

Supplementary Information

GraphPCA: a fast and interpretable dimension reduction algorithm for spatial transcriptomics data

Jiyuan Yang¹, Lu Wang^{1,2}, Lin Liu³ and Xiaoqi Zheng^{1,*}

¹ Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

² The Guangxi Key Laboratory of intelligent precision medicine, Nanning, Guangxi Zhuang Autonomous Region, China.

³ Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, CMA-Shanghai, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University and Shanghai Artificial Intelligence Laboratory, Shanghai, China.

*Corresponding author: xqzheng@shsmu.edu.cn (XZ).

This PDF file includes:

Supplementary Material

Supplementary Figures

Supplementary Tables

Supplementary Material

Algorithm for GraphPCA

Here, we describe the detailed algorithm for GraphPCA. Let X be an $n \times m$ gene expression matrix, where x_{ij} denotes the expression of the j th gene on the i th spot, $i = 1, \dots, n, j = 1, \dots, m$. The spatial coordinates of individual spots are denoted as $S = (s_1, \dots, s_N)^T, s_i \in R^2/R^3$. As mentioned in the Methods section, our aim is to find a k -dimensional projection $Z \in R^{n \times k}$ under the following objective function:

$$\min_{Z, W} \|X - ZW^T\|_2^2 + \frac{1}{2} \lambda \sum_{\{i, j \in A\}} A_{ij} \|Z_i - Z_j\|_2^2, s. t. W^T W = I, \quad (1)$$

where W is an orthonormal eigenvector matrix and λ is a hyperparameter that balances the reconstruction error and the smoothness of the projections over the graph. A is the adjacent matrix of the spatial neighborhood graph, which can be constructed by various distance metrics and graph construction strategies, such as the k -nearest neighbor (k NN) graph, shared nearest neighbor (SNN) graph, circular-neighborhood graph, and Delaunay triangulation. Moreover, the GraphPCA framework supports not only ordinary graphs with binary 0-1 adjacency matrices as input but also weighted graphs with arbitrary weighting schemes, including the Heat kernel and Dot-product. By default, adjacency matrix of k NN graph was employed as input.

The above objective function can be rewritten as

$$\min_{Z, W} \|X - ZW^T\|_2^2 + \lambda \text{tr}(Z^T L Z), s. t. W^T W = I, \quad (2)$$

where $L = D - A$ is the Laplacian matrix of the spatial neighborhood graph, and D is the diagonal matrix of node degrees. To solve the optimization problem, we fixed W as constant to solve the optimal Z^* , which becomes a quadratic and convex problem. Let J denote the objective function as given in equation (2). it is easy to derive the optimal solution by setting the derivative of J with respect to Z to be zero. Since

$$\begin{aligned} \frac{\partial J}{\partial Z} &= \frac{\partial}{\partial Z} \text{tr} [(X - ZW^T)^T (X - ZW^T) + \lambda Z^T L Z] \\ &= \frac{\partial}{\partial Z} \text{tr} [X^T X - 2WZ^T X + WZ^T ZW^T + \lambda Z^T L Z] \\ &= \frac{\partial}{\partial Z} \text{tr} [X^T X - 2WZ^T X + WZ^T ZW^T + \lambda Z^T L Z] \\ &= -2XW + 2Z + 2\lambda LZ. \end{aligned}$$

Thus the derivative equals to zero if and only if $Z^* = (I + \lambda L)^{-1} XW$. To simplify the notation, we denoted

$$K = (I + \lambda L)^{-1},$$

which is a symmetric and positive definite matrix. Then, we substituted Z in objective J with $Z^* = KXW$, and reduced the optimization problem as

$$\min_W ||X - KXWW^T||_2^2 + \lambda \text{tr}(W^T X^T K^T L K X W), s. t. W^T W = I. \quad (3)$$

Taking advantage of the trace properties, the equation (3) is equivalent to

$$\begin{aligned} & \min_{W, W^T W = I} \text{tr} [(X - KXWW^T)^T (X - KXWW^T) + \lambda (KXW)^T L K X W] \\ &= \min_{W, W^T W = I} \text{tr} (X^T X - 2X^T KXWW^T + WW^T X^T K^T K X W W^T + \lambda W^T X^T K^T L K X W) \\ &= \min_{W, W^T W = I} \text{tr} (X^T X - 2X^T KXWW^T + WW^T X^T K^T K X W W^T + \lambda L K X W W^T X^T K^T) \\ &= \min_{W, W^T W = I} \text{tr} [-2X^T KXWW^T + X^T K^T K X W W^T W W^T + (K^{-1} - I) K X W W^T X^T K^T] \\ &= \min_{W, W^T W = I} \text{tr} [-2X^T KXWW^T + X^T K^T K X W W^T + (K^{-1} - I) K X W W^T X^T K^T] \\ &= \min_{W, W^T W = I} \text{tr} (-2X^T KXWW^T + X^T K^T K X W W^T + X W W^T X^T K^T - K X W W^T X^T K^T) \\ &= \min_{W, W^T W = I} \text{tr} (-2X^T KXWW^T + X^T K^T K X W W^T + X^T K^T X W W^T - X^T K^T K X W W^T) \\ &= \min_{W, W^T W = I} \text{tr} (-X^T KXWW^T) \\ &= \min_{W, W^T W = I} \text{tr} (-W^T X^T K X W). \end{aligned} \quad (4)$$

It is easy to show that the matrix $X^T K X$ is symmetric and positive semi-definite. The optimal solution W^* of equation (4) is the combination of eigenvectors, associated with the largest k eigenvalues of the graph-revised covariance matrix $X^T K X$.

Integrating multiple graph regularizations in GraphPCA

Due to the flexibility of the GraphPCA framework, it is easy to integrate other types of information as graphical constraints, such as gene expression profiles or Hematoxylin and Eosin (H&E) images. For example, besides the spatial location information, we incorporated gene expression data as an additional graph constraint into the objective function of GraphPCA. Specifically,

$$\begin{aligned} & \min_{Z, W} ||X - ZW^T||_2^2 + \frac{1}{2} \lambda_1 \sum_{\{i, j \in A^1\}} A_{ij}^1 ||Z_i - Z_j||_2^2 + \frac{1}{2} \lambda_2 \sum_{\{i, j \in A^2\}} A_{ij}^2 ||Z_i - Z_j||_2^2, \\ & s. t. W^T W = I, \end{aligned} \quad (5)$$

where λ_1 and λ_2 control the strengths of the neighborhood graphs constructed by spatial location information and gene expression data, respectively. A^1 and A^2 represent the adjacent matrices of spatial neighborhood graph and gene

expression data-derived graph, respectively. Accordingly, we can also obtain the compact form of the objective function as (5)

$$\min_{Z,W} ||X - ZW^T||_2^2 + \lambda_1 \text{tr}(Z^T L^1 Z) + \lambda_2 \text{tr}(Z^T L^2 Z), s. t. W^T W = I, \quad (6)$$

where L^1 and L^2 are Laplacian matrices of the spatial neighborhood graph and gene expression graph, respectively. Analogy to the above optimization procedures, we can obtain the optimal $Z^* = (I + \lambda_1 L^1 + \lambda_2 L^2)^{-1} XW$. By substituting Z in objective function (6) with $Z^* = K' XW$, where $K' = (I + \lambda_1 L^1 + \lambda_2 L^2)^{-1}$, we can obtain

$$\min_W ||X - K' XW W^T||_2^2 + \lambda_1 \text{tr}(W^T X^T K'^T L^1 K' XW) + \lambda_2 \text{tr}(W^T X^T K'^T L^2 K' XW), \\ s. t. W^T W = I. \quad (7)$$

Making use of the trace properties of matrices, the equation (7) is equivalent to

$$\begin{aligned} & \min_{W, W^T W = I} \text{tr} [(X - K' XW W^T)^T (X - K' XW W^T) + \lambda_1 (K' XW)^T L^1 K' XW + \lambda_2 (K' XW)^T L^2 K' XW] \\ &= \min_{W, W^T W = I} \text{tr} (X^T X - 2X^T K' XW W^T + W W^T X^T K'^T K' XW W^T + \lambda_1 L^1 K' XW W^T X^T K'^T + \\ & \quad \lambda_2 L^2 K' XW W^T X^T K'^T) \\ &= \min_{W, W^T W = I} \text{tr} [-2W^T X^T K' XW + W^T X^T K'^T K' XW + \lambda_1 W^T X^T K'^T L^1 K' XW + \lambda_2 W^T X^T K'^T L^2 K' XW] \\ &= \max_{W, W^T W = I} \text{tr} [W^T X^T (2K' - K'^T K' - K'^T (\lambda_1 L^1 + \lambda_2 L^2) K') XW] \\ &= \max_{W, W^T W = I} \text{tr} \{W^T X^T [K' + (I - K'^T (I + \lambda_1 L^1 + \lambda_2 L^2)) K'] XW\} \\ &= \max_{W, W^T W = I} \text{tr} (W^T X^T K' XW). \end{aligned} \quad (8)$$

It can be observed that the optimal solution W^* of equation (8) is the combination of eigenvectors associated with the largest k eigenvalues of the graph-revised covariance matrix $X^T K' X$. Therefore, when incorporating an additional graph regularization, the optimal solutions for Z^* and W^* can be obtained by simply replacing λL in the matrix K with the product of $\lambda = [\lambda_1, \lambda_2]$ and the augmented Laplacian matrix $L = [L^1, L^2]$.

Supplementary Figures

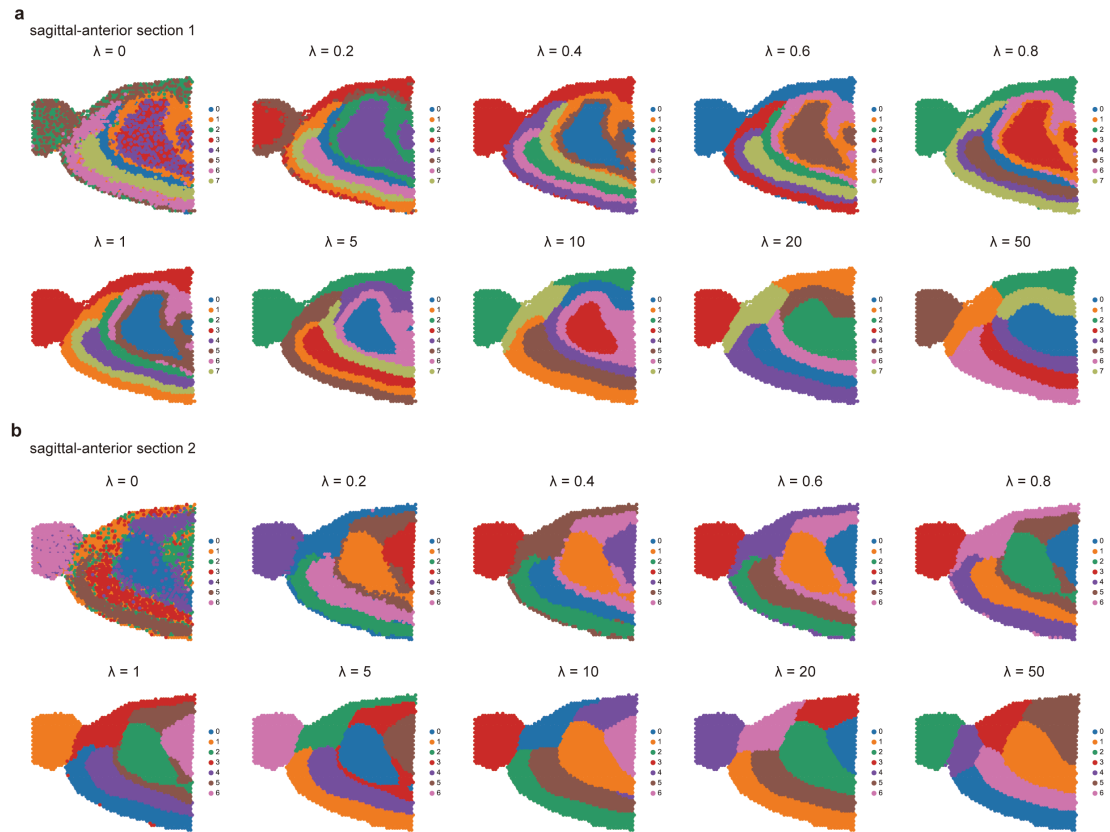


Fig. S1 Impact of the hyperparameter (λ) on the performance of GraphPCA in simulated mouse brain sagittal-anterior data.

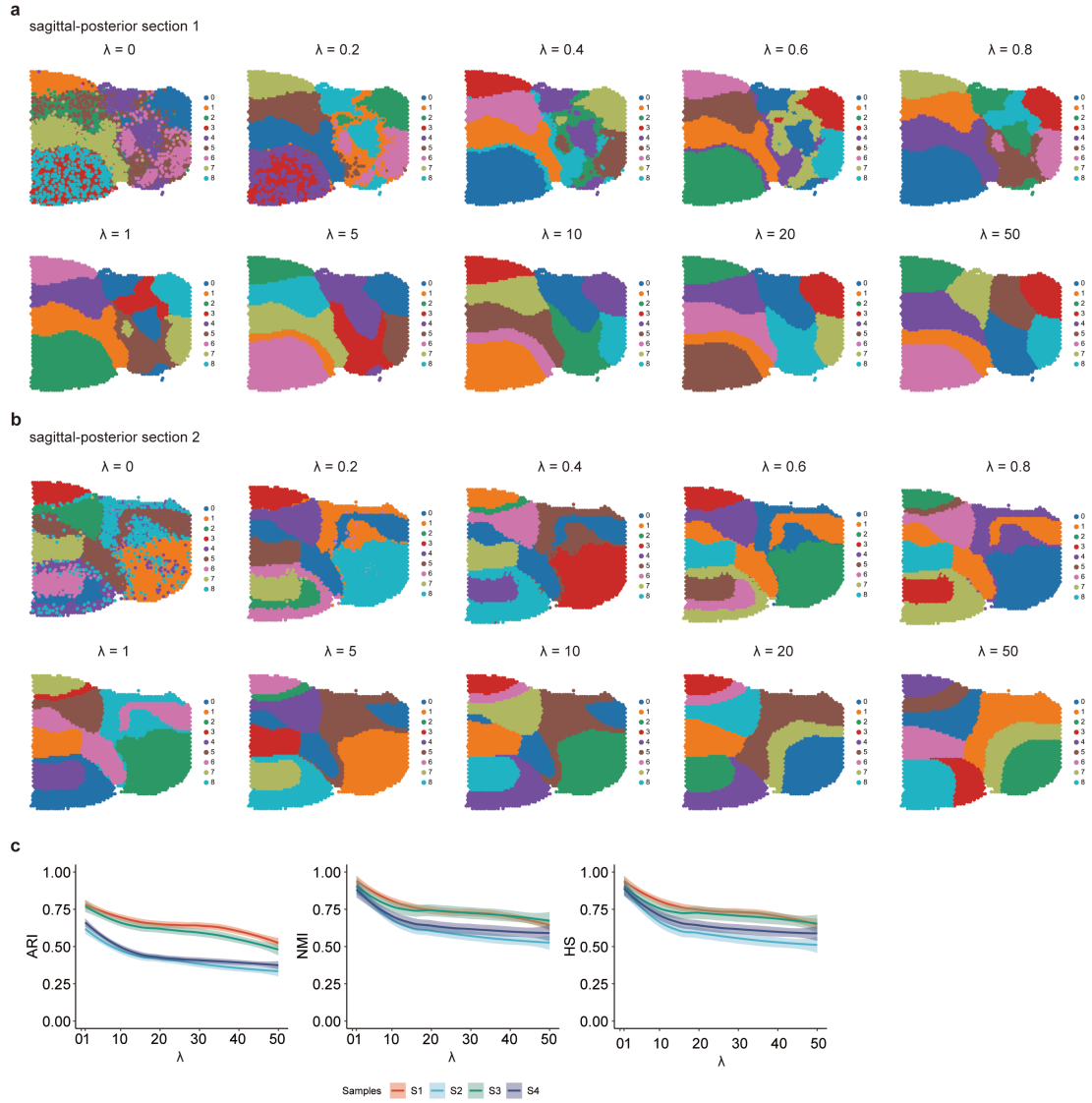


Fig. S2 Impact of the hyperparameter (λ) on the performance of GraphPCA in simulated mouse brain sagittal-posterior data. a,b Spatial domains identified by GraphPCA for sagittal-posterior sections 1 (a) and 2 (b). **c** Line plots depicting clustering accuracies (ARI, NMI, and HS) on simulated data across different values of the graph constraint parameter λ .

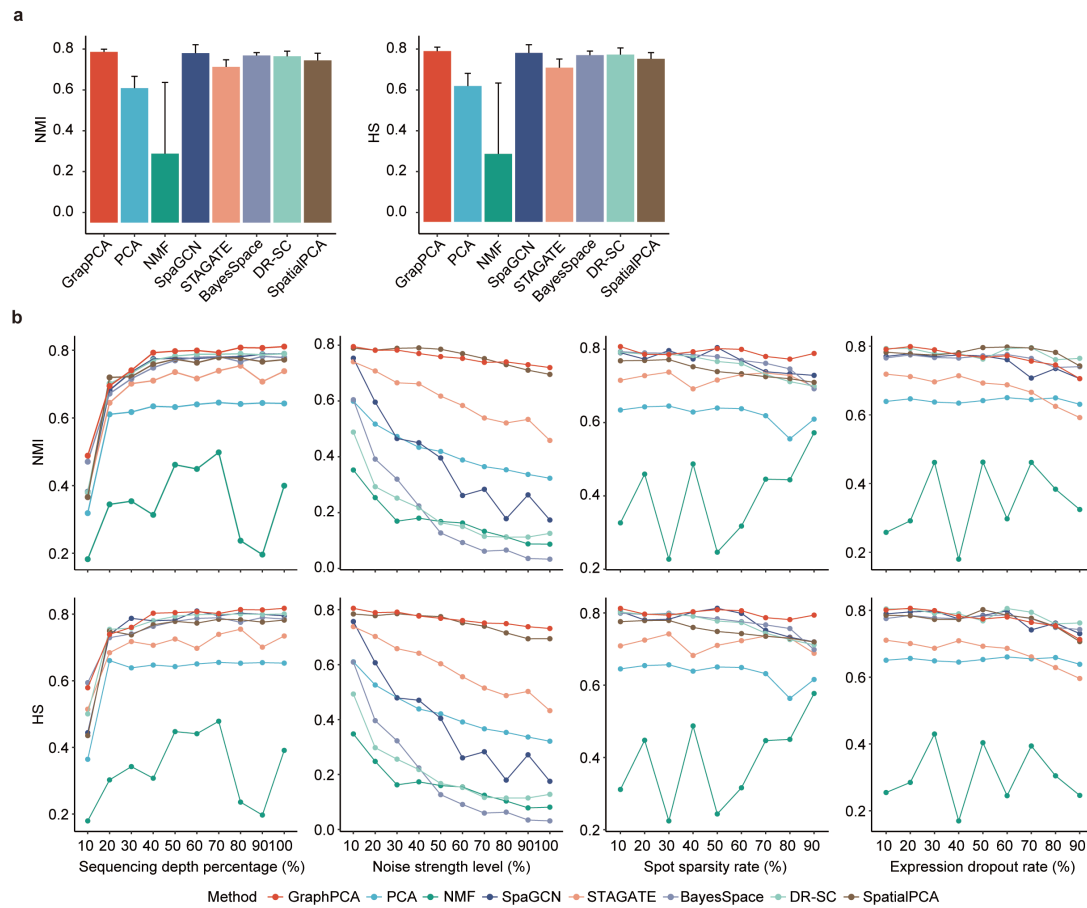


Fig. S3 Impact of the hyperparameter (λ) on the performance of GraphPCA in simulated mouse brain sagittal-posterior data. **a** Clustering accuracies (NMI and HS) of different methods on simulated data. Error bars indicate 95% confidence intervals across 20 replicates. **b** The robustness of GraphPCA and other methods under varying simulation scenarios including different sequencing depths, noise levels, spot sparsity, and expression dropout rates, evaluated by NMI (top panel) and HS (bottom panel). For each scenario and method, dots represent the mean accuracies calculated across 20 replicates.

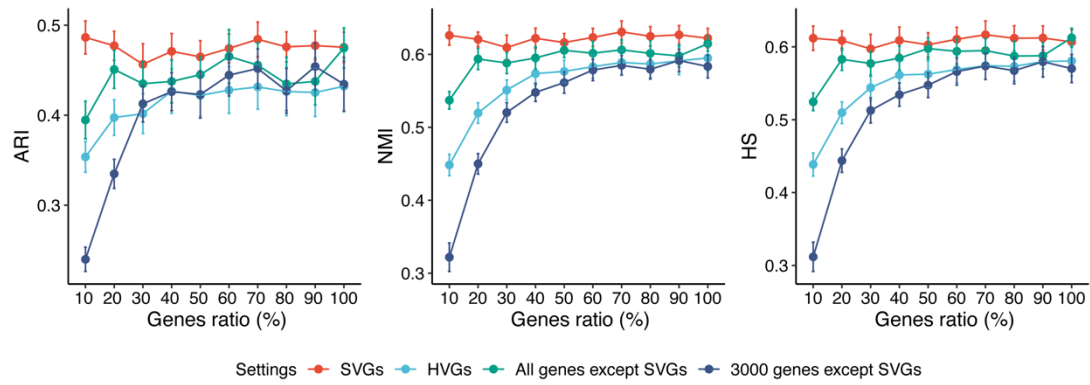
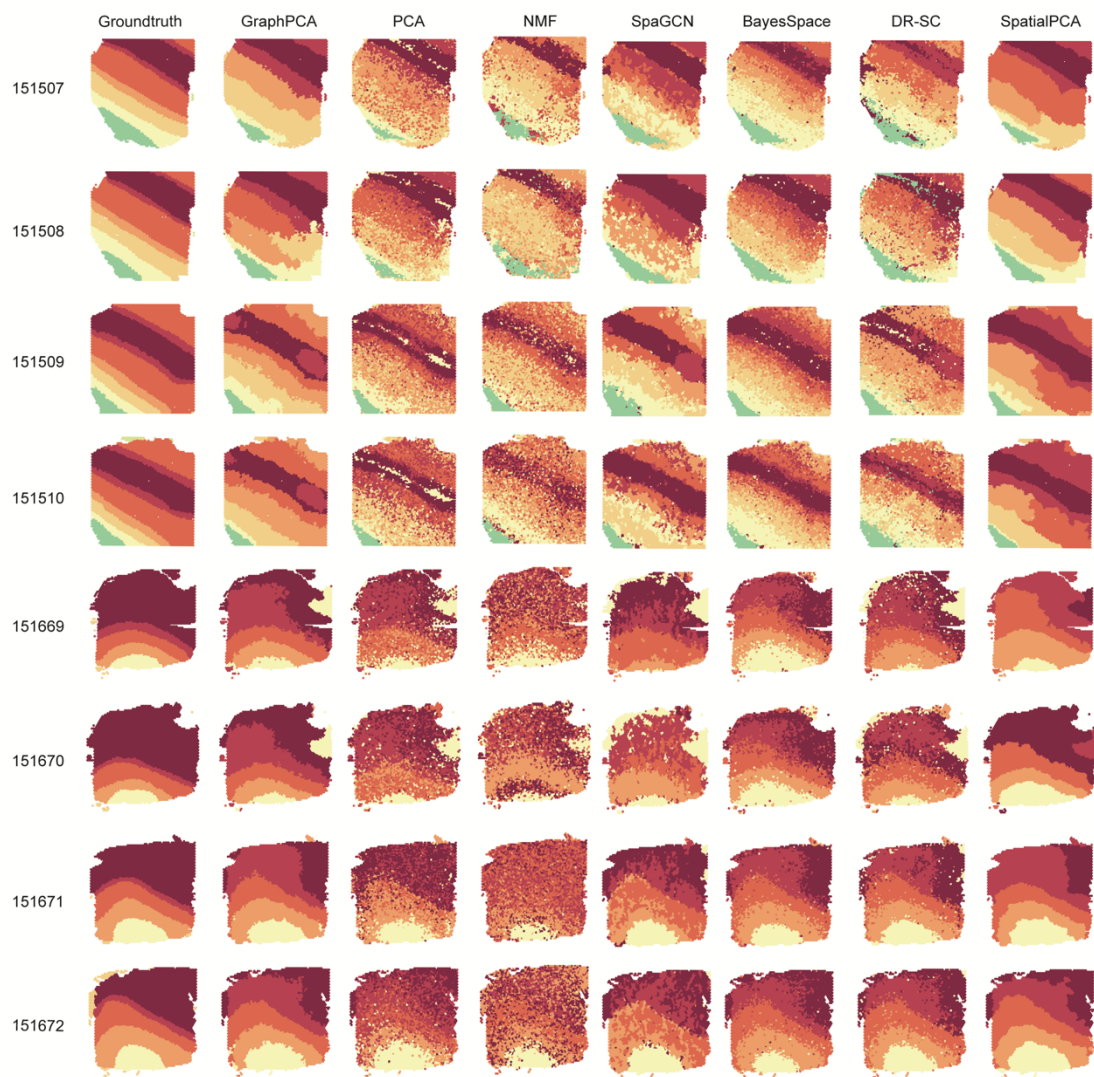


Fig. S4 Sensitivity analyses on the DLPFC dataset. Line plots illustrate the clustering accuracy of GraphPCA at different gene ratios. Dots and error bars in line plot represent means and 95% confidence intervals of clustering metrics, including ARI, NMI, and HS. Different colors indicate different simulation scenarios. Red corresponds to the top 3,000 spatially variable genes detected by SPARK. Sky blue corresponds to the top 3,000 highly variable genes detected by Scanpy. Green denotes all genes except the top 3,000 spatially variable genes detected by SPARK. Blue represents randomly selected 3,000 genes except the top 3000 spatially variable genes detected by SPARK.



See next page

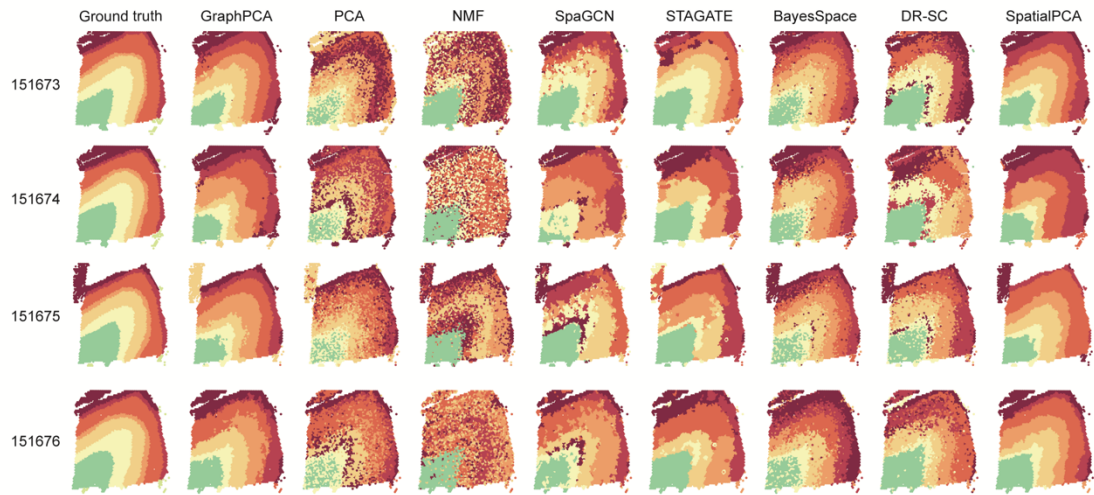
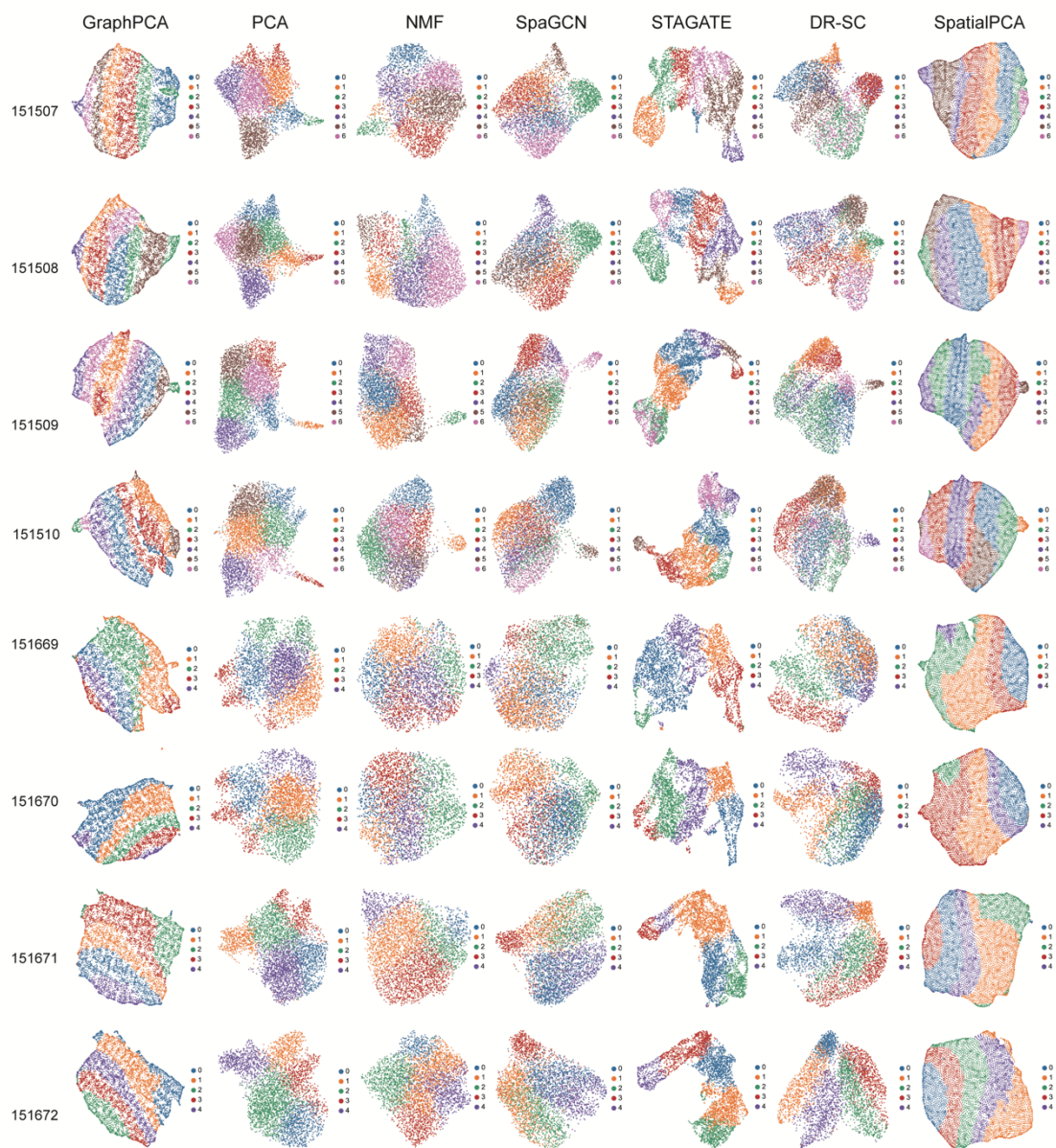


Fig. S5 Spatial domain detection results by different methods on the DLPFC dataset. The first column corresponds to the ground truth by pathological annotation, while the second to eighth columns represent clustering results by GraphPCA, PCA, NMF, SpaGCN, STAGATE, BayesSpace, DR-SC and SpatialPCA, respectively.



See next page

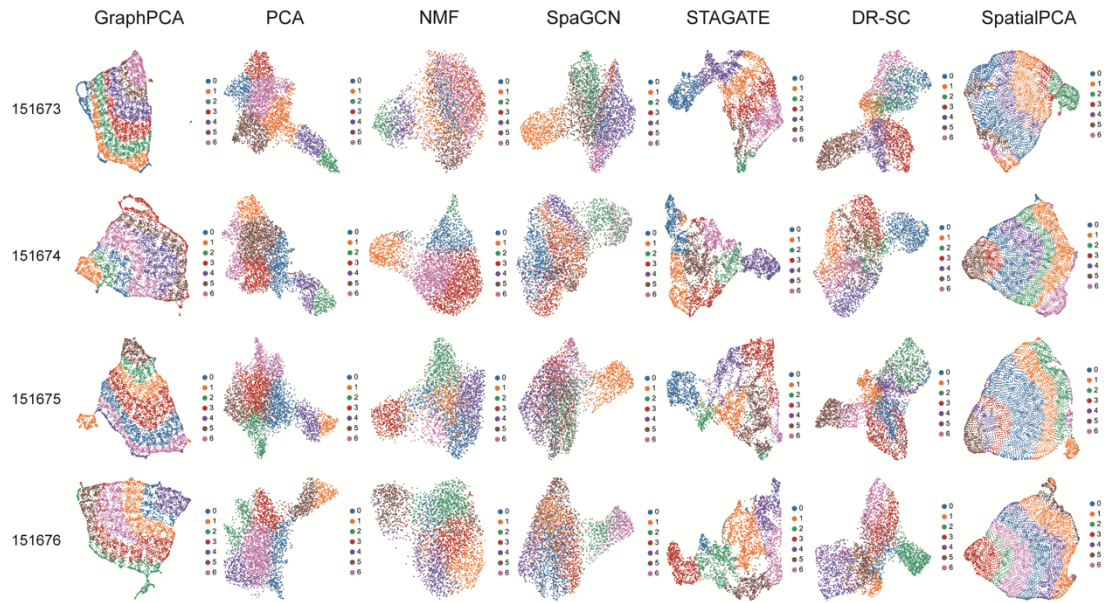


Fig. S6 UMAP visualization of all sections in the DLPFC dataset by low-dimensional embeddings obtained from different methods.

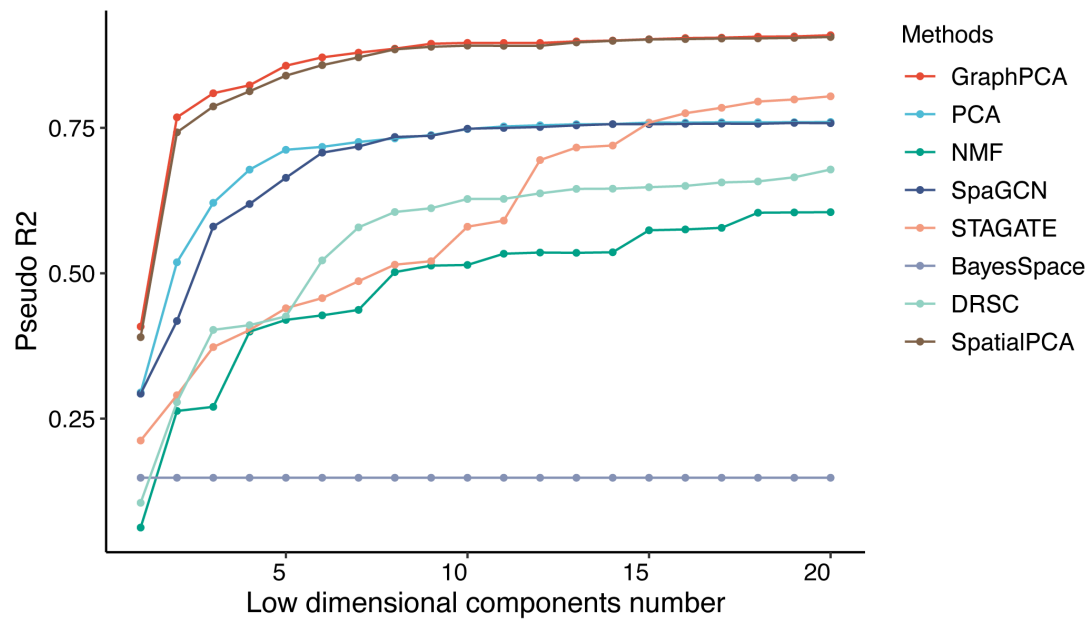


Fig. S7 Predictive performances of different methods at varying low dimensional components numbers on sample 151673 of the DLPFC dataset.

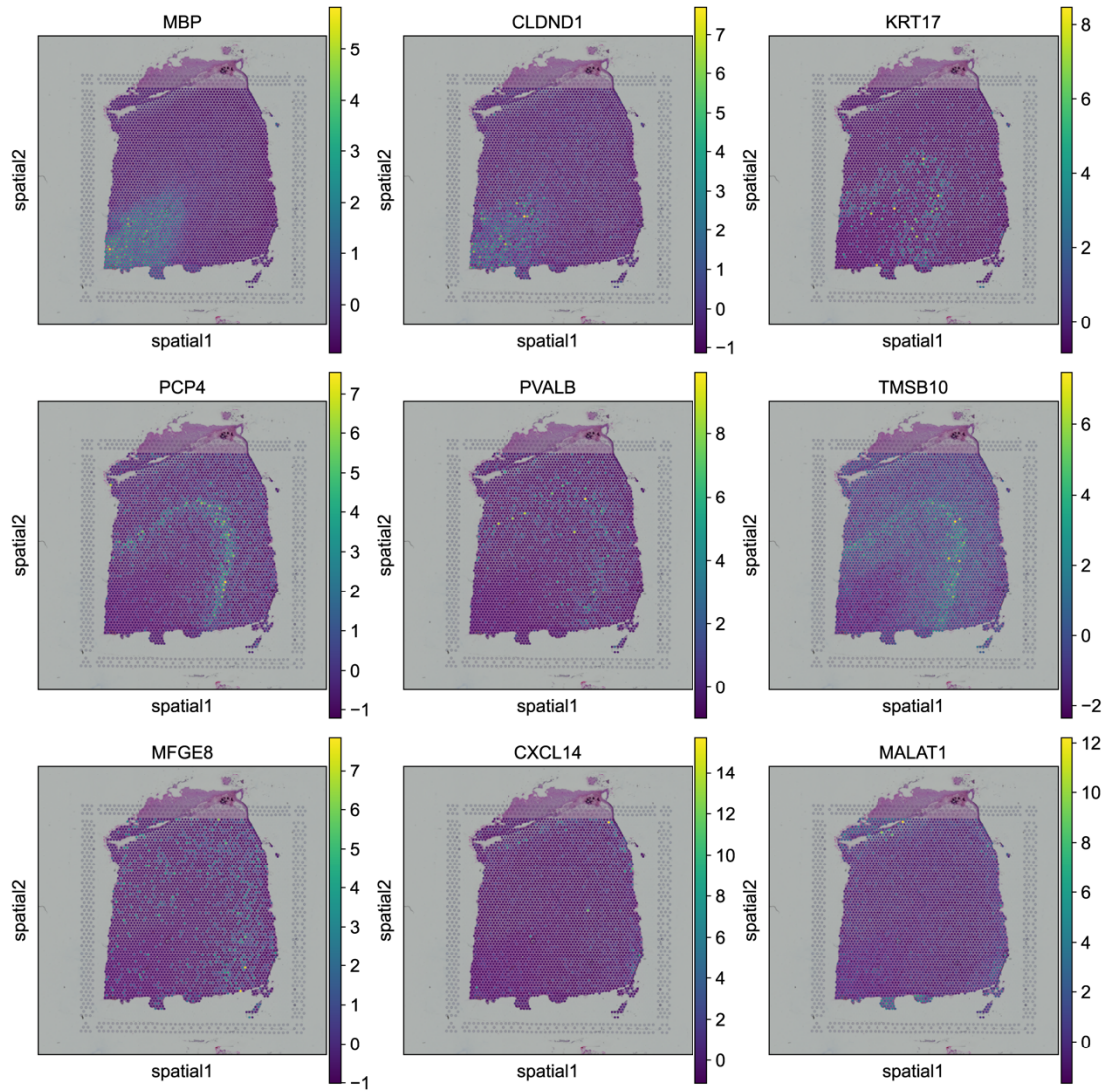


Fig. S8 Spatial expression patterns of layer marker genes for sample 151673 of the DLPFC dataset. Marker genes are detected for each spatial domain inferred from GraphPCA. Layer marker genes include *MBP* and *CLDND1* for white matter, *KRT17* for layer 6, *PCP4* for layer 5, *PVPLB* and *TMSB10* for layer 4, *MFGE8* for layer 3, *CXCL14* for layer 2, and *MALAT1* for layer 1.

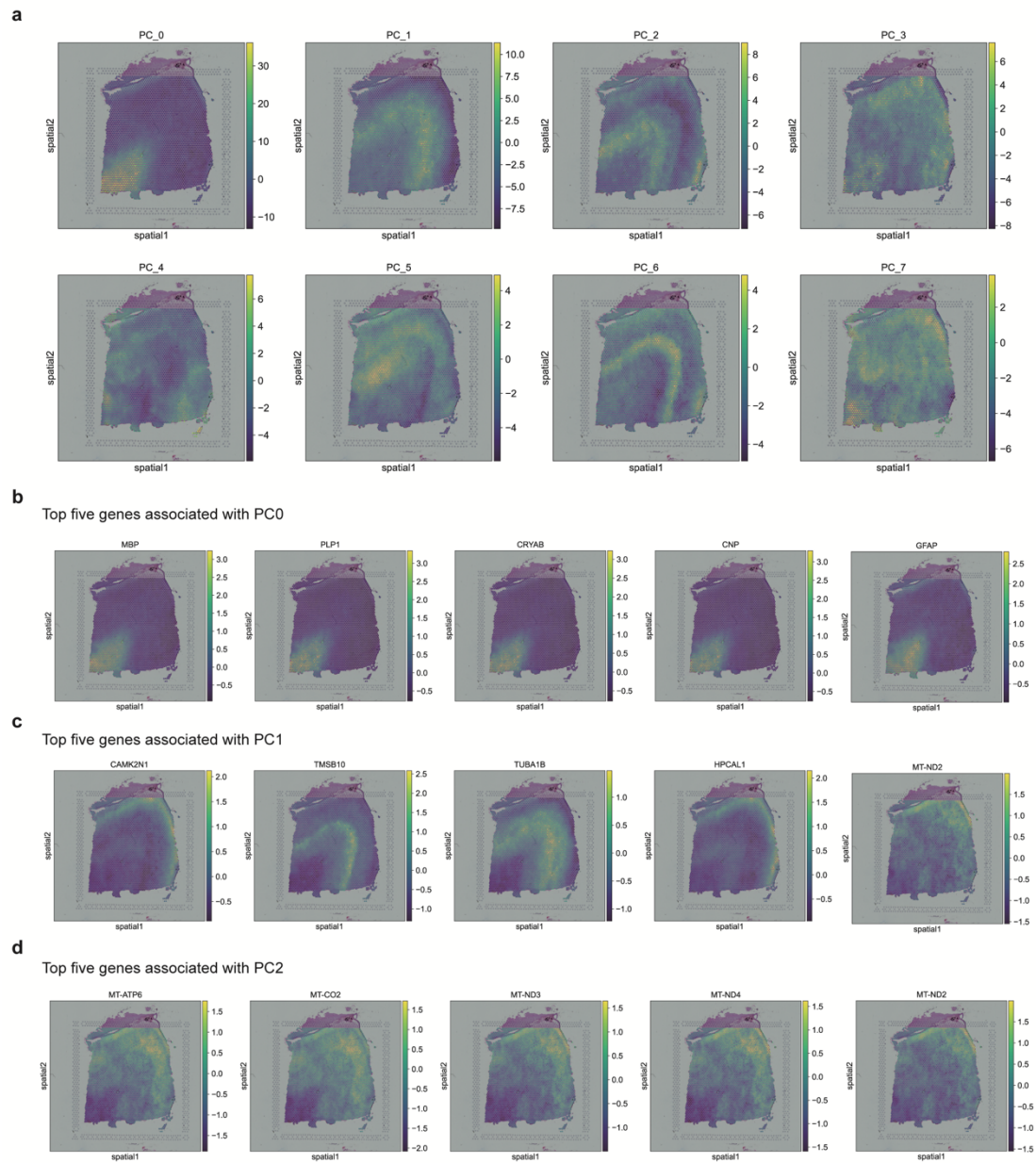


Fig. S10 Top principal components (PCs) generated by GraphPCA are associated with spatial expression patterns of the DLPFC dataset. **a** Visualization of top 8 PCs obtained from GraphPCA. **b,c,d** Gene expression patterns of 5 example genes associated with the top 3 PCs, respectively.

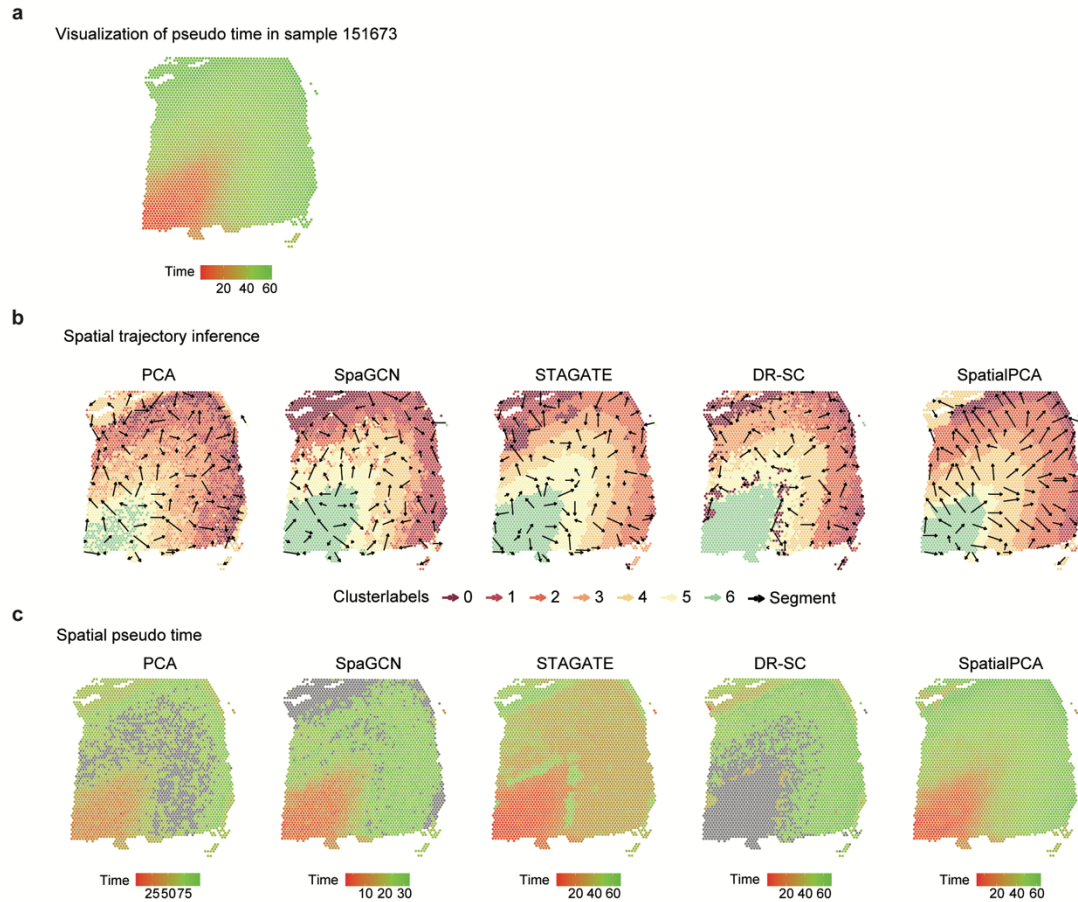


Fig. S11 Spatial trajectory and pseudo-time of the DLPFC dataset by GraphPCA, PCA, SpaGCN, STAGATE, DR-SC, and SpatialPCA. a Pseudo-time inferred from low-dimensional embeddings by GraphPCA. **b,c** Trajectory and pseudo-time inferred from low-dimensional embeddings by PCA, SpaGCN, STAGATE, DR-SC, and SpatialPCA. Arrows point from tissue locations with low pseudo-time to those with high pseudo-time. Color represents different tissue regions.

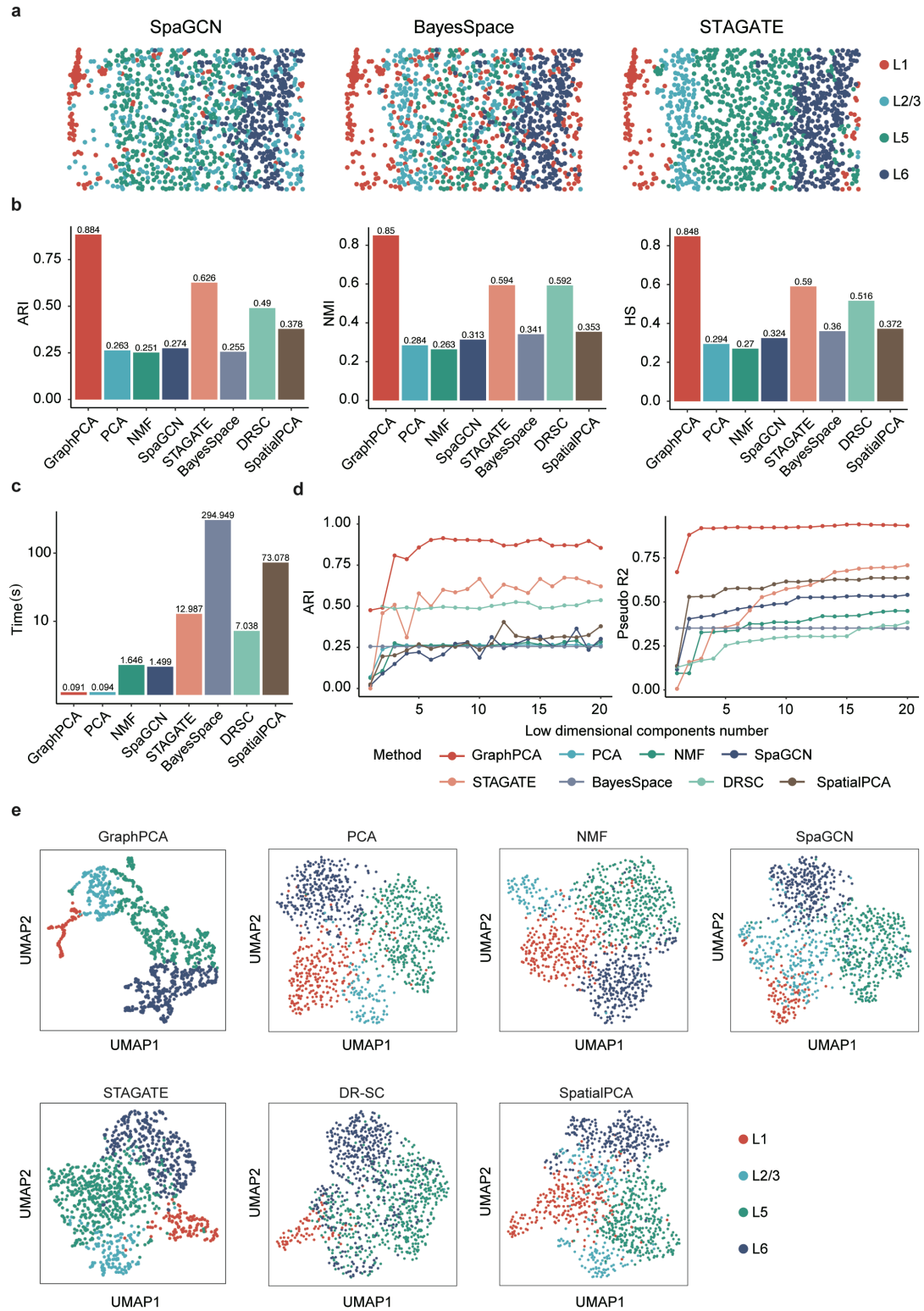


Fig. S12 Spatial domain detection results for the STARmap dataset. a Spatial domain detected by SpaGCN, BayesSpace, and STAGATE on section BZ5 of the STARmap dataset. **b** Bar plots comparing the runtimes by different methods for spatial domain detection. **c** Clustering accuracies (ARI, NMI, and

HS) of different methods for the STARmap dataset at domain-level. **d** Clustering accuracies in terms of ARI (left) and pseudo- R^2 (right) at varying low-dimensional components inferred by different methods. **e** UMAP visualization of low-dimensional embeddings inferred by seven methods.

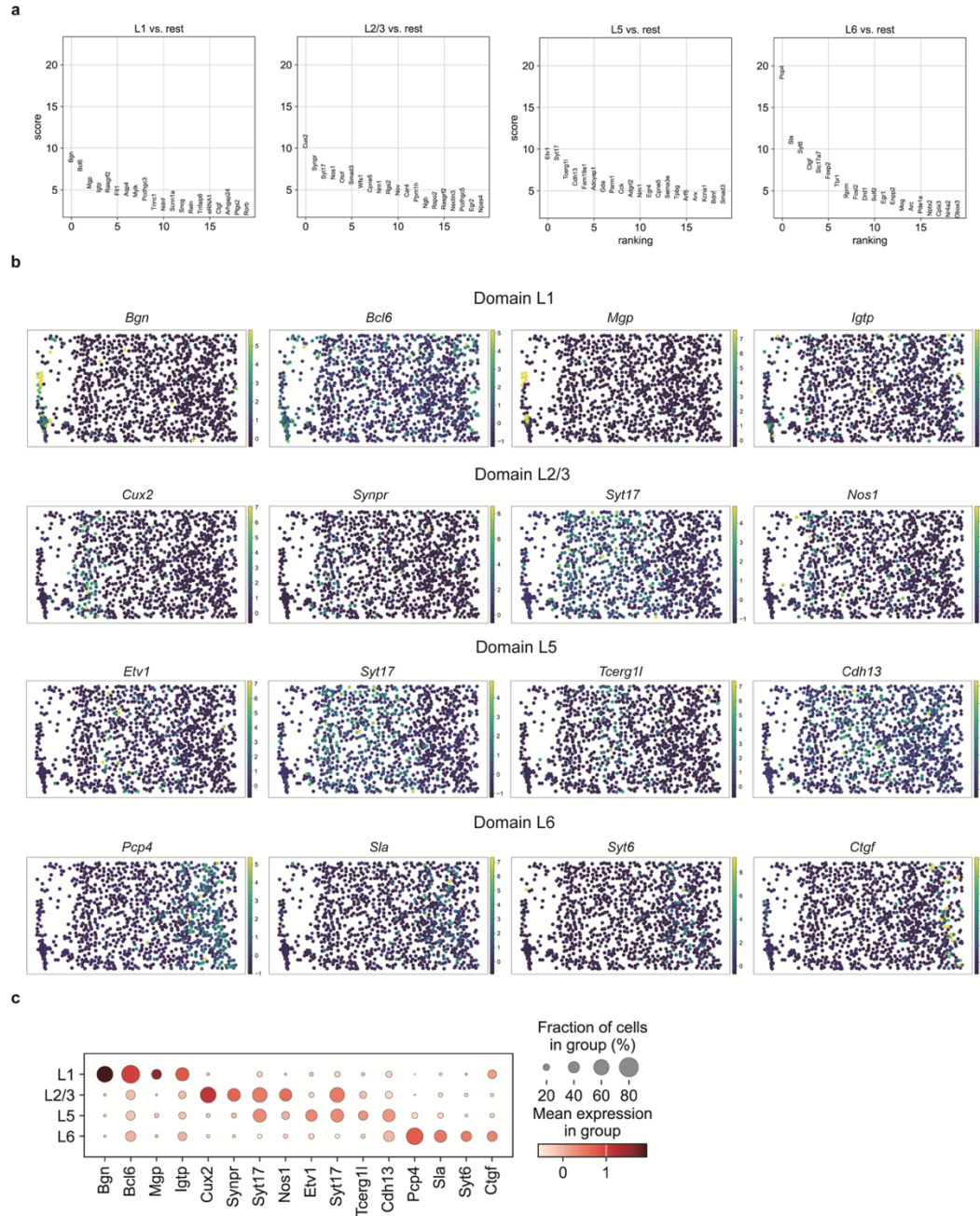


Fig. S13 Differential expression analyses of the STARmap data at domain-level. **a** Ranking of domain-specific DE genes identified by t-test using Scanpy. **b** Visualization of top 4 marker genes for each layer. **c** Dot matrix illustrating the mean expression of layer markers detected by GraphPCA. The color represents the mean expression within each of the domains and the dot size indicates the fraction of spots in the domains expressing a gene.

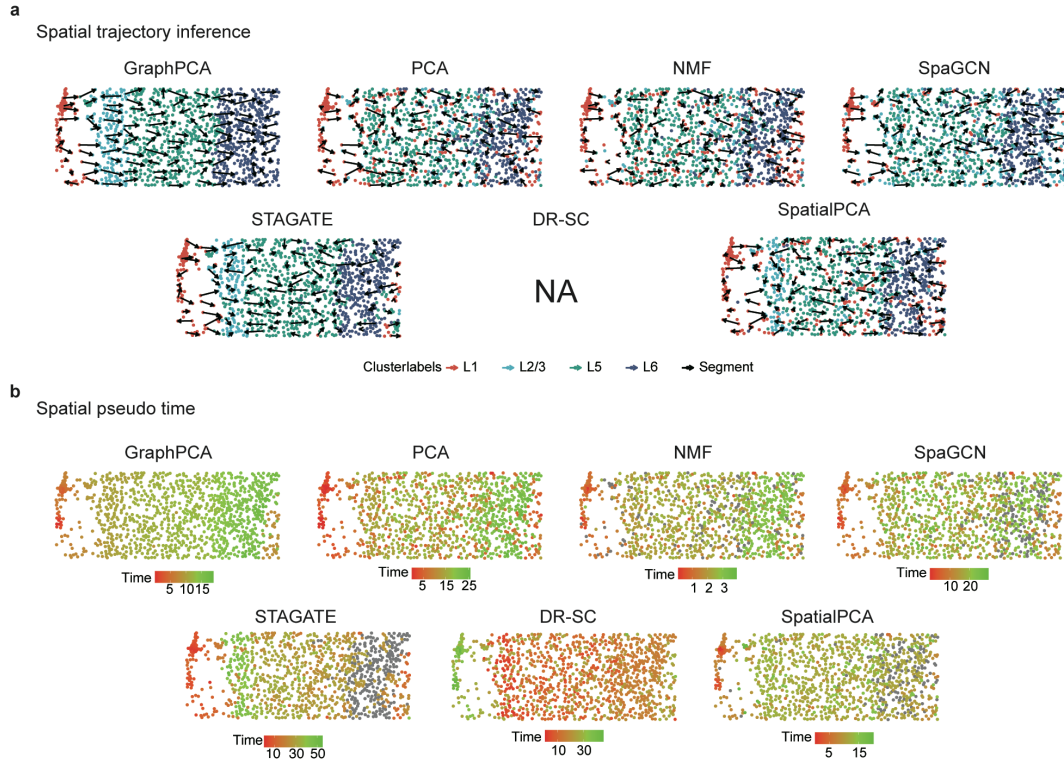


Fig. S14 Spatial trajectories and pseudo-time inferred by GraphPCA, PCA, NMF, SpaGCN, DR-SC and SpatialPCA for the STARmap data. a Trajectory inferred from low-dimensional embeddings by different methods. Arrows point from tissue locations with low pseudo-time to those with high pseudo-time. Color represents different tissue regions. Slingshot fails to get meaningful output on clustering result of DR-SC since there is only one cell detected as category of Layer L2/3. **b** Pseudo-time inferred from low-dimensional embeddings of different methods.

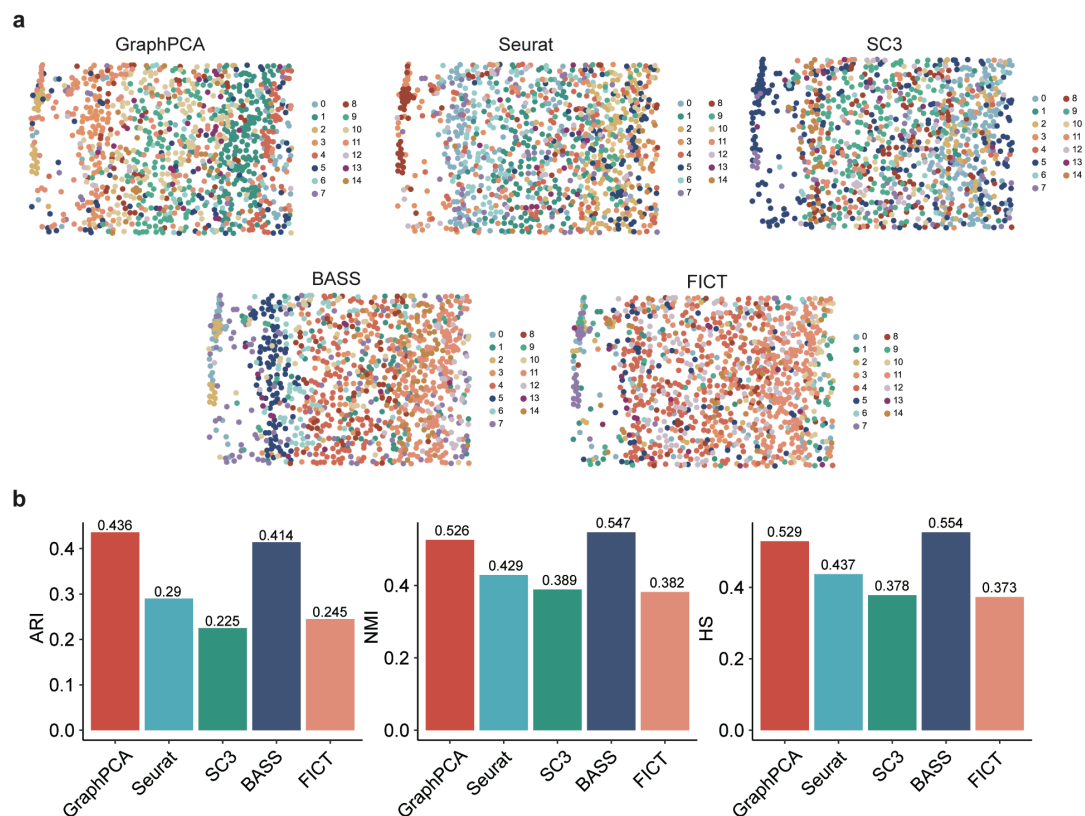


Fig. S15 Clustering results of different methods on the STARmap data at cell-type level. **a** Distribution of cell types detected by GraphPCA, Seurat, SC3, BASS, and FICT on section BZ5 of the STARmap data. **b** Clustering accuracies (ARI, NMI, and HS) by different methods at cell-type level.

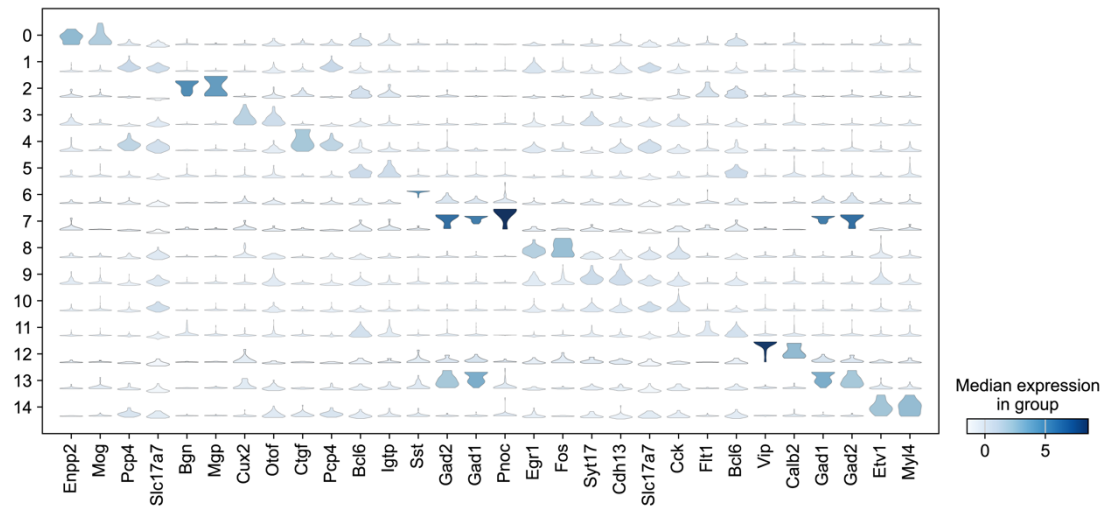


Fig. S17 Expressions of cell type marker genes in the STARmap data inferred by GraphPCA.

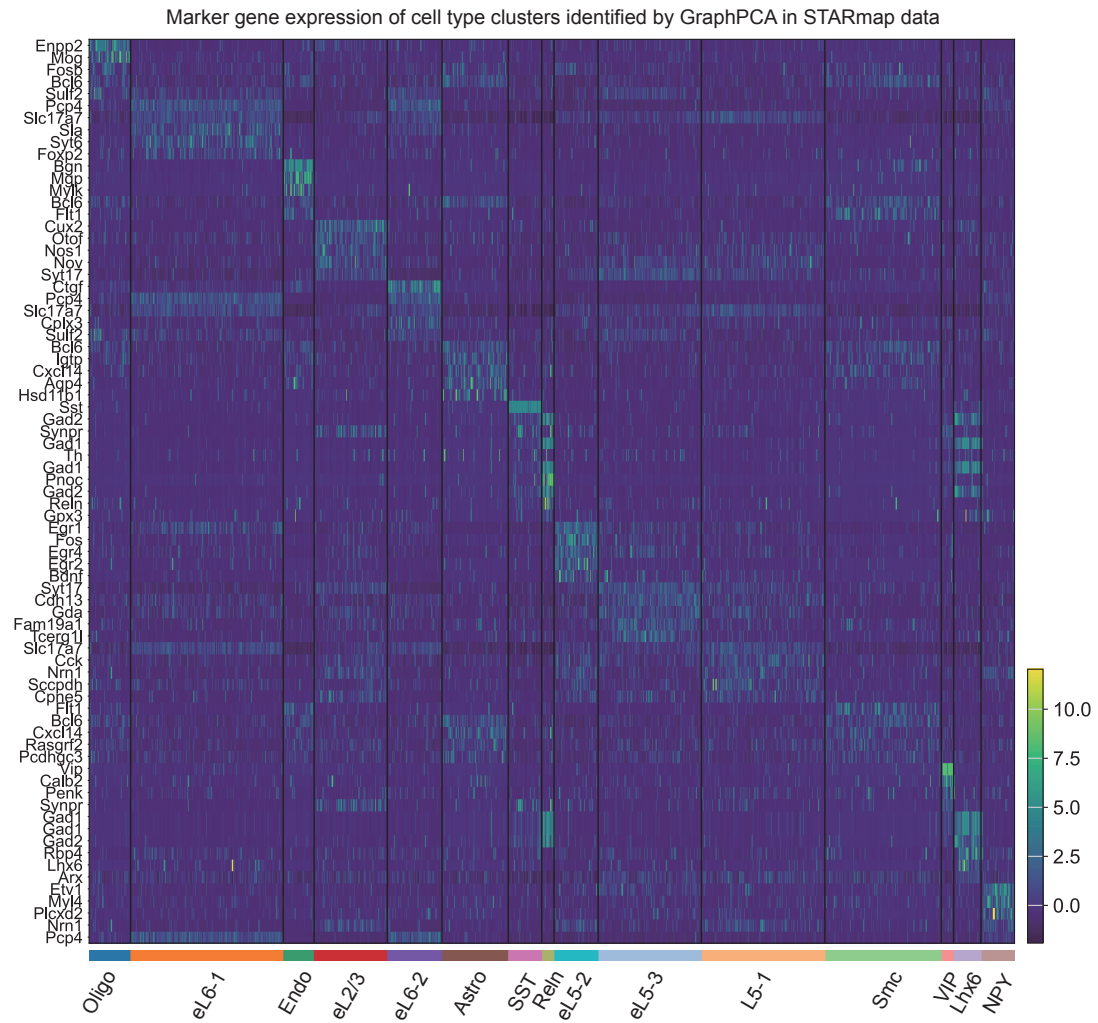


Fig. S18 Expression heatmap of cell type marker genes of the tissue section BZ5 inferred by GraphPCA.

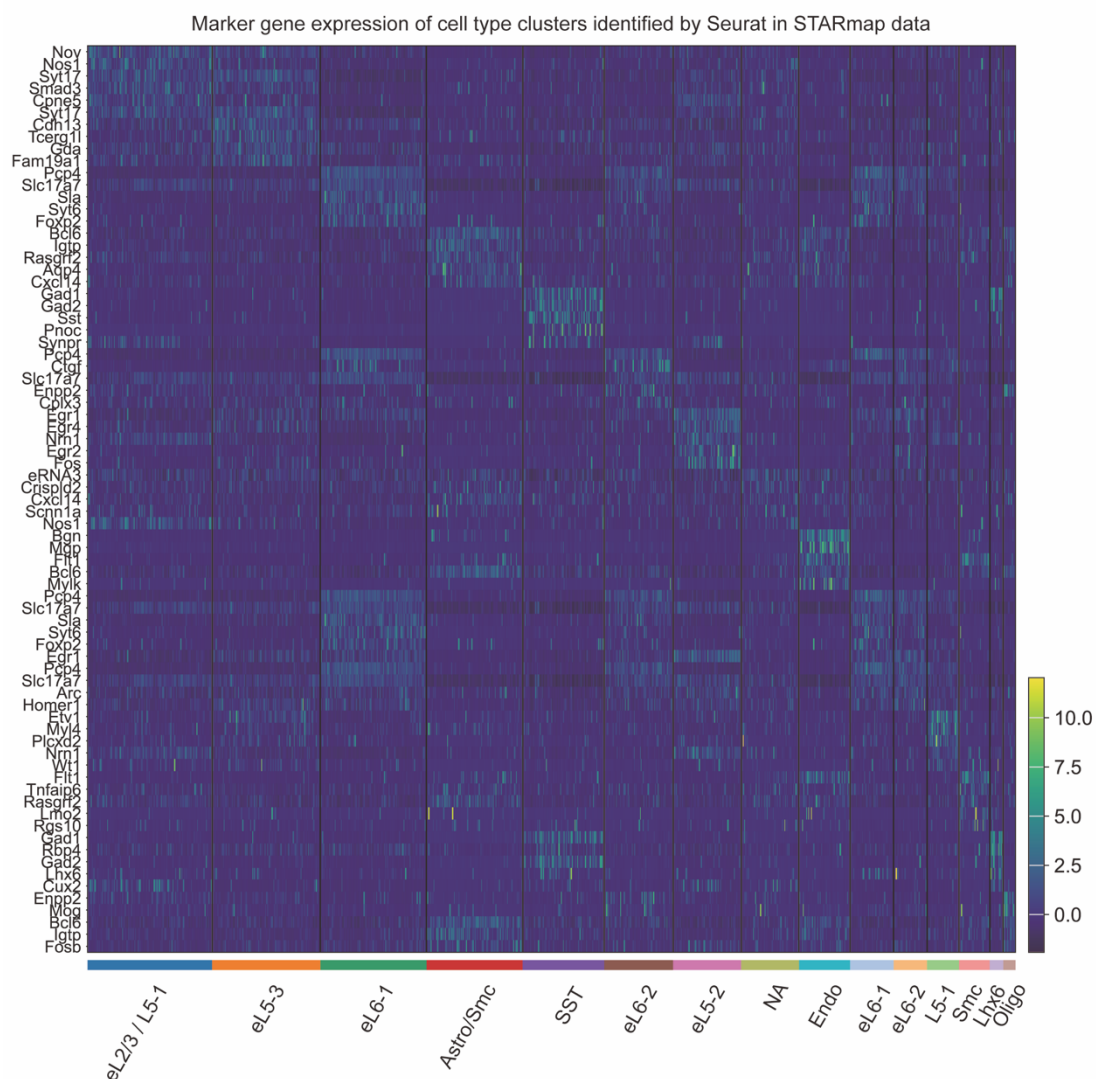


Fig. S19 Expression heatmap of cell type marker genes of the tissue section BZ5 inferred by Seurat.

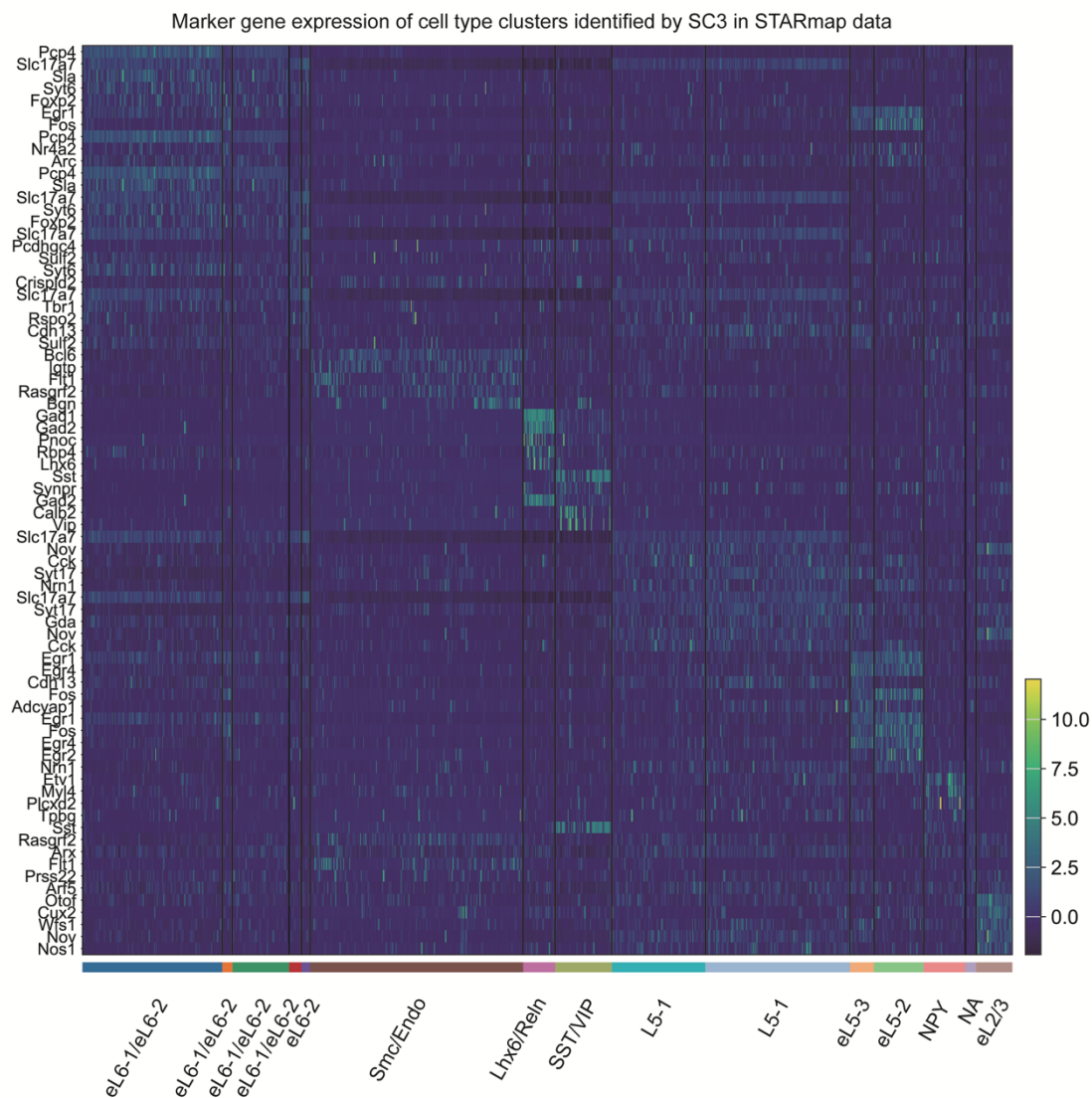


Fig. S20 Expression heatmap of cell type marker genes of the tissue section BZ5 inferred by SC3.

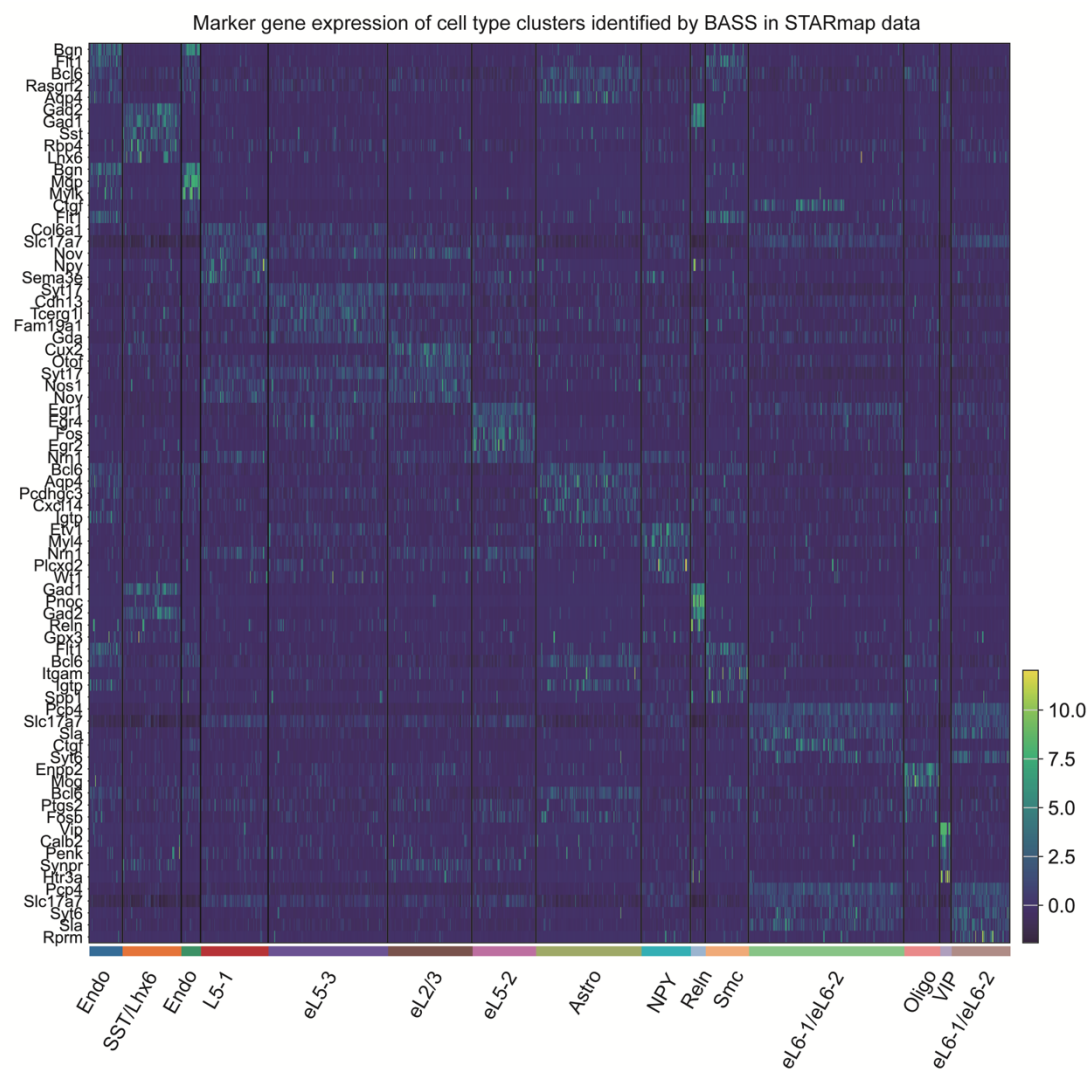


Fig. S21 Expression heatmap of cell type marker genes of the tissue section BZ5 inferred by BASS.

[illegible]

Fig. S22 Expression heatmap of cell type marker genes of the tissue section BZ5 inferred by FICT.

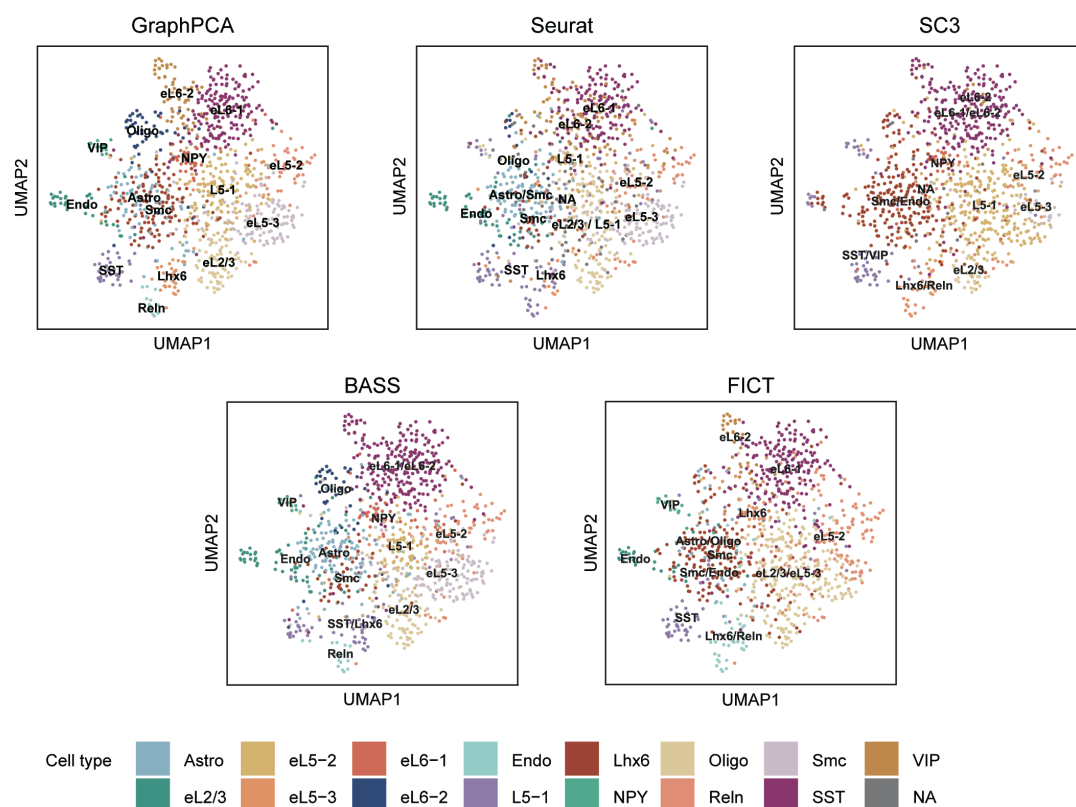


Fig. S23 UMAP visualization of cell type clustering results of the tissue section BZ5 inferred by GraphPCA, Seurat, SC3, BASS, and FICT.

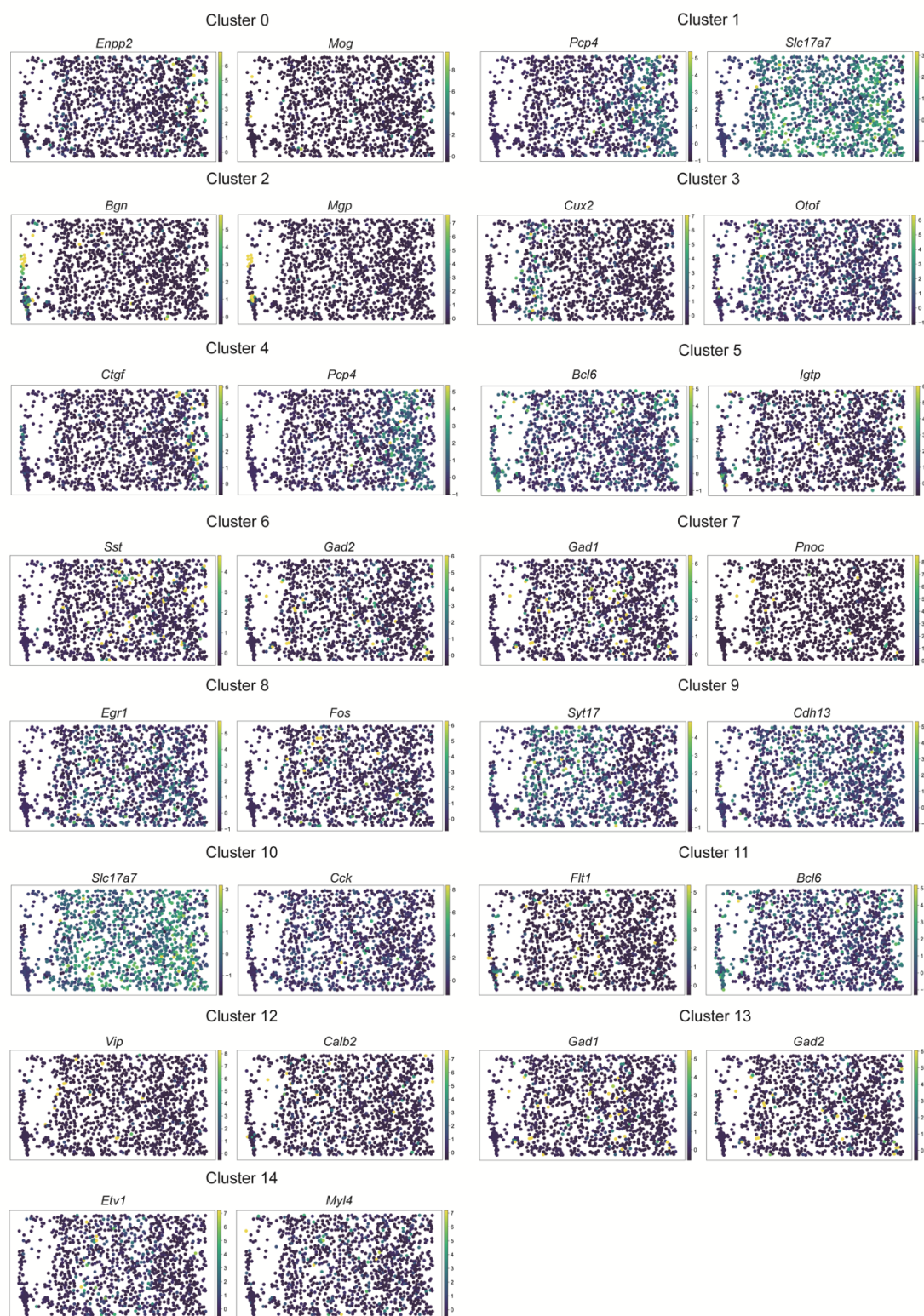


Fig. S24 Spatial expression patterns of cell-type marker genes in the STARmap data.

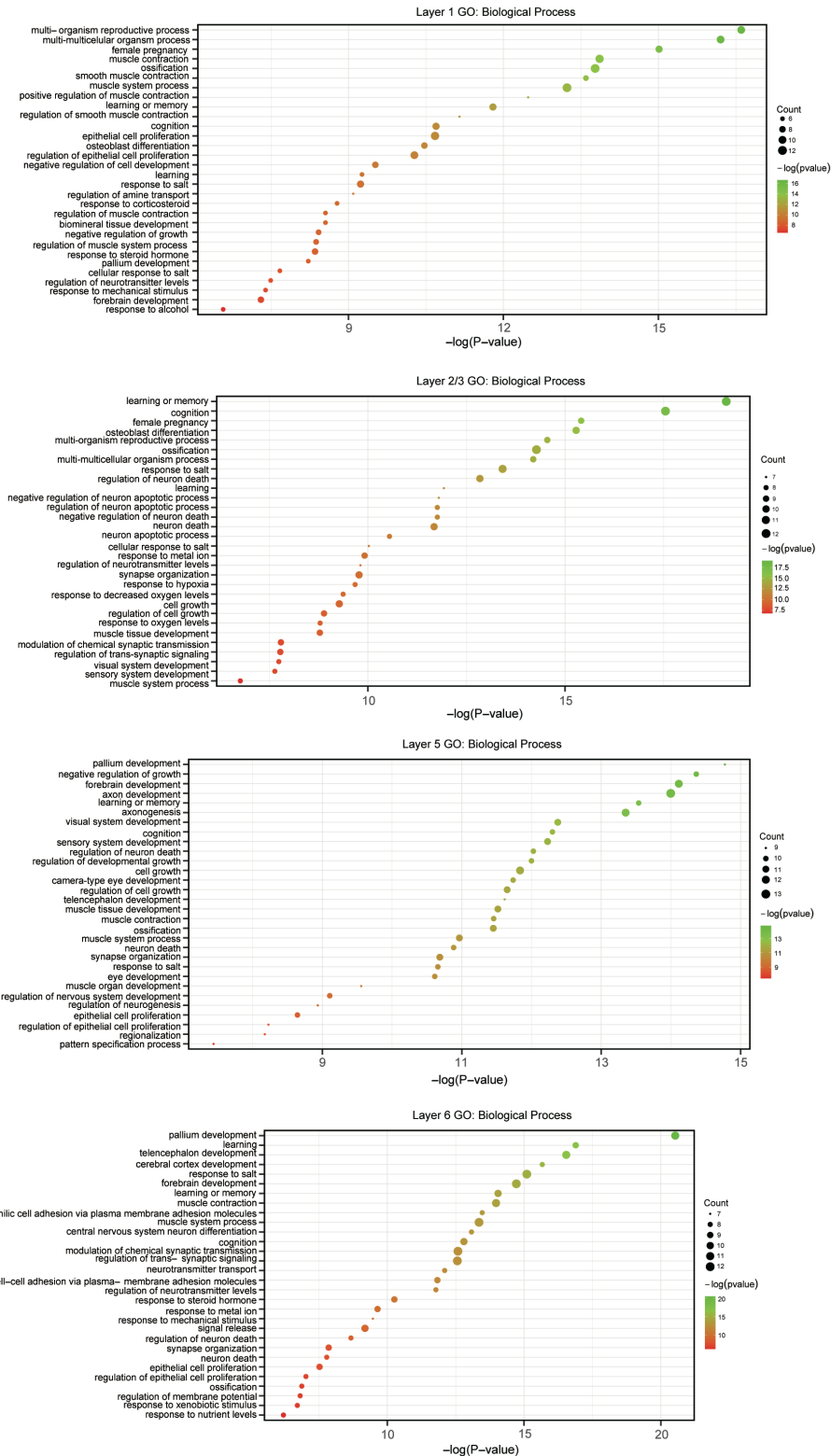


Fig. S25 Gene set enrichments of domain-specific DE genes in the STARmap data. The top 30 enriched gene sets are shown for each of the four detected tissue regions. Enrichment is calculated as $-\log_{10}$ adjusted p-value of the domain-specific DE genes.

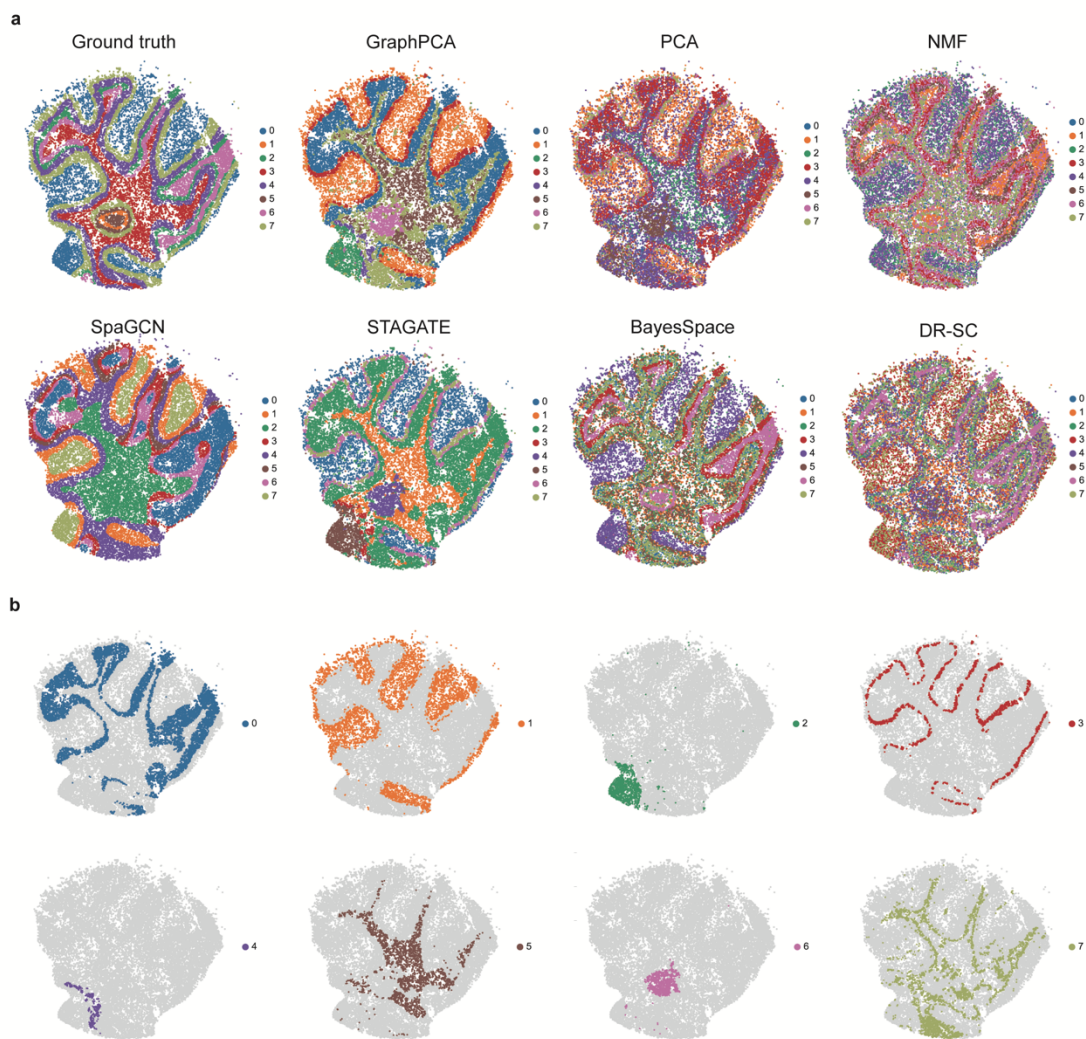


Fig. S26 Spatial domain detection results of the mouse cerebellum data by Slide-seq. a Annotated layer structure and spatial domains detected by GraphPCA, PCA, NMF, SpaGCN, STAGATE, BayesSpace, and DR-SC. We failed to run SpatialPCA on the Slide-seq data because it requires too much memory. **b** Visualization of each spatial domain detected by GraphPCA.

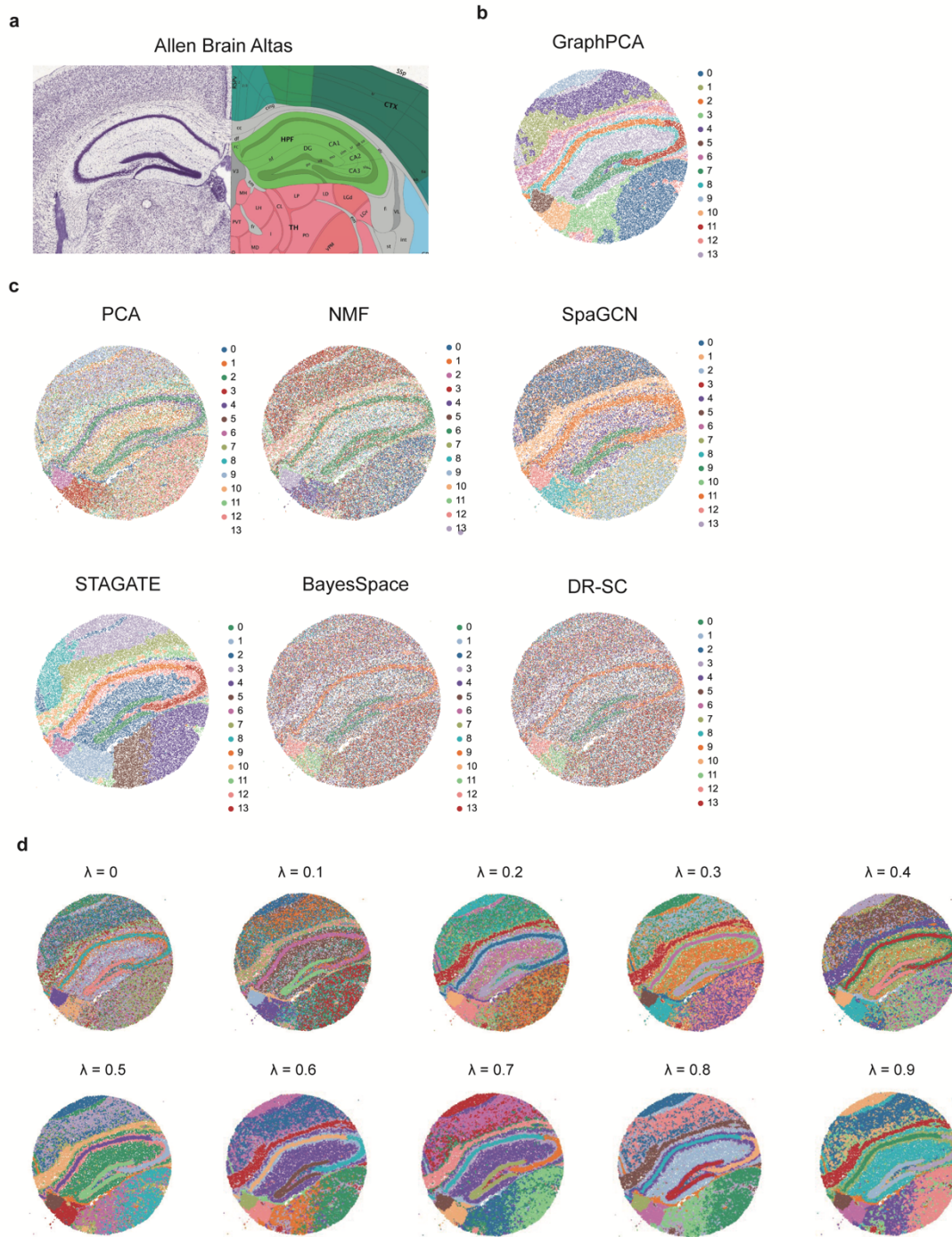


Fig. S27 Spatial domains detected in the mouse hippocampus data by Slide-seqV2. **a** Anatomical structure of the mouse hippocampus obtained from the Allen Brain Atlas. **b** Spatial domains detected by GraphPCA. **c** Spatial domains detected by PCA, NMF, SpaGCN, STAGATE, BayesSpace, and DR-SC, respectively. We failed to run SpatialPCA on the Slide-seqV2 data because it requires too much memory. **d** Impact of hyperparameter (λ) on detected spatial patterns using GraphPCA.

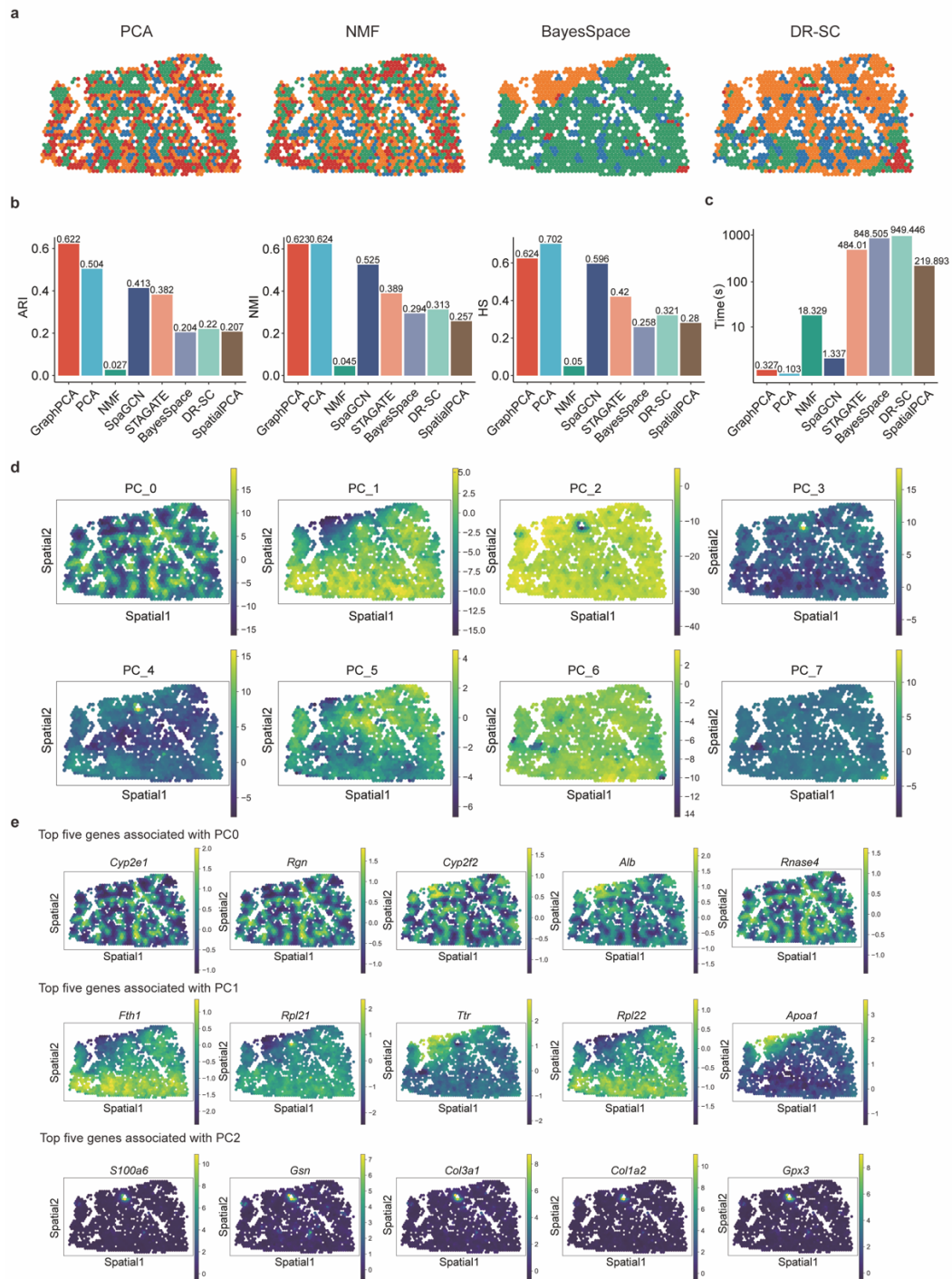


Fig. S28 Spatial domain detection results for the murine liver data by 10X Visium. a Spatial domains detected by PCA, NMF, BayesSpace, and DR-SC, respectively. **b** Clustering accuracies (ARI, NMI, and HS) of different methods for the liver data. **c** Bar plots of the runtimes spent by different methods in spatial domain detection. **d** Visualization of top 8 PCs obtained from GraphPCA. **e** Gene expression patterns of 5 example genes associated with the top 3 PCs, respectively.

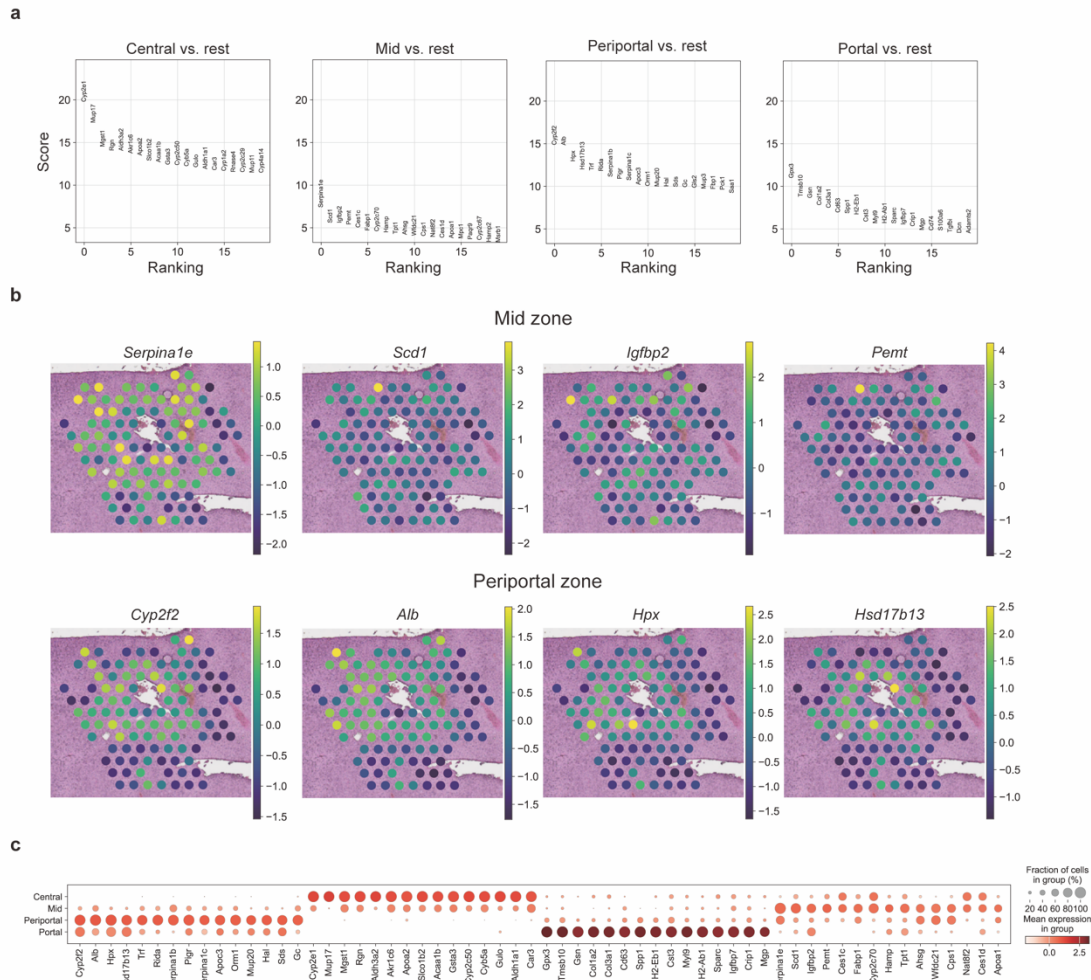


Fig. S29 Differential expression analyses of the murine liver data. a Ranking of domain-specific DE genes identified by t-test using Scanpy. **b** Visualization of top 4 marker genes for mid zone and periportal zone. **c** Dot matrix illustrating the mean expression of layer markers detected by GraphPCA. The color represents the mean expression within each of the domains and the dot size indicates the fraction of spots in the domains expressing a gene.

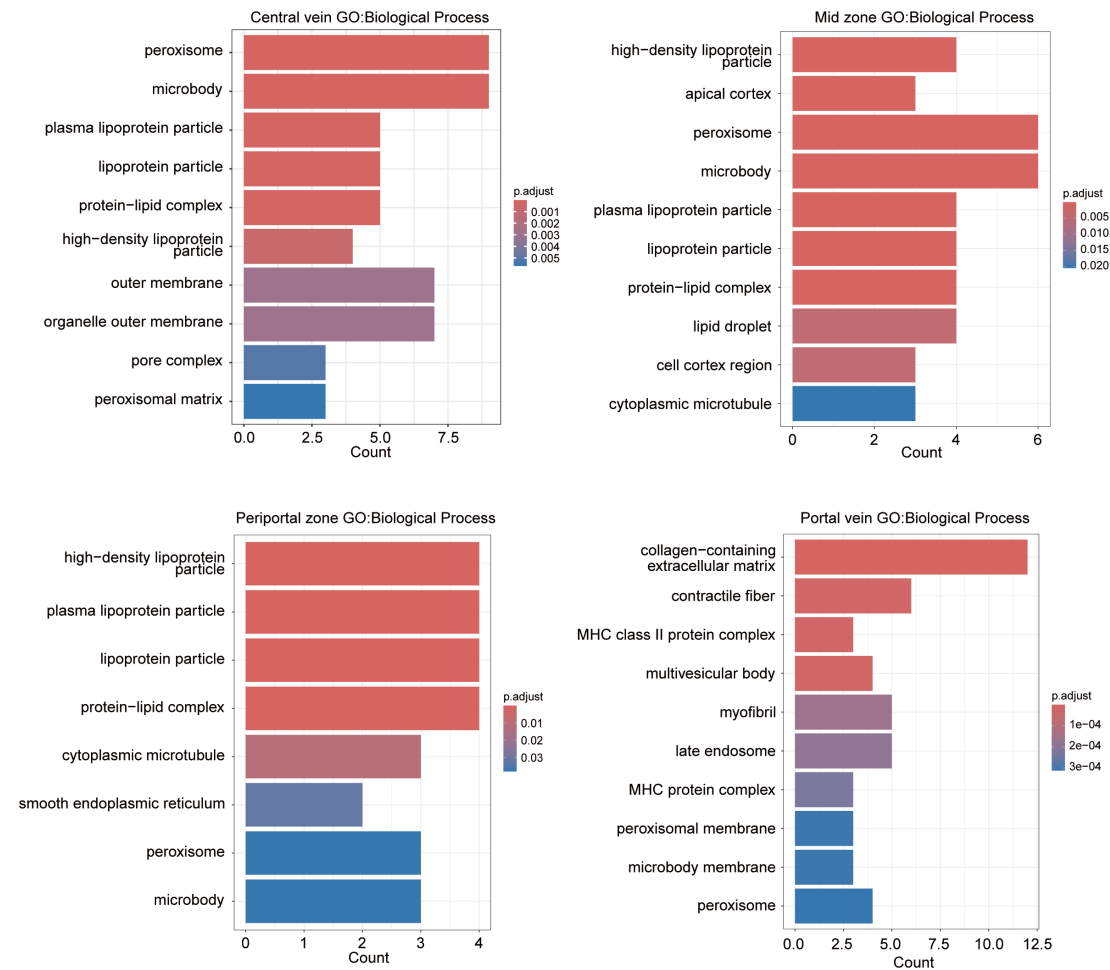


Fig. S30 Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathway analysis for the zonation marker genes detected by GraphPCA in the murine liver data. The top 10 enriched gene sets are shown for each of the four detected tissue regions. Enrichment is calculated as $-\log_{10}$ adjusted p-value of the zonation marker genes.

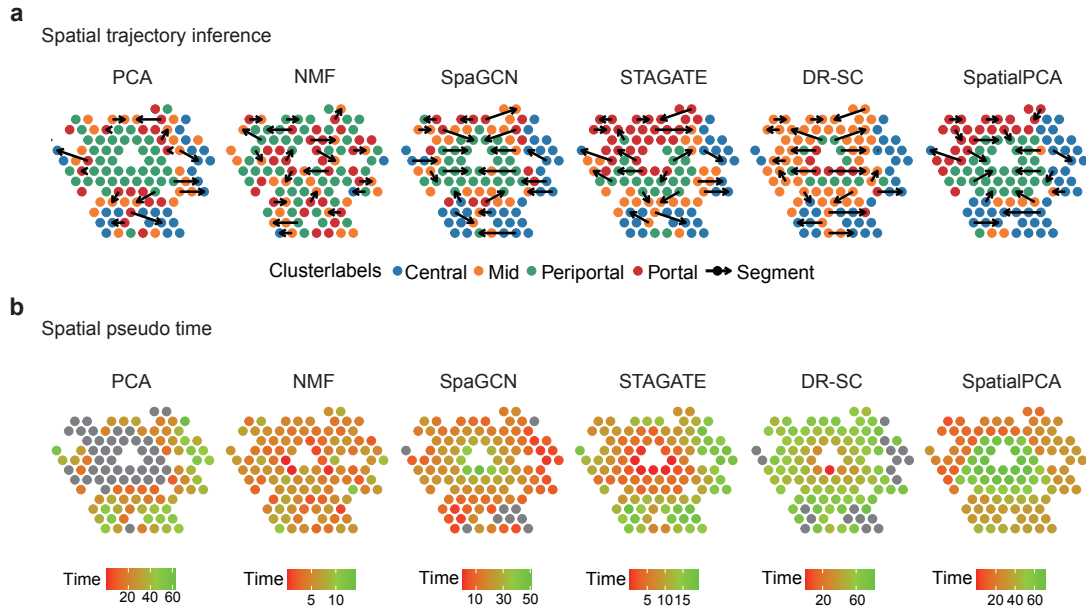


Fig. S31 Spatial trajectories and pseudo-time inferred by PCA, NMF, SpaGCN, STAGATE, DR-SC and SpatialPCA for the murine liver data. a Trajectory inferred from low-dimensional embeddings by different methods. Arrows point from tissue locations with low pseudo-time to those with high pseudo-time. Color represents different tissue regions. **b** Pseudo-time inferred from low-dimensional embeddings of different methods. The gray points indicate that the Slingshot algorithm detected more than one lineage. Consequently, for one of the lineages, the weights of points belonging to other lineages are set to zero, resulting in the corresponding pseudo-time value being NA.

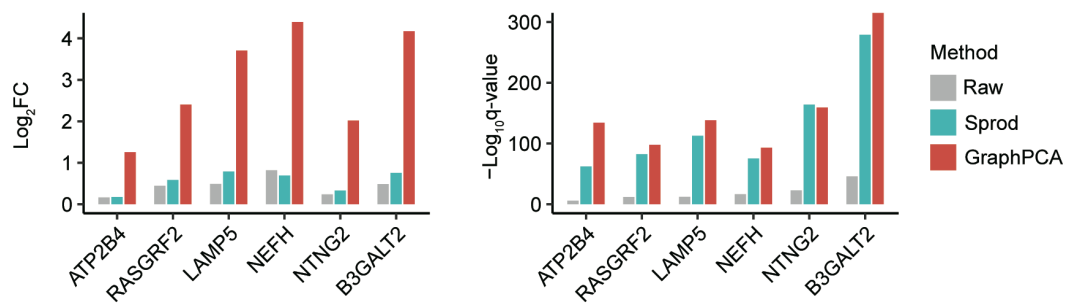


Fig. S32 Quantitative comparison of expression differences of layer marker genes using denoised data from GraphPCA and Sprod, as well as the raw data. Barplots of the log fold change (left) and negative log adjusted p-values (right). Higher log fold changes and negative log adjusted p-values (q-values) indicate a clearer layer-specific distribution of marker gene expression.

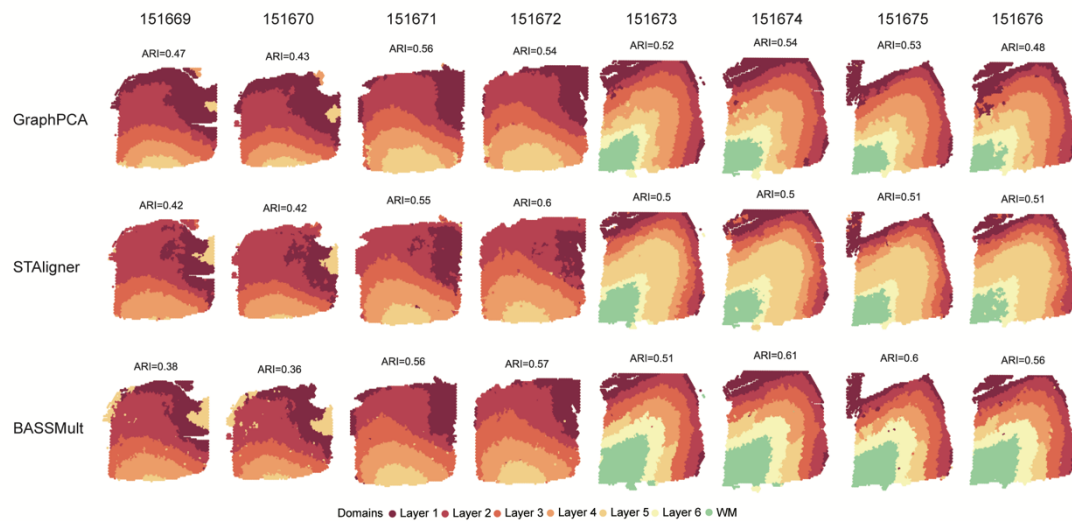


Fig. S33 Spatial domain detection results through multi-sample integration on the DLPFC dataset.

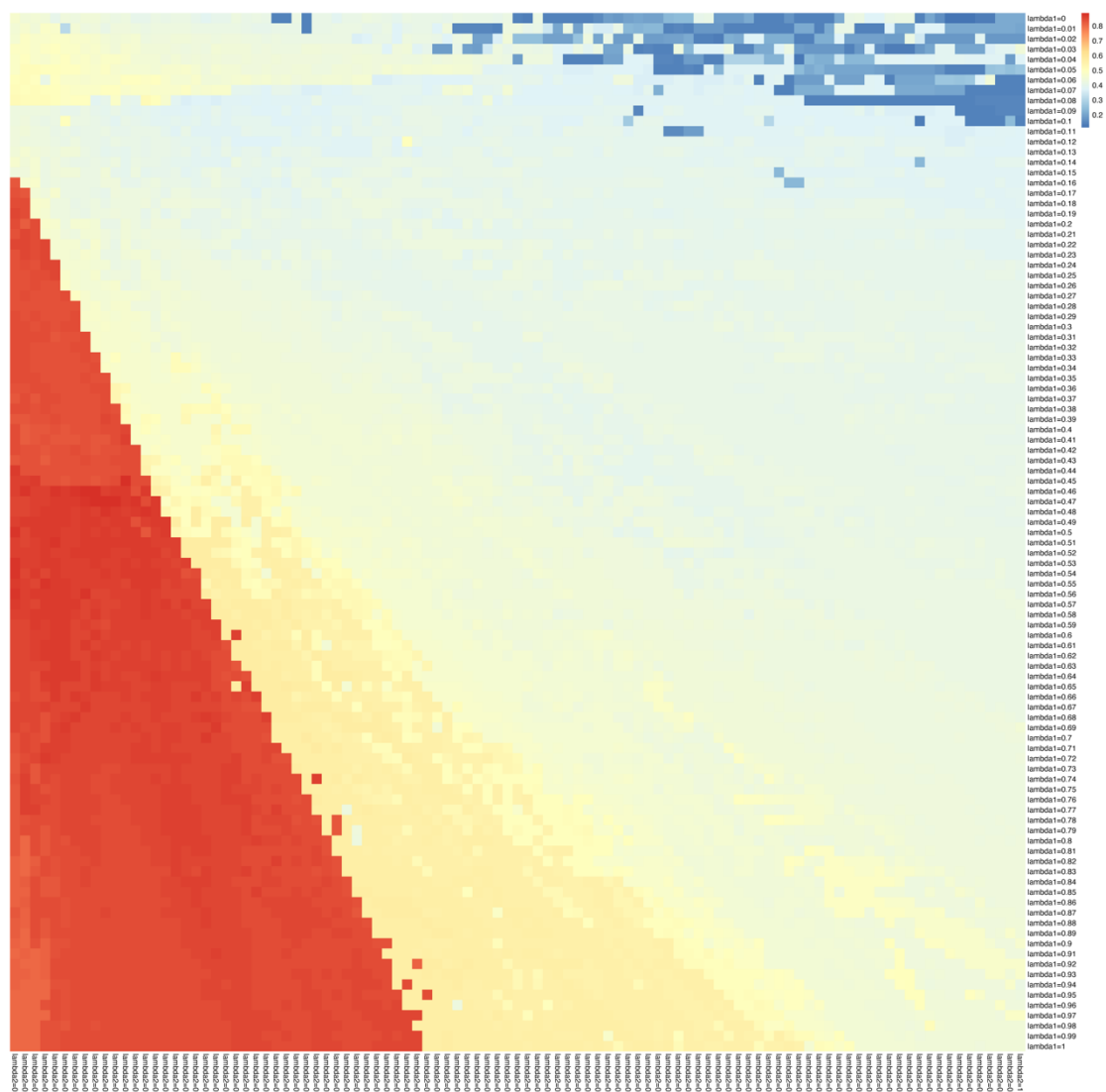


Fig. S34 Clustering accuracies at different hyperparameters in integration of both spatial location and gene expression data. Rows and columns correspond to graphical constraints of location and gene expression information, respectively.

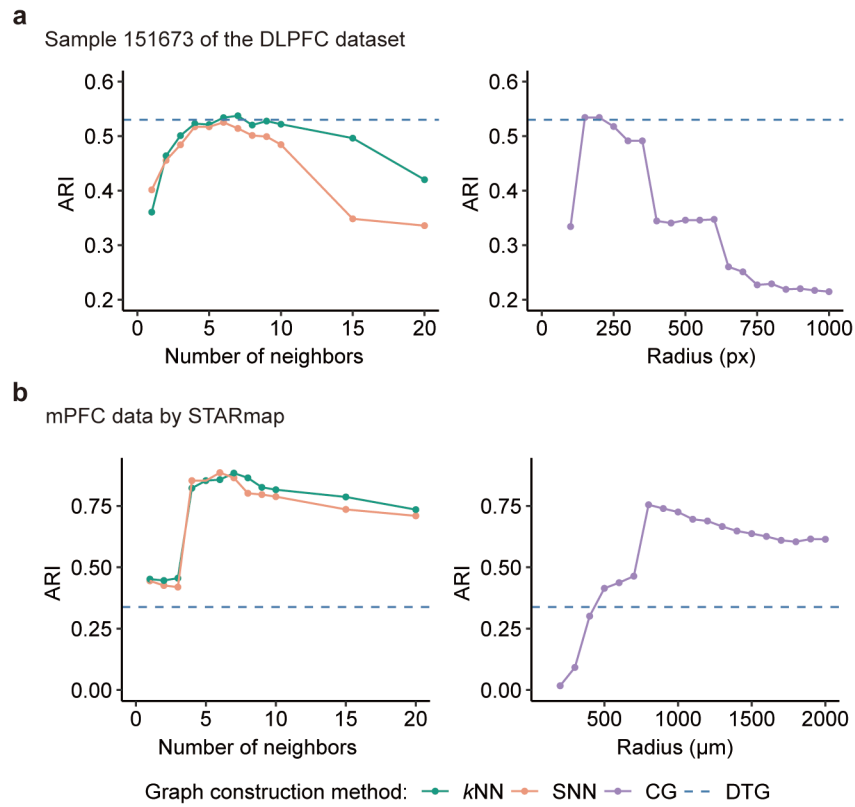


Fig. S35 Spatial domain detection accuracies by different graph construction methods and tunable hyperparameters. Clustering accuracies on sample 151673 of the DLPFC dataset (**a**) and mPFC data by STARmap (**b**) measured by ARI at different values of k for the k -nearest neighbor (k NN) and shared nearest neighbor (SNN), radius for the circular-neighborhood graph (CG), and the Delaunay triangulation graph (DTG).

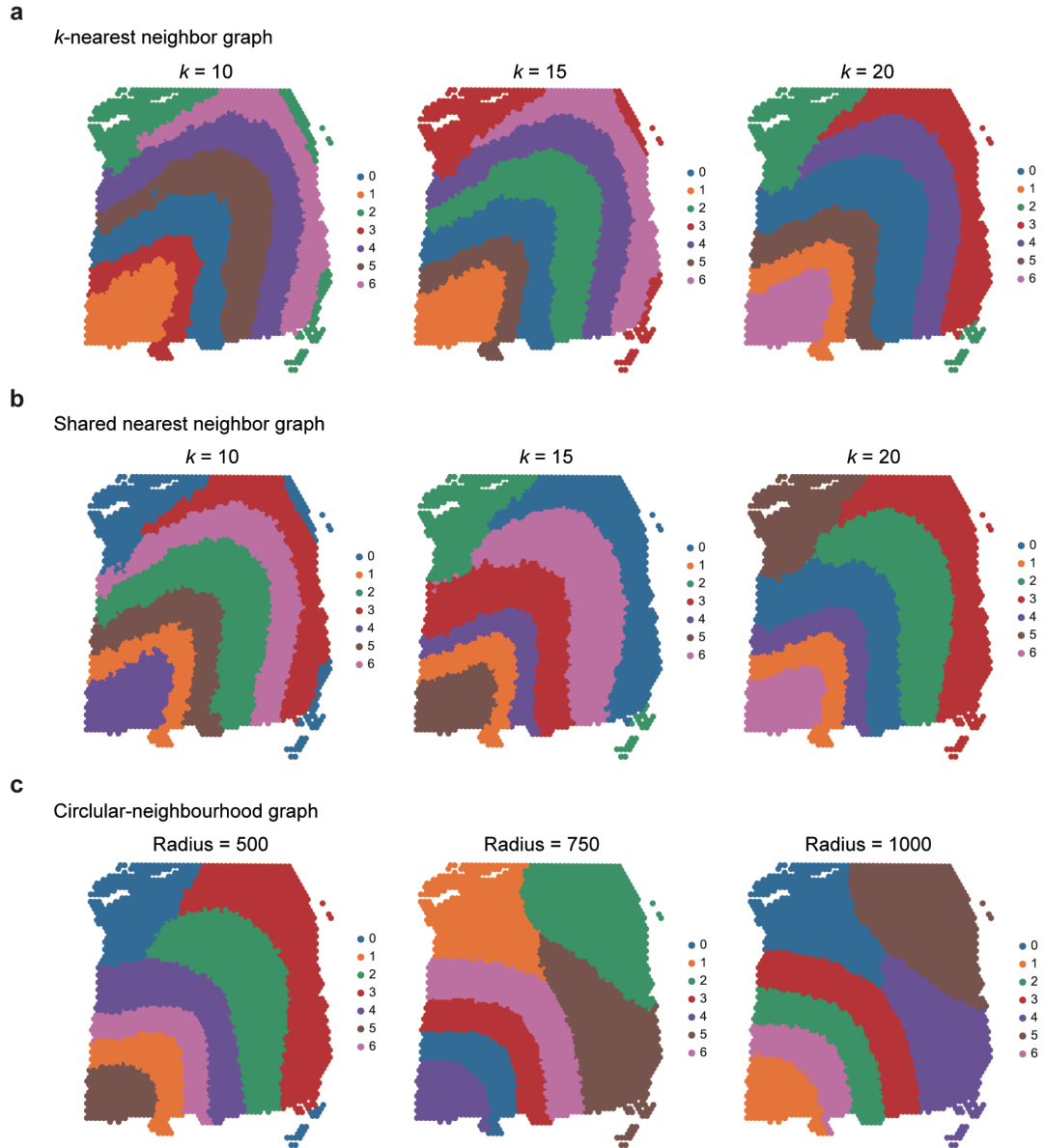


Fig. S36 Spatial domain detected by GraphPCA based on different values of parameters and graph construction methods on sample 151673 of the DLPFC dataset. The hyperparameter *k* in the *k*-nearest neighbor graph and shared nearest neighbor graph is set to be 10, 15, and 20. The hyperparameter radius in the circular-neighbourhood graph is set to be 500, 750, and 1,000.

Supplementary Tables

Table S1. Summary of the ST data used in this study.

Dataset	Platform	Slice id	# of spots
Human dorsolateral prefrontal cortex (DLPFC)	10X Visium	151507	4,226
		151508	4,384
		151509	4,789
		151510	4,634
		151669	3,661
		151670	3,498
		151671	4,110
		151672	4,015
		151673	3,639
		151674	3,673
		151675	3,592
		151676	3,460
Mouse medial prefrontal cortex (mPFC)	STARmap	BZ5	1,049
		BZ9	1,053
		BZ14	1,088
Murine liver	10X Visium	JB01	1,293
Mouse brain	10X Visium	sagittal anterior	2,695
		sagittal posterior	3,355
Mouse cerebellum	Slide-seq	SCP354	20,982
Mouse hippocampus	Slide-seqV2	SCP815	53,208

Table S2. Parameter settings used by GraphPCA across different datasets and data analysis tasks. For all data analyses, the k NN graph construction and K-means clustering algorithms were employed.

Dataset	Platform	Task	Parameters
Human dorsolateral prefrontal cortex (DLPFC)	10X Visium	Spatial domain detection	# of neighbors: 6 λ : 0.3
Mouse medial prefrontal cortex (mPFC)	STARmap	Spatial domain detection	# of neighbors: 6 λ : 0.3
		Cell type annotation	# of neighbors: 2 λ : 0.2
Murine liver	10X Visium	Spatial domain detection	# of neighbors: 2 λ : 0.2
Mouse cerebellum	Slide-seq	Spatial domain detection	# of neighbors: 8 λ : 0.3
Mouse hippocampus	Slide-seqV2	Spatial domain detection	# of neighbors: 8 λ : 0.3