Genome Biology

METHODOLOGY

Open Access

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

Jiyuan Yang[1], Lu Wang[1,2], Lin Liu[3] and Xiaoqi Zheng[1*]

*Correspondence:
xqzheng@shsmu.edu.cn

[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, Guangxi Zhuang Autonomous Region, Nanning, 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

Abstract

The rapid advancement of spatial transcriptomics technologies has revolutionized our understanding of cell heterogeneity and intricate spatial structures within tissues and organs. However, the high dimensionality and noise in spatial transcriptomic data present significant challenges for downstream data analyses. Here, we develop GraphPCA, an interpretable and quasi-linear dimension reduction algorithm that leverages the strengths of graphical regularization and principal component analysis. Comprehensive evaluations on simulated and multi-resolution spatial transcriptomic datasets generated from various platforms demonstrate the capacity of GraphPCA to enhance downstream analysis tasks including spatial domain detection, denoising, and trajectory inference compared to other state-of-the-art methods.

Keywords: Spatial transcriptomics, Dimension reduction, PCA, Spatial domain detection

Background

Spatial transcriptomics (ST) technologies have fundamentally reshaped the current research landscapes of cellular and molecular biology, and significantly deepened our understanding on cellular heterogeneity, gene expression-cellular microenvironment interaction, and spatial specificity of gene expression in complex tissues [1–4]. Unlike single cell RNA sequencing (scRNA-seq) which loses spatial information during cell dissociation [5–8], ST quantifies expression of single RNA molecules while preserving spatial location information through in situ or spatial barcoding approaches [9]. Currently, there are two main types of ST techniques depending on the underlying experimental protocols, i.e., imaging-based methods and next-generation sequencing-based (NGS-based) technologies. The first type of methods includes in situ hybridization (e.g., seqFISH [10, 11], MERFISH [12–14], osmFISH [15]) and in situ sequencing (STARmap [16] and FISSEQ [17]). These methods often provide finer cellular positional details (in single-cell or even subcellular resolution), but are limited to pre-selected encoding

probes, and thus fail to cover genes on the whole transcriptome scale [18]. The second type of techniques encodes location information into transcripts and quantifies both gene expression and spatial information through massively parallel sequencing, with representative technologies of ST [19], 10X Genomics Visium [20], Slide-seq [21], Slide-seqV2 [22], and Stereo-seq [23]. All these diverse technologies yield a data matrix that records the whole transcriptome expression levels at every spot, cell, or spatial pixel [3].

However, analyzing gene expression profiles obtained from various ST technologies faces enormous challenges due to the sparsity, ultra-high dimensionality, and low signal-to-noise ratio (SNR [24]) of the data matrix. Consequently, dimension reduction becomes a necessary preprocessing step to improve SNR and mitigate the curse-of-dimensionality [24, 25]. It is also a critical step in the analysis of scRNA-seq data, often followed by key downstream tasks such as cell type identification [26–28], trajectory inference [29–31], and batch effect removal [32–35]. With the rise of ST technologies, most researchers directly apply dimension reduction methods developed for bulk or scRNA-seq data to ST data (Seurat [36], Scanpy [37], STUtility [38]). However, those methods may fall short in terms of the capability of fully exploiting the location information in ST data, potentially leading to efficiency loss or even biased and erroneous results.

To overcome this limitation, dimension reduction methods crafted specifically for ST data have been proposed [39–41]. These methods typically assume that proximal spots exhibit similar gene expression patterns, such that the low-dimensional embedding can approximately preserve this spatial structure as much as possible. For instance, building upon the framework of probabilistic PCA, Shang and Zhou proposed SpatialPCA [39], which uses a kernel matrix to model the spatial correlation structure across locations. Also under the probabilistic PCA framework, DR-SC [40] integrates dimension reduction and clustering using a two-layer hierarchical Bayesian model, where the low-dimensional latent variables in it are assumed to follow a Gaussian mixture distribution. Both methods utilized computationally intensive iterative procedures, i.e., maximum likelihood-based optimization in SpatialPCA and MCMC in DR-SC, to approximate the locally optimal solution. There are also some recent approaches using deep learning algorithms to learn a nonlinear, low-dimensional embedding of the ST data. For example, SpaGCN [42] uses graph convolutional network to integrate multi-modal data including gene expression, spatial location of spots/cells, and histology data. STAGATE [43] obtains spatial representation of spots by a graph attention auto-encoder guided by a spatial neighbor network and cell type-aware network. However, deep learning-based methods suffer from lack of interpretability, high computational complexity in model training, and difficulty in tuning a large number of hyperparameters.

In this study, we develop GraphPCA, a novel graph-constrained, interpretable, and quasi-linear dimension-reduction algorithm tailored for ST data. GraphPCA learns the low-dimensional representation of ST data based on PCA with minimum reconstruction error, by incorporating spatial location information as constraints in the reconstruction step. By increasing the importance of the spatial network constraints, adjacent spots in the original dataset are more inclined to be positioned in nearby points in the low-dimensional embedding space. More importantly, the computationally efficient close-form solution of GraphPCA allows rapid embedding of massive single-cell or even

subcellular resolution spatial transcriptomic data generated from techniques including Slide-seq and Stereo-seq. Finally, we demonstrate the superiority of GraphPCA over competing methods, i.e., PCA, NMF, SpaGCN, BayesSpace [44], DR-SC, SpatialPCA, and STAGATE, through comprehensive synthetic experiments and real spatial transcriptomic data with a variety of resolutions, species, and tissue states.

Results

Overview of GraphPCA

Here, we developed GraphPCA, an interpretable and quasi-linear statistical algorithm for dimension reduction of spatial transcriptomics data (Fig. 1a, see "Methods" for details). Building upon the flexible PCA framework, GraphPCA enables the low-dimensional embeddings to effectively preserve location information by leveraging spatial neighborhood structure between spots/cells as graph constraints. The input data for GraphPCA includes a gene expression matrix along with spatial coordinates of spots, which are utilized to construct a spatial neighborhood graph (k NN graph by default [45]). In contrast to the classical PCA, GraphPCA infers an embedding matrix integrating both spatial location and gene expression information by solving an optimization problem with constraints determined by the constructed spatial neighborhood graph. In particular, we incorporated the constraints by introducing a penalty term (Fig. 1a, middle panel), with penalty strength controlled by a tunable hyperparameter λ. By design, the resulted constrained penalized optimization problem has a closed-form solution. As a consequence, GraphPCA can process ST data efficiently at vastly different scales. The low-dimensional spatial embeddings inferred by GraphPCA can be readily utilized for various downstream analysis tasks including spatial domain detection, visualization, denoising, and trajectory inference (Fig. 1a, right panel).

Synthetic experiments

To comprehensively evaluate the performance of GraphPCA and other competing algorithms, we generated a series of simulated data using scDesign3 [46]. With simulated datasets, one could use curated spatial domain labels as ground truth to evaluate the accuracy of the clustering results. Specifically, we downloaded the anatomical structure of mouse brain sagittal from the Allen Brain Atlas [47] as ground truth layer labels (Fig. 1b) and used scDesign3 to simulate gene expression data (see "Methods" for details). We compared GraphPCA with four popular algorithms for ST dimension reduction (PCA, NMF, DR-SC, and SpatialPCA) and three algorithms for spatial domain detection (SpaGCN, STAGATE, and BayesSpace). Clustering performance was evaluated using the adjusted Rand index (ARI [48]), normalized mutual information (NMI [49]), and homogeneity score (HS [50]).

We first examined the impact of the hyperparameter λ on spatial domain detection. By increasing λ from 0 to 1 with a step size of 0.01 across four synthetic datasets, we first observed an overall improvement in ARI as λ increases (Fig. 1c). This suggests that the integration of proper location information enhances spatial domain detection, with a larger graph constraint resulting in smoother and more contiguous clustering output (Additional file 1: Figs. S1–2). However, excessively large λ leads to the spatial constraints dominating the dimension reduction objective and results in deteriorated performance.

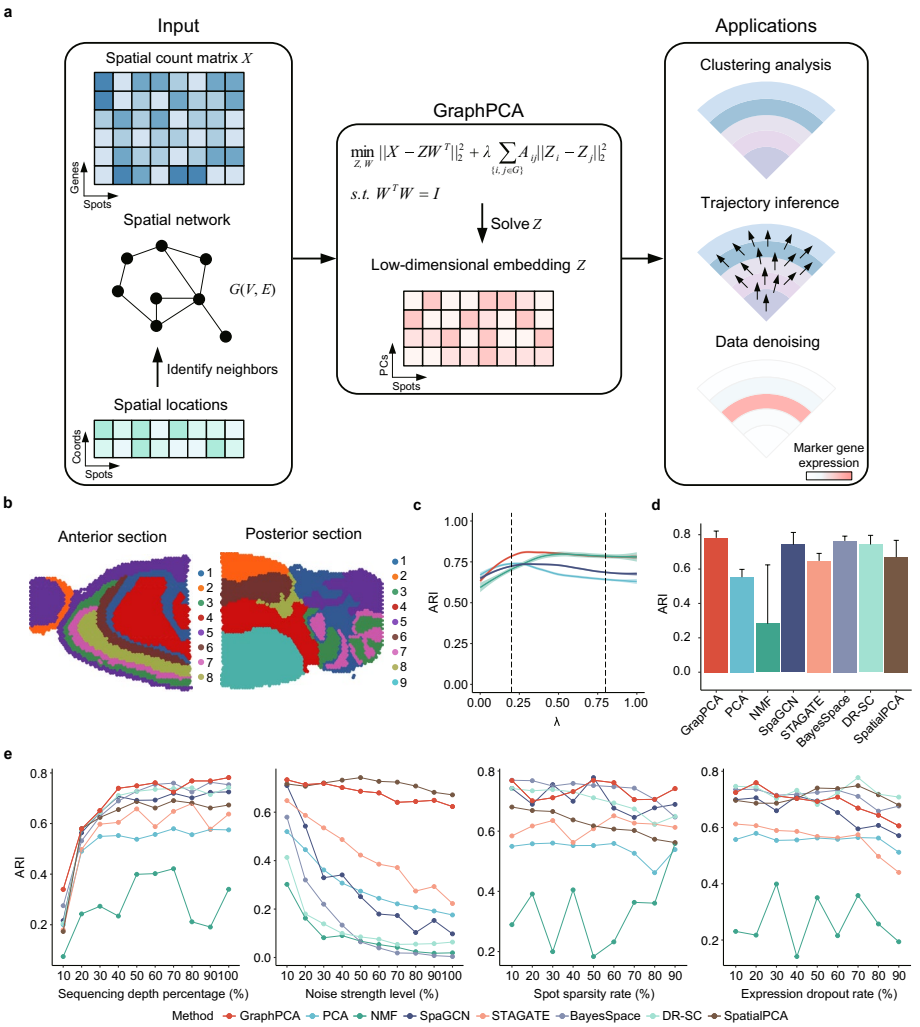

Fig. 1 Workflow of GraphPCA and synthetic experiments validation. **a** GraphPCA is a novel graph-constrained, interpretable, and quasi-linear dimension-reduction method tailored for ST data. It begins with input data including gene expression matrix ***X*** along with spatial coordinates ***S*** of spots. It first constructs a spatial neighborhood graph ***G*** using the spatial coordinates ***S*** and sets the graph constraint parameter λ to characterize the spatial relationships and dependences in the low-dimensional embedding ***Z***. Subsequently, GraphPCA infers the embedding matrix ***Z*** by integrating both spatial location ***S*** and gene expression information ***X*** by solving a non-convex optimization problem with graph constraints. The output of GraphPCA can be readily utilized for various downstream analysis tasks including spatial domain detection, trajectory inference, and denoising. **b** In simulation, we obtained the anatomical structure of mouse brain sagittal from the Allen Brain Atlas as ground truth layer labels and simulated ST data using scDesign3. **c** Clustering accuracies of GraphPCA on simulated datasets across varying values of the graph constraint parameter λ (x-axis). **d** Clustering accuracies of different methods on simulated data. Error bars indicate 95% confidence intervals across 20 replicates. **e** The robustness of GraphPCA and other methods under varying simulation scenarios including different sequencing depths, noise levels, spot sparsity, and expression dropout rates. For each scenario and method, dots represent the mean ARI calculated across 20 replicates

Based on the empirical observations from the synthetic experiments, we recommend choosing λ between 0.2 and 0.8 (default 0.3 unless stated otherwise) for tissue samples with evident spatial layered structure. Compared to other methods, GraphPCA demonstrates superior performance on the synthetic data (median ARI: 0.784), outperforming algorithms considering spatial information such as PCA (median ARI: 0.556)

and NMF (median ARI: 0.185) across all metrics (Fig. 1d, Additional file 1: Fig. S3a). Methods incorporating spatial information (DR-SC and SpatialPCA) and deep learning-based algorithms (SpaGCN and STAGATE) also achieve comparable clustering results, but exhibit relatively high variabilities (standard deviations across 4 replicates are 0.040, 0.053, 0.088, 0.071, and 0.044 for GraphPCA, DR-SC, SpatialPCA, SpaGCN, and STAGATE, respectively).

Next, we evaluated the robustness of GraphPCA under varying simulation scenarios including different sequencing depths, noise levels, spot sparsity, and expression dropout rates (Fig. 1e, Additional file 1: Fig. S3b). For sequencing depths, we down-sampled the simulated count matrix from 100 to 10% of the original depth and found that GraphPCA consistently outperformed other competing methods even with only 10% of the original sequencing depth (Fig. 1e, first column). By introducing different levels of Gaussian white noises with increased standard deviations to the raw data, GraphPCA performs consistently well despite the increasing SNR, demonstrating the robustness of GraphPCA to noise. In contrast, ST-tailored algorithms such as SpaGCN, DR-SC, and BayesSpace eventually fail to cluster cells correctly under low SNRs (Fig. 1e, second column). For spot sparsity, we randomly removed a fraction of spots to increase sparsity of the data (from 10 to 90%) and found that both GraphPCA and BayesSpace sustain relatively high performance under all sparsity ratios (Fig. 1e, third column). To examine the impact of expression dropout, we randomly set the expressions of a fraction of genes (from 10 to 90% of all genes) to 0 and found that GraphPCA still maintains high ARIs even at a dropout rate of 60% (Fig. 1e, fourth column). In summary, GraphPCA exhibits its superior accuracy and robustness in clustering performance across all four scenarios compared to competing methods.

Human dorsolateral prefrontal cortex data by Visium

We next assessed the clustering performance of GraphPCA against competing methods using real ST data with expert pathological annotations. These datasets encompass various sequencing technologies, resolutions, species, and tissue states. We first evaluated GraphPCA on the human dorsolateral prefrontal cortex data (DLPFC) obtained from the 10X Visium platform [51]. This dataset comprises twelve tissue sections from three adult donors, with a median depth of 291 million reads per sample. Each tissue section has a median of 3844 spots (Additional file 1: Table S1), and on average 1734 genes are recorded per spot. As an example, the slice 151673 assays 3639 spots across 33,538 genes, and each spot is annotated into one of six neuronal layers or white matter. We compared GraphPCA with other methods, using the true number of spatial domains and the same 3000 spatially variable genes (SVGs) identified by SPARK [52] as input.

As expected, dimension reduction algorithms that do not consider spatial information, i.e., PCA and NMF, fail to distinguish distinct layers of cortex, thus resulting in poor clustering accuracies (PCA: ARI 0.290, NMF: ARI 0.235) (Fig. 2a). Although incorporating additional histological imaging data as input, SpaGCN fails to accurately recover cortical layers 3–6. BayesSpace does not capture the boundary between white matter and layer 6 correctly. DR-SC identifies cortical layers that roughly match manual annotations in shape but display discrepancies in thickness, and a few spots adjacent to white matter are erroneously identified as layer 1 (Fig. 2a). In contrast, GraphPCA, STAGATE,

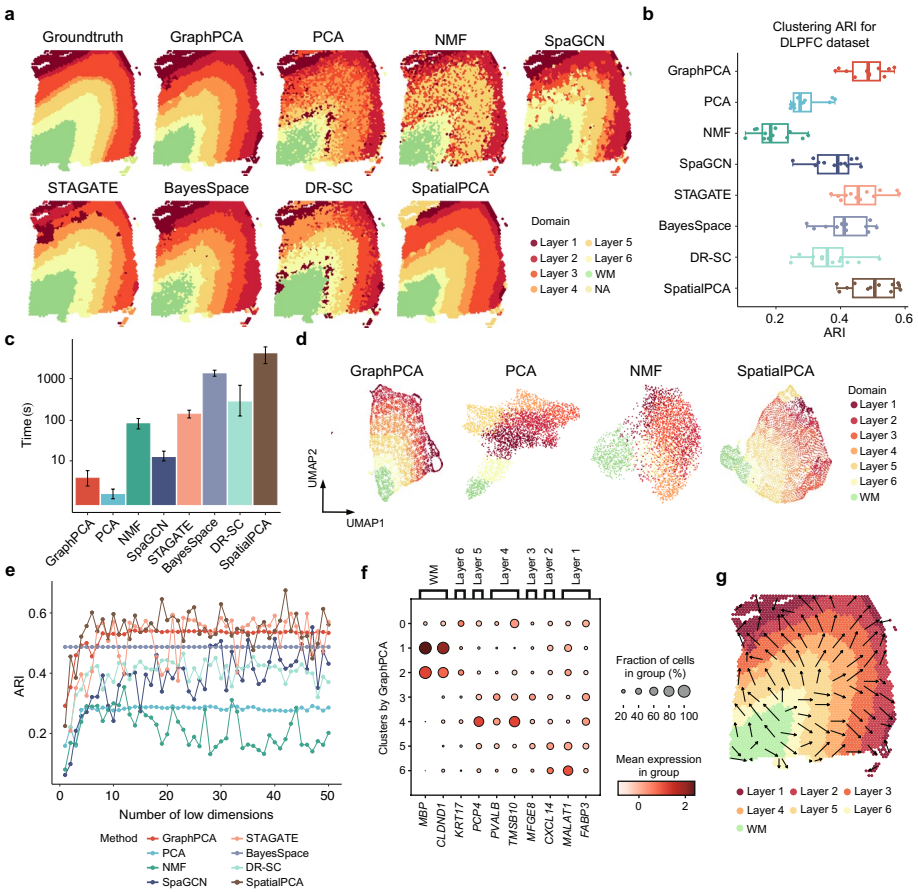

Fig. 2 Performance of GraphPCA on human dorsolateral prefrontal cortex (DLPFC) samples. **a** Clustering results of various methods on sample 151673 of the DLPFC dataset. The ground truth of tissue regions is the manual annotation of six cortex layers and white matter (WM), as provided by the original study. Manual annotations and clustering results of other DLPFC slices are shown in Additional file 1: Fig. S5. **b** Boxplots illustrating clustering accuracies across all 12 sections by eight methods. **c** Runtimes of different methods for spatial domain detection across all 12 sections. **d** UMAP visualizations of the sample 151673 generated by GraphPCA, PCA, NMF, and SpatialPCA embeddings, respectively. **e** Line plots showing the dynamic change in clustering accuracy at varying numbers of input low-dimensional components (*x*-axis) inferred by GraphPCA and other methods. **f** Scatter plot displaying the mean expression of layer-specific markers including *MBP* and *CLDND1* (white matter), *KRT17* (layer 6), *PCP4* (layer 5), *PVPLB* and *TMSB10* (layer 4), *MFGE8* (layer 3), *CXCL14* (layer 2), and *MALAT1* and *FABP3* (layer 1). Clusters in *y*-axis correspond to the spatial domains inferred by GraphPCA. **g** Trajectory inference on sample 151673 based on the inferred low-dimensional components of GraphPCA. Arrows point from tissue locations with low pseudo-time to those with high pseudo-time

and SpatialPCA effectively delineate cortical layers and accurately recover layer boundaries. GraphPCA achieves superior performance among compared methods, as indicated by higher ARI, NMI, and HS scores (0.536, 0.689, and 0.706, respectively), which nearly double those by space-unaware methods (PCA: ARI 0.290, NMI 0.452, HS 0.456; NMF: ARI 0.235, NMI 0.292, HS 0.287). Sensitivity analysis indicates that GraphPCA obtains stable spatial domain detection accuracy over varying numbers of SVGs for DLPFC data (Additional file 1: Fig. S4). While STAGATE (ARI 0.576) and SpatialPCA (ARI 0.561) obtain slightly higher ARI, they are significantly more time-consuming due to employing the iterative optimization procedures (Fig. 2c). More importantly, GraphPCA has the distinct advantage of providing an analytic solution, making it more interpretable and robust compared to deep learning-based algorithms like SpaGCN and STAGATE.

Similar results are also observed in other sections of the DLPFC data (Fig. 2b, Additional file 1: Fig. S5).

We then performed UMAP visualization of low-dimensional embeddings generated by each method. It is shown that different spatial domains inferred by GraphPCA are clearly separated, with their relative orders and global arrangements aligned well with histological images (Fig. 2d, Additional file 1: Fig. S6). This observation suggests that the GraphPCA approach effectively captures the local structure and relative positioning of the cells in the embedding space, which helps provide valuable insights into the spatial organization of the tissue. In contrast, spots associated with different layers are intermixed in UMAP by methods without explicitly considering spatial constraints (Fig. 2d, Additional file 1: Fig. S6). We next examined the impact of the number of principle components (PCs) on the clustering accuracy. Since GraphPCA provides a globally optimal solution without relying on iterative procedures and random initial values, it exhibits more stable clustering results across different numbers of PCs in terms of both ARI and McFadden-adjusted pseudo-R^2 (Fig. 2e, Additional file 1: Fig. S7). In contrast, NMF, SpaGCN, STAGATE, DR-SC, and SpatialPCA exhibit larger variabilities in clustering accuracies, indicating their high sensitivity to the embedding dimension and parameter initialization. Additionally, the seven spatial domains identified by GraphPCA are enriched with layer-specific marker genes reported previously [51] (Fig. 2f, Additional file 1: Fig. S8). We also performed gene set enrichment analysis (GSEA) to identify significant pathways that are highly enriched in each layer based on these layer-specific genes. The enriched pathways include neuron projection, cytoplasmic translation, and synapse-related, which represent common and crucial transcriptional programs underlying cortical development and maturation (Additional file 1: Fig. S9). Further analysis on the top PCs by GraphPCA reveals associations with spatial expression patterns of white matter, layer 3, and layer 4 (Additional file 1: Fig. S10a). The top 5 weighted genes by absolute values in each PC exhibit consistent spatial patterns, validating that the extracted patterns represent co-expression modules in corresponding regions (Additional file 1: Fig. S10b–d). Trajectory analysis by GraphPCA identifies an inward-outward spatial trajectory from white matter to cortical layers (Fig. 2g, Additional file 1: Fig. S11a), which agrees with previous studies that new neurons are generated in the ventricular zone and migrate outwards along radial glial fibers to integrate into existing layers [53, 54]. In contrast, the trajectory patterns produced by PCA, SpaGCN, STAGATE, and DR-SC appear to be random, failing to capture coordinated spatial-gene expression relationships (Additional file 1: Fig. S11b, c).

Mouse medial prefrontal cortex data by STARmap

Next, we applied GraphPCA to the high-resolution image-based ST data, specifically the mouse medial prefrontal cortex (mPFC) data generated by STARmap [16]. This dataset consists of 1049 cells measured across 166 genes, providing single-cell resolution for each spot. The mPFC is a vital cognitive control region that potentially involved in anxiety and fear regulation in human and mouse, and previous studies have divided into four laminar domains, i.e., L1, L2/3, L5, and L6 (Fig. 3a). We downloaded cell-type annotation of each spot from the original paper [16] and domain-level annotation in [55] as ground truth to evaluate clustering performance (Fig. 3b, g).

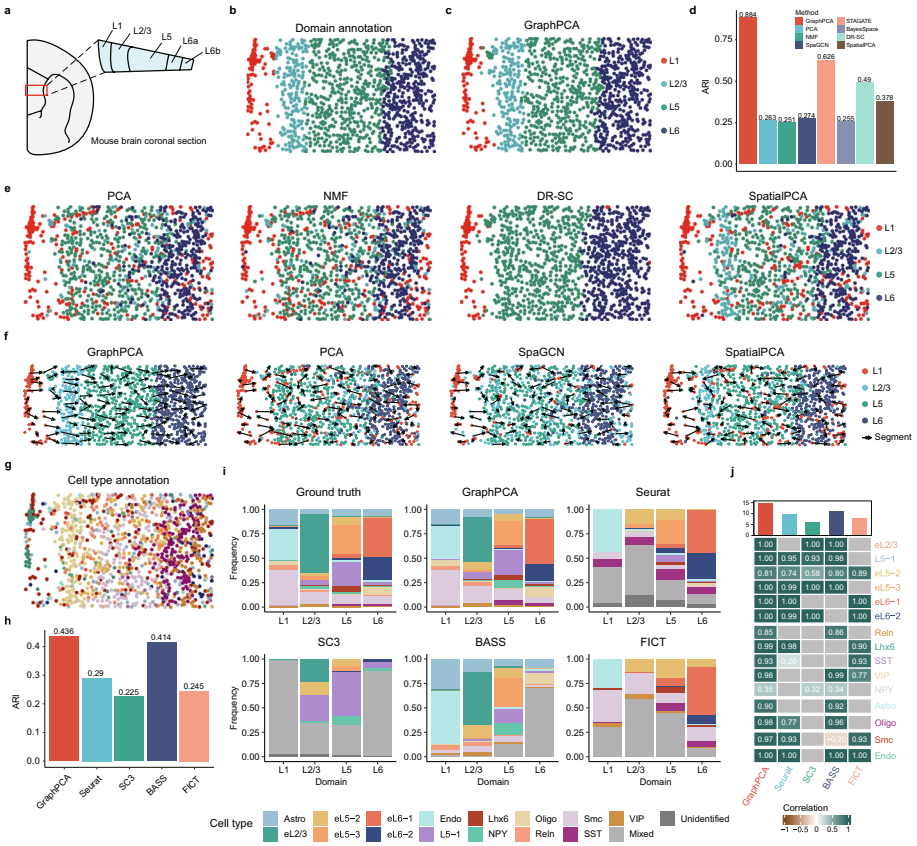

Fig. 3 Evaluation of GraphPCA on mouse medial prefrontal cortex (mPFC) data generated by STARmap. **a** An anatomic reference atlas displays the structure of the prelimbic area in the mouse prefrontal cortex. **b** Manual annotation provided by the original study. **c** Spatial domains detected by GraphPCA. **d** Barplots illustrating spatial domain detection ARI by different methods on the mPFC data. **e** Visualization of spatial domains identified by PCA, NMF, DR-SC, and SpatialPCA. The clustering results of SpaGCN, BayesSpace, and STAGATE are shown in Additional file 1: Fig. S12a. **f** Trajectories of the mPFC tissue based on low-dimensional embeddings obtained by GraphPCA, PCA, SpaGCN, and SpatialPCA. Trajectories are inferred by Slingshot under default parameters. **g** Annotated cell type labels for the mPFC data are available from the original study. **h** Barplots illustrating cell type clustering ARI by different methods on the mPFC data. **i** Spatial composition of cell types identified by different methods. Excitatory neurons: eL2/3, L5-1, eL5-2, eL5-3, eL6-1, and eL6-2; inhibitory neurons: Reln, VIP, SST, NPY, and Lhx6; Oligo: oligodendrocytes; Smc: smooth muscle cells; Astro: astrocytes; Endo: endothelia cells; Mixed: group of cells expressing marker genes associated with multiple cell types; and Unidentified: cells with no clear marker gene expression. **j** Cell type identification results by GraphPCA and other methods. Top: barplots displaying the number of cell types identified by each method. Bottom: heatmap showing correlations between estimated cell type proportions from each method and the ground truth

We first performed dimension reduction on the expression data and examined their ability to recover the spatial domain structure. As anticipated, GraphPCA accurately identifies four spatial domains, closely matching the ground truth and achieving the highest clustering scores (ARI: 0.886; NMI: 0.85; HS: 0.848) (Fig. 3c, d, Additional file 1: Fig. S12). The high performance of our method is consistent across different numbers of input low-dimensional components, as evidenced by ARI and McFadden-adjusted pseudo-R^2 metrics (Additional file 1: Fig. S12d). In contrast, clustering results reveal that PCA, NMF, SpaGCN, BayesSpace, and SpatialPCA struggle to effectively distinguish cortical layers L2/3 and L5, while incorrectly

identifying layer L1 as a discrete structure in space. Although DR-SC yields relatively compact clusters, it fails to identify L2/3 (Fig. 3e). STAGATE misidentified a few spots of layer L6 as layers L1 and L5 (Additional file 1: Fig. S12a). The precise layer detected by GraphPCA allows us to identify layer-specific marker genes. We then performed differential gene expression analysis and revealed known layer marker genes including *Bgn* (L1), *Cux2* (L2/3), *Tcerg1l* (L5), and *Pcp4* (L6) (Additional file 1: Fig. S13). Furthermore, GraphPCA takes the least runtime, even surpassing PCA, making it highly scalable for the incoming large-scale data (Additional file 1: Fig. S12c). Finally, trajectory inference using top PCs obtained from different dimension reduction methods indicates that GraphPCA correctly identifies a path from L1 to L6, aligned well with known cortical development, while other methods fail to do so (Fig. 3f, Additional file 1: Fig. S14).

We next performed cell type clustering on the same dataset, in which every cell is annotated as one of the 15 types including excitatory neurons: eL2/3, L5-1, eL5-2, eL5-3, eL6-1, and eL6-2; inhibitory neurons: Reln, VIP, SST, NPY, and Lhx6; Oligo: oligodendrocytes; Smc: smooth muscle cells; Astro: astrocytes; and Endo: endothelia cells (Fig. 3g). We compared GraphPCA with two popular cell type clustering methods for scRNA data (Seurat [36] and SC3 [56]) and two methods designed for single-cell resolution ST data (BASS [55] and FICT [57]). To obtain the clustering results in cellular level, we reduced the graph regularization strength of GraphPCA to increase the clustering resolution ($\lambda = 0.2$, parameter settings for different datasets and analysis tasks are listed as Additional file 1: Table S2). GraphPCA achieves the best clustering performance (ARI: 0.436), surpassing other methods (Seurat: 0.29; SC3: 0.225; BASS: 0.414; and FICT: 0.245) when the true number of cell types is specified in the subsequent K-means clustering (Fig. 3h). Similar conclusions can be drawn when evaluated by NMI and HS (Additional file 1: Fig. S15b).

Then, we analyzed the spatial distribution of cell types identified by various methods (Fig. 3i). GraphPCA successfully detects the known enriched cell types in their corresponding layers, such as Smc and Endo cells in L1, eL2/3 cells in L2/3, eL5-3 and L5-1 cells in L5, and eL6-1 and eL6-2 cells in L6. It accurately distinguishes the excitatory neurons at layer 6 (eL6) into two subtypes (eL6a and eL6b) and identifies inhibitory neurons VIP and its subtype Lhx6 with clear marker gene expression. Overall, GraphPCA identifies the highest number of correct cell types, while other methods produce multiple unidentified cell clusters without known marker genes, or mixed cell clusters with ambiguous marker gene expressions (Fig. 3j, Additional file 1: Figs. S16–23). In summary, the embeddings derived by GraphPCA effectively capture cell-identity differences at both domain and cell-type levels (Additional file 1: Figs. S24, 25).

We next extended the application to other NGS-based ST technologies with subcellular resolution. Based on the Slide-seq data of mouse cerebellum [21] and Slide-seqV2 data of mouse hippocampus [22], GraphPCA demonstrates superior clarity and continuity in representing the Purkinje layer in the cerebellum data (Additional file 1: Fig. S26). For the hippocampus data comprising expression of 23,264 genes in 53,208 spots, GraphPCA clearly represents the hippocampus, dentate gyrus, and

CA1 and CA3 layers, which are not clearly defined by other methods (Additional file 1: Fig. S27).

Murine liver data by 10X Visium

To validate the performance of GraphPCA on more complex tissues, we next applied it to the murine liver data generated by 10X Visium [58], which consists of 1293 cells measured across 31,053 genes. The murine liver is a heterogeneous tissue composed of hexagonal lobules that are radially polarized by blood flow and morphogens [59, 60]. As blood flows directionally toward the central vein, hepatocytes take up oxygen and nutrients, metabolize hormones, and create a gradient along the periportal-pericentral axis known as "liver zonation" [60–62] (Fig. 4a). According to the zonation annotation by Guilliams et al. [58], the murine liver lobule can be divided into four zones, i.e., central vein, mid zone, periportal zone, and portal vein (Fig. 4b).

We first performed dimension reduction and spatial domain detection using different methods on the murine liver data. GraphPCA accurately identifies each zone along the lobular axis in the correct sequential order: portal vein, periportal zone, mid zone, and central vein (Fig. 4c). However, other space-aware methods struggle to identify the portal vein within the periportal zone (Fig. 4d, Additional file 1: Fig.

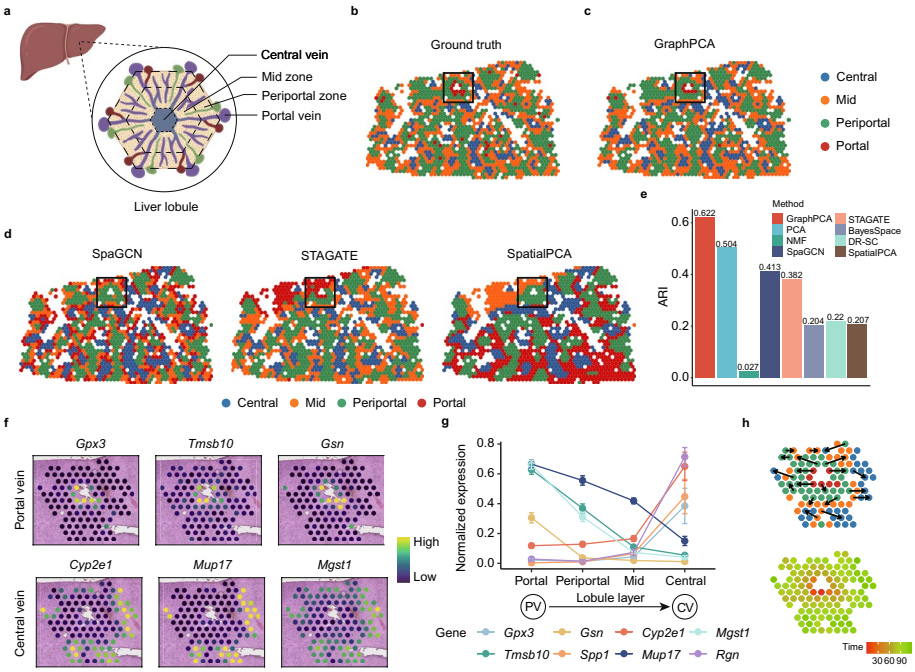

Fig. 4 Evaluation of GraphPCA on murine liver data generated by 10X Visium. **a** An anatomic reference atlas displays the structure of the murine liver lobule area. **b** Manual annotation provided by the original study. **c** Spatial domains detected by GraphPCA. **d** Visualization of spatial domains identified by SpaGCN, STAGATE, and SpatialPCA. The clustering results of PCA, NMF, BayesSpace, and DR-SC are shown in Additional file 1: Fig. S28a. **e** Barplots illustrating spatial domain detection ARI by different methods on the murine liver data. **f** Expression of zonation marker genes identified by GraphPCA in the ROI. **g** Line plots illustrating expression level of zonation marker genes identified by GraphPCA along the portal–central lobule axis (*x*-axis) in the ROI. Error bars indicate 95% confidence intervals across each layer spots. **h** Spatial trajectory and pseudo-time of the ROI based on low-dimensional embeddings obtained by GraphPCA. Trajectories are inferred by Slingshot under default parameters. Arrows point from tissue locations with low pseudo-time to those with high pseudo-time. Color represents different tissue regions

S28a). For instance, STAGATE and SpatialPCA misclassify the portal zone as part of the periportal zone, while BayesSpace produces overly smoothed results, leading to the emergence of two dominant clusters and failing to recognize the characteristic zones of the hepatic lobule. Overall, GraphPCA achieves the highest clustering performance (ARI: 0.622), which is at least 10% higher than those by other methods (PCA: 0.504; NMF: 0.027; SpaGCN: 0.413; STAGATE: 0.382; BayesSpace: 0.204; DR-SC: 0.22; and SpatialPCA: 0.207) (Fig. 4e).

Then, we selected the distinct zonal layers of the hepatic lobule in the tissue regions that were accurately detected by GraphPCA and performed differential expression analysis to identify zonation marker genes (Fig. 4f, Additional file 1: Fig. S29). The identified layers by GraphPCA are enriched with known marker genes (portal vein: *Gpx3*, *Tmsb10*, *Gsn*; central vein: *Cyp2e1*, *Mup17*, *Mgst1*) [59]. In particular, *Gpx3* is an antioxidant enzyme that plays a crucial role in protecting cells from oxidative stress by reducing hydrogen peroxide and lipid peroxides [63–65]. *Cyp2e1* is a member of the cytochrome P450 family involved in xenobiotic metabolism [66–68]. We noticed that *Cyp2e1*, *Mup17*, and *Mgst1* are not detectable near annotated portal veins, whereas *Cyp2f2*, *Alb*, and *Spp1* show high expression levels near the portal vein, with no signals in the central vein. Additionally, we visualized the spatial expression of zonation marker genes identified by GraphPCA for each zone (Fig. 4g), revealing a pronounced expression gradient along the lobular axis, consistent with a previous study [59].

We next performed GSEA to identify significant pathways that are significantly enriched in each layer based on these layer-specific genes (Additional file 1: Fig. S30). Pathway analysis of the portal vein marker genes reveals the strongest enrichment in genes associated with the process of "collagen-containing extracellular matrix." This pathway not only provides essential physical scaffolding for cellular constituents but also initiates crucial biochemical and biomechanical cues necessary for tissue morphogenesis, differentiation, and homeostasis [69, 70]. Pathways related to antigen processing and presentation, specifically the "MHC class II protein complex" and "multivesicular body," are also highly enriched within the portal vein, highlighting their role in influencing immune recognition and response [71]. In contrast, pathways "peroxisomes" and "microbodies" are associated with the central vein, underscoring their importance in maintaining the metabolic functions of hepatic tissue [72].

Finally, we performed trajectory inference by using top PCs obtained from different dimension reduction methods to identify one trajectory per method from portal vein to central vein (Fig. 4h, Additional file 1: Fig. S31). Specifically, GraphPCA captures the well-known spatial pattern of hepatocyte differentiation along the lobular axis [59, 73, 74]. As hepatocytes differentiate from the portal vein, they are gradually exposed to higher levels of oxygen, nutrients, and metabolites, creating a differentiation gradient that reflects the physiological maturation of hepatocytes in response to blood flow [60, 61, 74–76]. In contrast, the trajectories inferred from other space-aware methods appear to be random, with pseudo-time values intermingled across layers (Additional file 1: Fig. S31).

Taken together, the low-dimensional representation obtained from GraphPCA is accurate and robust in analyzing heterogeneous murine liver sample, enabling the exploration of transcriptional and functional heterogeneity across zones along the lobular axis between the portal and central veins.

Denoising gene expression profiles for deciphering gene spatial patterns

Leveraging the interpretability inherent in the PCA-based framework, GraphPCA is also capable of reconstructing the gene expression matrix from low-dimensional embeddings and projection vectors, thereby achieving denoising of the original expression data (see "Methods" for details). In this part, we aimed to unveil intrinsic spatial gene expression patterns in denoised DLPFC data using GraphPCA.

We conducted a comparative analysis involving GraphPCA and Sprod, state-of-the-art denoising methods for ST data. Specifically, we examined layer-marker genes of six cortical layers in sample 151673, as selected by Maynard et al. [51], and analyzed their spatial expression patterns before and after denoising by two methods (Fig. 5a). For both methods, the spatial expression of each layer-marker gene is revealed, with high expression regions matching the corresponding layers compared with raw data. Notably, in GraphPCA-denoised data, *LAMP5* and *NTNG2*, marker genes for temporal and visual cortices, respectively, exhibit differential expression between layers 5 and 6, which is obscured in the raw data. While Sprod also recovers these spatial patterns, gene expression signals are subtle, lacking sufficient differential expression between layers. We further assessed the quality

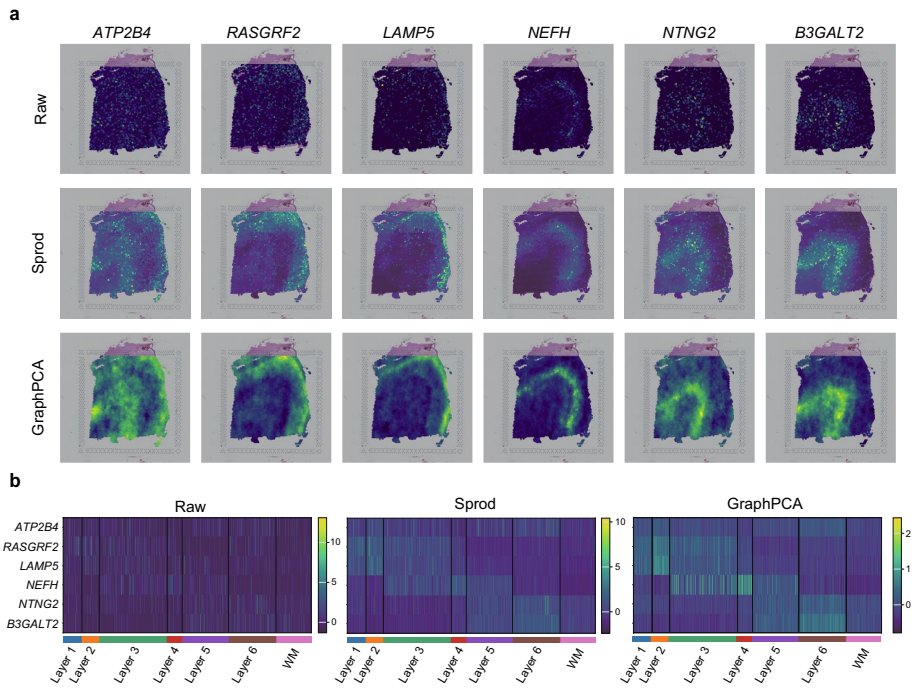

Fig. 5 GraphPCA deciphers gene spatial patterns of layer-specific marker genes in the DLPFC tissue. **a** Visualizations of raw and denoised spatial expressions of six layer-specific marker genes on sample 151673 of the DLPFC dataset. Benchmark methods include GraphPCA and state-of-the-art denoising tool Sprod. **b** Gene expression heatmaps of layer-specific marker genes on raw and denoised data generated by different methods

of the denoised data by two methods, as well as the raw data, measured by quantitative metrics including log fold change and log adjusted p values of known layer marker genes. GraphPCA achieves better performance compared to Sprod as indicated by higher log fold changes and negative log adjusted p values (Additional file 1: Fig. S32). In summary, GraphPCA improves the detection of spatial patterns and inter-layer expression differences of layer-marker genes (Fig. 5b), demonstrating its utility in denoising and elucidating intrinsic spatial gene expression patterns.

Integrating multiple ST datasets improves the performance of downstream analyses

As illustrated in previous sections, analysis based on single ST datasets can be inaccurate due to the high level of noise. In addition to denoising, another alternative strategy to enhance the signal is integrating multiple slices from the same sample or tissue type [55, 77]. To enable this function, we further designed GraphPCA_multi, a multi-sample extension of GraphPCA through aggregating multiple spatial neighborhood graphs and expression profiles into one matrix. The low-dimensional embedding per sample can be derived by optimizing the overall objective function across all samples (see "Methods" for details).

To validate the capability of our multi-sample integration model, we applied GraphPCA_multi to aggregate slice 151674 with three adjacent sections (slices 151673, 151675, 151676) and re-detected spatial domains using the joint embeddings. Clustering performance for slice 151674 significantly improves (ARI increased from 0.38 to 0.54), and the deteriorated pattern of layers 3 and 4 in single-slice estimation is recovered (Fig. 6a). Taking all samples, GraphPCA_multi outperforms single-sample analysis in terms of ARI (Fig. 6b) and surpasses state-of-the-art ST data integrating tools, BASSMult (a multi-sample version of BASS [55]) and STAligner [77] for most samples (Fig. 6c, Additional file 1: Fig. S33). In addition, GraphPCA_multi also achieves the lowest running time while exhibiting lower variance across DLPFC slices compared to BASSMult.

Next, we examined the performance of GraphPCA_multi in integrating samples from the same tissue type, not necessarily adjacent slides or from the same sample. To achieve this, besides the BZ5 sample used in pervious analyses, we obtained two additional tissue sections BZ9 and BZ14 generated by STARmap from different mice. Cells in all sections are annotated into four cortical layers (L1, L2/3, L5, and L6) according to the original publication [16]. As observed, GraphPCA_multi achieves superior spatial domain identification versus BASSMult and STAligner across all three sections. In detail, the four layers identified by GraphPCA_multi closely match the manual annotation in terms of tissue boundaries and thickness (Fig. 6d). In UMAP visualization using the top PCs from GraphPCA_multi, three tissue sections are evenly mixed (Fig. 6e), while different layers are clearly separated (Fig. 6f), indicating its potential in batch effect removal. Notably, this segregation accords with the functional similarity of adjacent layers and developmental trajectories. Similar to the DLPFC, the captured trajectory pattern reflected the developmental process of cortical layers, transitioning from existing layers to newer ones (from L1 to L6). In contrast, STAligner arranges cells in concentric clusters rather than matching organ developmental trajectory (Fig. 6f). These results validate the efficiency of GraphPCA_multi in integrating multiple samples to enable more accurate biological discovery.

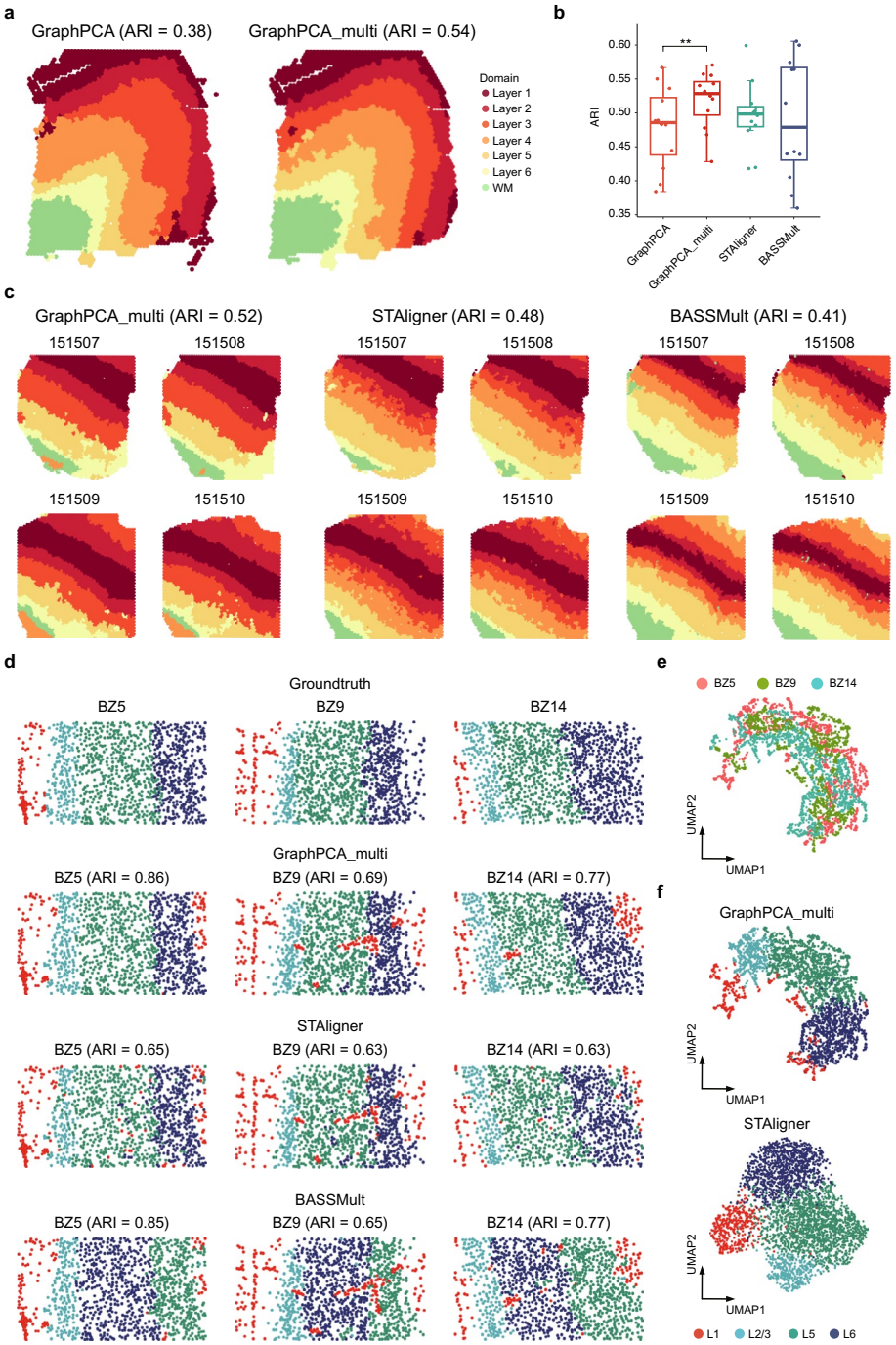

Fig. 6 Performance of GraphPCA in integration of multiple ST datasets. **a** Comparison of identified spatial domains by GraphPCA and GraphPCA_multi (the multi-sample version of GraphPCA) for sample 151674 of the DLPFC dataset. **b** Boxplot showing clustering accuracy of all 12 sections by GraphPCA, GraphPCA_multi, and two other multi-sample integration methods, i.e., STAligner and BASSMult. All three multi-sample methods take four neighboring sections for each section as input. Two stars (**) denote that the *p* value is less than 0.01 by paired one-sided *t*-test. **c** Inferred spatial domains on four slices (151507, 151508, 151509, and 151510) of DLPFC by GraphPCA_multi, STAligner, and BASSMult, respectively. **d** Inferred spatial domains on three sections (BZ5, BZ9, and BZ14) of the mPFC dataset by GraphPCA_multi, STAligner, and BASSMult, respectively. Pathological annotation by the original STARmap study is served as the ground truth. **e** UMAP visualization of low-dimensional embedding by GraphPCA_multi colored by three sections. **f** UMAP visualizations of low-dimensional embedding by GraphPCA_multi (top) and STAligner (bottom) colored by spatial domains

Discussion

In this study, we presented GraphPCA, a novel dimension-reduction method tailored for ST data by combining the strengths of graphical constraint and PCA-based framework. By leveraging the graphical constraint, GraphPCA ensures that projections of adjacent spots are closer in the low-dimensional space. Each embedding dimension is highly associated with specific spatial gene expression patterns, enabling the gene-component projection matrix to reflect the differential spatial distribution of co-expressed gene modules. Through extensive simulations and validations on real datasets, we demonstrated that the low-dimensional embeddings by GraphPCA can enhance performance of downstream analyses including spatial domain detection, trajectory inference, and denoising. The model flexibility of GraphPCA allows us to easily extend it for multi-sample integration, which further improves clustering accuracy by incorporating gene expression information from other slices. Importantly, the closed-form solution exploited in GraphPCA ensures great computational efficiency compared to deep learning-based approaches, madding it easily scalable to the escalating throughput of emerging spatial transcriptomic assays (e.g., Stereo-seq, Slide-seqV2) that involve tens of thousands of spots.

The hyperparameter λ in GraphPCA balances the relative importance of reconstruction error in dimensionality reduction and the spatial constraints, and thus a proper value of λ is critical. When $\lambda = 0$, GraphPCA reduces to the typical PCA; as λ increases, GraphPCA embeddings incorporate more spatial information, but excessively large λ may lead to less informative embeddings dominated by the spatial constraints. Therefore, we thus suggest that users try different values of λ to evaluate the consistency of the clustering results, and carefully check if the results are biologically plausible with background knowledge in downstream data analyses. In practical applications, it is also helpful to generate multiple embeddings across a range of λ and integrate all outcomes to improve robustness for downstream analyses. Other than spatial locations, the spot similarity network based on gene expression or H&E image has not been incorporated. In theory, we could impose different λ to adjust for other spatial-related constraints, in such case, our model still has a closed-form solution (see "Methods" for details). The preliminary result indicates that integrating both spatial location and gene expression data provides a modest clustering performance boost over the spatial position-only model on the mPFC data (Additional file 1: Fig. S34).

The construction of spot neighborhood graph is another pivotal factor determining the quality of the low-dimensional embedding. We primarily considered k-nearest neighbor (k NN) graphs based on the Euclidean distance, while the GraphPCA package also provides various other metrics to construct spatial neighborhood graph of spots/cells based on spatial coordinate information (e.g., circular-neighborhood graph (CG), Delaunay triangulation graph (DTG) [78], and shared nearest neighbor graph (SNN) [79], see "Methods" for details). For 10X Genomics Visium and Stereo-seq that exhibit grid-like spot arrangement patterns, these spatial graph construction methods can be made equivalent when setting tuning parameters to certain values. But for single-cell resolution ST techniques such as STARmap and MERFISH, different methods may produce different spatial neighborhood graphs and thus different low-dimensional embeddings. We also tested different spatial graph construction methods as well as their corresponding

hyperparameters on the performance of our method based on 10X Genomics Visium and single-cell resolution STARmap data (Additional file 1: Fig. S35). We could observe a clear improvement in clustering accuracies as the connectivity of spatial graphs increased (more neighbors in k NN and SNN and larger radius in circular-neighborhood graph), which suggests the important contribution of graphical constraints in enabling accurate spatial domain detection. However, excessively large spatial graphs can also be detrimental, as they may produce over-smoothed domain detection results (Additional file 1: Fig. S36).

Conclusions

In conclusion, we proposed an interpretable and quasi-linear dimension reduction algorithm, GraphPCA, that efficiently captures the biological signal from ST data. This scalable and extensible framework facilitates the learning of low-dimensional embeddings while integrating gene expression with spatial location, thereby enabling various downstream analyses. Through combining graph representation learning and the PCA framework, GraphPCA retains the advantage of linear embedding while capturing nonlinear spatial information underlying the tissue architecture. In the future, we plan to explore the adaptability of GraphPCA to other spatial techniques, such as the emergent spatially resolved proteomics, epigenomic, and metabolomic data, with the potential to unveil novel biological insights into the spatial orchestration of these crucial omics layers.

Methods

Data preprocessing

GraphPCA takes the gene expression matrix and spatial coordinate information from spatial transcriptomics data as input. 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, $i = 1, \cdots, n$, $j = 1, \cdots, m$. The spatial coordinates of individual spots are denoted as $S = (s_1, \ldots, s_n)^\intercal, s_i \in R^2/R^3$. Before performing GraphPCA, we assumed that the gene expression counts have already been preprocessed with analytic Pearson residuals proposed by Lause et al. [80] and further scaled for each gene to have zero mean and unit standard deviation. Then, we selected the top 3000 spatial variable genes using SPARK package.

Construction of spatial neighborhood graph

Under the assumption that spots in close spatial proximity often exhibit similar gene expression patterns, we constructed a sparse spatial neighborhood graph to capture the complex spatial interactions based on the spatial coordinate information S. By default, an undirected graph is built by employing the k-nearest neighbor (kNN) algorithm based on Euclidean distance, where k is the number of neighbors. Then, a sparse adjacent matrix A based on kNN graph can be calculated as:

$$A_{ij} = \begin{cases} 1, j \in N(i), \\ 0, otherwise, \end{cases}$$

where $N(i)$ represents the set of neighbors of spot i. To characterize distinct spatial distributions of spots across datasets, GraphPCA also accommodates other spatial

neighborhood graph construction approaches such as SNN graph, circular-neighborhood graph, and Delaunay triangulation [81].

Simulation design

We conducted a series of simulations to comprehensively evaluate the performance of GraphPCA and other dimension reduction and spatial domain detection algorithms. To achieve this, we downloaded mouse brain sagittal datasets from 10X Genomics and obtained the anatomical structure of the mouse brain sagittal from the Allen Brain Atlas as ground truth layer labels.

We utilized scDesign3, a multi-omics data simulator based on the generalized additive model for location, scale, and shape (GAMLSS), to generate the simulated gene expression matrix. scDesign3 models the joint distribution of multiple genes using the marginal distributions of individual genes, enabling precise capture of data properties and gene expression heterogeneity. In detail, we selected the top 3000 SVGs for each real dataset by SPARK and then generated simulated data by scDesign3 using the filtered gene expression matrix and layer labels as inputs, with the Gaussian process smoother parameter K = 300. Based on the simulated data, we established four scenarios to evaluate the robustness of GraphPCA and other dimension reduction algorithms with varying sequencing depths, noise levels, spot sparsity, and expression dropout rates. Spatial domain detection performance of each method was then evaluated using adjusted Rand index (ARI), normalized mutual information (NMI), and homogeneity score (HS).

GraphPCA model

The GraphPCA model aims to reduce the dimensionality of the gene expression data X while incorporating graph regularization constraints on the low-dimensional embedding. We first recalled the classical PCA, which strives to find a k-dimensional projections $Z \in R^{n\times k}$, which minimizes the reconstruction error of gene expression matrix X. The variational formulation of PCA is given as follows:

$$\min_{Z,W} ||X - ZW^{\intercal}||_2^2, s.t. W^{\intercal}W = I, \tag{1}$$

where W is an orthonormal eigenvector matrix. This is an optimization problem that can be addressed via the Lagrange multiplier method. The analytical solution to the optimization problem is $Z = XW$, where W represents the top k eigenvectors of $X^{\intercal}X$, where k is the dimensionality of Z. To ensure the low-dimensional embeddings retain spot location information, a natural idea is to impose graph-based constraints so that physically proximate spots have similar projections. Therefore, we considered the following objective function:

$$\min_{Z,W} ||X - ZW^{\intercal}||_2^2 + \frac{1}{2}\lambda \sum\nolimits_{\{i,j\in A\}} A_{ij}||Z_i - Z_j||_2^2, s.t. W^{\intercal}W = I, \tag{2}$$

where λ is a hyperparameter that balances reconstruction error and the smoothness of the projections over the graph. When $\lambda = 0$, the GraphPCA degenerates to classical PCA. We noted that the first term along with the constraint, corresponds to the objective of the standard PCA, and the second term is a graph regularization that encourages

nearby spot pairs in the graph to have similar projections. The above objective function can be rewritten as

$$\min_{Z,W} ||X - ZW^\intercal||_2^2 + \lambda tr(Z^\intercal LZ), s.t. W^\intercal W = I, \tag{3}$$

where $L = D - A$ is the Laplacian matrix of the spatial neighborhood graph, and D is the diagonal matrix of graph A. Similar to PCA, the objective function is also non-convex, but we could still derive an analytical solution via the Lagrange multiplier method to solve the Z and W (see Additional file 1: Supplementary Material for details). Let J denote the objective function as given in Eq. (3). We can derive the optimal solution by setting the derivative of J with respect to Z to be zero. Since.

$$\frac{\partial J}{\partial Z} = -2XW + 2Z + 2\lambda LZ$$

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

$$\min_{W} ||X - KXWW^\intercal||_2^2 + \lambda tr(W^\intercal X^\intercal K^\intercal LKXW), s.t. W^\intercal W = I. \tag{4}$$

Equation (4) is equivalent to

$$\min_{W} tr\left(-W^\intercal X^\intercal KXW\right), s.t. W^\intercal W = I.$$

It is easy to show that the matrix $X^\intercal KX$ is symmetric and positive semi-definite. The optimal solution W^* of Eq. (4) is the combination of eigenvectors, associated with the largest k eigenvalues of the graph-revised covariance matrix $X^\intercal KX$. X corresponding to the top k eigenvalues. It is noteworthy that the solution to GraphPCA can be interpreted as a ridge regularization of the classical PCA, regularized by the Laplacian matrix L. Thanks to this closed-form solution, GraphPCA is easy to implement and generates low-dimensional representations in linear time, enhancing integration performance into downstream analyses. As GraphPCA essentially imposes a graph prior on projections, it can be extended to dimensionality reduction tasks for any data with an underlying graph structure.

GraphPCA_multi model

We also developed a multi-sample integration extension called GraphPCA_multi. Suppose a spatial transcriptomics study measures L tissue slices. We followed the same preprocessing procedures as GraphPCA, taking the intersection of the SVGs of all the input slices. For each slice $l = 1, \cdots, L$, we constructed a spatial neighborhood graph A_l, with $l = 1, \cdots, L$, respectively. Then, we assembled these graphs into a large block diagonal matrix

$$A_{multi} = diag(A_1, A_2, \cdots, A_L),$$

where A_{multi} represents the integrated spot neighborhood graph. Then, we concatenated the gene expression matrices from all slices to form the input to GraphPCA_multi. Therefore, GraphPCA_multi imposes within-sample smoothness constraints through the block diagonal Laplacian matrix while identifying shared and distinct latent patterns across the multi-slice dataset.

Denoising process

The core principle of the PCA-based framework for denoising lies in the observation that noise typically resides in the lower eigenvalues, while the signal is concentrated in the higher eigenvalues. In the context of GraphPCA, the incorporation of spatial information allows for an enhanced embedding that retains the structural characteristics of the data. This embedding can effectively reconstruct the original gene expression matrix, thereby achieving denoising. In detail, the denoised gene expression matrix $\widetilde{X}$ is computed as the product of the embedding matrix Z and the projection matrix W. This approach not only filters out noise but also preserves significant biological signals inherent in ST data.

Downstream analyses

Clustering

We employed the K-means clustering algorithm to detecting spatial domains based on low-dimensional embedding obtained from GraphPCA, PCA, and NMF. Other methods used default embedded clustering algorithms, e.g., SpatialPCA uses a Walktrap algorithm and Louvain algorithm. For fair comparison, the true number of spatial domains was provided to all algorithms. We noted the radius hyperparameter r in DR-SC, defining the neighborhood radius, strongly influences embedding and clustering quality. Therefore, we generated a range of DR-SC embeddings (r from 5 to 100) and chose the best performing for comparisons as the oracle DR-SC version.

Differential gene expression analysis and spatial domain annotation

We employed the Wilcoxon test implemented in the sc.tl.rank_genes_groups function of the Scanpy package to identify differentially expressed genes for each spatial domain with a 1% FDR threshold (Benjamin-Hochberg adjustment). Then, spatial domains are annotated by marker genes and comparing expression spatial patterns against manual annotations.

Trajectory inference

After obtaining the clustering labels, we employed the Slingshot algorithm, the trajectory inference algorithm in scRNA data analysis [82], on the low-dimensional

embeddings to depict the spatial trajectory among locations on the tissue, setting white matter and layer 1 as start clusters for DLPFC and mPFC, respectively.

Gene enrichment analysis

We performed gene set enrichment analysis (GSEA [83]) on the top 100 differentially expressed genes sorted by adjusted p values using enrichGO function in the clusterProfiler [84] package, showing the top 30 enriched pathways. Gene sets are downloaded from the Molecular Signatures Database (MSigDB [85, 86], Broad Institute) including C2 (KEGG [87]) and C5 (GO BP: biological process, GO CC: cellular component, GO MF: molecular function).

Clustering performance evaluation metrics

We adopted four different metrics, i.e., adjusted Rand index (ARI), normalized mutual information (NMI), homogeneity score (HS), and McFadden-adjusted pseudo-R^2 [88], to measure clustering performance of different methods in spatial domain detection. Specifically,

(1) Adjusted Rand index (ARI)

Given two sets of clustering labels, the ARI is calculated as:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left(\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right)/\binom{n}{2}}{\frac{1}{2}\left(\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right) - \left(\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right)/\binom{n}{2}},$$

where n_{ij} is the number of spots overlapped by cluster i and cluster j. n_i and n_j are the number of spots in cluster i and j, respectively.

(2) Normalized mutual information (NMI)

NMI is calculated as:

$$NMI = \frac{\sum_{i,j} p_{ij} log \frac{p_{ij}}{p_i p_j}}{\left(-\sum_i p_i log p_i - \sum_j p_j log p_j\right)/2},$$

where $p_{ij} = \frac{n_{ij}}{n}$, $p_i = \frac{n_i}{n}$, and $p_j = \frac{n_j}{n}$.

(3) Homogeneity score (HS)

The homogeneity score measures the homogeneity or purity within each cluster of the clustering results, calculated as:

$$h = 1 - \frac{H(C|K)}{H(C)},$$

where C represents the ground truth. K represents the cluster labels predicted by the algorithm. $H(C|K)$ is the conditional entropy of the class distribution given cluster labels averaged over clusters, weighted by cluster size:

$$H(C|K) = -\sum_{C=1}^{|C|}\sum_{K=1}^{|K|}\frac{n_{c,k}}{n}log(\frac{n_{c,k}}{n_k}),$$

where $n_{c,k}$ is the size of cluster c spots assigned to cluster k, n_k is the size of spots in cluster k, and n is the total size of spots. $H(C)$ is the entropy of the class distribution:

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}log(\frac{n_c}{n}).$$

The homogeneity score ranges from 0 to 1, where HS = 1 indicates complete homogeneity, that each cluster only contains spots of a single class.

(4) McFadden-adjusted pseudo-R^2

The McFadden-adjusted pseudo-R^2 first constructs a multinomial regression model using low-dimensional features as predictors and the true spatial domains as the outcome, and then calculates the log-likelihood ratios between the regression model and the null model, where no predictors are used. Within the regression model, we calculated the McFadden-adjusted pseudo-R^2 to assess the predictive capacity of the predictor variables in projecting the ground truth. A higher pseudo-R^2 indicates that the method can effectively extract informative output for predicting the true spatial domains.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03429-x.

Additional file 1: Supplementary material describing the GraphPCA algorithm. Supplementary figures and tables.

Acknowledgements
We acknowledge the Bioinformatics Core in Center for Single-Cell Omics (CSCOmics), Shanghai Jiao Tong University School of Medicine, for providing the bioinformatics and high-performance computing services. L.L. is also affiliated with the Shanghai Artificial Intelligence Laboratory.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
J.Y. and X.Z. designed the study and conceived the dimension reduction algorithm. J.Y. and L.W. conducted and implemented experimental analyses with the guidance of X.Z.. X.Z. and L.L. supervised the work. J.Y., L.W., L.L., and X.Z. participated in writing the manuscript. All the authors have read and agreed to the submitted version of the manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (62372286 to X.Z., 12471274 and 12090024 to L.L.), Science and Technology Innovation Plan of Shanghai (23JC1403200 to X.Z.), Shanghai Science and Technology Commission Grant (21JC1402900, 2021SHZDZX0102 to L.L.), and Key Laboratory of Data Science and Intelligence Education (Hainan Normal University), Ministry of Education (DSIE202002 to X.Z.).

Data availability
GraphPCA is implemented as a Python package publicly available at Github [89] and Zenodo [90]. The source code is released under the MIT License. All analysis codes [89] for reproducing the results of the study are publicly available at https://github.com/YANG-ERA/GraphPCA. The human DLPFC dataset measured by 10X Visium is available at http://research.libd.org/spatialLIBD/ [91]; The STARmap dataset for mouse medial prefrontal cortex dataset is available at https://www.starmapresources.org/data [92]; The mouse sagittal posterior and anterior brain data measured by 10X Visium are available at https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior [93] and https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Anter

ior [94], respectively. The Slide-seq data are available at Broad Institute's single-cell repository (https://singlecell.broadinstitute.org/single_cell/) with ID SCP354 [95]. The Slide-seq V2 data are available at Broad Institute's single-cell repository with ID SCP815 [96]. Another version about Slide-seq V2 data, which is the pre-processed subset of the data that provides cell-type level annotation, is available in the Squidpy package [97]. The murine liver data measured by 10X Visium are obtained from the Liver Cell Atlas (https://www.livercellatlas.org/) [98]. The simulated data generated in this study are available at https://github.com/YANG-ERA/GraphPCA [89].

Declarations

Ethics approval and consent to participate
No ethical approval was required for this study. All utilized public datasets were generated by other organizations that obtained ethical approval.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 May 2024 Accepted: 29 October 2024
Published online: 07 November 2024

References

1. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome Med. 2022;14:68.
2. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods. 2022;19:534–46.
3. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021;596:211–20.
4. Palla G, Fischer DS, Regev A, Theis FJ. Spatial components of molecular tissue biology. Nat Biotechnol. 2022;40:308–18.
5. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241.
6. Durif G, Modolo L, Mold JE, Lambert-Lacroix S, Picard F. Probabilistic count matrix factorization for single cell expression data analysis. Bioinformatics. 2019;35:4011–9.
7. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9:284.
8. Kharchenko PV. The triumphs and limitations of computational methods for scRNA-seq. Nat Methods. 2021;18:723–32.
9. Tian L, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. Nat Biotechnol. 2023;41:773–82.
10. Shah S, Lubeck E, Zhou W, Cai L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron. 2016;92:342–57.
11. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods. 2014;11:360–1.
12. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348: aaa6090.
13. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc Natl Acad Sci U S A. 2016;113:11046–51.
14. Moffitt JR, Hao J, Bambah-Mukku D, Lu T, Dulac C, Zhuang X. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proc Natl Acad Sci U S A. 2016;113:14456–61.
15. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, Linnarsson S. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods. 2018;15:932–5.
16. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361: eaat5691.
17. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat Protoc. 2015;10:442–58.
18. Zhang L, Chen D, Song D, Liu X, Zhang Y, Xu X, Wang X. Clinical and translational values of spatial transcriptomics. Signal Transduction Targeted Ther. 2022;7:111.
19. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.
20. Rao N, Clark S, Habern O. Bridging genomics and tissue pathology. Genet Eng Biotechnol News. 2020;40:50–1.
21. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363:1463–7.
22. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, Chen F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat Biotechnol. 2021;39:313–9.

23. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, Qiu X, Yang J, Xu J, Hao S, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell. 2022;185:1777-1792.e1721.
24. Wang Y, Song B, Wang S, Chen M, Xie Y, Xiao G, Wang L, Wang T. Sprod for de-noising spatially resolved transcriptomics data based on position and image information. Nat Methods. 2022;19:950–8.
25. Liu Y, Wang T, Duggan B, Sharpnack M, Huang K, Zhang J, Ye X, Johnson TS. SPCS: a spatial and pattern combined smoothing method for spatial transcriptomic expression. Brief Bioinform. 2022;23: bbac116.
26. Yang L, Liu J, Lu Q, Riggs AD, Wu X. SAIC: an iterative clustering approach for analysis of single cell RNA-seq data. BMC Genomics. 2017;18:689.
27. 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;17:144.
28. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18:59.
29. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19:477.
30. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.
31. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, Dudoit S, Clement L. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun. 2020;11:1201.
32. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
33. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, Susztak K, Reilly MP, Hu G, Li M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11:2338.
34. Gan D, Li J. SCIBER: a simple method for removing batch effects from single-cell RNA-sequencing data. Bioinformatics. 2023;39: btac819.
35. Yu X, Xu X, Zhang J, Li X. Batch alignment of single-cell transcriptomics data using deep metric learning. Nat Commun. 2023;14:960.
36. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177:1888-1902.e1821.
37. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
38. Bergenstråhle J, Larsson L, Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics. 2020;21:482.
39. Shang L, Zhou X. Spatially aware dimension reduction for spatial transcriptomics. Nat Commun. 2022;13:7203.
40. Liu W, Liao X, Yang Y, Lin H, Yeong J, Zhou X, Shi X, Liu J. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. Nucleic Acids Res. 2022;50:e72–e72.
41. Xu H, Fu H, Long Y, Ang KS, Sethi R, Chong K, Li M, Uddamvathanak R, Lee HK, Ling J, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Med. 2024;16:12.
42. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021;18:1342–51.
43. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun. 2022;13:1739.
44. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SEB, Nghiem P, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol. 2021;39:1375–84.
45. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016;4:218.
46. Song D, Wang Q, Yan G, Liu T, Sun T, Li JJ. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat Biotechnol. 2024;42:247–52.
47. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445:168–76.
48. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
49. Strehl A, Ghosh J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3:583–617.
50. Priness I, Maimon O, Ben-Gal I. Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics. 2007;8:111.
51. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24:425–36.
52. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.
53. Gilmore EC, Herrup K. Cortical development: layers of complexity. Curr Biol. 1997;7:R231–4.
54. Larsen DD, Callaway EM. Development of layer-specific axonal arborizations in mouse primary somatosensory cortex. J Comp Neurol. 2006;494:398–414.
55. Li Z, Zhou X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. Genome Biol. 2022;23:168.
56. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6.
57. Teng H, Yuan Y, Bar-Joseph Z. Clustering spatial transcriptomics data. Bioinformatics. 2022;38:997–1004.
58. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, Bujko A, Martens L, Thone T, Browaeys R, et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell. 2022;185(379–396): e338.
59. Halpern KB, Shenhav R, Matcovitch-Natan O, Toth B, Lemze D, Golan M, Massasa EE, Baydatch S, Landen S, Moor AE, et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature. 2017;542:352–6.

60. Gebhardt R, Matz-Soja M. Liver zonation: novel aspects of its regulation and its impact on homeostasis. World J Gastroenterol. 2014;20:8491.
61. Cunningham RP, Porat-Shliom N. Liver zonation - revisiting old questions with new technologies. Front Physiol. 2021;12:732929.
62. Jungermann K, Keitzmann T. Zonation of parenchymal and nonparenchymal metabolism in liver. Annu Rev Nutr. 1996;16:179–203.
63. Ighodaro O, Akinloye O. First line defence antioxidants-superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPX): their fundamental role in the entire antioxidant defence grid. Alexandria J Med. 2018;54:287–93.
64. Pei J, Pan X, Wei G, Hua Y. Research progress of glutathione peroxidase family (GPX) in redoxidation. Front Pharmacol. 2023;14: 1147414.
65. Brigelius-Flohé R, Maiorino M. Glutathione peroxidases. Biochim Biophys Acta Gen Subj. 2013;1830:3289–303.
66. Guengerich FP. Cytochromes P450, drugs, and diseases. Mol Interv. 2003;3: 194.
67. Nagata K, Martin BM, Gillette JR, Sasame HA. Isozymes of cytochrome P-450 that metabolize naphthalene in liver and lung of untreated mice. Drug Metab Disposition. 1990;18:557–64.
68. Kang JS, Wanibuchi H, Morimura K, Wongpoomchai R, Chusiri Y, Gonzalez FJ, Fukushima S. Role of CYP2E1 in thioacetamide-induced mouse hepatotoxicity. Toxicol Appl Pharmacol. 2008;228:295–300.
69. Chapman G, Eagles D. Ultrastructural features of Glisson's capsule and the overlying mesothelium in rat, monkey and pike liver. Tissue Cell. 2007;39:343–51.
70. Malhotra V, Erlmann P. The pathway of collagen secretion. Annu Rev Cell Dev Biol. 2015;31:109–24.
71. Neefjes J, Jongsma ML, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11:823–36.
72. Novikoff AB, Goldfischer S. Visualization of peroxisomes (microbodies) and mitochondria with diaminobenzidine. J Histochem Cytochem. 1969;17:675–80.
73. Halpern KB, Shenhav R, Massalha H, Toth B, Egozi A, Massasa EE, Medgalia C, David E, Giladi A, Moor AE. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nat Biotechnol. 2018;36:962–70.
74. Braeuning A, Ittrich C, Köhle C, Hailfinger S, Bonin M, Buchmann A, Schwarz M. Differential gene expression in periportal and perivenous mouse hepatocytes. FEBS J. 2006;273:5051–61.
75. Wesley BT, Ross ADB, Muraro D, Miao Z, Saxton S, Tomaz RA, Morell CM, Ridley K, Zacharis ED, Petrus-Reurer S, et al. Single-cell atlas of human liver development reveals pathways directing hepatic cell fates. Nat Cell Biol. 2022;24:1487–98.
76. Wu B, Shentu X, Nan H, Guo P, Hao S, Xu J, Shangguan S, Cui L, Cen J, Deng Q, et al. A spatiotemporal atlas of cholestatic injury and repair in mice. Nat Genet. 2024;56:938–52.
77. Zhou X, Dong K, Zhang S. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. Nat Comput Sci. 2023;3:894–906.
78. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19:171–8.
79. Ertoz L, Steinbach M, Kumar V. A new shared nearest neighbor clustering algorithm and its applications. In: Workshop on Clustering High Dimensional Data and Its Applications at 2nd SIAM International Conference on Data Mining, Arlington, VA, USA. United States: Society for Industrial and Applied Mathematics; 2002.
80. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. 2021;22:258.
81. Lee DT, Schachter BJ. Two algorithms for constructing a Delaunay triangulation. Int J Comput Inform Sci. 1980;9:219–42.
82. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37:547–54.
83. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. the Mouse Genome Database G: Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 2019;47:D801–6.
84. Yu G, Wang L-G, Han Y. He Q-Y: clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. J Integrative Biol. 2012;16:284–7.
85. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov Jill P, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1:417–25.
86. Liberzon A. A description of the Molecular Signatures Database (MSigDB) Web site. Methods Mol Biol. 2014;1150:153–60.
87. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
88. Zeng H, Shen Elaine H, Hohmann John G, Oh Seung W, Bernard A, Royall Joshua J, Glattfelder Katie J, Sunkin Susan M, Morris John A, Guillozet-Bongaarts Angela L, et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. Cell. 2012;149:483–96.
89. Yang J, Wang L, Liu L, Zheng X. GraphPCA: a fast and interpretable dimension reduction algorithm for spatial transcriptomics data. Github; 2024. https://github.com/YANG-ERA/GraphPCA.
90. Yang J, Wang L, Liu L, Zheng X. GraphPCA: a fast and interpretable dimension reduction algorithm for spatial transcriptomics data. Zenodo; 2024. https://zenodo.org/records/13372627.
91. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. DLPFC dataset by 10X Visium. spatialLIBD; 2021. http://research.libd.org/spatialLIBD/.
92. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. mPFC dataset by STARmap. STARmap Resources; 2018. https://www.starmapresources.org/data.
93. Dataset. Mouse posterior brain dataset by 10X Visium. 10X Genomics Visium; 2020. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior.
94. Dataset. Mouse anterior brain dataset by 10X Visium. 10X Genomics Visium; 2020. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Anterior.
95. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Mouse cerebellum dataset by Slide-seq. Single Cell Portal; 2019. https://singlecell.broadinstitute.org/single_cell/study/SCP354/.

96. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, Chen F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Mouse hippocampus dataset by Slide-seq V2. Single Cell Portal; 2020. https://singlecell.broadinstitute.org/single_cell/study/SCP815/.
97. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, et al. Squidpy: a scalable framework for spatial omics analysis. Mouse hippocampus dataset by Slide-seq V2. Squipy python pacakage; 2022. https://squidpy.readthedocs.io/en/stable/notebooks/tutorials/tutorial_slideseqv2.html.
98. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, Bujko A, Martens L, Thone T, Browaeys R, et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Murine liver dataset by 10X Visium. Liver Cell Altas; 2022. https://www.livercellatlas.org/.

Publisher's Note