ARTICLES
https://doi.org/10.1038/s41592-021-01255-8

nature methods

SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network

Jian Hu[1]✉, Xiangjie Li[2], Kyle Coleman[1], Amelia Schroeder[1], Nan Ma[3], David J. Irwin[4], Edward B. Lee[5], Russell T. Shinohara[1] and Mingyao Li[1]✉

Recent advances in spatially resolved transcriptomics (SRT) technologies have enabled comprehensive characterization of gene expression patterns in the context of tissue microenvironment. To elucidate spatial gene expression variation, we present SpaGCN, a graph convolutional network approach that integrates gene expression, spatial location and histology in SRT data analysis. Through graph convolution, SpaGCN aggregates gene expression of each spot from its neighboring spots, which enables the identification of spatial domains with coherent expression and histology. The subsequent domain guided differential expression (DE) analysis then detects genes with enriched expression patterns in the identified domains. Analyzing seven SRT datasets using SpaGCN, we show it can detect genes with much more enriched spatial expression patterns than competing methods. Furthermore, genes detected by SpaGCN are transferrable and can be utilized to study spatial variation of gene expression in other datasets. SpaGCN is computationally fast, platform independent, making it a desirable tool for diverse SRT studies.

Recent technological advances in SRT have enabled gene expression profiling with spatial information in tissues[1]. Knowledge of the relative locations of different cells in a tissue is critical for understanding disease pathology because spatial information helps in understanding how the gene expression of a cell is influenced by its surrounding environment. Popular experimental methods for SRT can be broadly classified into two categories. The first category is in situ hybridization or sequencing-based technologies with single-cell resolution, which includes seqFISH[2,3], seqFISH+[4], MERFISH[5,6], STARmap[7] and FISSEQ[8] that measure the expression level for hundreds to thousands of genes in cells within their tissue context. The second category is in situ capturing-based technologies with spatial barcoding followed by sequencing, which includes spatial transcriptomics (ST)[9], SLIDE-seq[10], SLIDE-seqV2 (ref. [11]), HDST[12] and 10x Visium that measure the expression level for thousands of genes in captured locations, referred to as spots. These different SRT technologies have made it possible to uncover the complex transcriptional architecture of heterogeneous tissues and enhanced our understanding of cellular mechanisms in diseases[13,14].

In SRT studies, an important step is identifying spatial domains defined as regions that are spatially coherent in both gene expression and histology. Traditional clustering methods such as *K*-means and Louvain's method[15] only take gene expression data as input, and the resulting clusters may not be contiguous due to the lack of consideration of spatial information and histology. To account for spatial dependency of gene expression, new methods have been developed. For example, Zhu et al.[16] uses a Hidden-Markov random field (HMRF) approach to model spatial dependency of gene expression; stLearn[17] uses features extracted from histology image as well as expression of neighboring spots spatially to normalize gene expression data before clustering; BayesSpace[18] employs a Bayesian approach for clustering by imposing a prior that gives higher weight to physically close spots. Although these methods can cluster spots or cells into distinct groups, the lack of flexibility with different modalities has made them less versatile. As newer SRT technologies continue to be developed[19–22], it is desirable to have methods that are compatible with different SRT platforms.

To link spatial domains with biological functions, it is crucial to identify genes that show enriched expression in the identified domains. Methods such as Trendsceek[23], SpatialDE[24] and SPARK[25] have been developed to detect spatially variable genes (SVGs). These methods examine each gene independently and return a *P* value to represent the spatial variability of a gene. However, due to the lack of consideration of spatial domains, genes detected by these methods do not have guaranteed spatial expression patterns, making it difficult to utilize these genes for further biological investigations.

Rather than considering spatial domain and SVG identification as separate problems, we developed SpaGCN, a graph convolutional network (GCN)-based approach that considers these two problems jointly. SpaGCN first identifies spatial domains by integrating gene expression, spatial location and histology through the construction of an undirected weighted graph that represents the spatial dependency of the data. For each spatial domain, SpaGCN then detects

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]School of Statistics and Data Science, Nankai University, Tianjin, China. [3]Weitzman School of Design, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [5]Translational Neuropathology Research Laboratory, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
✉e-mail: jianhu@pennmedicine.upenn.edu; mingyao@pennmedicine.upenn.edu

SVGs that are enriched in the domain. By restricting the search space to spatial domains, the SVGs detected by SpaGCN are guaranteed to have spatial expression patterns. The spatial domains and the corresponding SVGs provide a comprehensive picture of the spatial gradients in gene expression in tissue. SpaGCN is versatile in analyzing many types of SRT data, including ST, 10x Visium, SLIDE-seqV2, STARmap, and MERFISH.

Results

Overview of SpaGCN and evaluation. We explain the workflow of SpaGCN using in situ capturing-based SRT data as an example, but the method can be easily modified to analyze other types of SRT data. As shown in Fig. 1a, SpaGCN first builds a graph to represent the relationship of all spots considering both spatial location and histology information. Next, SpaGCN utilizes a graph convolutional layer to aggregate gene expression information from neighboring spots. Then, SpaGCN uses the aggregated expression matrix to cluster spots using an unsupervised iterative clustering algorithm[26]. Each cluster is considered as a spatial domain from which SpaGCN then detects SVGs that are enriched in a domain by DE analysis (Fig. 1b). When a single gene cannot mark the expression pattern of a domain, SpaGCN will construct a meta gene, formed by the combination of multiple genes, to represent the expression pattern of the domain.

To showcase the strength of SpaGCN, we applied it to seven publicly available datasets (Supplementary Table 1). The spatial domains identified by SpaGCN agree better with known tissue structures than Louvain, stLearn, and BayesSpace. We also compared SVGs detected by SpaGCN with those detected by SpatialDE and SPARK, and found that the SpaGCN-detected SVGs have more coherent expression patterns and better biological interpretability than the other two methods. The specificity of spatial expression patterns revealed by SpaGCN-detected SVGs were further confirmed by Moran's *I* and Geary's *C* statistics[27], two commonly used metrics for quantifying spatial autocorrelation of gene expression[28,29].

Application to human primary pancreatic cancer ST data. To demonstrate the importance of incorporating histology information, we analyzed a human primary pancreatic cancer dataset generated using the ST technology[13]. This dataset includes 224 spots and 16,448 genes with three manually annotated tissue regions. The cancer region detected by Louvain based on gene expression alone did not closely match the pathologist-annotated cancer region (Fig. 2a). Spatial clustering methods such as stLearn and BayesSpace did not detect the cancer region either. SpaGCN revealed a similar pattern when using default parameters. As the histology image shows a clear difference between the cancer and noncancer regions, it suggests histology is informative for clustering. SpaGCN has the flexibility of modeling histology with a scaling parameter *s*, which controls the weight given to histology when detecting neighbors for each spot. By increasing the value of *s* from 1 to 2, SpaGCN detected a cluster that agrees well with the manually annotated cancer region. It is worth noting that when *s* was set at the default value of 1, SpaGCN detected the noncancer regions well. When *s* was increased to 2, SpaGCN not only maintained the ability to detect the noncancer regions but also detected the cancer region. This example showed that SpaGCN is flexible in incorporating histology information in clustering. Although stLearn can incorporate histology data, its use of histology information is pre-fixed by the radius when defining neighboring spots. The lack of flexibility in adjusting histology weight led to the discrepancy between their clustering and the pathologist's manual annotation.

Next, we detected SVGs using SpaGCN, SPARK and SpatialDE. In total, SpaGCN detected 12 SVGs, with three, eight and one SVGs for domains 0, 1 and 2, respectively (Fig. 2b; Supplementary Fig. 1). Furthermore, a meta gene using *KRT17*, *MMP11* and *SERPINA1* marked the cancer region better than the originally identified *KRT17* for domain 2 (Fig. 2c). *KRT17* functions as a tumor promoter and regulates proliferation in pancreatic cancer[30], and *MMP11* is a prognostic biomarker for pancreatic cancer[31]. Our identification of *KRT17* and *MMP11* as the two positive genes for the cancer region agrees well with pancreatic cancer biology. SPARK and SpatialDE detected 203 and 163 SVGs, with their *P* or *Q* values highly skewed towards 0 (Supplementary Figs. 2 and 3). However, the Moran's *I* and Geary's *C* values for their SVGs are much lower than those detected by SpaGCN, suggesting their lack of spatial patterns (Fig. 2d). Furthermore, genes with smaller *P* or *Q* values do not necessarily show better spatial expression patterns than those with larger *P* or *Q* values (Supplementary Figs. 4 and 5). More stringent filtering of spots and genes did not improve the spatial pattern for SpatialDE and SPARK-detected SVGs (Supplementary Fig. 6).

Application to human dorsolateral prefrontal cortex 10x Visium data. To show quantitatively that SpaGCN outperforms Louvain, stLearn and BayesSpace in spatial domain detection, we analyzed the LIBD human dorsolateral prefrontal cortex (DLPFC) data generated using 10x Visium[32]. This study sequenced 12 tissue slices that span six neuronal layers plus white matter from the DLPFC in three human brains. The manual annotation of the tissue layers provided by the original study allows us to evaluate the accuracy of spatial domain detection. Figure 3a shows that for the representative tissue slice 151673, both SpaGCN and BayesSpace revealed spatial domains that agree better with the manually annotated tissue layers than Louvain. Although stLearn utilized histology information, its performance is not much better than Louvain and is substantially worse than SpaGCN and BayesSpace. The relative performance of these methods remains the same when considering all 12 slices (Fig. 3b and Supplementary Table 2); the median ARI is 0.36 for stLearn, 0.42 for BayesSpace and 0.45 for SpaGCN.

To validate further the identified spatial domains, we detected SVGs for each domain in slice 151673. In total, SpaGCN detected 67 SVGs, with 53 of them being specific to domain 5, which corresponds to white matter (Supplementary Fig. 7). Patterns of SVGs for other domains are not very clear. These results indicate that gene expression profiles of spots from white matter are distinct from spots in the neuronal layers, while gene expression differences among the six neuronal layers are much smaller and more difficult to distinguish using individual marker genes. SVGs detected by SPARK and SpatialDE also suffered from the same problem. SPARK detected 3,187 SVGs with 1,131 of them having false discovery rate (FDR)-adjusted *P* values equal to 0, most of which only marked the white matter region (Supplementary Figs. 8 and 9). We also found that the SVGs detected by SPARK lack domain specificity (Supplementary Fig. 10). SpatialDE detected 3,654 SVGs with 806 of them having *Q* values equal to 0, but these genes do not necessarily show better spatial patterns than genes with larger *Q* values (Supplementary Fig. 11). Although SPARK and SpatialDE detected much larger numbers of SVGs than SpaGCN, the genes detected by these two methods cannot distinguish different degrees of spatial expression variability as their *P* or *Q* value distributions are highly skewed towards 0. Figure 3c shows that the Moran's *I* values for SpaGCN-detected SVGs are significantly higher than genes detected by SpatialDE and SPARK (median of 0.39 for SpaGCN against 0.09 for SPARK and 0.08 for SpatialDE). More stringent filtering of spots and genes did not improve the performance of SpatialDE and SPARK (Supplementary Fig. 12). For three out of the six neuronal layers, SpaGCN detected a single SVG to mark that region (Fig. 3d). For example, *CAMK2N1* is enriched in domain 0 (layers 1 and 2), *PCP4* is enriched in domain 1 (layer 4) and *NEFM* is enriched in domain 3 (layer 3).

To show that SpaGCN-detected SVGs are useful for downstream analysis, we performed *K*-means clustering on slice 151507, which

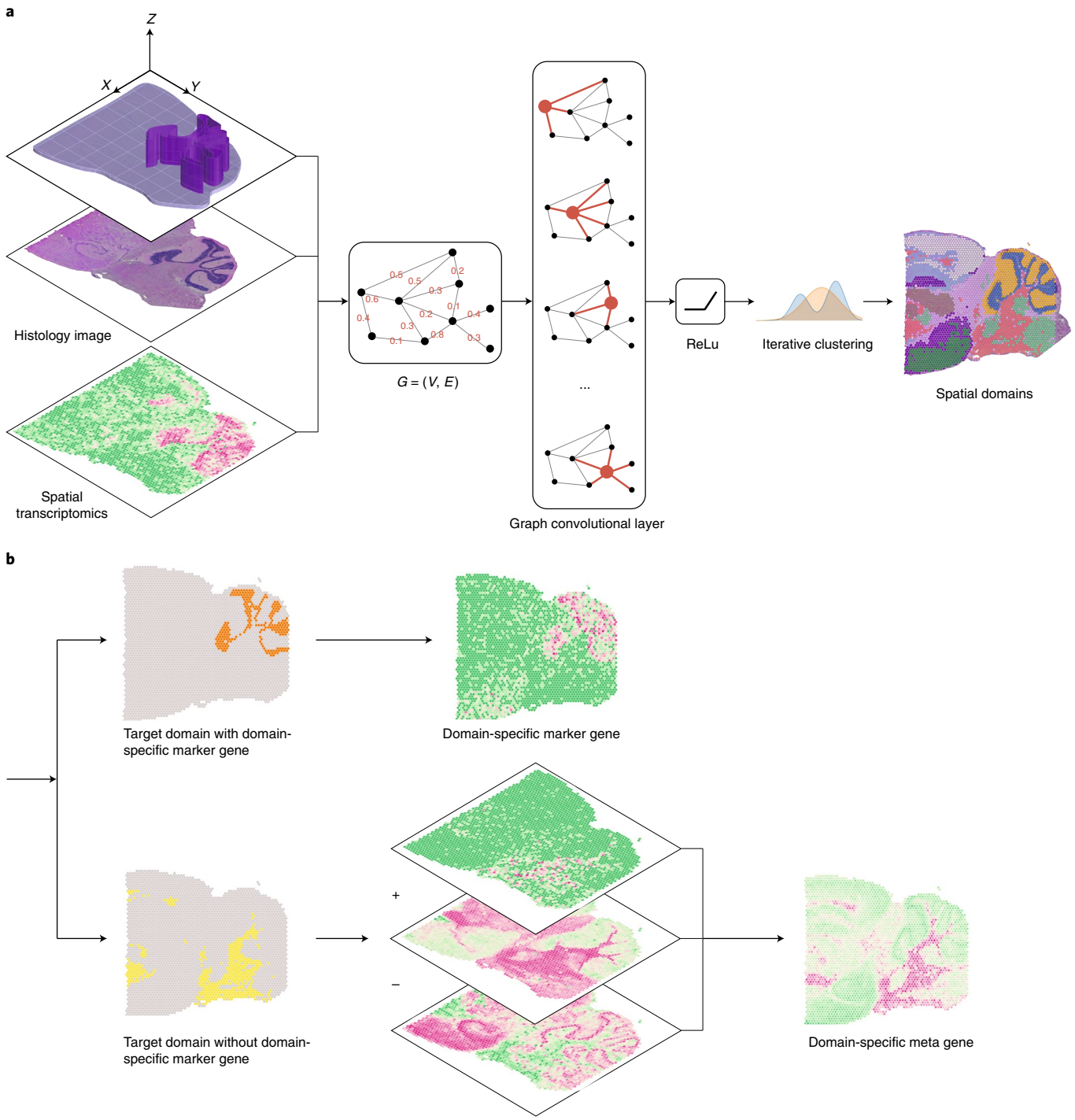


Fig. 1 | Workflow of SpaGCN. a, SpaGCN starts from integrating gene expression, spatial location, and histology information using a graph convolutional network (GCN), then separates spots into different spatial domains using unsupervised iterative clustering. The GCN is based on an undirected weighted graph in which the edge weight between every two spots is determined by Euclidean distance between the two spots, defined by the spatial coordinates (*x*,*y*) and the third dimensional coordinate *z*, obtained from the RGB values in the histology image. **b**, For each detected spatial domain, SpaGCN identifies SVGs or meta genes by domain guided DE analysis.

is from a different brain, using all 67 SVGs detected from slice 151673 by SpaGCN. Compared with manually curated layer assignment, this clustering analysis had a Adjusted Rand Index (ARI) of 0.23 (Fig. 3e,f). We performed similar analysis using SVGs detected by SpatialDE and SPARK. When randomly selecting 67 SVGs with 0 *P* or *Q* value from genes detected by SpatialDE/SPARK, the ARI is only 0.13 for SpatialDE and 0.14 for SPARK. The ARIs for SpatialDE and SPARK did not improve even with increased numbers of SVGs (Fig. 3e). These results further confirmed the lack of spatial patterns for genes detected by SPARK and SpatialDE.

Although it is difficult to identify single genes to mark certain neuronal layers, SpaGCN was able to find domain-specific meta genes. As shown in Fig. 3g, SpaGCN detected meta genes for domains 1, 2, 4 and 6. The meta gene for domain 2 is specific to

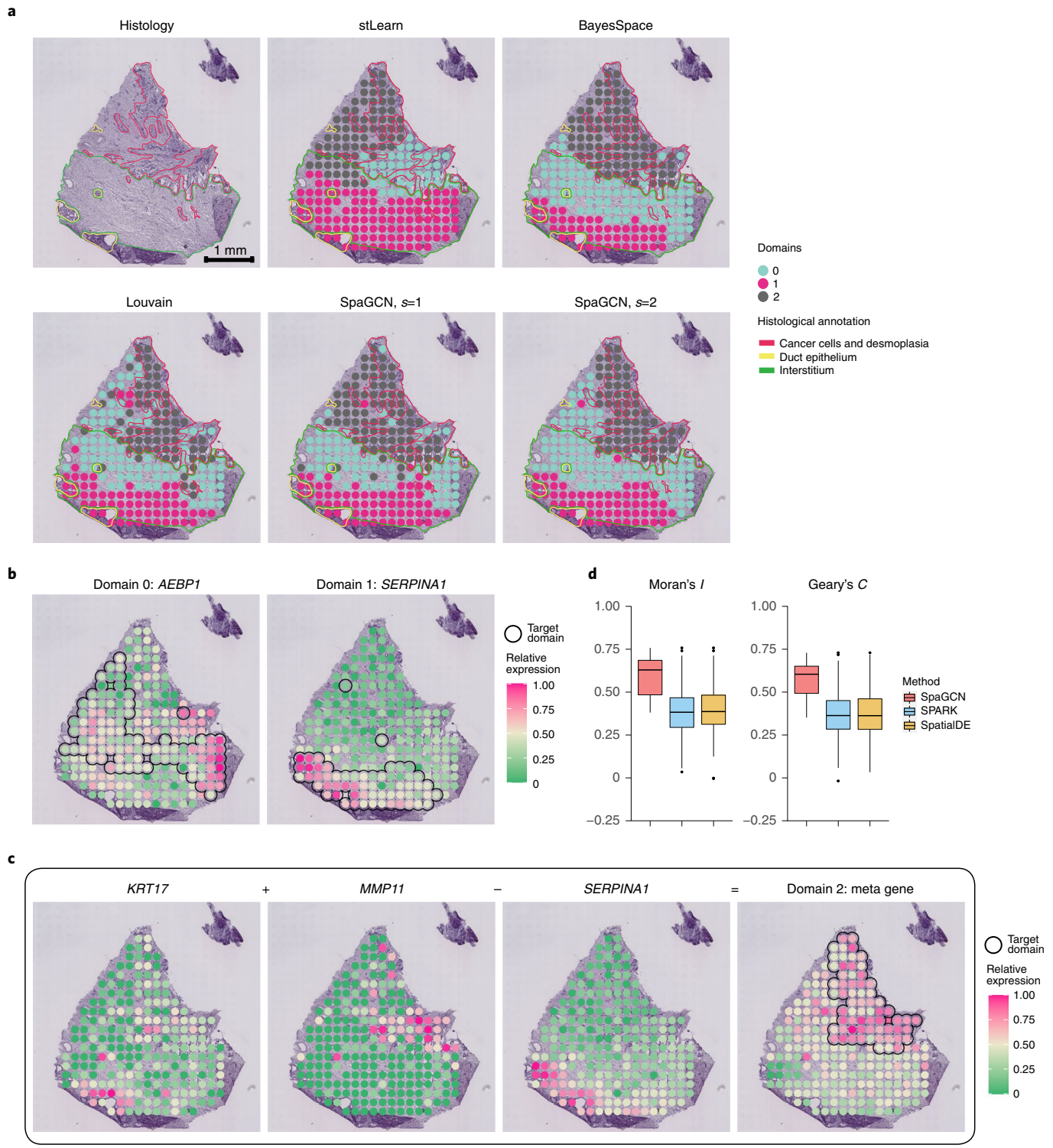


Fig. 2 | Spatial domains and SVGs detected in the human primary pancreatic cancer tissue data. a, Histology image of the tissue section with manually annotated regions from the original study[13], spatial domains detected by stLearn, BayesSpace, Louvain and SpaGCN. SpaGCN has a scaling parameter 's' that controls the weight given to histology when detecting neighbors for each spot. When $s = 2$, SpaGCN detected the pathologist-annotated cancer region. **b**, Spatial expression pattern of SVGs detected by SpaGCN for domain 0 (*AEBP1*) and domain 1 (*SERPINA1*). **c**, Spatial expression patterns of genes *KRT17*, *MMP11*, *SERPINA1*, which form the meta gene for domain 2 (*KRT17* + *MMP11* − *SERPINA1*). **d**, Boxplots of Moran's *I* and Geary's *C* values for SVGs detected by SpaGCN (n=12), SPARK (n=203), and SpatialDE (n=163). The lower and upper hinges correspond to the first and third quartiles, and the center refers to the median value. The upper (lower) whiskers extend from the hinge to the largest (smallest) value no further (at most) than 1.5 × interquartile range from the hinge. Data beyond the end of the whiskers are plotted individually.

layer 1. As layer 1 only has a few spots, it is difficult to find a highly enriched gene. However, by adding depleted genes such as *FTH1*, *MBP*, *MT-CO3* and *PLP1*, the expression pattern in this region is strengthened. Furthermore, the SVGs and meta genes detected by SpaGCN are transferrable to slice 151507 obtained from a different brain, in which the meta genes detected in slice 151673 mark the same layers in slice 151507 (Fig. 3g and Supplementary Fig. 13).

Application to mouse posterior brain 10x Visium data. Next, we analyzed a 10x Visium dataset generated from mouse posterior brain that includes 3,353 spots and 31,053 genes[33]. This dataset shows much more complex tissue structure than the previous two datasets. We compared the clustering result of SpaGCN with Louvain, stLearn and BayesSpace when the number of clusters was set at ten for all methods. Figure 4a shows that Louvain's clustering is similar to stLearn, BayesSpace and SpaGCN, but the spatial domains detected by the latter three methods are more spatially contiguous due to their ability to account for spatial dependency of gene expression.

We further investigated the ability of each method in detecting more refined tissue structure. Specifically, we performed subclustering analysis for spots in domain 5 detected by SpaGCN, which corresponds to the cortex (Fig. 4b). The subdomains detected by SpaGCN agree well with the Allen Brain Institute reference atlas diagram of the mouse cortex (Fig. 4c). The detected subdomains include layers 2/3, layers 4/5, layer 6, a hippocampal region (CA1) and the subiculum. Layers 2/3 are the 'external' cortical layers that are biologically responsible for local networks in which neurons in this subdomain communicate to other neurons in adjacent neocortical regions. Layers 4/5 are the 'internal' cortical layers that are biologically responsible for longer range neural networks. For example, the visual cortex, which corresponds to the neocortical region, is responsible for receiving visual information from the lateral geniculate nucleus that is far away. SpaGCN was able to separate the molecular (layer 1), external (layers 2/3), internal (layers 4/5) and the plexiform (6) layers. More importantly, SpaGCN outperformed Louvain and stLearn, which show combining of neocortical layers. SpaGCN also outperformed BayesSpace in distinguishing between the plexiform layer (subdomain 1) and the non-neocortical CA1 region of the hippocampus (subdomain 3). In contrast, BayesSpace combined layer 6 of the neocortex with the non-neocortical CA1 layer of the hippocampus.

Next, we compared SpaGCN with SPARK and SpatialDE for SVG detection. SpaGCN detected 1,028 SVGs for the ten spatial domains while SPARK and SpatialDE detected 9,678 and 12,676 SVGs, respectively (Supplementary Fig. 14). As shown in Fig. 4d, the Moran's *I* values of SpaGCN-detected SVGs are much higher than those detected by SPARK and SpatialDE (median of 0.54 for SpaGCN against 0.20 for SPARK and 0.16 for SpatialDE). More stringent filtering of spots and genes did not improve the performance of SPARK and SpatialDE (Supplementary Fig. 15). The *P* or *Q* value distributions of SpatialDE and SPARK are highly skewed towards 0 (Supplementary Fig. 16), and genes with similar *P* or *Q* values do not necessarily show similar spatial patterns and a smaller *P* or *Q* value does not guarantee a better spatial pattern (Supplementary Figs. 17 and 18). In contrast, multiple domain adaptive filtering criteria implemented in SpaGCN allow it to eliminate false positive SVGs and ensure all detected SVGs have clear spatial expression patterns.

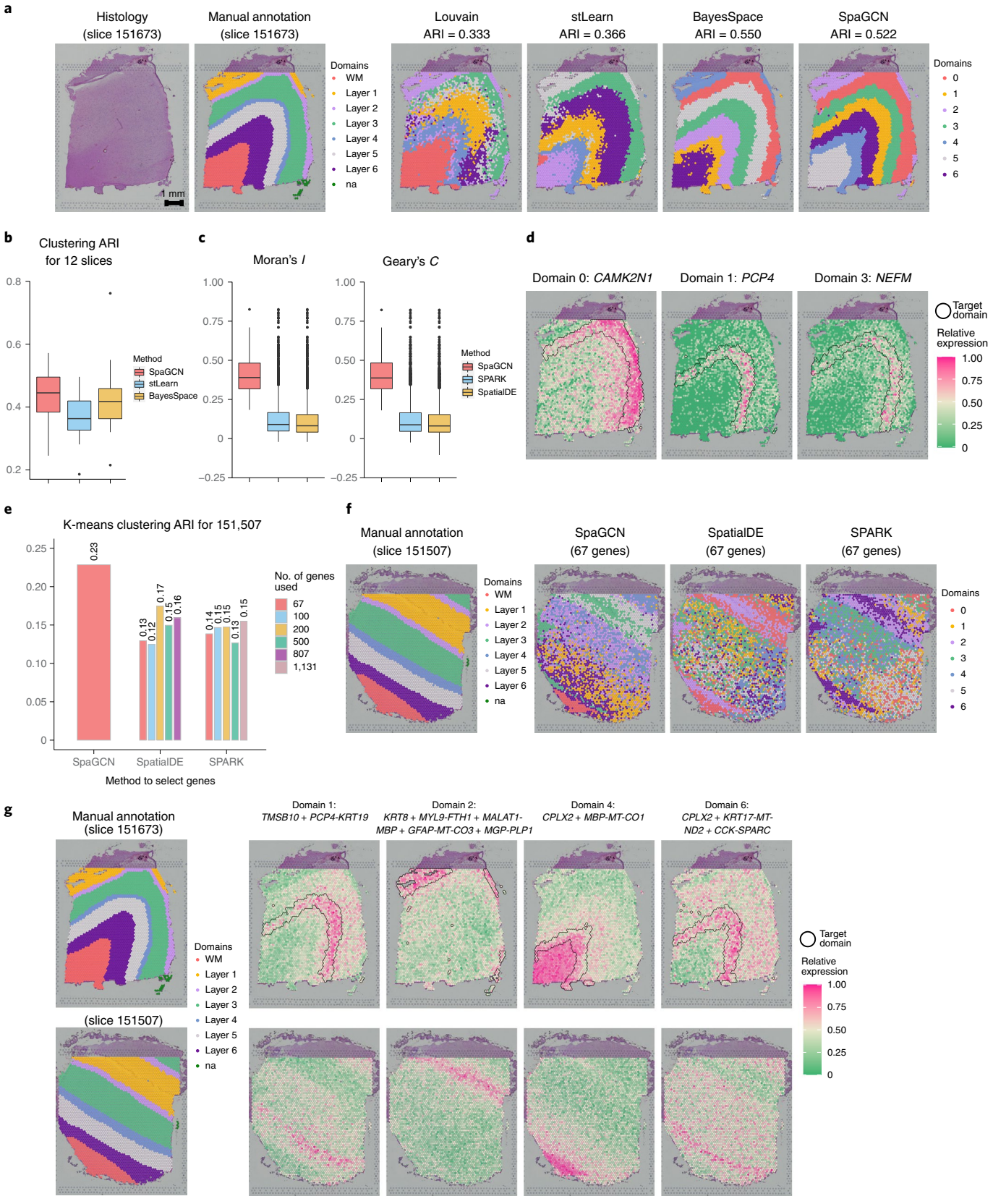
To illustrate how the filtering in SpaGCN works, we use domains 1, 5 and 8 as an example. For each of these domains, SpaGCN detected a single SVG enriched in that region. As shown in Fig. 4e, *PVALB* is enriched in domain 1 and *TRM62* is enriched in domain 8. Although domains 1 and 8 are adjacent to each other, these two SVGs can still well mark these domains. *NRGN* is a SVG that SpaGCN detected for domains 5 and 7. The high expression of *NRGN* in domains 5 and 7 also indicates that these two domains are neuroanatomically similar—both consisting of cortex and the pyramidal layer of the hippocampus. Both the cortex and hippocampus are regions that are on the curved surface of the brain. Domains 5 and 7, which would be contiguous in a three-dimensional (3D) reconstruction, are artifactually separated as a result of how the section was cut. Therefore, it is not surprising that in addition to *NRGN*, SpaGCN also detected many other SVGs for domains 5 and 7, some of which are highly expressed in both domains (Supplementary Fig. 19). The unique and powerful SVG detection procedure in SpaGCN ensures that genes such as these are not missed.

SpaGCN only identified four SVGs for domain 0. However, we reason that a meta gene, formed by the combination of multiple genes, may better reveal spatial patterns than any single genes. We used domain 0 as an example to show how SpaGCN can create informative meta genes to mark a spatial domain (Fig. 4f). First, by lowering the filtering thresholds, SpaGCN identified *KLK6* which is highly expressed in the lower part of domain 0. Using *KLK6* as a starting gene, SpaGCN used a novel approach to find a log-linear combination of gene expression of *KLK6*, *MBP* and *ATP1B1*, which accurately marked the spatial domain 0. In this meta gene, *KLK6* and *MBP* are considered as positive markers because they are highly expressed in some spots in domain 0, whereas *ATP1B1* is considered a negative marker as it is mainly expressed in regions other than domain 0. Previous studies have shown that *KLK6* and *MBP* expression is restricted to oligodendrocytes, while *ATP1B1* is mainly expressed in neurons and astrocytes[34]. This resonates with the fact that domain 0 represents white matter which is dominated by oligodendrocytes and has few neuronal cell bodies. Therefore, the genes that make up this meta gene have meaningful biological interpretations.

While we focused our analyses on one tissue section, SpaGCN can also jointly analyze multiple tissue sections. We show two examples using this mouse brain Visium data provided by 10x Genomics. Figure 5a shows SpaGCN clustering results for two mouse posterior sections. As these two tissue sections are from the same region, SpaGCN was able to infer cluster correspondence between the two tissue sections. Next, we used SpaGCN to analyze jointly two tissue

Fig. 3 | Spatial domains and SVGs detected in the LIBD human dorsolateral prefrontal cortex data. a, Histology and manually annotated layer structure for slice 151673 from the original study[32], and spatial domains detected by Louvain, stLearn, BayesSpace and SpaGCN. **b**, Boxplot of clustering ARIs for all 12 tissue slices for SpaGCN (*n*=12), stLearn (*n*=12) and BayesSpace (*n*=12). The lower and upper hinges correspond to the first and third quartiles and the center refers to the median value. The upper (lower) whiskers extend from the hinge to the largest (smallest) value no further (at most) than 1.5 × interquartile range from the hinge. Data beyond the end of the whiskers are plotted individually. **c**, Boxplot of Moran's *I* and Geary's *C* values for SVGs detected by SpaGCN (*n*=65), SPARK (*n*=3,187) and SpatialDE (*n*=3,654) for slice 151673. Boxplot hinges, median and whiskers are defined the same as in **b** and **d.** Spatial expression patterns of SVGs for domain 0 (*CAMK2N1*), domains 1 (*PCP4*) and domain 3 (*NEFM*) for slice 151673. **e**, ARIs between manually annotated layers and *K*-means' clustering using various numbers of SVGs detected by different methods. For SpaGCN, we used the 67 SVGs while for SPARK and SpatialDE, we used their top 67, 100, 200, 500 and all SVGs with the identical smallest FDR-adjusted *P* value or *Q* value. **f**, Manually annotated layer structure for slice 151507 from the original study. *K*-means clustering results for slice 151507 using the 67 SVGs detected by SpaGCN, and the top 67 SVGs with the identical smallest FDR-adjusted *P* value or *Q* value detected by SPARK and SpatialDE. **g**, Meta genes detected by SpaGCN in slice 151673 are transferrable to slice 151507 in a different brain.

sections with one from the mouse posterior brain and the other from the mouse anterior brain. As the anterior section and posterior section are adjacent in the brain, we modified the coordinates for spots in the posterior section such that the revised coordinates reflect the spatial adjacency of the two tissue sections. Using the modified coordinates as input, SpaGCN was able to produce clustering results that reflect the shared layer structure in the anterior and posterior brain (Fig. 5b).



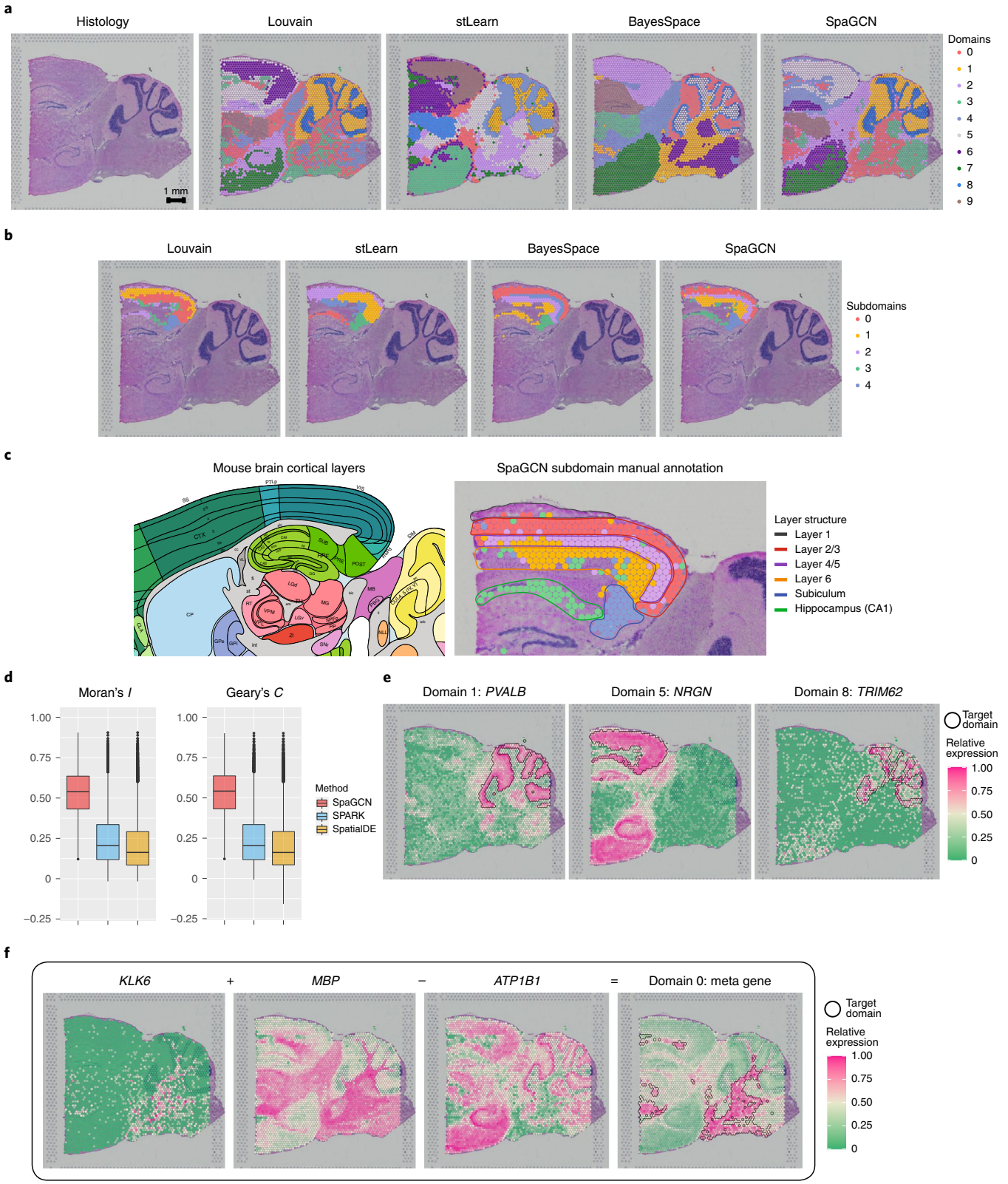


Fig. 4 | Spatial domains and SVGs detected in the mouse brain posterior brain data. a, Histology image of the tissue section and spatial domains detected by Louvain, stLearn, BayesSpace and SpaGCN. **b**, Spatial subdomains of the cortex region detected by Louvain, stLearn, BayesSpace and SpaGCN. **c**, Allen Brain Institute reference atlas diagram of the mouse cortex and the subdomain manual annotation of SpaGCN. **d**, Boxplot of Moran's *I* and Geary's *C* values for SVGs detected by SpaGCN (n=815), SPARK (n=9,678) and SpatialDE (n=12,676). The lower and upper hinges correspond to the first and third quartiles, and the center refers to the median value. The upper (lower) whiskers extend from the hinge to the largest (smallest) value no further (at most) than 1.5 × interquartile range from the hinge. Data beyond the end of the whiskers are plotted individually. **e**, Spatial expression patterns of SVGs detected by SpaGCN for domain 1 (*PVALB*), 5 (*NRGN*) and 8 (*TRIM62*). **f**, Spatial expression patterns of genes *KLK6*, *MBP*, *ATP1B1*, which form the meta gene for domain 0 (*KLK6* + *MBP* − *ATP1B1*).

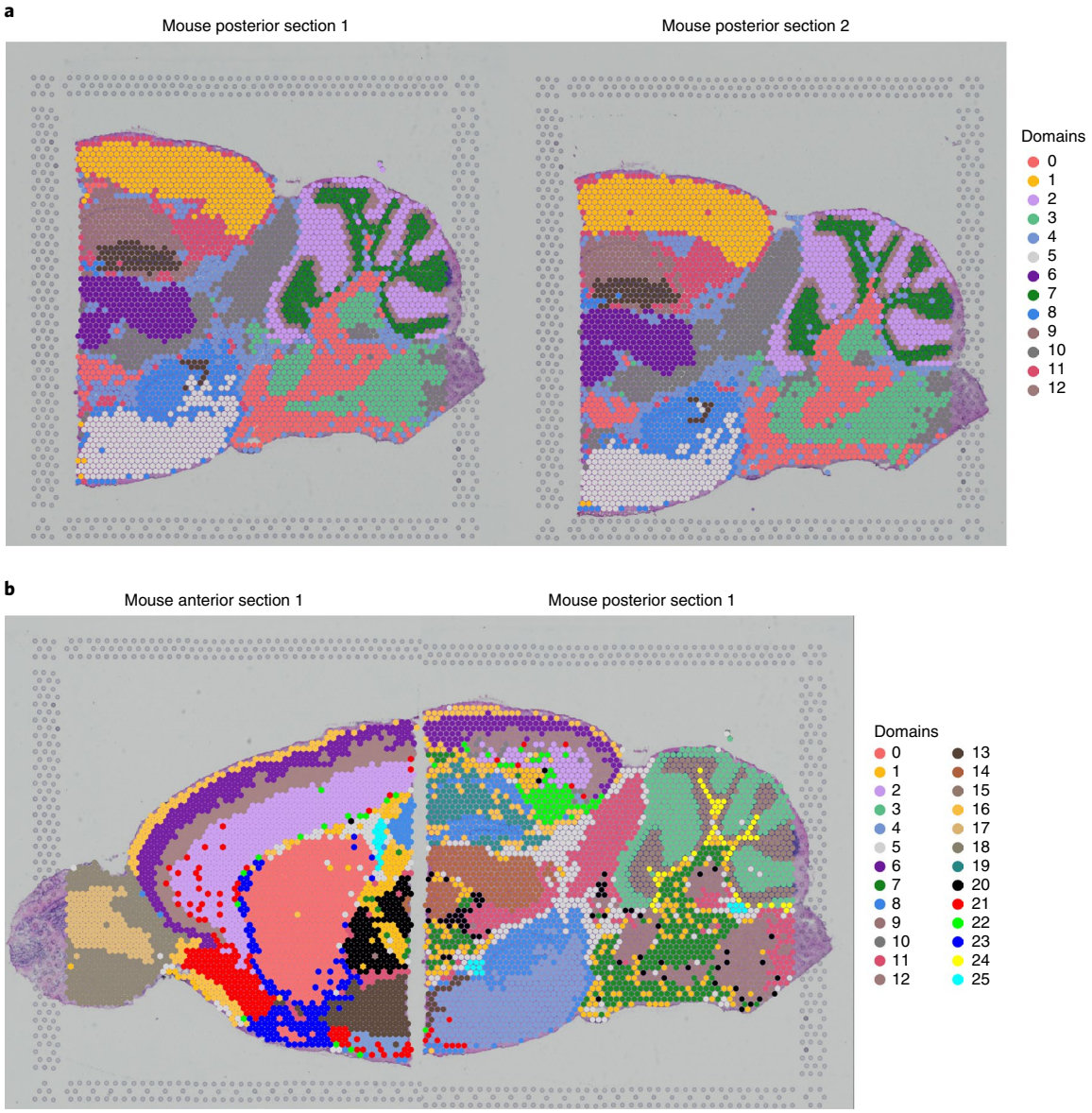


Fig. 5 | Joint spatial domain detection across multiple mouse brain tissue sections using SpaGCN. a, Joint analysis of two consecutive tissue sections from mouse posterior brain by SpaGCN. **b**, Joint analysis of a tissue section from mouse anterior brain and a tissue section from mouse posterior brain by SpaGCN.

Application to mouse visual cortex STARmap data. Finally, we analyzed a STARmap dataset that has single-cell resolution[7]. This dataset was generated from mouse visual cortex that spans from hippocampus to corpus callosum, and the six neocortical layers. In total, 1,020 genes were measured in 1,207 cells that include non-neuronal cells, excitatory and inhibitory neurons. The layer structure and cell type distribution of the tissue section provided by the original study are shown in Fig. 6a. As the tissue capture area of STARmap is much smaller than 10x Visium, we increased the contribution of neighboring cells from 0.5 to 1 when calculating the weighted gene expression of each cell in SpaGCN. Using this approach, SpaGCN detected spatial domains that agreed well with the annotated tissue structure (Fig. 6a,c), achieving an ARI of 0.51. By contrast, the ARIs of the other methods are much lower (0.30 for Louvain, 0.37 for BayesSpace and 0.03 for HMRF) (Fig. 6b). This example demonstrates that SpaGCN utilizes spatial information more efficiently than BayesSpace and HMRF. Using SpaGCN, we further detected 25 SVGs including genes *LAMP5*, *HPCAL1*, *CPLX1*, *PLP1*, *NRSN1*, *ATP1A2* and *BSG* that showed enriched expression patterns for domains 0 to 6 (Fig. 6e and Supplementary Fig. 20). Similar to previous analyses, SPARK and SpatialDE detected much larger number of SVGs but many of the SVGs lack spatial expression patterns (Fig. 6d and Supplementary Figs. 21–24).

Discussion

In this paper, we presented SpaGCN, a method that integrates gene expression, spatial location and histology to model spatial dependency of gene expression for the identification of spatial domains and domain enriched SVGs. SpaGCN has been extensively tested on datasets from different species, regions and tissues generated using diverse SRT technologies. Additional analyses on ST[9], SLIDE-seqV2 (ref. [11]) and MERFISH[5] data are shown in Supplementary Notes 1–3. Our results consistently showed that SpaGCN can identify spatial domains with coherent gene expression and histology, and

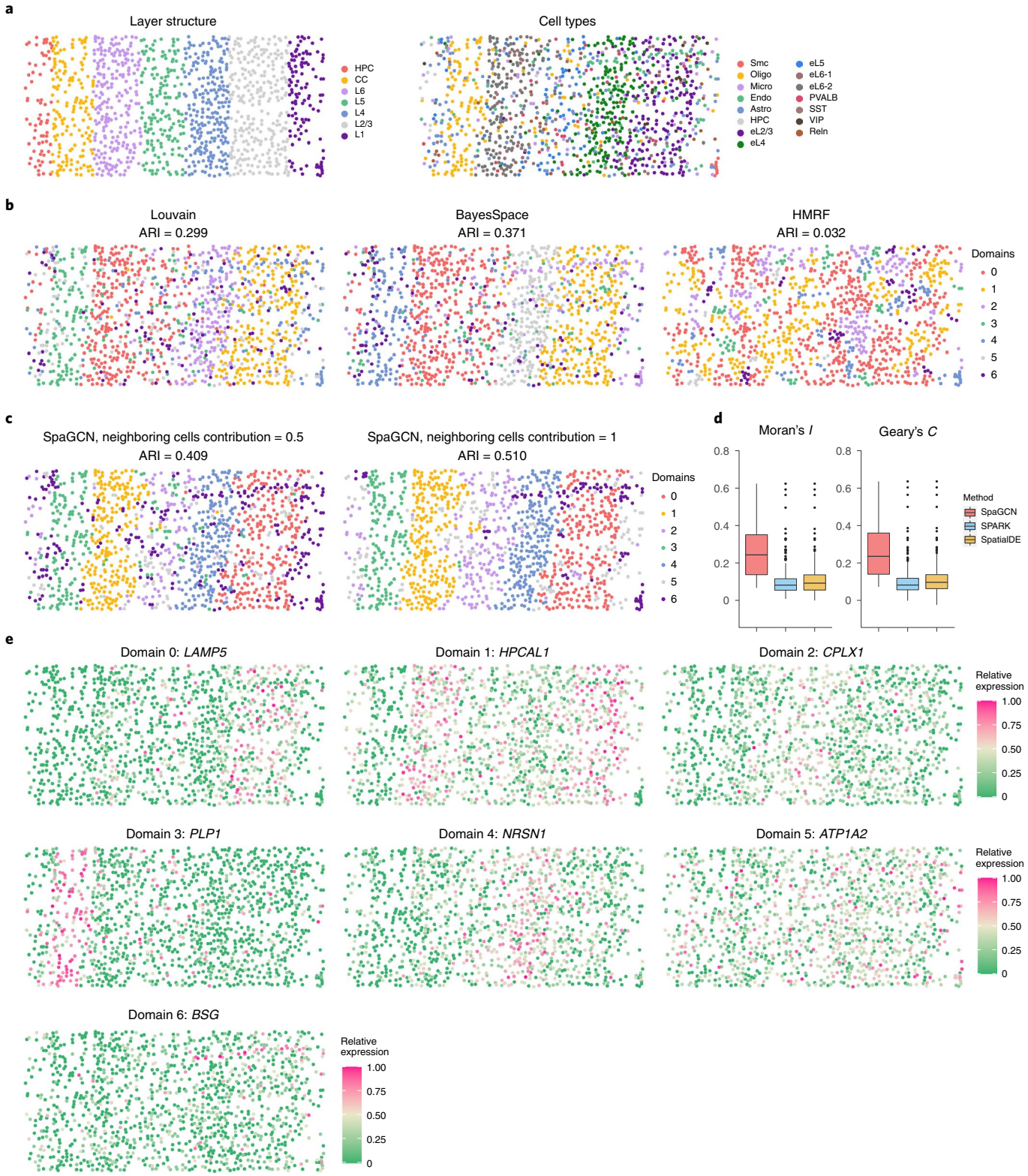


Fig. 6 | Spatial domains and SVGs detected in the mouse visual cortex STARmap data. a, Layer structure and cell type distribution of the tissue section from the original study[7]. **b**, Spatial domains detected by Louvain, BayesSpace and HMRF. **c**, Spatial domains detected by SpaGCN with different neighboring cell contributions. **d**, Boxplot of Moran's *I* and Geary's *C* values for SVGs detected by SpaGCN (n=22), SPARK (n=325) and SpatialDE (n=214). The lower and upper hinges correspond to the first and third quartiles, and the center refers to the median value. The upper (lower) whiskers extend from the hinge to the largest (smallest) value no further (at most) than 1.5 × interquartile range from the hinge. Data beyond the end of the whiskers are plotted individually. **e**, Spatial expression patterns of SVGs detected by SpaGCN for domains 0 (*LAMP5*), 1 (*HPCAL1*), 2 (*CPLX1*), 3 (*PLP1*), 4 (*NRSN1*), 5 (*ATP1A2*) and 6 (*BSG*).

detect SVGs and meta genes that have much clearer spatial expression patterns and biological interpretations than genes detected by SpatialDE and SPARK. Additionally, the SpaGCN-detected SVGs are transferrable and can be utilized for downstream analyses in independent tissue sections. SpaGCN is also computationally fast and memory efficient compared to SPARK and SpatialDE (Supplementary Note 4).

The spatial domain detection step in SpaGCN is flexible. First, SpaGCN can adjust the weight of histology in gene expression smoothing. For datasets with clear tissue structure in histology, higher weight led to clearer separation of cancer versus noncancer regions. Second, during the GCN fitting procedure, the graph weights are updated, which allows SpaGCN to learn an efficient way to aggregate gene expression from neighboring spots for each gene. For data generated from different platforms, the spatial dependency between spots/cells is different as the size of the captured tissue area varies. The flexibility in modeling spatial dependency makes SpaGCN versatile for different types of SRT data.

A limitation of SpaGCN is that the spatial domain detection is mainly driven by gene expression, which may lead to the discrepancy between the detected domains and the underlying tissue anatomical structure. This is a general problem for gene expression-based clustering methods. Another limitation of SpaGCN is the lack of separation of spatial variation and cell type variation in gene expression patterns for the detected SVGs. To address these limitations, methods that can jointly consider gene expression and histological features in clustering are needed. Further, cell type-specific gene expression needs to be estimated to tease out the contribution of cell types and spatial location in gene expression variation. We anticipate that methods development along these directions is warranted for future research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01255-8.

Received: 1 December 2020; Accepted: 29 July 2021;
Published online: 28 October 2021

References

1. Asp, M., Bergenstrahle, J. & Lundeberg, J. Spatially resolved transcriptomes-next generation tools for tissue exploration. *Bioessays* **42**, e1900221 (2020).
2. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
3. 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* **92**, 342–357 (2016).
4. Eng, C. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
5. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
6. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
7. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
8. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
9. Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
10. Rodriques, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
11. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2020).
12. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**, 987–990 (2019).
13. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).
14. Chen, W. T. et al. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991 e919 (2020).
15. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
16. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G. C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **36**, 1183–1190 (2018).
17. Pham, D. et al. stLearn: Integrating spatial location, tissue morphology and gene expression to find cell types, cell–cell interactions and spatial trajectories within undissociated tissues. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.31.125658 (2020).
18. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00935-2 (2021).
19. Fu, X. et al. Continuous polony gels for tissue mapping with high resolution and RNA capture efficiency. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.17.435795 (2021).
20. Chen, A. et al. Large field of view-spatially resolved transcriptomics at nanoscale resolution. *bioRxiv.* https://doi.org/10.1101/2021.01.17.427004 (2021).
21. Liu, Y. et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665–1681 e1618 (2020).
22. Cho, C. S. et al. Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* **184**, 3559–3572 e3522 (2021).
23. Edsgard, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342 (2018).
24. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: Identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
25. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
26. Xie, J., Girshick, R. & Farhadi, A. Unsupervised deep embedding for clustering analysis. In *Proc. 33rd International Conference on Machine Learning* Vol. 48 (JMLR: W&CP, 2016).
27. Li, H., Calder, C. A. & Cressie, N. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geographical Anal.* **39**, 357–375 (2007).
28. Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M. J. T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.* **48**, e107 (2020).
29. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00830-w (2021).
30. Li, D. et al. KRT17 Functions as a tumor promoter and regulates proliferation, migration and invasion in pancreatic cancer via mTOR/S6k1 pathway. *Cancer Manag. Res.* **12**, 2087–2095 (2020).
31. Lee, J., Lee, J. & Kim, J. H. Identification of matrix metalloproteinase 11 as a prognostic biomarker in pancreatic cancer. *Anticancer Res.* **39**, 5963–5971 (2019).
32. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
33. Dataset. Mouse posterior brain data. 10x Genomics https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior (2020).
34. Zhang, Y. et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).

Methods

Data preprocessing. SpaGCN takes spatial gene expression and histology image data (when available) as input. For ease of explanation, we will use in situ capturing-based SRT data to illustrate the method. The spatial gene expression data are stored in an $N \times D$ matrix of unique molecular identifier (UMI) counts with N spots and D genes, along with the (x,y) two-dimensional (2D) spatial coordinates of each spot. Genes expressed in fewer than three spots are eliminated. The gene expression values in each spot are normalized such that the UMI count for each gene is divided by the total UMI count across all genes in a given spot, multiplied by 10,000, and then transformed to a natural log scale.

Conversion of SRT data into graph-structured data. After preprocessing, SpaGCN converts the gene expression and histology image data into a weighted undirected graph, $G(V,E)$. In this graph, each vertex $v \in V$ represents a spot and every two vertices in V are connected via an edge with a specified weight. We note that spage2vec[35] also employed a graph-based approach, but with the goal of clustering messenger RNA molecules in which each node represents a mRNA. Such a segmentation-free approach may offer advantages over methods that require segmented cells as cell segmentation is still one of the hardest problems in single-cell analysis.

Calculation of distance between two vertices. The distance between any two vertices u and v in the graph reflects the relative similarity of the two corresponding spots. This distance is determined by two factors: (1) the physical locations of spots u and v in the tissue slice, and (2) the corresponding histology information of these two spots. Although some spots are physically close to each other in the tissue, the histology image may reveal them belonging to different tissue layers. Therefore, SpaGCN considers two spots to be close if and only if (1) the two spots are physically close, and (2) they have similar histological features as shown in the histology image. To define a distance metric considering both aspects, SpaGCN extends the 2D space in the tissue slice into a 3D space that incorporates histology information. For spot v, its physical location in the tissue slice is represented by 2D coordinates (x_v, y_v). To determine the corresponding pixel in the histology image for spot v, SpaGCN maps spot v to the histology image according to its pixel coordinates (x_{pv}, y_{pv}). Instead of using the color of the pixel at (x_{pv}, y_{pv}), SpaGCN draws a square centered on (x_{pv}, y_{pv}) containing 50×50 pixels and calculates the mean color value for the RGB channels, (r_v, g_v, b_v), of all pixels that fall in the square. This step smooths the color value and ensures that the color is not dominated by a single pixel. To derive a single value to represent the histology image features, SpaGCN uses a weighted sum of the RGB values as follows,

$$z_v = \frac{r_v \times V_r + g_v \times V_g + b_v \times V_b}{V_r + V_g + V_b},$$

where $V_r = \text{variance}(r_v)$, $V_g = \text{variance}(g_v)$ and $V_b = \text{variance}(b_v)$ for all $v \in V$ in this transformation, higher weight is given to the channel with larger variance so that this combined value z_v captures an accurate representation of the patterns in the histology image.

Next, SpaGCN rescales z_v as

$$z_v^* = \frac{z_v - \mu_z}{\sigma_z} \times \max(\sigma_x, \sigma_y) \times s,$$

where μ_z is the mean of z_v, σ_x, σ_y, σ_z are the standard deviations of x_v, y_v and z_v, respectively, for $v \in V$, and s is a scaling factor. By default, s is set at 1 to ensure that z_v^* has the same scale variance as x_v and y_v, and we set s to a value larger than 1 when the goal is to increase the weight of histology. The coordinates of spot v are set to be (x_v, y_v, z_v^*) in the extended 3D space. Finally, the Euclidean distance between every two spots u and v is calculated as

$$d(u,v) = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u^* - z_v^*)^2}.$$

When histology information is not available, the Euclidean distance between every two spots will be calculated based on spatial location information only.

Calculation of weight for each edge and construction of graph. The weight of each edge (u,v) measures the degree of relatedness between spots u and v and is negatively associated with their distance. The graph structure G is stored in an $N \times N$ adjacency matrix $A = [w(u,v)]$, where the edge weight between spot u and spot v is defined as

$$w(u,v) = \exp\left(-\frac{d(u,v)^2}{2l^2}\right).$$

The hyperparameter l, also known as the characteristic length scale, determines how rapidly the weight decays as a function of distance. A similar function has been employed in SpatialDE[24]. Let I denote the identity matrix. For spot v, the corresponding row sum of $A - I$, denoted by a_v, can be interpreted as the relative contribution of other spots to its gene expression. By default, we choose the value of l such that the average of a_v across all spots is equal to a prespecified value, for example, 0.5. For data generated from SRT platforms with small tissue capture areas, for example, SLIDE-seqV2, STARmap and MERFISH, we suggest choosing the value of l such that neighboring spots/cells contribute more information in gene expression aggregation.

Graph convolutional layer. SpaGCN reduces the dimension of the preprocessed gene expression matrix using principal component analysis (PCA). The top 50 principal components are used as input, which work well for all datasets analyzed in this paper. Next, utilizing the power of a graph convolutional network, SpaGCN concatenates the gene expression information and edge weights in G to cluster the nodes. Following Kipf and Welling[36], the graph convolutional layer can be written as

$$f(X, A) = \delta(AXB),$$

where X is the $N \times 50$ embedding matrix obtained from PCA, B is a 50×50 matrix representing filter parameters of the convolutional layer, and $\delta(\cdot)$ is a nonlinear activation function such as ReLU. The graph convolutional layer ensures that a corresponding row of parameters in B will control the aggregation of neighborhood information for each feature in X, thus offering the flexibility of feature-specific aggregation of information provided by neighboring spots. The filter parameters in B are shared across all vertices in the graph and are automatically updated during an iterative training progress. Through graph convolution, SpaGCN has aggregated the gene expression information according to the edge weights specified in G. The output of this layer is an aggregated matrix that includes information on gene expression, spatial location and histology. The graph convolutional layer was implemented based on Kipf and Welling[36], in which the backpropagation is operated via a localized first-order approximation of spectral graph convolution.

Spatial domain identification by clustering. Next, based on the output from the above graph convolutional layer, SpaGCN employs an unsupervised clustering algorithm iteratively to cluster the spots into different spatial domains[26]. Each cluster identified from this analysis is considered to be a spatial domain, which contains spots that are coherent in gene expression and histology. To initialize cluster centroids, we use Louvain's method[15] on the aggregated output matrix from the graph convolutional layer. If the number of domains in the tissue is known, the resolution parameter in Louvain will be set to generate the same number of spatial domains. Otherwise, we vary the resolution parameter from 0.2 to 1.0 and select the resolution that gives the highest Silhouette score[37].

To update the cluster assignments iteratively, we define a metric to measure the distance from a spot to a cluster centroid using the Student's t-distribution as a kernel. The distance between the embedded point h_i for spot i and centroid μ_j for cluster j

$$q_{ij} = \frac{\left(1 + h_i - \mu_j^2\right)^{-1}}{\sum_{j'=1}^{K}\left(1 + h_i - \mu_{j'}^2\right)^{-1}},$$

can be interpreted as the probability of assigning cell i to cluster j.

Next, we iteratively refine the clusters by defining an auxiliary target distribution P based on q_{ij}

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^{N} q_{ij}}{\sum_{j'=1}^{K}\left(q_{ij'}^2 / \sum_{i=1}^{N} q_{ij'}\right)},$$

which upweights spots assigned with high confidence, and normalizes the contribution of each centroid to the overall loss function to prevent large clusters from distorting the hidden feature space. Now that we have the soft assignment q_{ij} and the auxiliary distribution p_{ij}, we can define the objective function as a Kullback–Leibler (KL) divergence loss,

$$L = \text{KL}(P||Q) = \sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\log\frac{p_{ij}}{q_{ij}}.$$

The network parameters and cluster centroids are simultaneously optimized by minimizing L using stochastic gradient descent with momentum. This unsupervised iterative clustering algorithm has previously been utilized for scRNA-seq analysis and showed superior performance over Louvain's method[38,39].

After clustering, SpaGCN also provides an optional refinement step for the clustering result. In this step, SpaGCN examines the domain assignment of each spot and its surrounding spots. For a given spot, if more than half of its surrounding spots are assigned to a different domain, this spot will be relabeled to the same domain as the major label of its surrounding spots. As this refinement step only relabels a few spots, it has little impact on the downstream SVG detection. We performed cluster refinement only for the human dorsolateral prefrontal cortex 10x Visium data and the STARmap data when comparing to their manual annotations with clear domain boundaries.

Detection of SVGs. We are interested in detecting SVGs that are enriched in each spatial domain. We note that some genes may be expressed in multiple but disconnected domains. Although they are not uniquely expressed in a particular domain, these genes are still useful for understanding spatial variation of gene expression and can be used to form meta genes that are uniquely expressed in a specific domain. Therefore, rather than doing DE analysis using spots from a target domain versus all other spots, we first select spots to form a neighboring set of the target domain. The goal is to detect genes that are highly expressed in the target domain but are not expressed or are expressed at low levels in the neighboring spots. Supplementary Fig. 25 illustrates how a neighboring domain is identified. Briefly, we draw a circle with a prespecified radius around each spot in the target domain, and spots from nontarget domains but residing in the circle are considered its neighbors. The radius is set such that each spot in the target domain has approximately ten neighbors on average. Next, neighbors of all spots in the target domain are collected and form a neighboring set. For each nontarget domain, if more than 50% (default) of its spots are in the neighboring set, this domain is then selected as a neighboring domain. This criterion is set to avoid the situation in which a domain is selected as a neighboring domain, but only a small proportion of its spots are adjacent to the target domain.

After neighboring domains are determined, SpaGCN then performs DE analysis between spots in the target domain and the neighboring domain(s) using the Wilcoxon rank-sum test. Genes with a FDR-adjusted P value <0.05 are selected as SVGs. To ensure only genes with enriched expression patterns in the target domain are selected, we further require a gene to meet the following three criteria: (1) the percentage of spots expressing the gene in the target domain, that is, in-fraction, is >80%; (2) for each neighboring domain, the ratio of the percentages of spots expressing the gene in the target domain and the neighboring domain(s), that is, in/out fraction ratio, is >1; and (3) the expression fold change between the target and neighboring domain(s) is >1.5. If a user is interested in finding SVGs for a particular combination of spatial domains, SpaGCN offers the option to do so.

Detection of spatially variable meta genes. The spatial domain-specific DE analysis described above typically detects SVGs with enriched expression for the majority of the domains. For domains in which no such SVGs are detected, we aim to identify a set of genes that, when combined to form a meta gene, shows an enriched expression pattern in the given domain. To identify genes to form a meta gene, we employ a multi-step approach. First, we lower the thresholds for SVG filtering, for example, change the minimum fold change threshold from 1.5 to 1.2, to identify genes showing a weaker enriched expression pattern in the target domain. In the presence of multiple such weaker SVGs, we randomly select one of them as the base gene and denote it as $gene_0$. Second, we aim to aggregate expression from other genes to the base gene to enhance the spatial pattern for the target domain. To achieve this goal, we first calculate the mean expression level of $gene_0$ for spots in the target domain as e_0. Then, all spots from nontarget domains with the expression level higher than e_0 of $gene_0$ are extracted to form a control group. Next, we perform DE analysis using spots from the target domain against spots in the control group using the Wilcoxon rank-sum test. The gene with the smallest FDR-adjusted P value and higher expression in the target domain is selected as $gene_{0+}$. Similarly, we perform DE analysis using spots from the control group against those from the target domain and select a gene with the smallest FDR-adjusted P value and higher expression in the control group as $gene_{0-}$. The expression of the meta gene is calculated as

$$\log(meta\_gene_1) = \log(gene_0) + \log(gene_{0+}) - \log(gene_{0-}) + C_0,$$

where C_0 is a constant to make $\log(meta\_gene_1)$ non-negative. The log transformation is used to rescale expression and make the expression levels comparable across different genes. We have found that including negative genes can strengthen the spatial expression pattern for domains that do not have enriched positive marker genes. This algorithm can be used iteratively to find additional genes to form an updated meta gene with a clearer spatial pattern for the target domain. For the $(t+1)$th iteration, the meta gene expression is calculated as

$$\log(meta\_gene_{t+1}) = \log(meta\_gene_t) + \log(gene_{t+}) - \log(gene_{t-}) + C_t$$

In the $(t+1)$th iteration, after adding $gene_{t+}$ and subtracting $gene_{t-}$, SpaGCN will select the $(t+1)$th control group based on $meta\_gene_{t+1}$. The size of the new control group, which is the number of spots not in the target domain but have higher expression of $meta\_gene_{t+1}$ than spots in the target domain, should be smaller than the size of the tth control group, to ensure that $meta\_gene_{t+1}$ has a clearer spatial pattern than $meta\_gene_t$. Also, $meta\_gene_{t+1}$ is expected to have a larger difference of mean expression between the target and control groups than $meta\_gene_t$. Therefore, at each iteration, SpaGCN checks whether both criteria are met, and the search of additional genes will stop otherwise. An illustration of this iterative meta gene search is shown in Supplementary Fig. 26.

Evaluation of SVGs using Moran's *I* and Geary's *C* statistics. Gene expressions at different locations may not be independent. For example, the expression levels of a gene at nearby locations may be closer in value than expression levels at locations that are farther apart. This phenomenon is called spatial autocorrelation, which measures the correlation of a variable with itself through space. To evaluate whether the detected SVGs exhibit an organized spatial expression pattern, we used Moran's I and Geary's C, two commonly used statistics to quantify the degree of spatial autocorrelation of gene expression. Spatial autocorrelation can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another. Moran's I metric[27] is a correlation coefficient that measures the overall spatial autocorrelation of a dataset. Intuitively, for a given gene, it measures how one spot is similar to other spots surrounding it. If the spots are attracted (or repelled) by each other, it implies the spots are not independent. Thus, the presence of autocorrelation indicates the spatial pattern of gene expression. The Moran's I value ranges from −1 to 1, where a value close to 1 indicates a clear spatial pattern, a value close to 0 indicates random spatial expression, and a value close to −1 indicates a chess board-like pattern. To evaluate the spatial variability of a given gene, we calculate the Moran's I using the following formula,

$$I = \frac{N}{W}\frac{\sum_i\sum_j[w_{ij}(x_i-\bar{x})(x_j-\bar{x})]}{\sum_i(x_i-\bar{x})^2},$$

where x_i and x_i are the gene expression of spots i and j, $\bar{x}$ is the mean expression of the gene, N is the total number of spots, w_{ij} is spatial weight between spots i and j calculated using the 2D spatial coordinates of the spots, and W is the sum of w_{ij}. For each spot, we select the k nearest neighbors using spatial coordinates. Moran's I statistic is robust to the choice of k and is set at 4 in our analysis. We assign w_{ij}=1 if spot j is in the nearest neighbors of spot i, and w_{ij}=0 otherwise.

Geary's C is another commonly used statistic for measuring spatial autocorrelation. It is calculated as

$$C = \frac{N}{2W}\frac{\sum_i\sum_j[w_{ij}(x_i-x_j)^2]}{\sum_i(x_i-\bar{x})^2},$$

The value of Geary's C ranges from 0 to 2. To make it on the same scale as Moran's I, we convert it to the $[-1,1]$ scale by

$$C^* = 1 - C,$$

where 1 indicates perfect positive autocorrelation, 0 indicates no autocorrelation and −1 indicates perfect negative autocorrelation. For each gene, the values of Geary's C are similar but not identical to Moran's I.

Transferability of SpaGCN detected SVGs. A genuine SVG should show transferability, that is, the spatial expression pattern should be similar across different datasets collected from the same tissue type. To show the transferability of SpaGCN-detected SVGs, first, we show that the SVGs detected in one dataset also show similar spatial expression patterns in an independent dataset. Second, we show that SVGs detected in one dataset can be used to cluster spots in an independent dataset and achieve relatively high clustering accuracy.

Comparison with other methods. To evaluate the performance of SpaGCN in identifying spatial domains, we compare with Louvain, BaysSpace, stLearn and HMRF. We use the default parameter settings for all methods and the same number of clusters in clustering. To evaluate the performance of SpaGCN in detecting SVGs, we compare with SPARK and SpatialDE. We initially detect SVGs using their default parameter settings. By default, both SPARK and SpatialDE filter out spots with total UMI counts less than ten. In SPARK, genes expressed in fewer than 10% of the spots are filtered out. In SpatialDE, genes expressed in fewer than three spots are filtered out. To evaluate the impact of filtering criteria on SPARK and SpatialDE, we further eliminate genes expressed in fewer than 20% of the spots. The specificity of spatial expression patterns of detected SVGs is evaluated by Moran's I and Geary's C statistics.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors analyzed seven publicly available SRT datasets. The data were acquired from the following websites or accession numbers: (1) human primary pancreatic cancer ST data (GSE111672); (2) LIBD human dorsolateral prefrontal cortex, dorsolateral prefrontal cortex 10x Visium data (http://research.libd.org/spatialLIBD/); (3) mouse posterior brain 10x Visium data (https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior); (4) mouse cortex SLIDE-seqV2 data (https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2); (5) mouse visual cortex STARmap data (https://www.starmapresources.com/data); (6) mouse olfactory bulb ST data (https://drive.google.com/drive/folders/1C4l3lBaYl7uuV2AA2o0WDzO_mkc_b0pv?usp=sharing); (7) mouse hypothalamus MERFISH data (https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248). Details of the datasets analyzed in this paper are described in Supplementary Table 1.

Code availability

An open-source implementation of the SpaGCN algorithm can be downloaded from https://github.com/jianhuupenn/SpaGCN.

References

35. Partel, G. & Wahlby, C. Spage2vec: Unsupervised representation of localized spatial gene expression signatures. *FEBS J.* **288**, 1859–1870 (2021).
36. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *Proc. International Conference on Learning Representations*. arXiv:1609.02907 (2017).
37. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretaion and validation of cluster analysis. *Computational Appl. Math.* **20**, 53–65 (1987).
38. Lakkis, J. et al. A joint deep learning model enables simultaneous batch effect correction, denoising and clustering in single-cell transcriptomics. *Genome Res.* https://doi.org/10.1101/gr.271874.120 (2021).
39. Li, X. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 2338 (2020).

Acknowledgements

This work was supported by the following grants: R01GM125301, R01EY030192, R01EY031209, R01HL113147 and R01HL150359 (to M.L.), and P01AG066597 (to D.J.I. and E.B.L.). We thank R. Moncada and I. Yanai for sharing the human pancreatic cancer histology image data, and R. Stickles, E. Murray, E. Macosko and F. Chen for sharing the SLIDE-seqV2 data.

Author contributions

This study was conceived of and led by M.L. J.H. designed the model and algorithm. J.H. implemented the SpaGCN software and led the data analysis with input from M.L., X.L., K.C., A.S., N.M., D.I., E.L. and R.T.S. N.M. contributed to figure design and generation. J.H. and M.L. wrote the paper with feedback from all other coauthors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01255-8.

Correspondence and requests for materials should be addressed to Jian Hu or Mingyao Li.

Peer review information *Nature Methods* thanks Andrew Jaffe, Kristen Maynard and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Lin Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

nature research

Corresponding author(s): Mingyao Li, Jian Hu

Last updated by author(s): Jul 6, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	SpaGCN v1.0.0 (https://github.com/jianhuupenn/SpaGCN), stLearn v0.3.1(https://stlearn.readthedocs.io/en/latest/), BayesSpace v1.0.0 (https://github.com/edward130603/BayesSpace), HMRF v1.3.3 (https://bitbucket.org/qzhudfci/smfishhmrf-py/src/master/), spage2vec(https://github.com/wahlby-lab/spage2vec) were used for spatial domain detection. SpaGCN v1.0.0, SpatialDE v1.1.3 (https://github.com/Teichlab/SpatialDE) and SPARK v1.0.2 (https://github.com/xzhoulab/SPARK) were used for spatially variable gene detection. Scany v1.5.1 was used for data per-processing, Louvain's clustering, differential expression analysis and t-SNE visualization. Sklearn v0.22.1 was used for K-mean's clusterinng.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We analyzed multiple spatial transcriptomics datasets. Publicly available data were acquired from the following websites or accession numbers:
(1) Human primary pancreatic cancer data (Moncada et al. 2020).
Figure 2. The count matrix and spatial data can be downloaded from GEO (accession GSE111672).

(2) LIBD human dorsolateral prefrontal cortex (Maynard et al. 2021).
Figure 3. The count matrix and spatial data can be downloaded from http://research.libd.org/spatialLIBD/;
(3) Mouse posterior brain (10X Genomics).
Figures 4, 5. The count matrix and spatial data can be downloaded from https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Posterior;
(4) Mouse cortex SLIDE-seqV2 data (Stickels, R.R., et al. 2020).
Supplementary Note 2. The count matrix and spatial data can be downloaded from https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2#study-summary;
(5) Mouse visual cortex STARmap data (Wang, X., et al. 2018).
Figure 6. The count matrix and spatial data can be downloaded from https://www.dropbox.com/sh/f7ebheru1lbz91s/AADm6D54GSEFXB1feRy6OSASa/visual_1020/20180505_BY3_1kgenes?dl=0&subfolder_nav_tracking=1
(6) Mouse olfactory bulb (Ståhl et al. 2016).
Supplementary Note 1. The count matrix and spatial data can be downloaded from https://drive.google.com/drive/folders/1C4l3lBaYl7uuV2AA2o0WDzO_mkc_b0pv?usp=sharing
(7) MERFISH mouse hypothalamus data (Moffitt et al. 2018).
Supplementary Note 3. The count matrix and spatial data can be downloaded from GEO (accession GSE71585).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No data collection was involved in the present study. We analyzed 7 publicly available spatially resolved transcriptomics datasets, including two datasets generated by Spatial Transcriptomics, two datasets generated by 10x Visium, one dataset generated by SLIDE-seqV2, one dataset generated by STARmap, and one dataset generated by MERFISH. These 7 datasets represent data generated from a wide range of platforms and show that SpaGCN is compatible with different types of spatially resolved transcriptomics data. No statistical methods were used to predetermine the sample size or the number of datasets.
Data exclusions	All spots and genes were used, no exclusion was done prior to analysis.
Replication	We did not perform replication. Instead, we confirmed our findings by comparing to other published molecular biology results.
Randomization	Not relevant to this study since each tissue section was analyzed separately. The analyses reported in this paper do not involve between sample comparisons.
Blinding	Not relevant since no data collection was involved in the present study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging