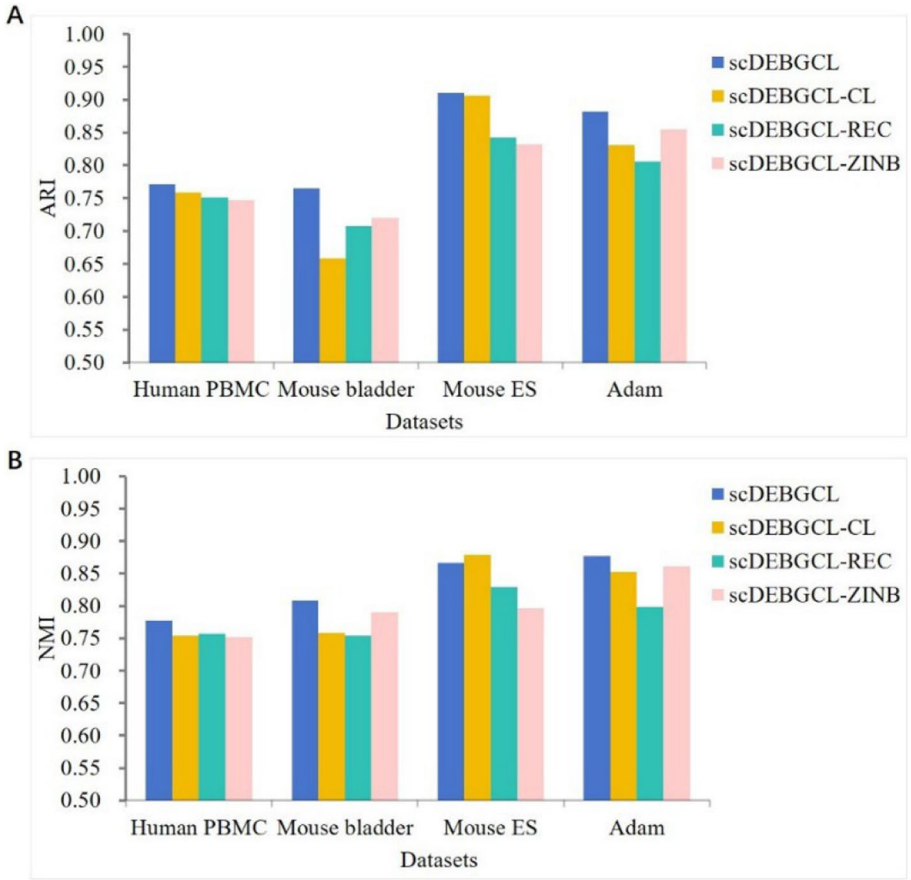

Fig. 6 Clustering performance comparison of scDEBGCL and its three variants. scDEBGCL-CL (with the contrastive learning module removed), scDEBGCL-REC (with the bipartite graph reconstruction module removed), and scDEBGCL-ZINB (with the ZINB module removed)

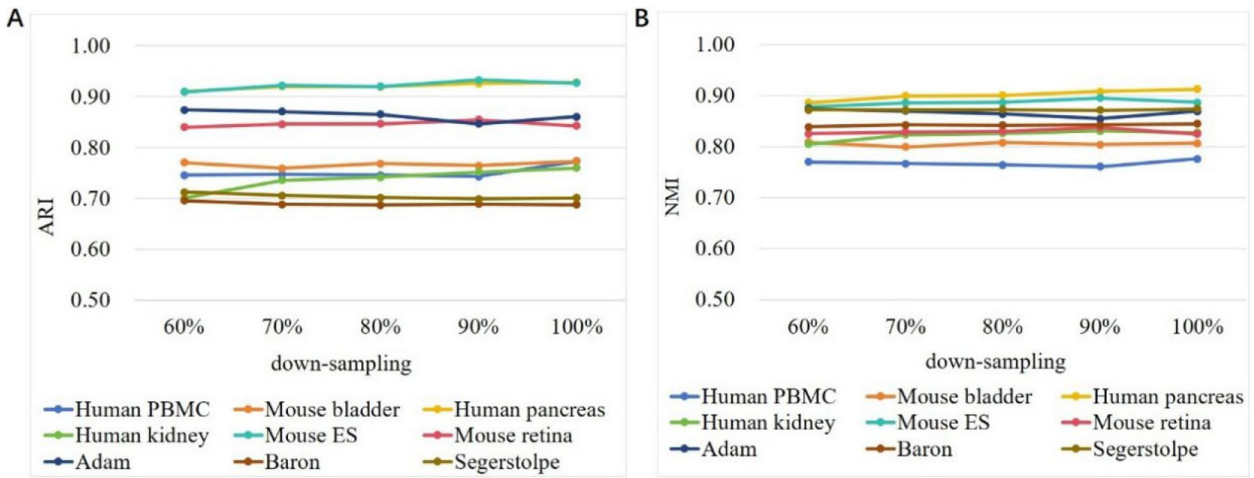

Fig. 7 Robustness experiment on nine data sets. The figure shows the changes in ARI (**A**) and NMI (**B**) values when the data is down-sampled to 60%, 70%, 80%, and 90% of the original data set size

sets. The experimental results (Fig. 7 and Additional file 1: Table S13) show that the clustering performance of scDEBGCL on the down-sampled data is not significantly different from that on the complete data, indicating that our scDEBGCL algorithm can learn effective information from incomplete data and exhibits excellent stability and robustness.

Biological analysis

To further explore the effectiveness of the cell embedding representations learned by our scDEBGCL algorithm, we apply it to the task of cell trajectory inference. Cell trajectory inference aims to construct continuous paths representing dynamic cellular processes. Here we select Slingshot [45] with default parameters to construct minimum spanning trees on clustered cells (obtained by the Louvain method in Scanpy) for lineage reconstruction. One simulated dataset (DPT) and a real scRNA-seq dataset (Yan) are selected for this analysis. The simulated dataset DPT has five branches, and each branch represents a developmental state. The Yan dataset covers multiple key developmental stages from fertilized eggs to blastocysts. The pseudotime ordering score (POS) [46] and Kendall's rank correlation (Cor) [47] coefficient are used to quantitatively assess the similarity between computationally inferred pseudotime orders and true time points.

Specifically, DPT contains 500 cells and 10,000 genes with ground truth and pseudotime orders. This simulated dataset models five cell states along a developmental trajectory. Figure 8A presents the cell trajectories inferred by different methods; the curves represent the inferred cell trajectories by Slingshot based on the obtained cell embedding. It can be seen that the cell embedding representations learned by our scDEBGCL algorithm performed the best on the simulated dataset, achieving high scores of Cor = 0.84 and POS = 0.99, and accurately reconstructing the

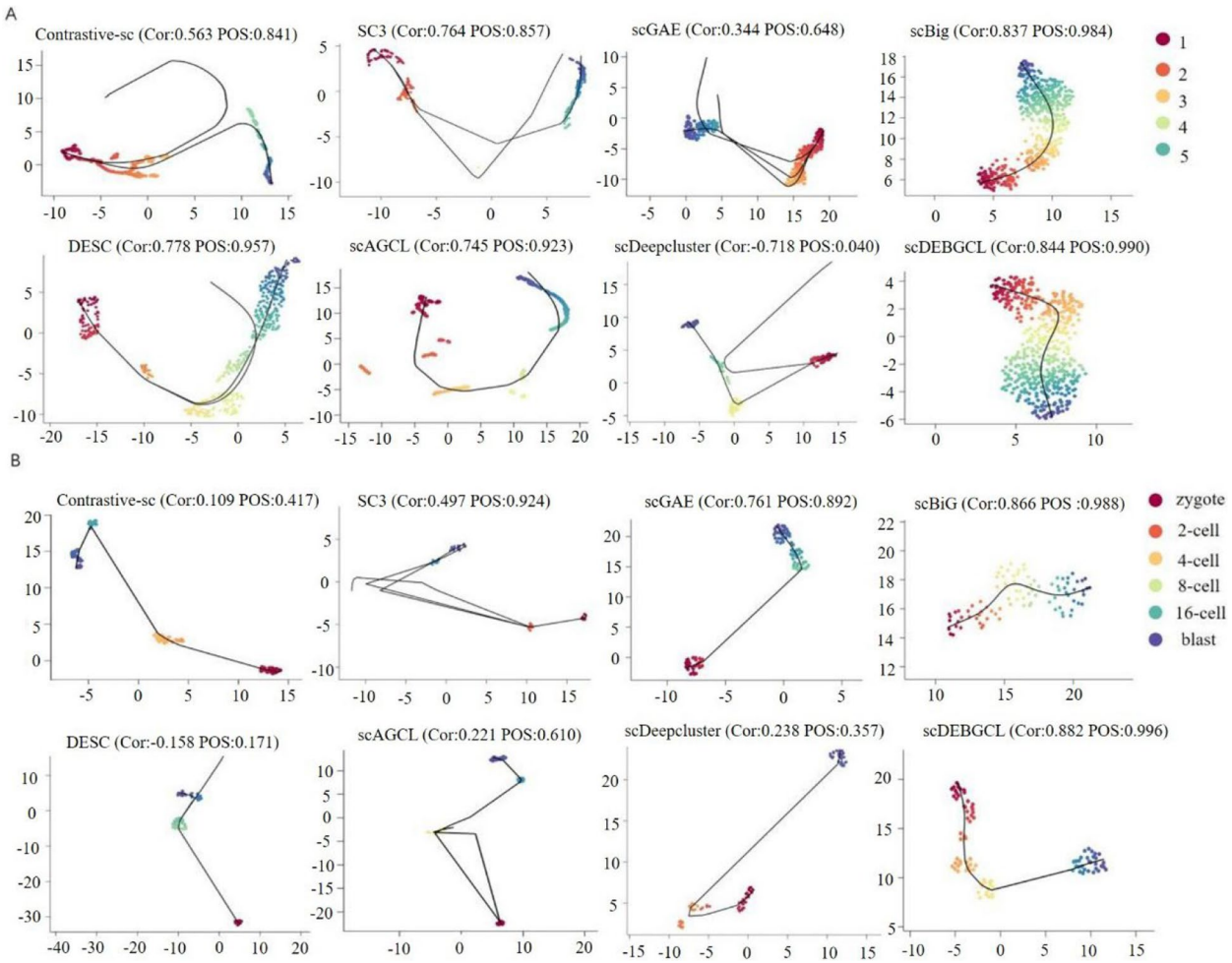

Fig. 8 Cell trajectories inferred based on different dimensionality reduction methods. The curves represent the inferred cell trajectories by Slingshot based on the obtained cell embeddings. **A** Results on the simulated dataset. **B** Results on the human embryo dataset

continuous developmental path of cells from early embryos to late embryos. The trajectory lines are smooth with reasonable directions and can well pass through cell clusters at different developmental stages. In comparison, scBiG (Cor = 0.84, POS = 0.98) and scAGCL (Cor = 0.75, POS = 0.92) also achieved excellent trajectory inference results. However, methods such as scGAE exhibited issues of trajectory deviation or stage confusion, while the trajectories generated by methods like Contrastive-sc and scDeepcluster deviated significantly from the actual biological structure and failed to reflect real biological processes. Similarly, the cell embeddings learned by scDEBGCL achieved the optimal performance on the Yan dataset with Cor = 0.88 and POS = 1.00. The visualized trajectories clearly run through all cell types, demonstrating its precise inference ability for continuous developmental states. scBiG also achieves relatively high scores (Cor = 0.87, POS = 0.99), but its trajectory paths are slightly curved and the stage boundaries are not clear enough. In comparison, methods such as scGAE and scDeepcluster exhibit trajectory jumping phenomena, while Contrastive-sc and scAGCL perform even worse in this task.

In parallel, to verify the clustering accuracy of scDEBGCL on scRNA-seq data and explore whether these clustering results are conducive to downstream biological analysis, we select Adam to perform marker gene identification. Specifically, we use COSG to identify the marker genes of each cell cluster predicted by scDEBGCL on the Adam dataset. Figure 9A and C shows the expression dot plots and gene expression heat maps of the top three marker genes identified by COSG for each cluster. As shown, the marker genes exhibit distinct expression profiles unique to their respective clusters. To further verify the consistency of these marker genes with ground-truth annotations, we also use COSG [48] to identify marker genes for each cell type (Fig. 9B and D). Figure 9 demonstrates that the marker genes identified from scDEBGCL-predicted cell clusters are highly consistent with those derived from true cell labels. For instance, the three marker genes (Slc12a1, Cldn10, and

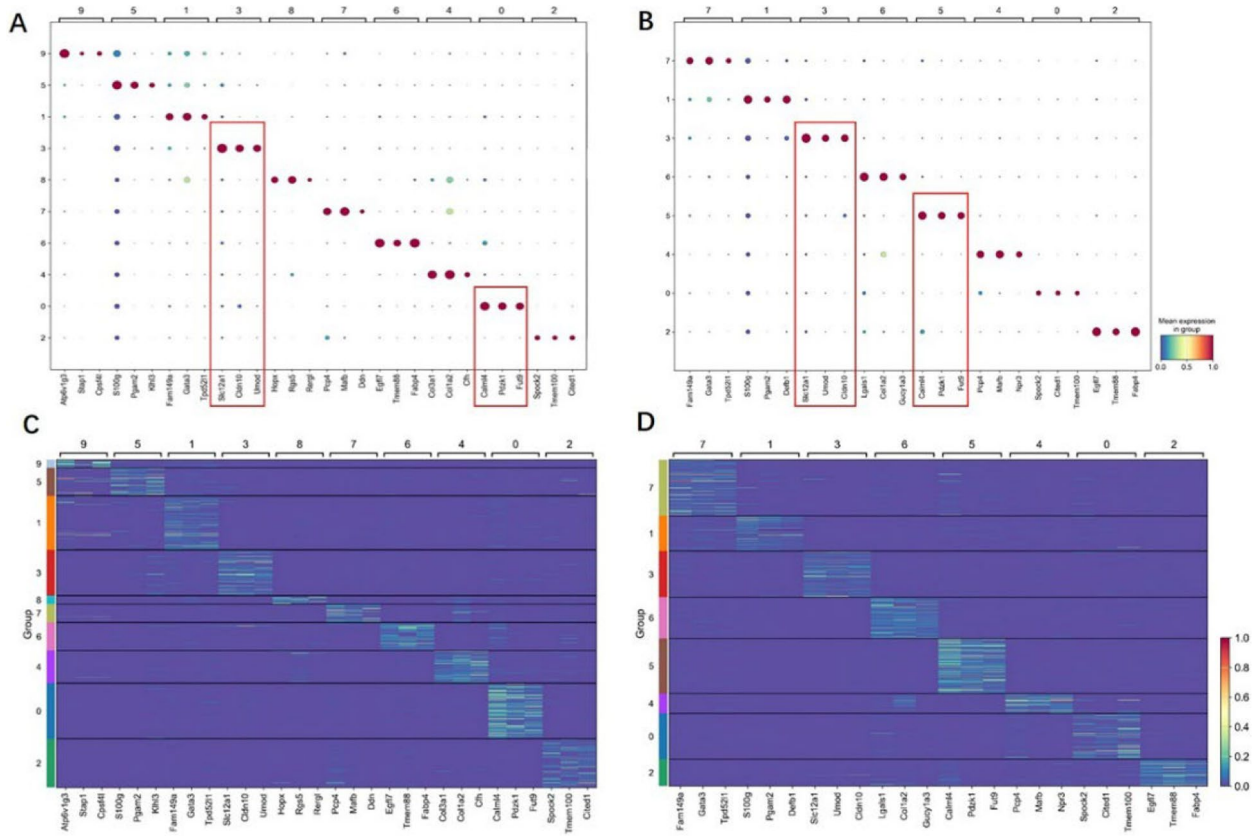

Fig. 9 Expression dot plots and heat maps of the top three marker genes identified by COSG on Adam dataset. **A** Expression dot plots of the top three marker genes identified by COSG for each cluster; **B** Expression dot plots of the top three marker genes identified by COSG for each cell type; **C** Expression heat maps of the top three marker genes identified by COSG for each cluster; **D** Expression heat maps of the top three marker genes identified by COSG for each cell type

Umod) identified for Cell Cluster 3 based on the cell labels predicted by scDEBGCL exactly match those identified based on the true cell labels. A similar situation is also observed in predicted cell cluster 0 and true cell type 5. Furthermore, Fig. 9C and D shows that the marker genes identified for each cell cluster based on the predicted cell labels not only exhibit specific expression patterns but also demonstrate high similarity in expression to those identified based on the true cell labels. This consistency indicates that the cell clusters predicted by our model closely match true cell types, which is conducive to the identification of marker genes.

In addition, we calculated the similarities between the top 50 DEGs of the predicted cell clusters and those of the real cell types. As shown in the DEGs similarity heat map for the Adam dataset (Fig. 10A), each cell type consistently aligns with a specific cell cluster exhibiting the highest DEG similarity. Specifically, six out of eight cell clusters exhibit a similarity higher than 0.9, which demonstrates that the marker genes identified from the cell clusters predicted by scDEBGCL are accurate and reliable. We also performed Gene Ontology (GO) enrichment analysis of the top 50 DEGs on one of these cell clusters (cluster 3), as shown in Fig. 10B. The results reveal significant enrichment in the processes of regulation of body fluid levels and kidney development, consistent with established literature. For instance, the Slc12a1 gene is closely associated with the overall survival of clear cell renal cell carcinoma [49]. Meanwhile, Umod, which encodes uromodulin, plays a crucial role in the pathogenesis of kidney diseases [50].

Conclusions

In this article, we introduce scDEBGCL, a graph representation learning method for learning single-cell embeddings. scDEBGCL converts the intercellular expression profiles of genes into a bipartite graph and employs Singular Value Decomposition (SVD) to enhance the bipartite graph for extracting global key gene-cell interaction information. By contrasting the primary and enhanced views, scDEBGCL effectively learns discriminative cellular feature representations. In addition, scDEBGCL jointly optimizes the model by integrating contrastive learning loss, bipartite graph reconstruction loss, and reconstruction loss based on the Zero-Inflated Negative Binomial (ZINB) distribution, so as to learn more accurate low-dimensional embedding representations of cells. Experiments on real-world data demonstrate that, compared with various existing methods, the cell embedding representations learned by scDEBGCL capture meaningful cell–cell relationships and can improve the accuracy of downstream tasks such as cell clustering, cell trajectory inference, and marker gene identification.

Although scDEBGCL can effectively learn valid feature representations of cells in scRNA-seq data and provide a reliable basis for downstream analyses, scDEBGCL has its own limitations. For instance, the model relies solely on gene expression matrices and does not yet incorporate external biological information or multi-omics data. In addition, balancing multiple loss functions during training increases computational costs. Therefore, in the future, we will consider fusing gene-cell bipartite graphs with protein–protein interaction networks [51]

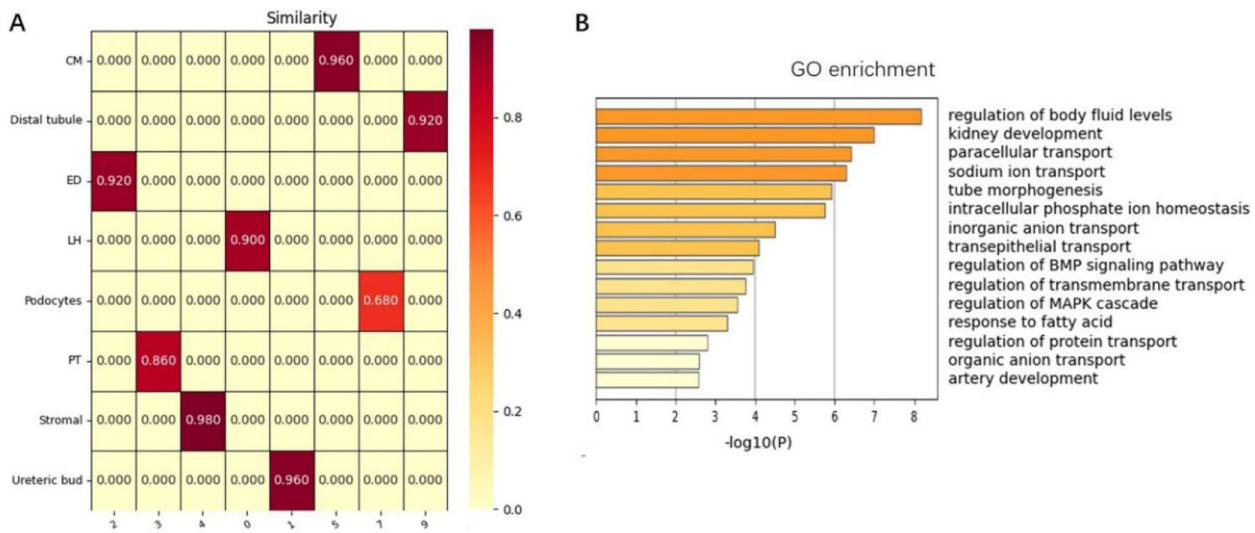

Fig. 10 Biological analysis of DEGs based on Adam dataset. **A** Comparison of similarity of DEGs in eight scDEBGCL clusters with golden standard eight cell types; **B** GO enrichment of the top 50 DEGs on one cell cluster

and leveraging multi-view graph contrastive learning to learn the embedding representations of cells and genes. To improve the efficiency of the model, we plan to adopt lightweight GNN architectures, further optimizing model for the analysis of complex biological processes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-026-02598-4.

Additional file 1: Supplementary tables. Table S1 The scRNA-seq datasets used in this study. Table S2 Clustering performance accessed by ARI on nine scRNA-seq data sets. Table S3 Clustering performance accessed by NMI on nine scRNA-seq data sets. Table S4 Clustering performance accessed by Q_{local} on nine scRNA-seq data sets. Table S5 Clustering performance accessed by Q_{global} on nine scRNA-seq data sets. Table S6 ARI values of scDEBGCL on nine data sets with different parameter *q*. Table S7 ARI values of scDEBGCL on nine data sets with different highly variable genes. Table S8 Clustering performance of scDEBGCL under different weight combinations.ARI scores on Human pancreas dataset;ARI scores on Mouse ES dataset. Table S9 ARI values of scDEBGCL on nine data sets with different parameter τ. Table S10 ARI values of scDEBGCL on nine data sets with different parameter layer. Table S11 ARI values of scDEBGCL on nine data sets with different parameter dim. Table S12 Performance comparison of scDEBGCL and its three variants. Table S13 Clustering performance of scDEBGCL on whole data and down-sampling data.NMI scores of scDEBGCL on nine datasets.ARI scores of scDEBGCL on nine datasets.

Acknowledgements
Not applicable.

Authors' contributions
J.W. conceived and designed the study. J.X. contributed to methodology development. J.W., A.L., and D.K. performed validation. J.W. drafted the original manuscript, while J.X. revised and edited it. Y.S., J.X., and C.Z. supervised the study. J.W. carried out the investigation, and A.L. curated the data. All authors read and approved the final manuscript.

Funding
This work was supported by grants from the National Natural Science Foundation of China (Nos. 62402002, 62532017, 62433001, 62322301, and 62302002) and Joint Fund Project of the Jointly-Constructed National Key Laboratory of Pathogenesis and Prevention of High-Incidence Diseases in Central Asia-Anhui University Workstation (SKL-HIDCA-2024-AH6).

Data availability
All datasets analyzed in this paper are publicly available. Specifically, Adam[[36](.)] and Diaphragm[[38](.)] are downloaded from [https://github.com/xuebaliang/scziDesk]. Yan [[16](.)] is available at https://hemberglab.github.io/scRNA.seq.datasets. Human Dendritic[[37](.)] consists of human blood dendritic cells (DC) scRNA-seq data from GEO accession GSE80171. The remaining 6 data sets [[30–35](.)] come from the website [https://zenodo.org/records/10460509]. The data and source code is available at [https://zenodo.org/records/19028601]. All data generated or analyzed during this study are included in this published article, its supplementary information files and publicly available repositories.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, China. [2]School of Life Sciences and Medical Engineering, Anhui University, Hefei, China. [3]College of Mathematics and System Sciences, Xinjiang University, Urumqi, China. [4]School of Artificial Intelligence, Anhui University, Hefei, Anhui 230601, China.

Received: 28 September 2025 Accepted: 8 April 2026
Published online: 14 April 2026

References

1. Cortese R, Adams TS, Cataldo KH, et al. Single-cell RNA-seq uncovers cellular heterogeneity and provides a signature for paediatric sleep apnoea. Eur Respir J. 2023;61:2201465.
2. Cheng C, Chen W, Jin H, et al. A review of single-cell rna-seq annotation, integration, and cell–cell communication. Cells. 2023;12:1970.
3. Saelens W, Cannoodt R, Todorov H, et al. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37:547–54.
4. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566:496–502.
5. Ke D, Cao M, Ni J, et al. Macrophage and fibroblast trajectory inference and crosstalk analysis during myocardial infarction using integrated single-cell transcriptomic datasets. J Transl Med. 2024;22:560.
6. He S, Li C, Lu M, et al. Comprehensive analysis of scrna-seq and bulk RNA-seq reveals the non-cardiomyocytes heterogeneity and novel cell populations in dilated cardiomyopathy. J Transl Med. 2025;23:17.
7. Ma P, Amemiya HM, He LL, et al. Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. Cell. 2023;186:877-891. e814.
8. Tan Z, Chen X, Zuo J, et al. Comprehensive analysis of scrna-seq and bulk RNA-seq reveals dynamic changes in the tumor immune microenvironment of bladder cancer and establishes a prognostic model. J Transl Med. 2023;21:223.
9. Zhou Y, Yu M, Tie C, et al. Tumor microenvironment-specific chemical internalization for enhanced gene therapy of metastatic breast cancer. Research. 2021:9760398.
10. Zhang L, Li Z, Skrzypczynska KM, et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. Cell. 2020;181:442-459. e429.
11. Jiang R, Sun T, Song D, et al. Statistics or biology: the zero-inflation controversy about scrna-seq data. Genome Biol. 2022;23:31.
12. Kiselev VY, Kirschner K, Schaub MT, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6.
13. Tian T, Wan J, Song Q, et al. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat Mach Intell. 2019;1:191–8.
14. Li X, Wang K, Lyu Y, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11:2338.
15. Luo Z, Xu C, Zhang Z, et al. A topology-preserving dimensionality reduction method for single-cell RNA-seq data using graph autoencoder. Sci Rep. 2021;11:20028.
16. Li T, Qian K, Wang X, et al. scBiG for representation learning of single-cell gene expression data based on bipartite graph embedding. NAR Genomics Bioinform. 2024;6:lqae004.
17. Ciortan M, Defrance M. Contrastive self-supervised clustering of scrna-seq data. BMC Bioinformatics. 2021;22:280.
18. Van Vinh L, Nhat Quang T, Hoang Hiep L, et al. Deep clustering of single-cell RNA-seq using adversarial graph contrastive learning. Brief Bioinform. 2025;26:bbaf423.
19. Yu Z, Su Y, Lu Y, et al. Topological identification and interpretation for single-cell gene regulation elucidation across multiple platforms using scMGCA. Nat Commun. 2023;14:400.
20. Qi R, Zheng CH, Ji CM et al. Cell classification based on stacked autoencoder for single-cell RNA sequencing. In: International Conference on Intelligent Computing. Cham: Springer; 2022, p. 245–259.
21. Hu H, Li Z, Li X, et al. ScCAEs: deep clustering of single-cell RNA-seq via convolutional autoencoder embedding and soft K-means. Brief Bioinform. 2022;23:bbab321.

22. Song Q, Su J, Zhang W. ScGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. Nat Commun. 2021;12:3826.
23. Min W, Wang Z, Zhu F et al. Scasdc: attention enhanced structural deep clustering for single-cell rna-seq data. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). USA: IEEE; 2024, p. 1092–1097.
24. Sun H, Qu H, Duan K, et al. Scmgcn: a multi-view graph convolutional network for cell type identification in scrna-seq data. Int J Mol Sci. 2024;25:2234.
25. You Y, Chen T, Sui Y, et al. Graph contrastive learning with augmentations. Adv Neural Inf Process Syst. 2020;33:5812–23.
26. Zhu Y, Xu Y, Yu F et al. Graph contrastive learning with adaptive augmentation. In: Proceedings of the web conference 2021. 2021, p. 2069–2080.
27. Cai X, Huang C, Xia L et al. LightGCL: simple yet effective graph contrastive learning for recommendation. 2023. arXiv preprint arXiv:2302.08191.
28. Rajwade A, Rangarajan A, Banerjee A. Image denoising using the higher order singular value decomposition. IEEE Trans Pattern Anal Mach Intell. 2012;35:849–62.
29. Rangarajan A. Learning matrix space image representations. In: International workshop on energy minimization methods in computer vision and pattern recognition. 2001, p. 153–168. Springer.
30. Zheng GX, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
31. Han X, Wang R, Zhou Y, et al. Mapping the mouse cell atlas by microwell-seq. Cell. 2018;172:1091-1107. e1017.
32. Baron M, Veres A, Wolock SL, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Syst. 2016;3:346-360. e344.
33. Young MD, Mitchell TJ, Vieira Braga FA, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science. 2018;361:594–9.
34. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.
35. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
36. Adam M, Potter AS, Potter SS. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. Development. 2017;144:3625–32.
37. Villani A-C, Satija R, Reynolds G, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356:eaah4573.
38. Schaum N, Karkanias J, Neff NF, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: the Tabula Muris Consortium. Nature. 2018;562:367.
39. Yan L, Yang M, Guo H, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20:1131–9.
40. Halko N, Martinsson P-G, Tropp JA. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. 2011;53:217–88.
41. He X, Liao L, Zhang H et al. Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. 2017, p. 173–182.
42. Strehl A, Ghosh J. Cluster ensembles–-a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3:583–617.
43. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
44. Klimovskaia A, Lopez-Paz D, Bottou L, et al. Poincaré maps for analyzing complex hierarchies in single-cell data. Nat Commun. 2020;11:2966.
45. Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19:477.
46. Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016;44:e117-e117.
47. Kendall MG. A new measure of rank correlation. Biometrika. 1938;30:81–93.
48. Dai M, Pei X, Wang X-J. Accurate and fast cell marker gene identification with COSG. Brief Bioinform. 2022;23:bbab579.
49. Peng L, Cao Z, Wang Q, et al. Screening of possible biomarkers and therapeutic targets in kidney renal clear cell carcinoma: evidence from bioinformatic analysis. Front Oncol. 2022;12:963483.
50. Devuyst O, Bochud M, Olinger E. UMOD and the architecture of kidney disease. Pflugers Arch Eur J Physiol. 2022;474:771–81.
51. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2012;41:D808–15.

Publisher's Note