BMC Genomics

RESEARCH

Open Access

scAMZI: attention-based deep autoencoder with zero-inflated layer for clustering scRNA-seq data

Lin Yuan[1,2,3], Zhijie Xu[1,2,3], Boyuan Meng[1,2,3] and Lan Ye[4*]

Abstract

Background Clustering scRNA-seq data plays a vital role in scRNA-seq data analysis and downstream analyses. Many computational methods have been proposed and achieved remarkable results. However, there are several limitations of these methods. First, they do not fully exploit cellular features. Second, they are developed based on gene expression information and lack of flexibility in integrating intercellular relationships. Finally, the performance of these methods is affected by dropout event.

Results We propose a novel deep learning (DL) model based on attention autoencoder and zero-inflated (ZI) layer, namely scAMZI, to cluster scRNA-seq data. scAMZI is mainly composed of SimAM (a Simple, parameter-free Attention Module), autoencoder, ZINB (Zero-Inflated Negative Binomial) model and ZI layer. Based on ZINB model, we introduce autoencoder and SimAM to reduce dimensionality of data and learn feature representations of cells and relationships between cells. Meanwhile, ZI layer is used to handle zero values in the data. We compare the performance of scAMZI with nine methods (three shallow learning algorithms and six state-of-the-art DL-based methods) on fourteen benchmark scRNA-seq datasets of various sizes (from hundreds to tens of thousands of cells) with known cell types. Experimental results demonstrate that scAMZI outperforms competing methods.

Conclusions scAMZI outperforms competing methods and can facilitate downstream analyses such as cell annotation, marker gene discovery, and cell trajectory inference. The package of scAMZI is made freely available at https://doi.org/10.5281/zenodo.13131559.

Keywords Clustering scRNA-seq data, Autoencoder, SimAM, Zero-inflated layer, ZINB model

*Correspondence:
Lan Ye
sdeyyelan@email.sdu.edu.cn
[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), 3501 Daxue Road, Jinan 250353, China
[2] Shandong Engineering Research Center of Big Data Applied Technology, Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), 3501 Daxue Road, Jinan 250353, China
[3] Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, 3501 Daxue Road, Jinan 250353, China
[4] Cancer Center, The Second Hospital of Shandong University, 247 Beiyuan Street, Jinan 250033, China

Background

The large amount of single-cell RNA sequencing (scRNA-seq) data provides researchers with an unprecedented opportunity to characterize different cell states and types in multicellular organisms [1]. Clustering plays a vital role in scRNA-seq data analysis, and its results affect downstream analyses, such as cell type identification [2], tumor heterogeneity [3], and cell lineage analysis [4]. However, the high drop rate and sparsity of scRNA-seq data bring huge challenges to scRNA-seq data clustering. Researchers have proposed a large number of computational methods to cluster scRNA-seq data.

Researchers apply traditional clustering algorithms to scRNA-seq data clustering. SAIC [5] utilizes an iterative *k*-means clustering to separate cells into distinct clusters. The predefined *k* can affect clustering results, and the k-means-based method is sensitive to outliers, resulting in failure to detect rare cell types. CIDR [6] is one of the representative hierarchical clustering algorithms. However, CIDR has high time complexity and is difficult to deal with large-scale scRNA-seq datasets. SSRE [7] uses sparse subspace representation and similarity enhancement strategy to cluster scRNA-seq data. Louvain [8], one of most widely used community detection algorithm for clustering scRNA-seq data, recursively merges communities into a single node and performs modular clustering. SCANPY [9] is a scRNA-seq data analysis toolkit with a clustering method based on the Louvain algorithm. Seurat [10] also uses the Louvain algorithm to cluster cell types. Community-detection-based clustering may not find small communities and rare cell types. DBSCAN [11] is the most commonly used density-based clustering algorithm. GiniClust [12] uses DBSCAN and adaptive parameter to cluster scRNA-seq data and find rare cell types. However, the adaptive parameters may lead to unreasonably large cell clusters.

Recently, many DL-based methods have achieved remarkable results on scRNA-seq data clustering. For example, DCA [13], one of the earliest DL-based algorithms, proposes a deep count autoencoder to cluster scRNA-seq data. scDeepCluster [14] maps data into a low-dimensional space via a ZINB-based autoencoder and performs clustering using kullback–leibler (KL) divergence. scGMAI [15] is an autoencoder-based Gaussian model that utilizes autoencoder to reconstruct gene expression values and uses fast independent component analysis (FastICA) [16] for dimensionality reduction. scDCC [17] adds prior knowledge as additional terms into the loss function and uses an autoencoder to cluster scRNA-seq data. DREAM [18] combines Gaussian mixture model and variational autoencoder to identify cell types. scGAE [19] clusters data by using a multi-task-oriented graph autoencoder combines with topological information and feature information. scDSSC [20], an impressive method for clustering scRNA-seq data, combines the Self-Expressiveness Property of data with autoencoders to perform deep sparse subspace clustering. SCEA [21] uses a graph attention autoencoder and an MLP-based encoder to perform clustering.

These methods have achieved remarkable results. However, there are several limitations of these methods. First, they do not fully exploit cellular features. Second, they are developed based on gene expression information and lack of flexibility in integrating intercellular relationships. Finally, the performance of these methods is affected by dropout event.

In this paper, we integrates SimAM (a Simple, parameter-free Attention Module) [22] into the modeling process to guide deep neural network to simultaneously learn meaningful cellular features and latent relationships between cells. Different from traditional autoencoder, we add ZI layer to the decoder to eliminate the impact of dropout event. We present a more flexible clustering form and is more effective in clustering scRNA-seq data. Here, we name the proposed method based on SimAM and ZI layer as scAMZI. A schematic overview of scAMZI is presented in Fig. 1. First, scAMZI constructs a data preprocessing mechanism for data standardization and quality control. Second, we introduce the autoencoder and SimAM based on ZINB model [23] to reduce dimensionality of data and learn feature representations of cells and relationship between cells. Meanwhile, a ZI layer is added to the decoder to handle zero values in the data. Finally, scAMZI uses spectral clustering combined with low-dimensional embedding features to cluster scRNA-seq data. We compare the performance of scAMZI with nine methods (three shallow learning algorithms and six state-of-the-art DL-based methods) on fourteen benchmark datasets of various sizes (from hundreds to tens of thousands of cells) with known cell types. Experimental results not only demonstrate that scAMZI outperforms competing methods, but also show that scAMZI can facilitate downstream analyses such as cell annotation, marker gene discovery, and cell trajectory inference.

Results

Ablation experiment

To evaluate the impact of SimAM and ZI layer on model performance, we constructed two variants of scAMZI, (w/o) SimAM and (w/o) ZI layer. (w/o) SimAM represents scAMZI without SimAM, and (w/o) ZI layer represents scAMZI without ZI layer. After deleting SimAM module or ZI layer, the traditional autoencoder can still calculate the results. We trained these two models using the same parameters and compared the performance of

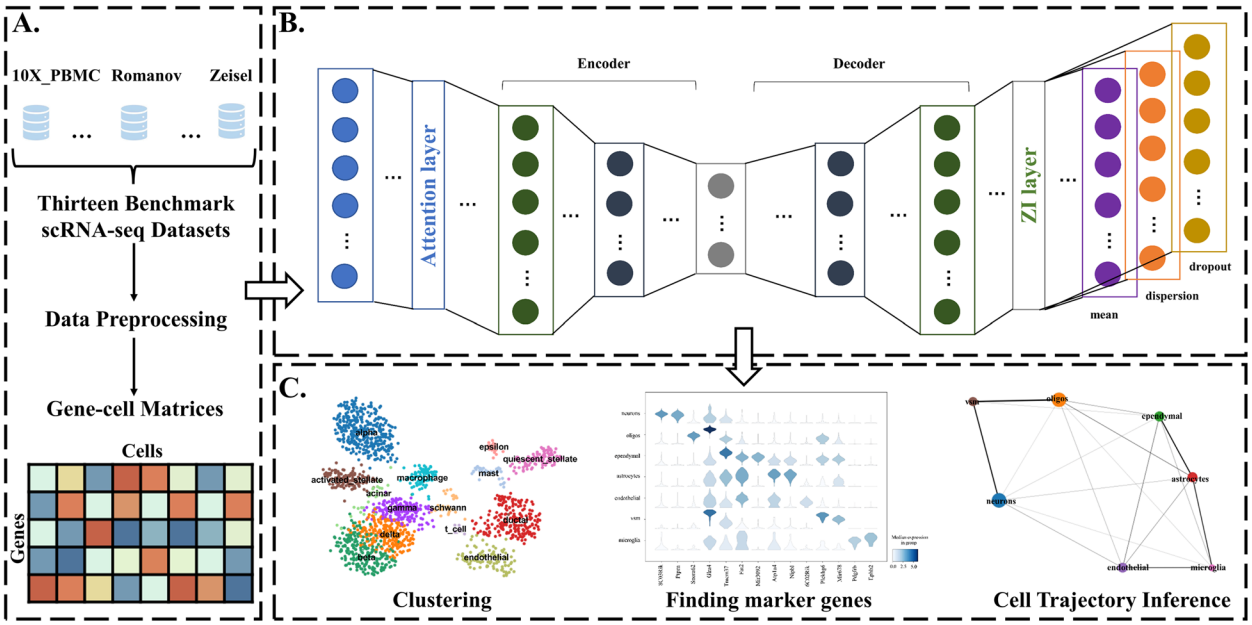

Fig. 1 Schematic overview of scAMZI. **A** The input is a gene-cell matrix from a benchmark scRNA-seq dataset. **B** The neural network architecture of scAMZI is mainly composed of SimAM, autoencoder, ZINB model and ZI layer. Based on ZINB model, we introduce autoencoder and SimAM to reduce dimensionality of data and learn feature representations of cells and relationships between cells. ZI layer is used to handle zero values. **C** scAMZI is used to cluster scRNA-seq data, find marker genes and infer cell trajectory

these models with scAMZI using fourteen benchmark scRNA-seq datasets (see 'Data preprocessing'). As shown in Fig. 2A, removing the SimAM or ZI layer results in a 9.21% and 6.79% drop in average ARI (Adjusted Rand Index), respectively. As shown in Fig. 2B, removing the SimAM or ZI layer results in an 8.41% and 6.80% drop in average NMI (Normalized Mutual Information), respectively. The detailed results of the ablation experiment were listed in S1 Table.

Experimental results show that the performance of scAMZI decreases significantly after removing the SimAM or ZI layer, and both SimAM and ZI layer are

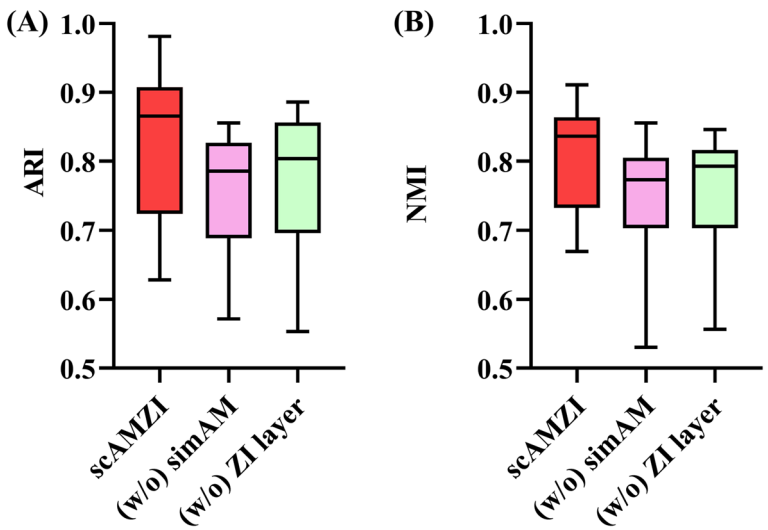

Fig. 2 **A** Performance comparison of different neural network architectures in terms of ARI in ablation experiments. **B** Performance comparison of different neural network architectures in terms of NMI in ablation experiments. (w/o) SimAM represents scAMZI without SimAM, (w/o) ZI layer represents scAMZI without ZI layer

beneficial to scAMZI. The scAMZI and ZI layer play a vital role in clustering scRNA-seq data.

Simulated dataset experiment

scRNA-seq clustering usually suffer from dropout event and cell-type imbalanced dataset in practical applications. We evaluated the clustering ability of scAMZI using simulated imbalanced datasets with different dropout rates. We used splatter [24], a commonly used tool, to generate these simulated datasets. We generated six datasets containing five cell types with a dropout rate of 0.05 and six datasets containing two cell types with a dropout rate of 0.25. The twelve datasets were generated from two different batches and contained cell-type imbalanced datasets.

As shown in Fig. 3 A and B, in terms of NMI and ARI, our proposed scAMZI achieved good performance on the twelve datasets. The NMIs and ARIs of the twelve datasets were listed in S2 Table. We selected sim3 from the simulated dataset with a dropout rate of 0.05 for visualization. As shown in Fig. 3C, scAMZI can accurately classify these five cell types. In addition, we performed differential expression analysis on the sim3 dataset to find marker gene for each cell type. Figure 3D shows that scAMZI can accurately find the marker gene for each cell type. Experimental results show that scAMZI can effectively eliminate the impacts of dropout event and cell-type imbalanced dataset.

Batch effect correction experiment

In this research, the quality of the cell representation projected to the latent space directly exerts influence on clustering performance.

To investigate whether scAMZI clusters the same cell types together in that latent space, we retained the hidden layers of scAMZI and compared the cluster results of the same cell types (endothelial cell and macrophage cell) in the original space and latent space on the integrated

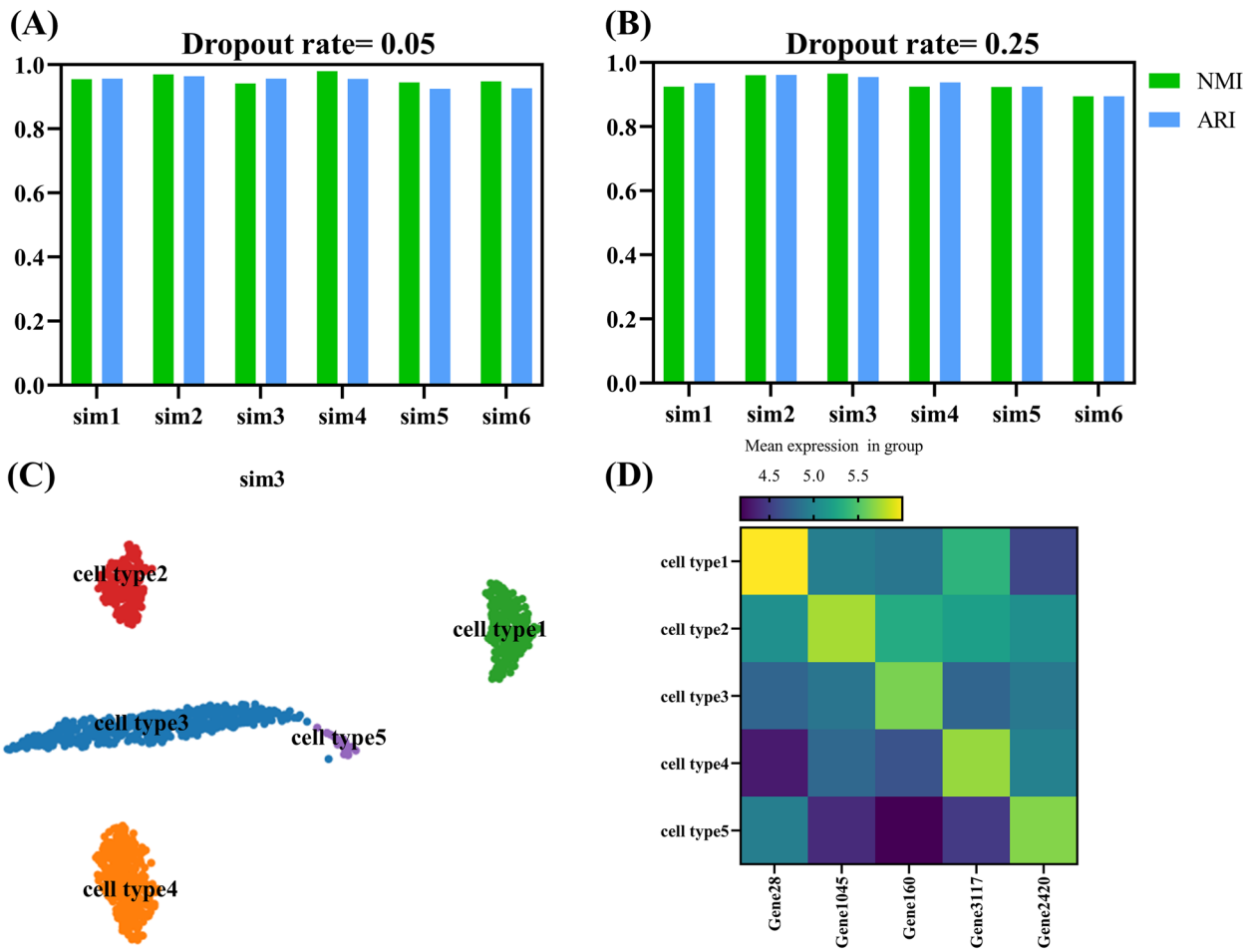

Fig. 3 **A** The performance of scAMZI on six simulated datasets containing five cell types with a dropout rate of 0.05. **B** The performance of scAMZI on six simulated datasets containing two cell types with a dropout rate of 0.25. **C** Clustering result of scAMZI on sim3 from the simulated dataset with a dropout rate of 0.05. **D** Marker genes for each cell type found by scAMZI on the first simulation data

datasets (Human1 [25], HumanLiver2 [26], Marshall [27], Zhao [28], Siletti [29], and Horeth [30]). Human1 and HumanLiver2 come from the fourteen benchmark datasets. We collected four datasets (Marshall, Zhao, Siletti, and Horeth) that contain both endothelial cell and macrophage cell. These six datasets come from different experiments, and we constructed an integrated dataset using these six datasets. The detailed information of these four datasets was listed in S3 Table. Figure 4A shows the clustering results of endothelial cells and macrophage cells in original space, and Fig. 4B shows the clustering results of endothelial cells and macrophage cells in latent space. As shown in Fig. 4 A and B, due to the batch effect, endothelial cells from six different datasets are clustered into six clusters in the original space. scAMZI projected the same cell types come from different datasets into the same latent space and clustered these cells together. For macrophage cell, the same thing was observed. The results show that scAMZI can effectively correct the batch effect.

We compared the NMIs and ARIs of scAMZI with competing methods using Human2 dataset. Figure 4C shows that scAMZI outperforms competing methods in terms of NMI and ARI. The detailed results were listed in S4 Table. We used the Human2 dataset for cell visualization. Figure 4D shows that scAMZI accurately classified the fourteen cell types in Human2 dataset. In addition, we performed differential expression analysis on the Human2 dataset to find marker gene for each cell type. Figure 4E shows that scAMZI can accurately find the marker gene for each cell type.

Performance of scAMZI on fourteen benchmark scRNA-seq datasets

In this section, we compared scAMZI with SCANPY [9], Seurat [10], SSRE [7], DCA [13], scDeepCluster [14], scDCC [17], scGAE [19], scDSSC [20] and SCEA [21] to test the clustering ability of scAMZI. These methods include three shallow learning algorithms (SCANPY, Seurat, SSRE) and six DL-based methods (DCA, scDeepCluster, scDCC, scGAE, scDSSC, SCEA). The ten methods were tested on the fourteen benchmark datasets. The evaluation metrics are ARI and NMI. Figures 5 and 6 show the ARIs and NMIs of scAMZI and competing methods on fourteen datasets, respectively.

As shown in Fig. 5, in 11 of the 14 datasets, scAMZI achieved the highest ARI values. scAMZI performed slightly worse than scDeepCluster and scGAE on Human_kidney dataset, slightly worse than scDeepCluster on CITE_CMBC dataset, and slightly worse than scDCC and scDeepCluster on Zeisel. The detailed results were listed in S5 Table. As shown in Fig. 6, in 10 of the 14 datasets, scAMZI achieved the highest NMI values. The detailed results were listed in S6 Table. In Human_kidney dataset, scAMZI performed slightly worse than scDeepCluster, scGAE and SCANPY. In CITE_CMBC dataset, scAMZI performed slightly worse than SCANPY, Seurat, scDCC, scDeepCluster and scGAE. In Human3 dataset, scAMZI performed slightly worse than SCANPY. In Zeisel dataset, scAMZI performed slightly worse than scDCC and scDeepCluster. The reason may be that ZINB cannot approximate the true distribution, or scAMZI has difficulty learning the optimal low-dimensional embedding feature representation. The above results demonstrate that scAMZI outperforms state-of-the-art methods and improves the performance of clustering methods.

The comparison methods have some shortcomings in cell feature representation and dropout processing. Specifically, scDSSC failed to fully extract cell feature representation and failed to effectively handle dropout. DCA used traditional autoencoders for encoding, without considering the impact of noise and redundant information during the decoding process, resulting in the failure of reconstructing cell features. scDCC ignored the topological relationship between cells, which affects the accuracy of clustering results. scDeepCluster focused on the characteristics of cells and failed to fully utilize the relationships between cells, which are crucial for revealing the potential similarities between cells. scGAE did not consider the impact of dropout events on clustering, which may lead to unstable and uninterpretable clustering results. SSRE enhanced the learning of similarities between cells, but its computational complexity is too high, resulting in its poor performance in large-scale data. Although SCANPY and Seurat are widely used analysis tools, they still need to be improved in terms of cell feature expression and calculation of topological relationships between cells. SCEA failed to effectively integrate the topological relationships between cells and avoid the influence of dropout, which limits its application in cell clustering analysis.

scAMZI's performance is mainly due to its unique architectural innovation. SimAM and ZI layer were introduced into the autoencoder. The introduction of SimAM attention mechanism in the coding process can fully extract the cell feature representation, so that the model can more accurately capture the key feature information of the cell and provide more accurate data for subsequent cluster analysis. The introduction of ZI layer in the decoding process can better handle the impact of dropout on clustering, and enhance the stability and robustness of the model. Compared with existing methods, scAMZI showed obvious advantages in the experiment and achieved better clustering results.

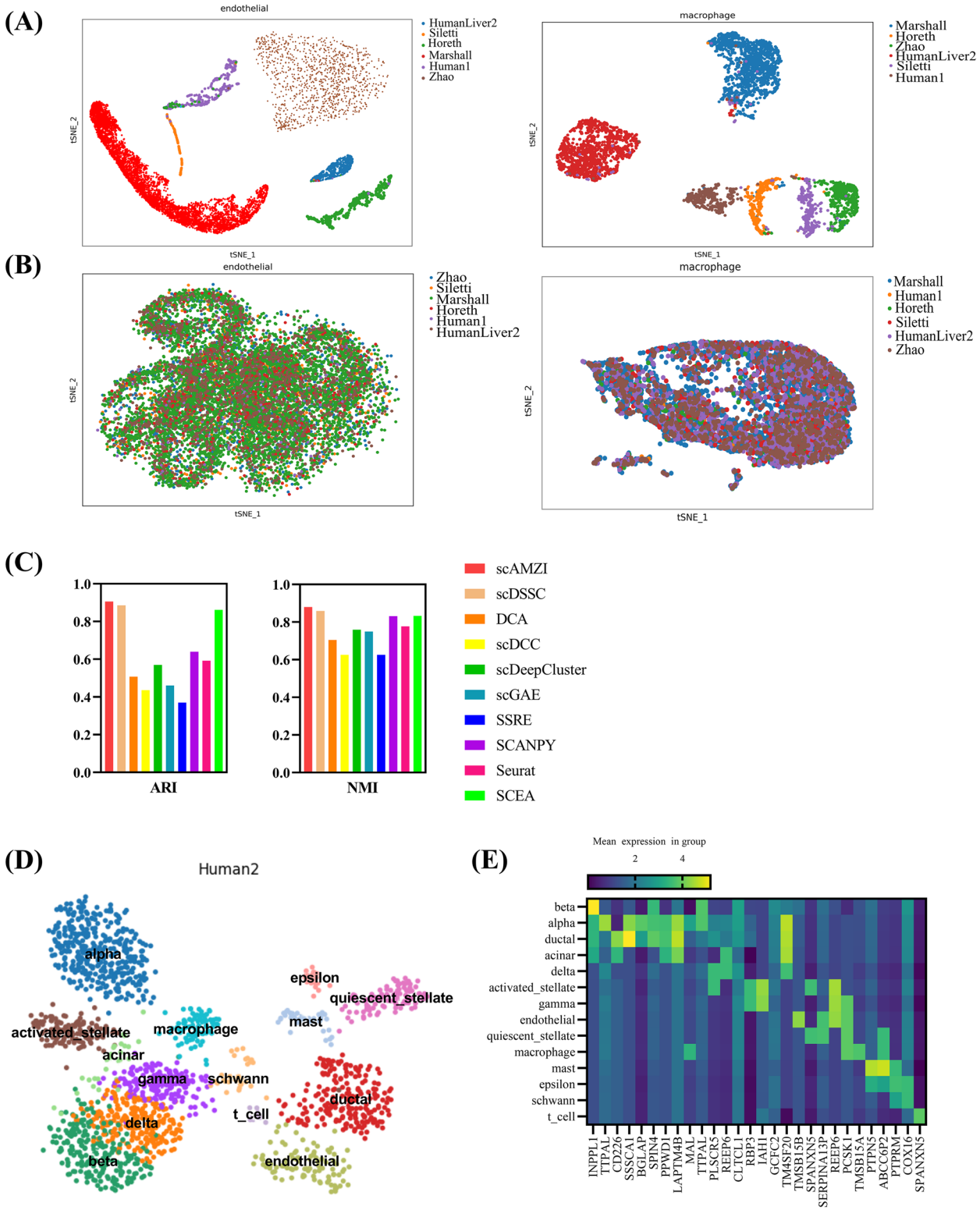

Fig. 4 **A** The clustering results of endothelial cells and macrophage cells from six datasets in original space. **B** The clustering results of endothelial cells and macrophage cells from six datasets in latent space. **C** Performance comparison of scAMZI and competing methods in terms of ARI and NMI on Human2 dataset. **D** Clustering result of scAMZI on Human2 dataset. **E** Marker genes for each cell type found by scAMZI on Human2 dataset

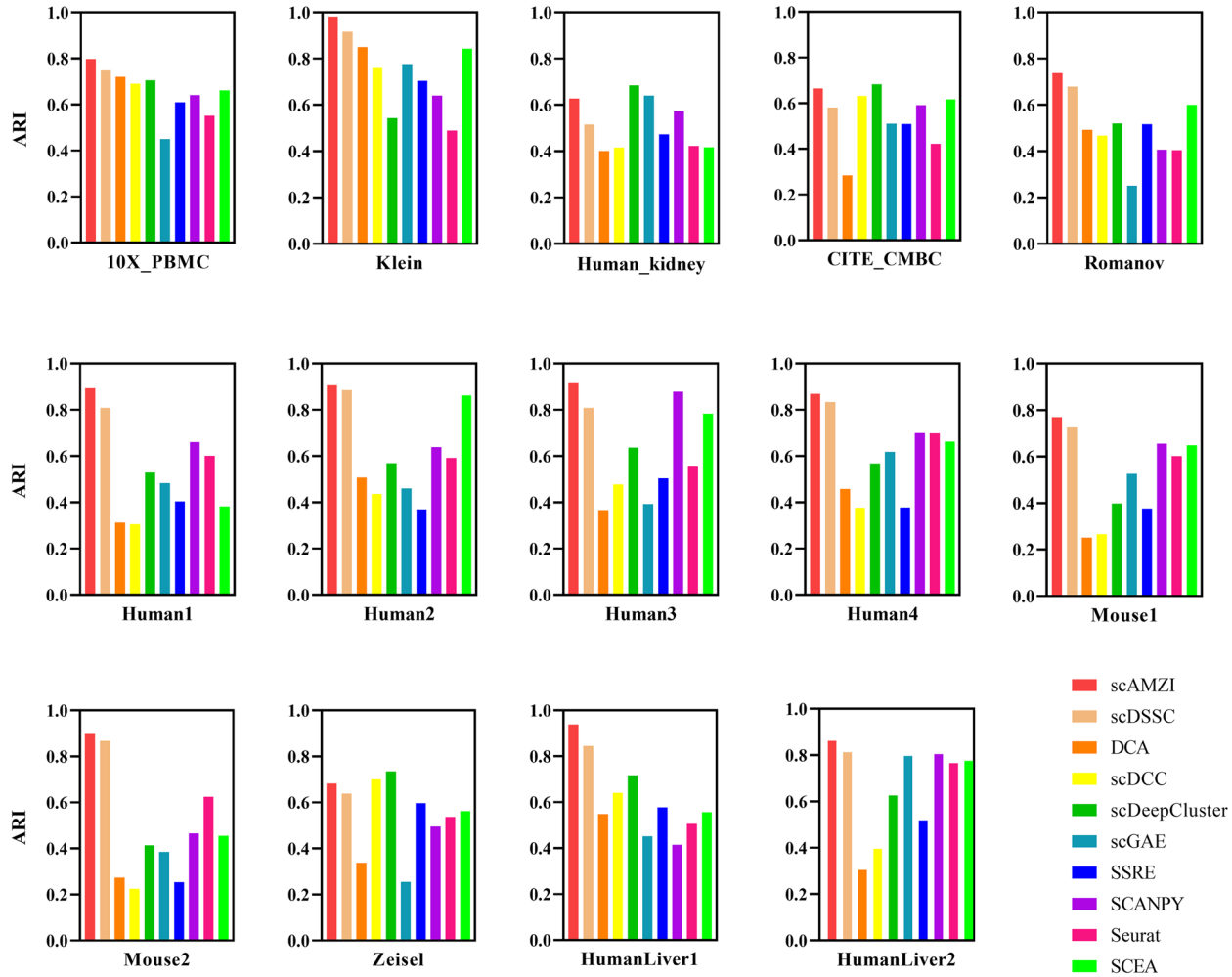

Fig. 5 Performance comparison of scAMZI and competing methods in terms of ARI on fourteen benchmark scRNA-seq datasets

Cell annotation and visualization

After clustering scRNA-seq data, determining the cell type is crucial to understanding the functions and interactions of cells in organisms. Therefore, accurate cell annotation is crucial for scRNA-seq data research.

In this section, after obtaining the clustering results of scAMZI, we used SCANPY to perform cell annotation on the Romanov and Human1 datasets with ground-truth cell type labels. Firstly, we used preprocessing methods to remove noise and redundant information. Secondly, we calculated the nearest neighbors of each cell and used UMAP [31] for dimensionality reduction. Finally, we used SCANPY and dimensionality reduction data for cell annotation and cell visualization. As shown in Fig. 7A, scAMZI accurately clusters the seven cell types in Romanov dataset, except that the same type of cells in neurons are slightly dispersed. Figure 7B shows that scAMZI accurately clusters the fourteen cell types in the dataset, and there are clear distinctions among these fourteen cell types.

We compared the visualization and annotation results with the nine competing methods on Romanov and Human1 datasets. The cell annotation and visualization results of SCEA are obtained by running the code pro-vided by the method, and the results of the remaining eight methods are from the literature [20]. For Human1, scAMZI, scDSSC, SCANPY and scDCC achieved good clustering results. Comparatively, Seurat, SSRE, DCA, scDeepCluster and scGAE showed overlap between different cell types. SCEA did not accurately cluster the fourteen cell types in the dataset. For Romanov, SCANPY and Seurat gave the worst results. The number of cell types far exceeds seven. SSRE, DCA, scDeepCluster, scGAE and SCEA showed overlap between different cell types. Comparatively,

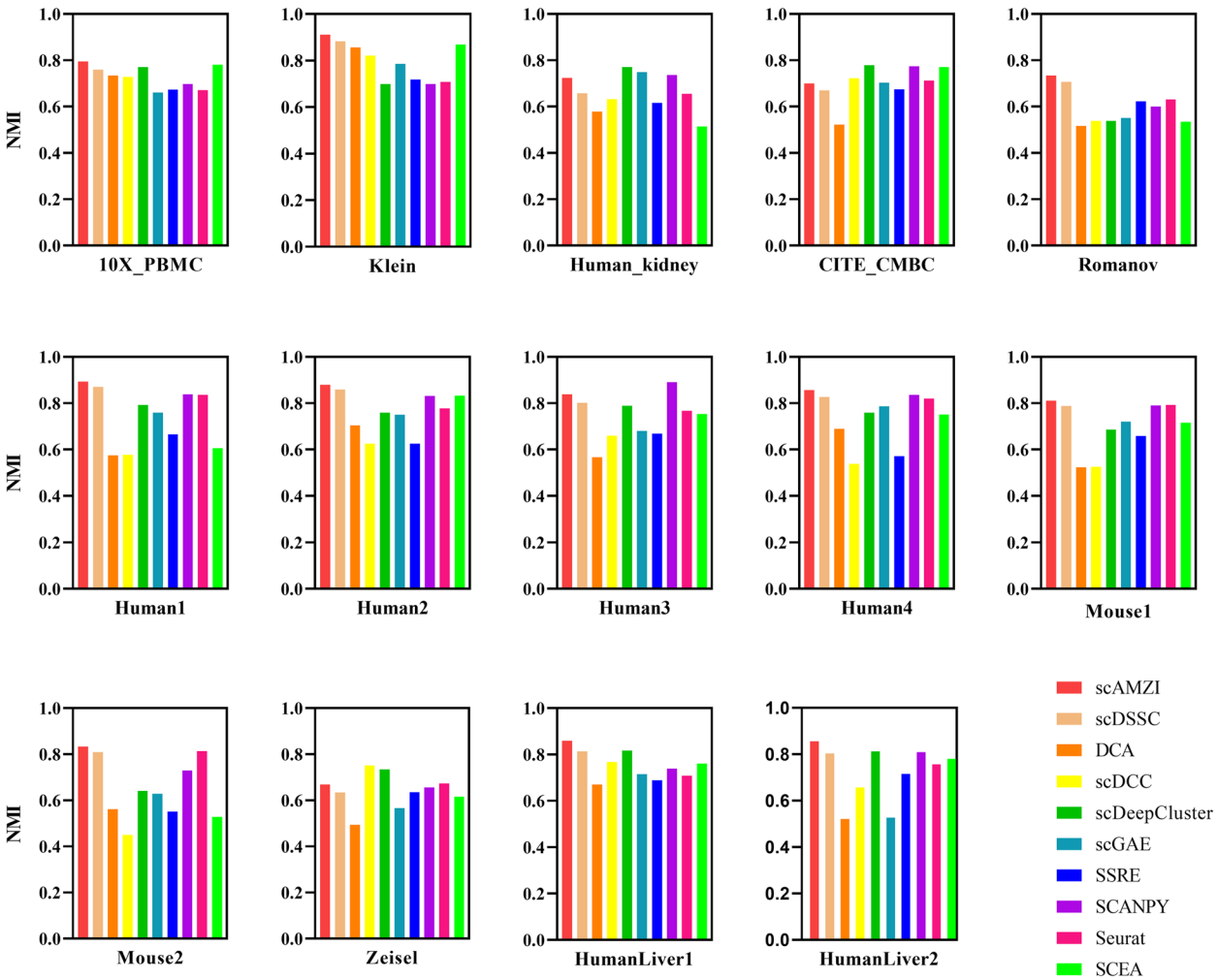

Fig. 6 Performance comparison of scAMZI and competing methods in terms of NMI on fourteen benchmark scRNA-seq datasets

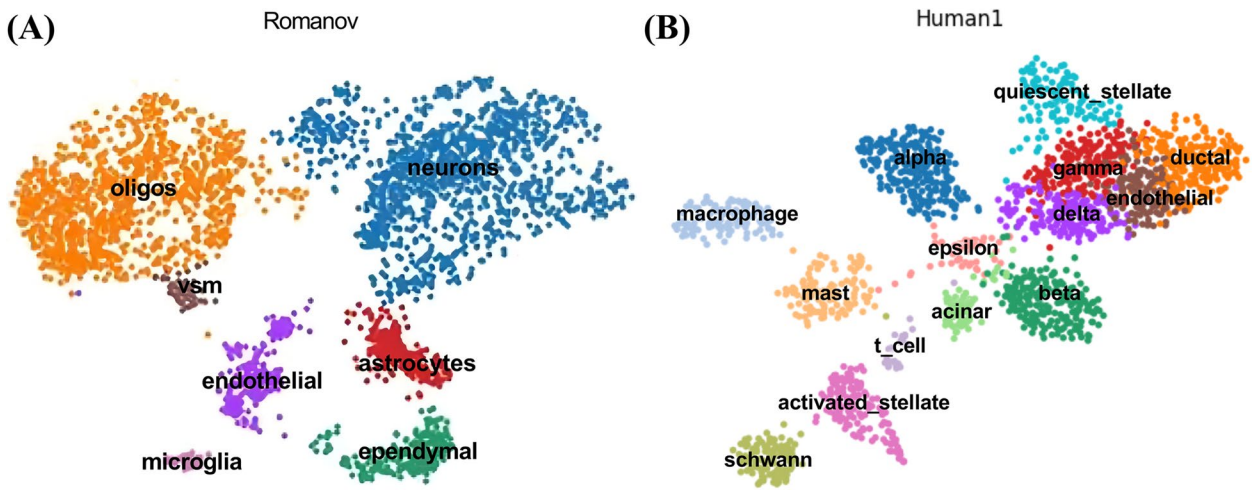

Fig. 7 **A** Clustering results of scAMZI on Romanov dataset. **B** Clustering results of scAMZI on Human1 dataset

scAMZI, scDSSC and scDCC achieved good clustering results. More details are shown in S1 Figure.

Finding marker genes and cell trajectory inference

Marker genes are genes that can mark specific cell types. By marking specific cell types or cell states, researchers can identify and separate different cell types and study their functions, relationships, and biological processes. Cell trajectory inference can help researchers track and infer the dynamic changes of cells. By analyzing cell trajectory, researchers can reveal key processes such as cell migration, differentiation, and proliferation, and gain a deeper understanding of complex biological systems [32].

In this section, we demonstrated that scAMZI can provide support for finding marker genes and inferring cell trajectory. In the Romanov dataset, we used the same steps as in 'Cell annotation and visualization' to obtain the dimensionality reduction data. Then we used the *t*-test to calculate the ranking of genes with large variations in each cluster and selected top 2 gene as the marker genes for visualization. Figure 8A shows the marker genes and their expression levels for each cluster. ATP1A4 and NIPBL showed significantly higher average expression levels in the astrocytes cluster, while their expression levels were relatively low in other cell clusters. This difference fully demonstrates the excellent performance of scAMZI in accurately identifying cell type-specific marker genes. Figure 8A also shows that the maker genes found by scAMZI vary greatly between different cell types.

Furthermore, Fig. 8B strongly demonstrates the important value of scAMZI clustering results in revealing the dynamic process of cell differentiation. By analyzing the spatial distribution and mutual relationship of different cell clusters, the path of cell differentiation can be preliminarily inferred. In Fig. 8B, we observed that some adjacent cell clusters showed a gradual transition trend in gene expression profiles, suggesting that they may be in the continuous stages of cell differentiation. We speculate that these cell clusters gradually differentiate into different states of mature cells. This cell trajectory inference provides intuitive and valuable clues for studying the mechanism of cell differentiation. Although traditional methods can also identify some major cell types, they have certain limitations in revealing the dynamic process of cell differentiation. With its unique model architecture, scAMZI can more accurately capture subtle differences in gene expression between cells and successfully identify more transitional cell clusters, thereby more accurately depicting the trajectory of cell differentiation and providing richer information for a deeper understanding of the complex process of cell differentiation.

Conclusion

In this paper, we proposed a novel DL-based scAMZI with SimAM and zero-inflated layer for clustering scRNA-seq data. scAMZI is mainly composed of SimAM,

autoencoder, ZINB and ZI layer. First, scAMZI uses the attention module SimAM to simultaneously learn meaningful cellular features and latent relationships between cells from the preprocessed scRNA-seq data. Next, the features are fed into the autoencoder. Then, in the last layer of the decoder, scAMZI uses ZI layer to process the feature information to obtain the three parameters for estimating the ZINB distribution. Finally, scAMZI clusters scRNA-seq data using spectral clustering combined with low-dimensional embedding feature representations from trained autoencoders. We.

introduce autoencoder and SimAM to reduce dimensionality of data and learn feature representations of cells and relationships between cells. Meanwhile, ZI layer is used to handle zero values in the data. We compare the performance of scAMZI with nine methods on fourteen benchmark scRNA-seq datasets of various sizes (from hundreds to tens of thousands of cells) with known cell types. Experimental results not only demonstrate that scAMZI outperforms competing methods, but also show that scAMZI can facilitate downstream analyses such as cell annotation, marker gene discovery, and cell trajectory inference.

As shown in S7 Table, we compared the computational cost of scAMZI and competing methods on the HumanLiver2 dataset. The computational cost of scAMZI is shorter than that of all competing methods. Excellent performance on multiple datasets and multiple experiments demonstrated the scalability of scAMZI.

Materials and methods

Data preprocessing

The number of cells in the fourteen benchmark scRNA-seq datasets ranges from hundreds to tens of thousands. The fourteen datasets are 10X_PBMC [33], Klein [34], Human_kidney [35], CITE_CMBC [33], Romanov [36], Human1 [25], Human2 [25] Human3 [25], Human4 [25], Mouse1 [25], Mouse2 [25], Zeisel [37], HumanLiver [38] and HumanLiver2 [26]. The cell number, gene number and cell type number of these datasets are shown in Table 1. These data are available at https://doi.org/10.5281/zenodo.13131559.

The scRNA-seq data is a matrix where columns indicate genes and rows indicate cells. We pre-process the data as follows. First, we filter out genes with zero values and select genes that are expressed in all cells. Next, we calculate the factor for each cell and use these factors to normalize the read counts. Before calculating the factors,

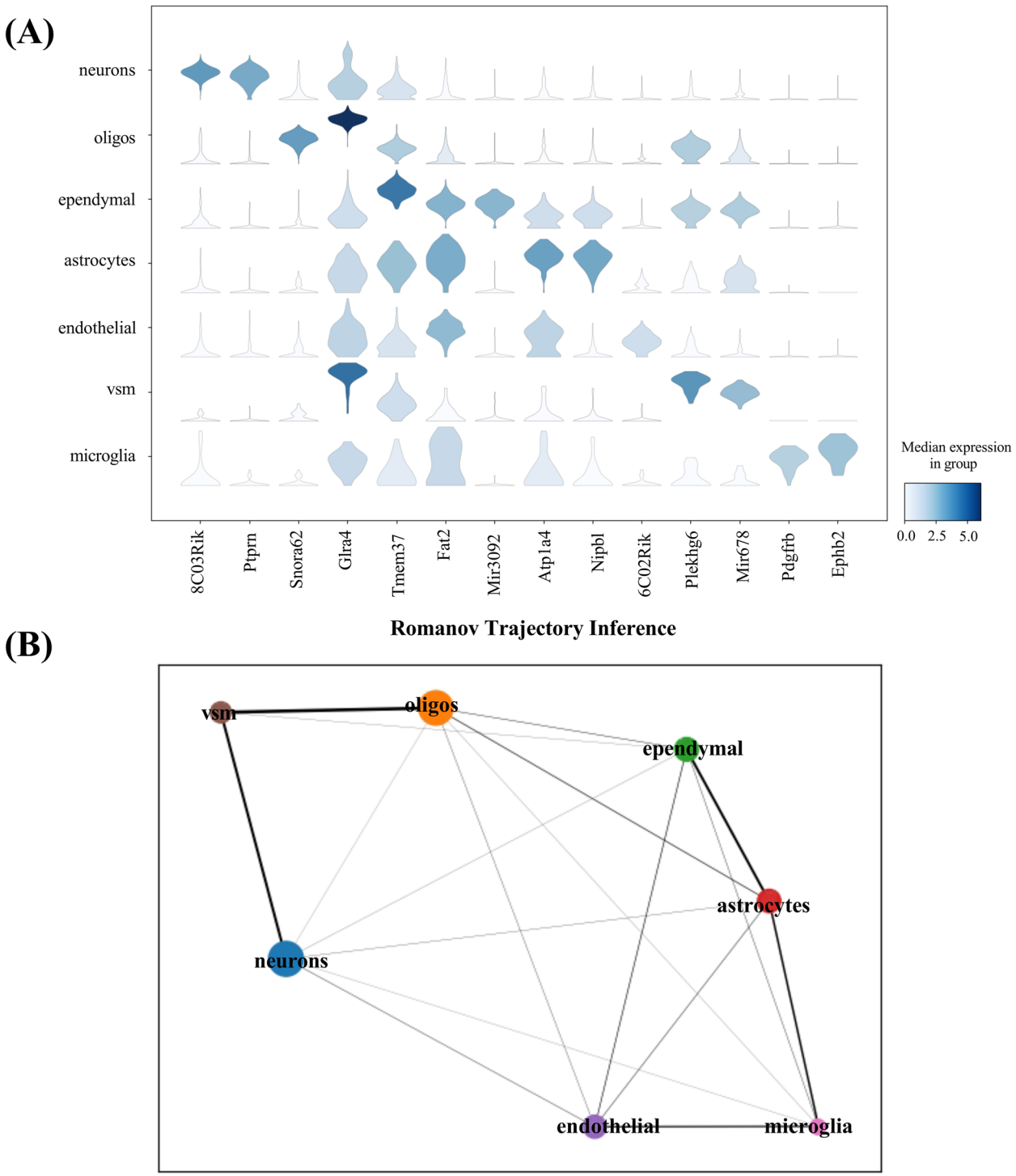

Fig. 8 **A** Marker genes and their expression levels for each cluster found by scAMZI on Romanov dataset. **B** Cell trajectory inference result of Romanov dataset provided by scAMZI

we calculate average expression value for each gene. The size factor for each cell is the median of the ratio of expression value of each gene in that cell divided by the mean of the gene. Then we can obtain normalized count data by dividing read count of the cell by size factor of the cell.

Table 1 Summary of fourteen benchmark scRNA-seq datasets

Dataset	Cell number	Gene number	Cell type number	Source
10X_PBMC	4271	16653	8	[33]
Klein	2717	24175	4	[34]
Human_kidney	5685	25215	11	[35]
CITE_CMBC	8617	2000	15	[33]
Romanov	2881	24341	7	[36]
Human1	1937	20125	14	[25]
Human2	1724	20125	14	[25]
Human3	3605	20125	14	[25]
Human4	1303	20125	14	[25]
Mouse1	822	14878	13	[25]
Mouse2	1064	14878	13	[25]
Zeisel	3005	19972	9	[37]
HumanLiver	8444	5000	11	[38]
HumanLiver2	12494	20939	10	[26]

$$\tilde{x}=\frac{x-\mu}{\sigma} \tag{1}$$

where x represents expression profile of the cell, μ is the mean value of all genes, σ is the standard deviation, and $\tilde{x}$ represents the normalized count data.

Next, we log-transform the read counts, scaling the counts to ensure they have unit variance and zero mean. This will help better understand and compare the count data. During the transformation, top 2000 highly variable genes are selected as the basis for initial noise reduction [17, 39, 40]. We use SCANPY [9] to pre-process the raw read count data.

$$x_{\log}=\log_2(\tilde{x}+1) \tag{2}$$

The scAMZI framework

We design a novel DL model based on attention autoencoder and ZI layer, namely scAMZI, to cluster scRNA-seq data. scAMZI aims to optimize the latent space and learn important cellular features for accurate clustering of scRNA-seq data. As shown in Fig. 1, scAMZI uses pre-processed scRNA-seq data as input data. scAMZI is mainly composed of SimAM, autoencoder, ZINB model and ZI layer. First, scAMZI uses the attention module SimAM to simultaneously learn meaningful cellular features and latent relationships between cells from the pre-processed scRNA-seq data. Next, the features are fed into autoencoder. Then, in the last layer of the decoder, scAMZI uses ZI layer to process the feature information to obtain the three parameters for estimating ZINB distribution. Finally, scAMZI clusters scRNA-seq data using spectral clustering combined with low-dimensional embedding feature representations from trained autoencoders. Based on ZINB model, we introduce autoencoder and SimAM to reduce dimensionality of data and learn feature representations of cells and relationships between cells. Meanwhile, ZI layer is used to handle zero values in the data.

Attention mechanism network

In this section, SimAM is introduced in the autoencoder to increases the weight of important input information and reduces the attention to irrelevant input information. SimAM obtain the importance of each neuron in the neural network structure through a fast closed-form solution of the energy function [22]. First, we obtain the energy tensor by calculating the difference between the dimension of the input data and mean value $\hat{\mu}$ on the dimension, and squaring the difference $\hat{\sigma}$. Then, the sigmoid function is used to control the energy tensor in the range of 0–1 as the attention weight e_t for each input data dimension. The autoencoder combined with SimAM can assign different weights e_t to the input data according to its importance, improving the ability to obtain important features and reducing the impact of redundant information.

$$\begin{cases} e_t = \frac{4(\hat{\sigma}^2+\lambda)}{(X_t-\hat{\mu}^2)+2\hat{\sigma}^2+2\lambda} \\ \hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} X_i \\ \hat{\sigma}^2 = \frac{1}{M}\sum_{\mathrm{i}=1}^{M}\left(X_{\mathrm{i}}-\hat{\mu}\right)^2 \end{cases} \tag{3}$$

where $\hat{\mu}$ represents mean, $\hat{\sigma}^2$ represents variance, λ is the regularization factor, and t represents the t-th cell.

The autoencoder based on the ZINB model

Since the high technical noise causes overfitting of deep learning models, we construct an autoencoder based on the ZINB model using fully connected (FC) layers. The mapping function can be defined as follows:

$$\begin{cases} X' = X + \text{noise} \\ H = F(X') \end{cases} \tag{4}$$

where X indicates the input expression matrix, and noise indicates the Gaussian noise added to each layer of the encoder. By adding noise to the inputs of each layer, the model's robustness and generalization ability can be enhanced.X' represents the expression matrix after adding noise, F represents the mapping function of the encoder, and H represents the output feature vector.

The decoder reconstructs the original input data from the low-dimensional feature representation H. The decoding mapping function is defined as follows:

$$\hat{X} = F'(H) \tag{5}$$

where H is the input matrix of the decoder,F' is the decoding mapping function, and $\hat{X}$ is the reconstructed data after the decoding mapping function.

Zero-inflated layer

The extremely high dropout rate impedes downstream analyses such as cell type identification and cell trajectory inference. To solve this problem, we introduce a zero-inflated layer into the decoding network to eliminate the impact of dropout event. We use Gaussian random sampling to simulate dropout events and use simulate anneal arithmetic (SAA) [41] strategy to adaptively adjust the parameters of the zero-inflated layer. The sampling probability of the model can be gradually reduced during training, making the decoder more robust. The model can be more stable when encountering different samples, reducing the risk of overfitting.

$$R = ZILayer(\hat{X}) \tag{6}$$

where $\hat{X}$ is the reconstructed data after the decoding mapping function (see 'The autoencoder based on the ZINB model') and R is output.

Loss function

To accurately obtain the cell and gene features in scRNA-seq data, we use a loss function based on ZINB autoencoder model to characterize original counting data. ZINB models loss events based on a combination of zero components and NB distributions.

$$\begin{cases} ZINB(\overline{X}|\pi,\mu,\theta) = \pi\delta_0(\overline{X}) + (1-\pi) \times NB(\overline{X}|\mu,\theta) \\ NB(\overline{X}|\mu,\theta) = \frac{\Gamma(\overline{X}+\theta)}{\Gamma(\overline{X}+1)\Gamma(\theta)} \times \left(\frac{\theta}{\theta+\mu}\right)^{\theta} + \left(\frac{H}{\theta+\mu}\right)^{\overline{X}} \end{cases} \tag{7}$$

where $\overline{X}$ is the original matrix, π is the probability of the dropout event, μ and θ represent the mean and dispersion of the NB distribution, respectively. The decoder network is designed with three output layers for computing three sets of parameters. The calculation formulas for these parameters are defined as follows:

$$\begin{cases} X'' = f^N(R) \\ \Pi = \text{sigmoid}\left(W_\pi X''\right) \\ M = \text{diag}(s_i)\exp\left(W_\mu X''\right) \\ \Theta = \exp(W_\theta X'') \end{cases} \tag{8}$$

where X'' is the output data processed by the ZI layer, Π ,M and Θ represent the estimation matrices of π,μ and θ, respectively. Since π represents a probability value between 0 and 1, we use the sigmoid activation function. Since both π and θ are non-negative, we use exponential activation function. The size factor s_i is precomputed. The reconstruction loss function of the decoder is defined as follows:

$$L_{ZINB}(\pi,\mu,\theta|X) = -\log(ZINB(\overline{X}|\pi,\mu,\theta)) \tag{9}$$

Implementation and parameters setting

In this study, we select top 2000 highly variable genes as the input of the autoencoder network, which consists of a three-layer fully connected structure with 256–32-10 neurons. 2000 is a widely used parameter, and using parameter 2000 can capture sufficient biological information and ensure the efficiency of analysis [17, 39, 40]. The parameter value of 256–32-10 can both extract key features and maintain good clustering effect [20]. The settings of the decoder are reversed from those of the encoder. During the network training process, we adopt pre-training and fine-tuning strategy, using Adam optimizer with learning rates of 0.002 and 0.001 to update the autoencoder, respectively. These two learning rates are a common and effective combination, which not only avoids gradient problems in training, but also achieves rapid convergence and stable optimization.

Evaluation strategies

In the experiments, we adopt two widely used methods NMI (Normalized Mutual Information) and ARI (Adjusted Rand Index) to evaluate model performance. Let $\mathrm{U} = \left\{\mathrm{U}_1,\mathrm{U}_2,...,\mathrm{U}_{\mathrm{c_u}}\right\}$ and $\mathrm{V} = \left\{\mathrm{V}_1,\mathrm{V}_2,...,\mathrm{V}_{\mathrm{c_v}}\right\}$ be the predicted and ground-truth clusters. NMI and ARI are defined as follows.

$$\begin{cases} \mathrm{NMI} = \frac{I(U,V)}{\max\{H(U),H(V)\}} \\ I(U,V) = \sum_{\mathrm{p}=1}^{\mathrm{c_u}}\sum_{\mathrm{q}=1}^{\mathrm{c_v}}\left|U_\mathrm{p} \cap V\mathrm{q}\right| \log\frac{n\left|U_\mathrm{p}\cap V\mathrm{q}\right|}{\left|U_\mathrm{p}\right|\times\left|V\mathrm{q}\right|} \\ H(U) = -\sum_{p=1}^{C_U}\left|U_p\right|\log\frac{\left|U_p\right|}{n} \\ H(V) = -\sum_{q=1}^{C_V}\left|V_q\right|\log\frac{\left|V_q\right|}{n} \end{cases} \tag{10}$$

where $I(U,V)$ represents mutual information, $H(U)$ and $H(V)$ are entropy values.

$$\mathrm{ARI} = \frac{\binom{\mathrm{n}}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \tag{11}$$

where a represents the number of pairs of two objects in the same group in U and V, b represents the number of pairs of two objects in different groups in U and V, c represents the number of pairs of two objects in the same group in U but in different groups in V, and d represents the number of pairs of two objects in different groups in U but in the same group in V.

In this paper, we compare scAMZI with SCANPY [9], Seurat [10], SSRE [7], DCA [13], scDeepCluster [42], scDCC [17], scGAE [19], scDSSC [20] and SCEA [43] to test the performance of scAMZI. SCANPY (2018) and Seurat (2015) are the most commonly used tools for analyzing scRNA-seq data. Their clustering methods are based on the Louvain algorithm. SSRE (2021) uses sparse subspace representation and similarity enhancement strategy to cluster scRNA-seq data. DCA (2019) proposes a deep count autoencoder to cluster scRNA-seq data. scDeepCluster (2019) maps scRNA-seq data into a low-dimensional space via a ZINB-based autoencoder and performs clustering based on KL divergence. scDCC (2021) adds prior knowledge as additional terms into the loss function and uses an autoencoder to cluster scRNA-seq data. scGAE (2021) clusters scRNA-seq data by using a multi-task-oriented graph autoencoder combines with topological information and feature information. scDSSC (2022) combines the Self-Expressiveness Property of data with autoencoders to perform deep sparse subspace clustering. SCEA (2023) uses a graph attention autoencoder and an MLP-based encoder to perform clustering [44]. We use the codes and recommended optimal model parameters provided by these methods. We train and evaluate scAMZI and competing methods using the same training and test datasets. All experiments are conducted on NVIDIA RTX 3090 GPU.

Concrete mathematical proof of scAMZI

For cell t, scAMZI calculates the attention weight e_t through the energy function (Eq. (3)). Features are enhanced by attention weight $X_t = X_t \cdot e_t$. When X_t is far away from the mean $\hat{\mu}$(i.e., high-discrimination features), $(X_t - \hat{\mu})^2 \gg 2\sigma^2 + 2\lambda \Rightarrow e_t \approx \frac{4(\hat{\sigma}^2+\lambda)}{(X_t-\hat{\mu}^2)}$, e_t decreases exponentially as $|X_t - \hat{\mu}|$ increases, but due to the large value of X_t, the value of $X_t = X_t \cdot e_t$ is less affected. When X_t is close to the mean $\hat{\mu}$(i.e., low-discrimination features), $(X_t - \hat{\mu})^2 \ll 2\sigma^2 + 2\lambda \Rightarrow e_t \approx \frac{4(\hat{\sigma}^2+\lambda)}{2\hat{\sigma}^2+2\lambda} = 2$. scAMZI enhances high-discrimination features and suppresses low-discrimination features by minimizing e_t. scAMZI dynamically adjusts feature weights so that the weights of high-discrimination features are significantly higher than those of low-discrimination features.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11511-2.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.
Supplementary Material 8.
Supplementary Material 9.

Acknowledgements
Not applicable.

Authors' contributions
LY conceived the method. LY, ZJX and LY designed the method. LY, BYM and LY conducted the experiments and wrote the main manuscript text. All authors reviewed the manuscript.

Funding
This work was supported by grants from the National Natural Science Foundation of China, Nos. 62472239 and 62002189, and supported by the Natural Science Foundation of Shandong Province, China (Nos. ZR2024MF011 and ZR2021MH104), and supported by the Young Taishan Scholars Program, Shandong, China (No. tsqn201909178), and supported by the Cultivation Fund of the Second Hospital of Shandong University (No. 2023JX16), and supported by the Ability Improvement Project of Science and Technology SMES in Shandong Province (2023TSGC0279), and supported by the Youth Innovation Team of Colleges and Universities in Shandong Province (2023KJ329), and supported by the Shandong Medical Association Qilu medical special project (YKH2022K02112), and supported by the Shandong Province Key Research and Development Program-International Scientific and Technological Cooperation Project (2024KJHZ029).

Data availability
The package of scAMZI is made freely available at https://doi.org/10.5281/zenodo.13131559.

Declarations

Ethics approval and consent to participant
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 3 September 2024 Accepted: 20 March 2025
Published online: 07 April 2025

References

1. Kolodziejczyk Aleksandra A, Kim JK, Svensson V, Marioni John C, Teichmann Sarah A. The technology and biology of single-Cell RNA sequencing. Mol Cell. 2015;58(4):610–20.
2. Yu L, Cao Y, Yang JY, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. 2022;23(1):49.
3. Lee D, Park Y, Kim S. Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. Brief Bioinform. 2021;22(3):bbaa188.
4. Wang R, Dang M, Harada K, Han G, Wang F, Pool Pizzi M, Zhao M, Tatlonghari G, Zhang S, Hao D. Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. Nat Med. 2021;27(1):141–51.
5. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 2017;8(1):1–13.

6. Lin P, Troup M, Ho JW. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18:1–11.
7. Liang Z, Li M, Zheng R, Tian Y, Yan X, Chen J, Wu F-X, Wang J. SSRE: cell type detection based on sparse subspace representation and similarity enhancement. Genomics Proteomics Bioinformatics. 2021;19(2):282–91.
8. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech: Theory Exp. 2008;2008(10):P10008.
9. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
10. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
11. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd. 1996;1996:226–31.
12. 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.
13. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390.
14. Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat Mach Intell. 2019;1(4):191–8.
15. Yu B, Chen C, Qi R, Zheng R, Skillman-Lawrence PJ, Wang X, Ma A, Gu H. scGMAI: a Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder. Briefings in bioinformatics. 2021;22(4):bbaa316.
16. Oja E, Yuan Z. The FastICA algorithm revisited: Convergence analysis. IEEE Trans Neural Networks. 2006;17(6):1370–81.
17. Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. Nat Commun. 2021;12(1):1873.
18. Jiang J, Xu J, Liu Y, Song B, Guo X, Zeng X, Zou Q. Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. Brief Bioinform. 2023;24(3):bbad152.
19. Luo Z, Xu C, Zhang Z, Jin W. A topology-preserving dimensionality reduction method for single-cell RNA-seq data using graph autoencoder. Sci Rep. 2021;11(1):20028.
20. Wang H, Zhao J, Zheng C, Su Y. scDSSC: deep sparse subspace clustering for scRNA-seq data. PLoS Comput Biol. 2022;18(12):e1010772.
21. Abadi SAR, Laghaee SP, Koohi S. An optimized graph-based structure for single-cell RNA-seq cell-type classification based on non-linear dimension reduction. BMC Genomics. 2023;24(1):1–13.
22. Yang L, Zhang R-Y, Li L, Xie X. Simam: A simple, parameter-free attention module for convolutional neural networks. In: International conference on machine learning. PMLR; 2021. pp.11863–11874.
23. Yau KK, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biom J. 2003;45(4):437–52.
24. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017;18(1):174.
25. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell systems. 2016;3(4):346-360. e344.
26 Abravanel DL, Klughammer J, Blosser T, Goltsev Y, Jiang S, Bai Y, Murray E, Alon S, Cui Y, Goodwin DR. Abstract PD6–03: Spatio-molecular dissection of the breast cancer metastatic microenvironment. Cancer Res. 2022;82(4_Supplement):PD6-03-PD06-03.
27. Marshall JL, Noel T, Wang QS, Chen H, Murray E, Subramanian A, Vernon KA, Bazua-Valenti S, Liguori K, Keller K. High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. Iscience. 2022;25(4):104097.
28. Zhao Q, Pedroza A, Sharma D, Gu W, Dalal A, Weldy C, Jackson W, Li DY, Ryan Y, Nguyen T. A cell and transcriptome atlas of the human arterial vasculature. bioRxiv 2024;2024-09.
29. Siletti K, Hodge R, MossiAlbiach A, Lee KW, Ding S-L, Hu L, Lönnerberg P, Bakken T, Casper T, Clark M. Transcriptomic diversity of cell types across the adult human brain. Science. 2023;382(6667):eadd7046.
30. Horeth E, Bard J, Che M, Wrynn T, Song E, Marzullo B, Burke M, Popat S, Loree T, Zemer J. High-resolution transcriptomic landscape of the human submandibular gland. J Dent Res. 2023;102(5):525–35.
31. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
32. Pang F, Shi D, Yuan L. Screening and identification of key genes for cervical cancer, ovarian cancer and endometrial cancer by combinational bioinformatic analysis. Curr Bioinform. 2023;18(8):647–57.
33. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):14049.
34. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
35. Young MD, Mitchell TJ, Vieira Braga FA, Tran MG, Stewart BJ, Ferdinand JR, Collord G, Botting RA, Popescu D-M, Loudon KW. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science. 2018;361(6402):594–9.
36. Romanov RA, Zeisel A, Bakker J, Girach F, Hellysaz A, Tomer R, Alpar A, Mulder J, Clotman F, Keimpema E. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nat Neurosci. 2017;20(2):176–88.
37. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138–42.
38. MacParland SA, Liu JC, Ma X-Z, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N, Echeverri J, Linares I. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018;9(1):4383.
39. Chaffin M, Papangeli I, Simonson B, Akkad A-D, Hill MC, Arduini A, Fleming SJ, Melanson M, Hayat S, Kost-Alimova M. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature. 2022;608(7921):174–80.
40. Nguyen AT, Wang K, Hu G, Wang X, Miao Z, Azevedo JA, Suh E, Van Deerlin VM, Choi D, Roeder K. APOE and TREM2 regulate amyloid-responsive microglia in Alzheimer's disease. Acta Neuropathol. 2020;140:477–93.
41 Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671–80.
42 Patop IL, Wüst S, Kadener S. Past, present, and future of circRNAs. EMBO J. 2019;38(16):e100836.
43. Abadi SAR, Laghaee SP, Koohi S. An optimized graph-based structure for single-cell RNA-seq cell-type classification based on non-linear dimension reduction. BMC Genomics. 2023;24(1):227.
44. Yuan L, Zhao L, Jiang Y, Shen Z, Zhang Q, Zhang M, Zheng C-H, Huang D-S. scMGATGRN: a multiview graph attention network–based method for inferring gene regulatory networks from single-cell transcriptomic data. Brief Bioinform. 2024;25(6):bbae526.

Publisher's Note