SECE: accurate identification of spatial domain by incorporating global spatial proximity and local expression proximity

Yuanyuan Yu[1], Yao He[1*], Zhi Xie[1,2*]

[1] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

[2] Center for Precision Medicine, Sun Yat-sen University, Guangzhou, China

[*] Email: Z.X: xiezhi@gmail.com; Y.H: scheyao@hotmail.com

Abstract

Motivation: Accurate identification of spatial domains is essential for analyzing spatial transcriptomics data to elucidate tissue microenvironments and biological functions. Existing methods utilize either local or global spatial relationships between spots to aid domain segmentation. A method that can concurrently capture both local and global spatial information may improve identification of spatial domains.

Results: In this article, we propose SECE, a deep learning-based method that captures both local and global relationships among spots and aggregates their information using expression similarity and spatial similarity. We benchmarked SECE against eight state-of-the-art methods on six real spatial transcriptomics datasets spanning four different platforms. SECE consistently outperformed other methods in spatial domain identification accuracy. Moreover, SECE produced spatial embeddings that exhibited clearer patterns in low-dimensional visualizations and facilitated more accurate trajectory inference.

Availability and implementation: SECE is implemented and provided as a pip installable Python package which is available on GitHub https://github.com/xie-lab/SECE.

1 Introduction

Spatial transcriptomics (ST) captures gene expression profiles with spatial information, providing novel insights into tissue molecular heterogeneity. Applying ST technology plays a crucial role in identifying cell-cell interactions, signaling pathways within the tissue microenvironment, and has enabled groundbreaking discoveries across fields such as neuroscience (Maynard, et al., 2021), developmental biology (Wang, et al., 2022), and cancer biology (Hunter, et al., 2021). A variety of ST platforms have been developed with varying throughput and resolution. Image-based ST platforms, including STARmap (Wang, et al., 2018), seqFISH (Lubeck, et al., 2014; Shah, et al., 2016), seqFISH+ (Eng, et al., 2019), MERFISH (Moffitt, et al., 2018) and FISSEQ (Lee, et al., 2014), provide highly accurate gene expression measurement at single cell resolution but only for a limited number of targeted genes (Moses and Pachter, 2022). On the other hand, sequencing-based ST platforms, such as Spatial transcriptomics (Stahl, et al., 2016) and its commercial version 10X Genomics Visium, Slide-seqV2 (Stickels, et al., 2021), HDST (Vickovic, et al., 2019), Seq-Scope (Cho, et al., 2021) and Stereo-seq (Chen, et al., 2022), can perform high-throughput sequencing on a genome-wide scale with increasing spatial resolution. The spatial resolution of sequencing-based technology continues to improve, with Seq-Scope and Stereo-seq capable of merging subcellular spots based on cellular location to achieve single-cell resolution.

In ST data, spatial domains refer to regions exhibiting consistent patterns in both gene expression and physical location, each with specific anatomical structures (Maynard, et al., 2021;

Ortiz, et al., 2020). Accurately identifying spatial domains is crucial for various downstream analyses, including trajectory inference, cell type deconvolution, and cell-cell communications, as well as their biological interpretation. Spatial domains are distinct from cell types that have been extensively studied in single-cell data. Cell types can be obtained by clustering transcriptional information, and their spatial distribution patterns are uncertain. They may be spatially concentrated, such as excitatory neurons, or discretely distributed, such as astrocytes (Zeisel, et al., 2018). Spatial domains are continuous in space, so relying solely on gene expression is insufficient to capture them. Integrating spatial location information is necessary to ensure continuity of domains, which presents a new challenge.

Several methods have been developed to address this challenge. Many existing techniques utilize spatial location information to find neighboring spots for each spot and enhance the similarity between neighbors to ensure the spatial continuity of the domain. Among them, BayesSpace (Zhao, et al., 2021) and BASS (Li and Zhou, 2022) perform latent variable modeling of regional labels and use Bayesian methods for inference. SpaGCN (Hu, et al., 2021) and STAGATE (Dong and Zhang, 2022) employ graph convolutional networks and graph attention networks to aggregate neighbor information, respectively. GraphST (Long, et al., 2023) , SpaceFlow (Ren, et al., 2022) , and conST (Zong, et al., 2022) utilize self-supervised graph embedding learning strategies. However, spatial adjacency relationships only represent local information, neglecting to consider the global structure and patterns may result in a lack of comprehensive understanding of the data. In contrast, SpatialPCA (Shang and Zhou, 2022) introduces a spatially aware dimension reduction method, which leverages global spatial relationship by measuring pairwise spots similarity. Nonetheless, focusing solely on global

similarities may overlook subtle spatial details. Moreover, the simultaneous capture of both local and global information, harnessing the advantages of each, remains an area needing further exploration.

To this end, we developed SECE, an accurate method for identifying spatial domains in ST data by incorporating both global and local proximity. Global proximity is quantified by physical distance, and it is considered that the pairwise similarity between spots decreases with longer spatial distances. Local proximity, on the other hand, is determined by expression similarity, aggregating neighbor information based on the similarity of gene expression between spots. SECE utilize graph neural networks for spatial embedding (SE) learning, it aggregates local expression similarity through an attention mechanism while simultaneously constraining the global spatial similarity using a Gaussian kernel function. Subsequently, by performing clustering on the SE, we can derive the spatial domain to which each spot belongs. SE also facilitate downstream analyses like visualization and trajectory inference. We demonstrated SECE's versatility across diverse ST platforms, including high-resolution methods like STARmap/Slide-seqV2/Stereo-seq and lower-resolution platforms like Visium. SECE's accurate spatial representations in brain and tumor datasets highlight its ability to gain biological insights from complex ST data.

2 Materials and Methods

SECE is versatile for modeling ST data across resolutions including subcellular, single-cell, near-single-cell, and multicellular (Fig. 1A). It takes gene expression matrix and spatial coordinate matrix of ST data as input, and outputs spatial domains and embeddings for each spot. First, SECE uses an autoencoder (AE) module with count distribution assumption to compress the expression

matrix into expression features. Next, it converts spatial coordinates into local and global position relationships, storing them in the adjacency matrix (ADM) and spatial similarity matrix (SSM), respectively. Then, the graph attention network (GAT) module is utilized to balance the expression similarity of local neighbors and the global spatial similarity to obtain SE (Fig. 1B). Finally, spatial domains are identified by clustering SE using mclust (Scrucca, et al., 2016). Additionally, downstream analyses, including low-dimensional visualizations (McInnes, et al., 2018; Van der Maaten and Hinton, 2008) and trajectory inference (Cao, et al., 2019; Wolf, et al., 2019), are derived from SE (Fig. 1C).

2.1 Extracting expression features with AE module

Given ST data with N spots and G genes, the dimensions of gene expression matrix X and spatial coordinate matrix Y are $N \times G$ and $N \times 2$, respectively. The raw counts X are normalized by library size and then log-transformed to obtain the normalized expression matrix $\tilde{X}$. We first employ AE with zero-inflated negative binomial (ZINB) or negative binomial (NB) distribution (Eraslan, et al., 2019; Lopez, et al., 2018) to compress $\tilde{X}$ into low-dimensional features Z. Let x_{ng} denotes the count value of gene g in spot n, the likelihood function of x_{ng} under ZINB and NB distributions is given by:

$$ZINB\left(x_{ng};\pi_{ng},r_{ng},p_g\right) = \pi_{ng}\delta_0\left(x_{ng}\right) + (1-\pi_{ng})\frac{\Gamma\left(x_{ng}+r_{ng}\right)}{x_{ng}!\,\Gamma\left(r_{ng}\right)}{p_g}^{r_{ng}}(1-p_g)^{x_{ng}},$$

and

$$NB\left(x_{ng};r_{ng},p_g\right) = \frac{\Gamma\left(x_{ng}+r_{ng}\right)}{x_{ng}!\,\Gamma\left(r_{ng}\right)}{p_g}^{r_{ng}}(1-p_g)^{x_{ng}},$$

where δ_0 is the Dirac delta function, π_{ng} is the probability of true gene expression being 0, Γ is the gamma function, (r_{ng}, p_g) is the standard parametrization for the NB distribution.

The AE module takes $\tilde{X}$ as input, and outputs distribution parameters. The formulation is:

$$Z = f_e(\tilde{X})$$

$$Z' = f_{d1}(Z)$$

$$(\Pi, R, P) = f_{d2}(Z')$$

where f_e represents an encoder, while f_{d1} and f_{d2} constitute the decoder. Specifically, f_e consists of two nonlinear layers, each utilizing Rectified Linear Unit ($ReLU$) activation functions. These layers reduce the feature dimension G into m' and m, respectively, yielding the expression feature Z. Subsequently, f_{d1} decodes Z into Z' with a feature dimension of m'. f_{d2} comprises three output layers, which take Z' as input and output three parameter matrices (Π, R, P) of the ZINB distribution, each consisting of elements (π_{ng}, r_{ng}, p_g). The activation functions of these three output layers are exponential, sigmoid, and exponential functions respectively. f_{d2} employs two layers to learn (R, P) under the NB distribution assumption.

The goal is to minimize the reconstructed loss by minimizing the negative log-likelihood (NLL) function, that is, $Loss_{pre} = NLL_{ZINB}(\mathrm{X}; \Pi, \mathrm{R}, \mathrm{P})$ or $Loss_{pre} = NLL_{NB}(\mathrm{X}; \mathrm{R}, \mathrm{P})$. We employed ZINB for highly sparse data like Stereo-seq and Slide-seq, and NB for less sparse data including STARmap and Visium. Both are implemented in SECE package.

2.2 Capturing both local and global relationships

To incorporate physical location information, we construct an ADM and SSM from the spatial coordinates Y. ADM summarizes local spatial relationships by storing neighbors for each spot. This local neighborhood information is later employed to adaptively aggregate features in GAT module based on expression similarity. In contrast, SSM captures global spatial proximity by

providing a spatial similarity measure for all pairs of spots, not just neighbors. The SSM is subsequently utilized to constrain the global similarity of SE.

ADM A is a $N \times N$-dimensional symmetric matrix, where elements are assigned values of 1 or 0 to indicate neighboring spots or non-neighboring spots, respectively. More precisely, the element A_{ij} denotes the adjacent relationship between spot i and spot j:

$$A_{ij} = \begin{cases} 0, & v_i \in \mathcal{N}(j) \\ 1, & v_i \notin \mathcal{N}(j). \end{cases}$$

Here, $\mathcal{N}(j)$ represents the set of spatial neighbors of spot j, which can be determined based on coordinates Y by employing K-Nearest Neighbor (KNN) or applying a distance cutoff. By default, we utilize KNN, with the number of neighbors set to 6 for Visium datasets and 10 for other datasets.

SSM Σ is also a $N \times N$-dimensional symmetric matrix, wherein elements decrease as the distance between spots increases, exhibiting an exponential decay tendency. For spot i and spot j with coordinates y_i and y_j, the correspond element is expressed as:

$$\Sigma_{ij} = \exp\left(-\frac{\left|y_i - y_j\right|^2}{\gamma}\right),$$

where the bandwidth parameter γ controls the spatial influence. By default, γ is set as the 0.05 quantile distance. A larger γ results in a greater spatial influence.

2.3 Learning SE with GAT module

After capturing the expression features and extracting local and global position relationships, we employ the GAT module to learn SE, subsequently clustering SE to delineate spatial domains. The GAT module consists of two GAT layers.

We first introduce the GAT layer. It takes feature matrix and ADM as input, and outputs a new feature matrix after aggregating neighbor information. Let $H = (h_1, h_2, \dots, h_N)$ denotes the input feature matrix, which has dimensions $N \times m_0$, with N samples and m_0 features. The output of GAT layer is denoted as $H' = (h'_1, h'_2, \dots, h'_N)$ with dimensions $N \times m'_0$. The GAT layer performs aggregation for each sample adaptively based on the normalized attention scores. For sample j, the output feature h'_j can be formulated as

$$h'_j = \sigma\left(\sum_{v_i \in \mathcal{N}(j)} \alpha_{ij} W\, h_i\right),$$

where W is a weight matrix with dimensions $m'_0 \times m_0$, $\mathcal{N}(j)$ represents the set of neighboring samples of sample j, α_{ij} is the normalized attention coefficient matrix using the SoftMax function:

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})},$$

where $e_{ij} = a^T(Wh_i \parallel Wh_j)$, a is learnable vector and $\parallel$ is the concatenation operation. We used the Exponential Linear Unit (ELU) as activation function σ in the GAT layer.

The GAT module in SECE consists of two GAT layers. It takes expression features Z and ADM A as input, and outputs SE matrix U, which is a $N \times m$-dimensional matrix:

$$U = GAT_2(GAT_1(Z, A), A).$$

During neighbor information aggregation using GAT, local expression similarities are captured via attention and adaptively aggregated. To preserve as much information as possible in expression features, the local learning target is the reconstruction loss $L_{local} = MSE(U, Z)$. We further constrain pairwise correlation using SSM that including global information, that is, the correlation of each SE at N positions UU^T is close to Σ, $L_{global} = MSE(UU^T, \Sigma)$. The objective function balances the two similarities by λ_{global} and λ_{local}:

$$Loss = \lambda_{global} * L_{global} + \lambda_{local} * L_{local}.$$

We kept $\lambda_{local} = 1$ unchanged and adjusted λ_{global} based on the ST platform. After obtaining the final SE U, we utilized the clustering method mclust (Scrucca, et al., 2016) to cluster the U and determine the spatial domain for each spot.

2.4 Architecture of SECE

In the AE module, the dimension of the hidden layer m' and bottleneck layer m are 128 and 32, respectively. Adaptive Moment Estimation (Adam) optimizer is used to minimize $Loss_{pre}$ with a learning rate of 1e-3 and dropout rate of 0.1. For the GAT module, the dimensions of the two GAT layers are both 32. The Adam optimizer is employed to minimize $Loss$ with a learning rate of 1e-2 and dropout rate of 0.2. The default number of iterations for the AE and GAT modules are set to 40 and 50, respectively.

The hyperparameters λ_{global} and λ_{local} control the contributions of global and local similarities, respectively, and a larger λ_{global} giving greater global influence. Each ST platform has different resolutions and fields of view. For example, spots of ST arrays contain dozens of cells, while Stereo-seq only contain single cell. Stereo-seq can sequence the hemibrain while

STARmap can only detect a minor region of the visual cortex. We choose smaller λ_{global} value for platforms with more spots and higher resolution. Specifically, for Stereo-seq and Slide-seqV2 data including over 10,000 spots with approximately single-cell resolution, we used $\lambda_{global} = 0.08$. For Visium data with several thousand spots, λ_{global} was set to 0.3. For STARmap data with only 1207 cells, we set λ_{global} to 2. Users could also select λ_{global} following this standard.

3 Results

3.1 Application to STARmap data

We first tested SECE on the mouse visual cortex data generated by STARmap (Wang, et al., 2018) with ground truth layers. The 1,207 cells were divided into 7 layers, including Layer(L)1, L2/3, L4, L5, L6, as well as the corpus callosum (CC) and hippocampus (HPC) (Fig. 2A).

First, we compared the spatial domain identification results of SECE with eight existing methods, including BASS, SpaceFlow, GraphST, STAGATE, SpatialPCA, SpaGCN, BayesSpace, and conST (Fig. 2B, C). The consistency of spatial domains with the ground truth layers were evaluated using Adjusted Rand Index (ARI) and Accuracy (ACC) based on Kuhn-Munkres (Munkres, 1957). SECE achieved the highest consistency, with an ARI value of 0.65 and an ACC value of 0.79 (Fig. 2D). It was closely followed by BASS with ARI of 0.63, however, BASS incorrectly combined HPC and L5 (domain 5), which prevents the utilization of the Kuhn-Munkres algorithm to reassign clusters and compute the accuracy (ACC). Furthermore, other algorithms also exhibited weak spatial aggregation, as indicated by the Local inverse Simpson's index (LISI) (Korsunsky, et al., 2019) (Supplementary Figure S1A). Additionally, these methods mistakenly classified aggregated endothelial cells (Endo) in L1 and L2/3 into the

same domain, such as domain 2 in SpaceFlow, domain 3 in STAGATE and BayesSpace, domain 5 in GraphST and SpatialPCA, and domain in SpaGCN (Fig. 2C, Supplementary Figure S1B).

Next, we compared SE of SECE with SpaceFlow, GraphST, STAGATE, and SpatialPCA. Due to the difficulty of conST in generating clear spatial domains, its SE learning comparison was excluded. SE of SECE explained the ground truth layers most effectively (ASW=0.16), followed by SpaceFlow (ASW=0.12), GraphST (ASW=0.07), STAGATE (ASW=0.07), and SpatialPCA (ASW=0.05) (Fig. 2D, right). Moreover, SECE had a clearer and continuous pattern in UMAP visualization compared to the other methods (Fig. 2E). When comparing trajectory inference results, we selected the cortex part in ground truth, namely L1, L2/3, L4, L5 and L6. The PAGA(Wolf, et al., 2019) analysis based on SECE revealed a linear and continuous relationship between these layers (Fig. 2F). Conversely, SpaceFlow mistakenly made the connection between L1 versus L4 and L5, while the patterns in the other three methods were more unclear. For individual cells, SECE exhibited a sequential increase of pseudo-time from L6 to L1 (Supplementary Figure S1C, D) based on Monocle3(Cao, et al., 2019), and the similar trend was also observed in SpaceFlow. However, GraphST, STAGATE, and SpatialPCA ordered L1 and L2/3 incorrectly (Supplementary Figure S1E, F).

3.2 Application to Slide-seqV2 data

We further tested SECE on the hippocampal dataset generated by Slide-seqV2 platform (Stickels, et al., 2021). SECE identified 14 distinct domains, and we annotated them according to the known structure of Allen Brain Atlas (ABA) (Fig. 3A). The domains were hippocampus (Cornu Ammonis (CA)1, CA2, CA3, Dentate gyrus (DG), and CA slm/so/sr), cortex (Layers 4, 5a, 5b, 6),

third ventricle, CC, and three subregions of the thalamus (Fig. 3B, left). We further meticulously verified the subtle spatial domains. For example, the four important components of the hippocampal region, CA1, CA2, CA3, and DG, were clearly demarcated, as evidenced by the high expression of their known markers *Wsf1*, *Rgs14*, *Nptxr* and *C1ql2* (Saunders, et al., 2018), respectively (Fig. 3C, Supplementary Figure S2A). Layer 5 in cortical regions was identified as two sublayers, Layer 5a and 5b, with different gene expression levels (Supplementary Figure S2A, B). The composition of cell types in each domain further supported the delineation (Supplementary Figure S3).

For comparison, we evaluated the performance of existing methods for spatial domain identification (Fig. 3D). Notably, SECE was the only approach that could accurately detect subregions in both the cortex and CA. Specifically, for cortical areas, BASS failed to distinguish between Layer 4 and Layer 5, SpaceFlow mixed Layer 5 and Layer 6, SpatialPCA exhibited suboptimal division smoothness, and the remaining algorithms were unable to generate clear cortical regions. In the case of CA, except for GraphST, none of the methods succeeded in identifying the CA2 region. Furthermore, SECE exhibited the highest spatial aggregation performance as it had the smallest LISI values (Supplementary Figure S4A).

We also compared the performance of SE. UMAP generated from SECE clearly displayed clustering patterns for the hippocampus, cortex, thalamus, third ventricle, and their subregions (Fig. 3B, right). In addition, the sublayers of them like Layer 4 and Layer 6 were arranged in a sequence, while UMAP of STAGATE and GraphST missed them (Supplementary Figure S4B). Moreover, we conducted trajectory inference for the cortex due to its continuous relationship between sublayers. We selected the domains corresponding to the cortex in each method and

started with the deepest clusters, such as Layer 6 for SECE, domain 1 for SpaceFlow, domain 2 for GraphST, and domain 3 for SpatialPCA and STAGATE (Fig. 3E, F). For SECE, the pseudo-time consistently increased with decreasing cortical depth, clearly capturing the pseudo-time relationship between different layers. In contrast, SpaceFlow exhibited less distinguishable differences between the learned layers. SpatialPCA reversed the relationship between Layer 5b and Layer 6, while GraphST and STAGATE failed to identify the relationship within the cortical region.

3.3 Application to Stereo-seq data

In this section, we assessed on the mouse olfactory bulb (Fu, et al., 2021) Stereo-seq(Chen, et al., 2022) data. Spatial domains identified by SECE were annotated based on known olfactory bulb layers, including rostral migratory stream (RMS), granule cell layer (GCL), inner plexiform layer (IPL), mitral cell layer (MCL), external plexiform layer (EPL), glomerular layer (GL) and olfactory nerve layer (ONL) from inside to outside (Fig. 4A; Supplementary Figure S5A). These domains were validated using known markers (Saunders, et al., 2018) for each layer (Fig. 4B). Notably, besides the known seven layers, we made a finer division of GCL and ONL, and the sublayers were named GCL-Inner, GCL-Outer, ONL-Inner and ONL-Outer, respectively. Several evidences confirmed these sublayers. First, marker gene expression differed between them, where the GCL-inner highly expressed *Nrgn* and the GCL-outer highly expressed *Pcp4*. Markers in the ONL-outer also exhibited higher expression levels compared to those in the ONL-inner. Second, sublayers exhibited distinct cell type composition (Supplementary Figure S5B to D). GCL-Outer was almost composed of GC, while GCL-Inner contained a certain amount of Oligo, and the GC

subtypes in GCL-Inner and GCL-Outer also differed. Moreover, the ONL-Outer contained almost exclusively OEC, while the ONL-Inner contained a fraction of OEC and more Astro. This supported previous findings that the ONL, as a part of the olfactory bulb blood-brain barrier, had fine internal and external subregions with different cell types (Beiersdorfer, et al., 2020). Our study provided further support for this finding at the spatially resolved single-cell level.

We also compared the performance of SECE to the existing methods (Fig. 4C). BASS failed to separate the RMS from the GCL layer and could not distinguish GL from EPL. It also identified some irrelevant domains with few cells, such as domain5, 8, and 9. STAGATE and SpaGCN mixed GL and EPL. SpaceFlow, GraphST, and SpatialPCA encountered challenges in accurately dividing the GCL layer. BayesSpace and conST did not yield a clear domain identification. Additionally, BASS had the most spatial aggregation performance with lowest LISI value, while SECE ranked second (Supplementary Figure S5E). Furthermore, SECE showed robustness when tested with different number of clusters (7, 8, and 10), consistently providing well-bounded ring stratification (Supplementary Figure S6). Moreover, we assessed the SE learning capabilities. The UMAP visualization based on SECE exhibited a continuous low-dimensional pattern with nine layers arranged in order of spatial position from inside to outside (Fig. 4D). The trajectory inference for these layers demonstrated approximately linear relationship (Fig. 4E). SpatialPCA also displayed linear patterns except for domains 6 and 7, while results of SpaceFlow, GraphST, and STAGATE exhibited many false positive connections between domains.

We further applied SECE onto mouse hemibrain dataset with a more complex anatomic structure (Chen, et al., 2022). SECE achieved spatial domain annotations that were in highly consistent with ABA anatomy (Fig. 5A, B, left). We could clearly separate different regions,

including cortical regions, hippocampal regions, midbrain regions, thalamic regions, and fiber tracts (FT). The cortex included five layers: L1, L2/3, L4, L5, L6, LVC and CAA, which were supported by their cell type composition (Supplementary Figure S7, 8). In contrast, other methods failed to accurately identify these cortical layers (Fig. 5C, Supplementary Figure S9). Specifically, all of them encountered difficulties in identifying Cortex L5 with high smoothness. In addition, LISI showed that spatial domains identified by SECE had the strongest spatial pattern (Supplementary Figure S9A). Furthermore, low-dimensional visualization of SE using t-SNE (Van der Maaten and Hinton, 2008) showed the effect of aggregation in the same region and separation in different regions (Fig. 5B, right).

3.4 Application to Visium data

Finally, we tested the applicability of SECE on Visium dataset. The breast cancer dataset was divided into four phenotypic regions according to pathological images: ductal carcinoma in situ/lobular carcinoma in situ (DCIS/LCIS), healthy tissue (Healthy), invasive ductal carcinoma (IDC), and tumor surrounding regions with low features of malignancy (Tumor edge), as well as 20 annotated subdivisions (Fu, et al., 2021) (Fig. 6A; Supplementary Figure S10A).

We initially segmented 20 domains using each algorithm and performed phenotype annotations, which were then compared with image-based manual annotations (Fig. 6B; Supplementary Figure S10B, C). We found that SECE divided individual tumor regions into wrapped layers more clearly. Specifically, cluster 20 and cluster 15 divided the IDC-5 into internal and external layers, while cluster 11 and cluster 10 split the DCIS/LCIS-1. (Fig. 6B; Supplementary Figure S10A). To investigate the biological significance of the refined

stratification, we named clusters 20,15,11 and 10 as IDC-inner, IDC-outer, DCIS/LCIS-inner and DCIS/LCIS-outer, respectively. We also focused on the parts of clusters 4 and 5 that were located near the tumor edge, referred to as IDC-edge and DCIS/LCIS-edge. (Fig. 6C). We analyzed the cell type composition of each spot in these domains by integrating the annotated breast cancer scRNA-seq data (Wu, et al., 2021) and deconvoluting each spot using cell2location (Kleshchevnikov, et al., 2022) (Supplementary Figure S11, S12). The proportion of tumor cells gradually decreased as the tumor moves towards the edge, while the that of immune cells and stromal cells gradually increased (Fig. 6D), confirming the correctness of subregions.

We further explored the characteristics of these subregions (-edge, -outer and -inner) of DCIS/LCIS and IDC (Fig. 6E). Differential expression analysis(Hao, et al., 2021) showed that DCIS/LCIS-edge had a high level of humoral immunity, which could be confirmed by the enrichment of B cell receptor signaling pathway related genes (*IGLC2*, *IGLC3*, *IGHG1*, *IGHG3*, *IGHA1*). DCIS/LCIS-outer overexpressed *KRT14* of the keratin family, which was a key regulator of metastasis, suggesting invasive potential (Bilandzic, et al., 2019; Cheung, et al., 2016). In contrast, DCIS/LCIS-inner showed non-metastatic traits but also had tumor-promoting abilities, indicated by a high expression of *LDHA* (Martinez-Ordonez, et al., 2021). For the IDC subregions, IDC-edge showed elevated expression of biomarkers linked to tumor proliferation, invasion, and migration including tumor-associated macrophage (*APOE*), complement components (*C1qA*, *C1qB*), cathepsin (*CTSD*), and apolipoprotein (*APOC1*) (Nalio Ramos, et al., 2022; Revel, et al., 2022; Seo, et al., 2022; Zhang, et al., 2018; Zhang, et al., 2022). IDC-outer had a high level of immunity and some transferability, derived from the high expression of MHC class I-related genes (*HLA-A*, *HLA-B*, *B2M*) (Noblejas-Lopez, et al., 2019; Nomura, et al., 2014). Besides, there were

increased expression of genes such as *MGST1* and *MRPS30-DT*, which were known to promote breast carcinoma cell growth and metastasis (Wu, et al., 2019; Zeng, et al., 2020). In IDC-inner, there were higher levels of tumor activity and lower levels of immune response. *LINC00052*, known to promote breast cancer cell proliferation by increasing signals of epidermal growth factor receptor (EGFR) such as *HER3* (Huang, et al., 2022; Salameh, et al., 2017; Xiong, et al., 2021), was overexpressed. Upregulation of *COX6C* and *FAM234B* implied higher levels of cellular respiration (Grzybowska-Szatkowska and Slaska, 2014) and lower immune response function (Lyu, et al., 2020), respectively.

Furthermore, we inferred developmental trajectories for the IDC and DCIS/LCIS regions. The starting points for calculating pseudo-time in IDC and DCIS/LCIS were the interior of the tumor regions (Fig. 6F). There were increased pseudo-time values as the region of the tumor moves outward, effectively mimicking the gradual progression of tumor development.

Moreover, to further evaluate the power of SECE in spatial domain identification, we tested 12 Human dorsolateral prefrontal cortex (DLPFC) datasets generated from the Visium platform. The original study (Maynard, et al., 2021) had manually annotated the spatial domains of these datasets, encompassing white matter and six cortical layers (Supplementary Figure S13, S14). Spatial domain identified by SECE exhibited the highest agreement with the original annotations. The median ARI for SECE was 0.58, surpassing the second and third ranked algorithms, STAGATE and SpatialPCA, which achieved median ARI values of 0.55 and 0.54, respectively. These findings highlighted the superior performance of SECE in accurately delineating the spatial domains within the low-resolution data.

4 Discussion

SECE captures both local and global relationships among spots and aggregates their information using expression similarity and spatial similarity, respectively. This approach enables precise spatial domain division and facilitates interpretable spatial embedding learning across diverse ST datasets. Moreover, The AE module that explicitly models gene expression counts enhances SECE's ability to handle noisy data. Current methods use scaled values of highly variable genes (HVG) (e.g., GraphST, SpatialPCA, STAGATE, SpaceFlow) or employ dimensionality reduction techniques like principal component analysis (PCA) (e.g., BASS, SpaGCN, BayesSpace, conST) for expression features.

With the increasing captured area of ST data and advancements in resolution, there is a growing demand for computational methods to exhibit higher efficiency and scalability. We recorded the runtime and GPU memory consumption for each dataset (Supplementary Figure S15). For Slide-seqV2 hippocampus data and the Stereo-seq hemibrain data, which contained over 50,000 cells, SECE achieved a running time of less than 4.2 minutes while utilizing less than 5GB of GPU memory. These results demonstrated the superior computational efficiency and scalability of SECE when dealing with large-scale datasets.

While SECE has demonstrated notable performance, there are still several aspects that can be further enhanced. Firstly, we employ a pre-defined SSM to characterize global similarity, but exploring more flexible global correlation patterns could be advantageous. Secondly, we only utilized ST data as input, but incorporating matching histology data may provide additional benefits (Hu, et al., 2021). Although matching histology image data is currently only available on specific platforms like Visium, we can still leverage such images as optional supplementary

information when available. Finally, integrating single-cell data with ST data can enhance the data quality of the latter, increase the throughput, or reduce the noise in the gene expression (Biancalani, et al., 2021). Therefore, incorporating single-cell data is another approach to improve spatial representation capabilities.

Conflict of interest

None declared.

Funding

This project was supported by National Key R&D Program of China (2019YFA0904400, Z.X.) and Guangzhou Science and Technology Project (202201020336, Z.X.).

Data availability

The datasets are available at Zenodo https://zenodo.org/record/8130682.

References

Beiersdorfer, A., *et al*. Sublamina-specific organization of the blood brain barrier in the mouse olfactory nerve layer. *Glia* 2020;68(3):631-645.

Biancalani, T., *et al*. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* 2021;18(11):1352-1362.

Bilandzic, M., *et al*. Keratin-14 (KRT14) Positive Leader Cells Mediate Mesothelial Clearance and Invasion by Ovarian Cancer Cells. *Cancers (Basel)* 2019;11(9).

Cao, J., *et al*. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;566(7745):496-502.

Chen, A., *et al*. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;185(10):1777-1792 e1721.

Cheung, K.J., *et al*. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc Natl Acad Sci U S A* 2016;113(7):E854-863.

Cho, C.S., *et al*. Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* 2021;184(13):3559-3572 e3522.

Dong, K. and Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 2022;13(1):1739.

Eng, C.L., *et al*. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 2019;568(7751):235-239.

Eraslan, G., *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;10(1):390.

Fu, H., *et al*. Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics. *bioRxiv* 2021:2021.2006.2015.448542.

Fu, H., *et al*. Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics. *bioRxiv* 2021.

Grzybowska-Szatkowska, L. and Slaska, B. Mitochondrial NADH dehydrogenase polymorphisms are associated with breast cancer in Poland. *J Appl Genet* 2014;55(2):173-181.

Hao, Y., *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184(13):3573-3587 e3529.

Hu, J., *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;18(11):1342-1351.

Huang, X., *et al.* Long Non-Coding RNA LINC00052 Targets miR-548p/Notch2/Pyk2 to Modulate Tumor Budding and Metastasis of Human Breast Cancer. *Biochem Genet* 2022.

Hunter, M.V., *et al*. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat Commun* 2021;12(1):6278.

Kleshchevnikov, V., *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* 2022;40(5):661-671.

Korsunsky, I., *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;16(12):1289-1296.

Lee, J.H., *et al*. Highly multiplexed subcellular RNA sequencing in situ. *Science* 2014;343(6177):1360-1363.

Li, Z. and 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(1):168.

Long, Y., *et al*. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat Commun* 2023;14(1):1155.

Lopez, R., *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;15(12):1053-1058.

Lubeck, E., *et al.* Single-cell in situ RNA profiling by sequential hybridization. *Nat Methods* 2014;11(4):360-361.

Lyu, L., *et al*. Overexpression of FAM234B Predicts Poor Prognosis in Patients with Luminal Breast Cancer. *Cancer Manag Res* 2020;12:12457-12471.

Martinez-Ordonez, A., *et al*. POU1F1 transcription factor induces metabolic reprogramming and breast cancer progression via LDHA regulation. *Oncogene* 2021;40(15):2725-2740.

Maynard, K.R., *et al*. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;24(3):425-436.

McInnes, L., Healy, J. and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In.; 2018. p. arXiv:1802.03426.

Moffitt, J.R., *et al*. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 2018;362(6416).

Moses, L. and Pachter, L. Museum of spatial transcriptomics. *Nat Methods* 2022;19(5):534-546.

Munkres, J. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics* 1957;5(1):32-38.

Nalio Ramos, R., *et al.* Tissue-resident FOLR2(+) macrophages associate with CD8(+) T cell infiltration in human breast cancer. *Cell* 2022;185(7):1189-1207 e1125.

Noblejas-Lopez, M.D.M., *et al.* Expression of MHC class I, HLA-A and HLA-B identifies immune-activated breast tumors with favorable outcome. *Oncoimmunology* 2019;8(10):e1629780.

Nomura, T., *et al.* β2-Microglobulin-mediated signaling as a target for cancer therapy. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)* 2014;14(3):343-352.

Ortiz, C., *et al.* Molecular atlas of the adult mouse brain. *Sci Adv* 2020;6(26):eabb3446.

Ren, H., *et al.* Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nat Commun* 2022;13(1):4076.

Revel, M., *et al.* C1q+ macrophages: passengers or drivers of cancer progression. *Trends Cancer* 2022;8(7):517-526.

Salameh, A., *et al.* HER3 and LINC00052 interplay promotes tumor growth in breast cancer. *Oncotarget* 2017;8(4):6526-6539.

Saunders, A., *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* 2018;174(4):1015-1030.e1016.

Scrucca, L., *et al.* mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R j* 2016;8(1):289-317.

Seo, S.U., *et al.* Cathepsin D as a potential therapeutic target to enhance anticancer drug-induced apoptosis via RNF183-mediated destabilization of Bcl-xL in cancer cells. *Cell Death Dis* 2022;13(2):115.

Shah, S., *et al.* In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* 2016;92(2):342-357.

Shang, L. and Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nat Commun* 2022;13(1):7203.

Stahl, P.L., *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;353(6294):78-82.

Stickels, R.R., *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* 2021;39(3):313-319.

Van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* 2008;9(11).

Vickovic, S., *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;16(10):987-990.

Wang, M., *et al.* High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and larvae. *Dev Cell* 2022;57(10):1271-1283 e1274.

Wang, X., *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;361(6400).

Wolf, F.A., *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* 2019;20.

Wu, B., *et al.* MRPS30-DT Knockdown Inhibits Breast Cancer Progression by Targeting Jab1/Cops5. *Front Oncol* 2019;9:1170.

Wu, S.Z., *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* 2021;53(9):1334-1347.

Xiong, D., Wang, D. and Chen, Y. Role of the long non-coding RNA LINC00052 in tumors. *Oncol Lett* 2021;21(4):316.

Zeisel, A., *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* 2018;174(4):999-1014 e1022.

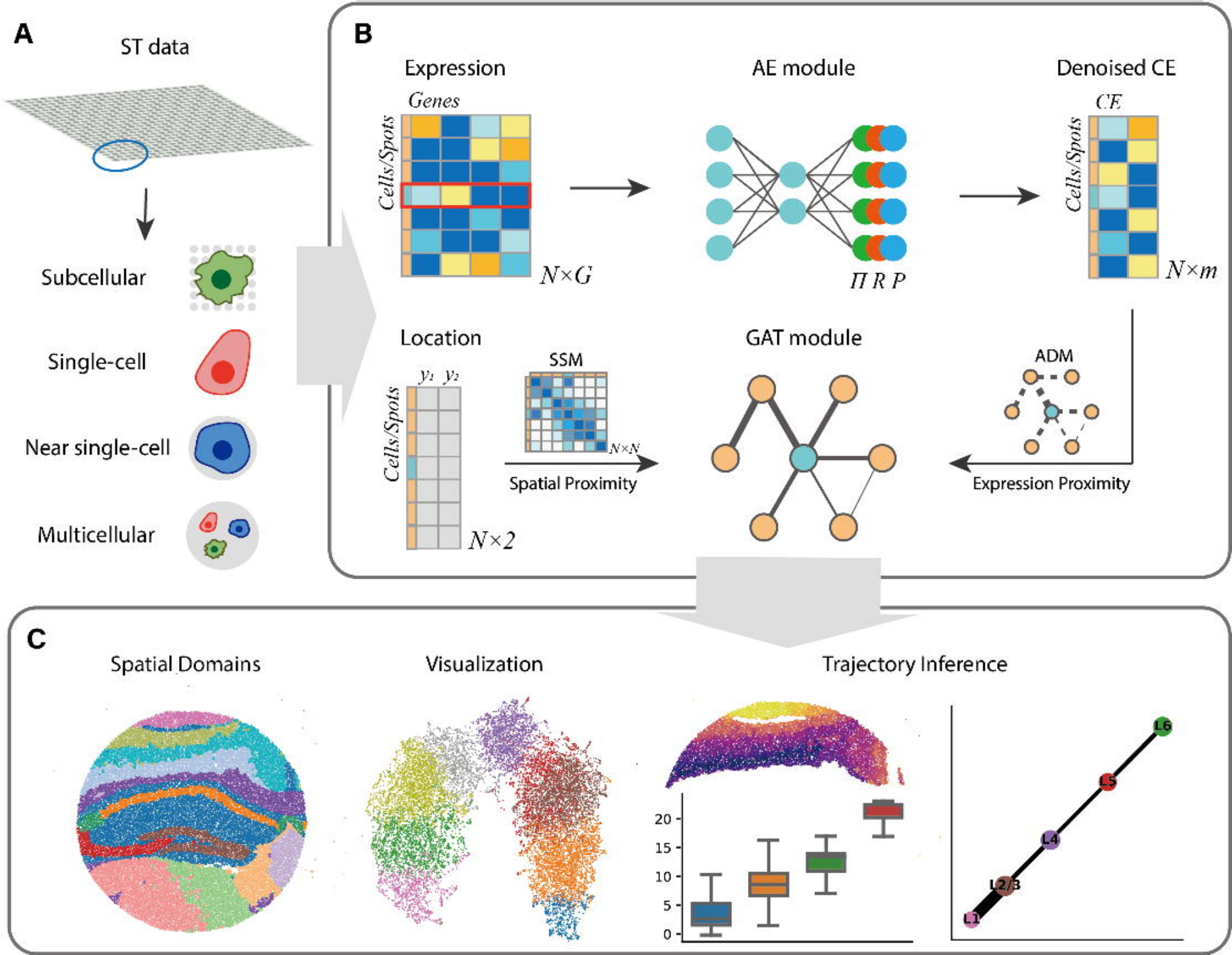
Zeng, B., *et al*. Knockdown of microsomal glutathione S-transferase 1 inhibits lung adenocarcinoma cell proliferation and induces apoptosis. *Biomed Pharmacother* 2020;121:109562.

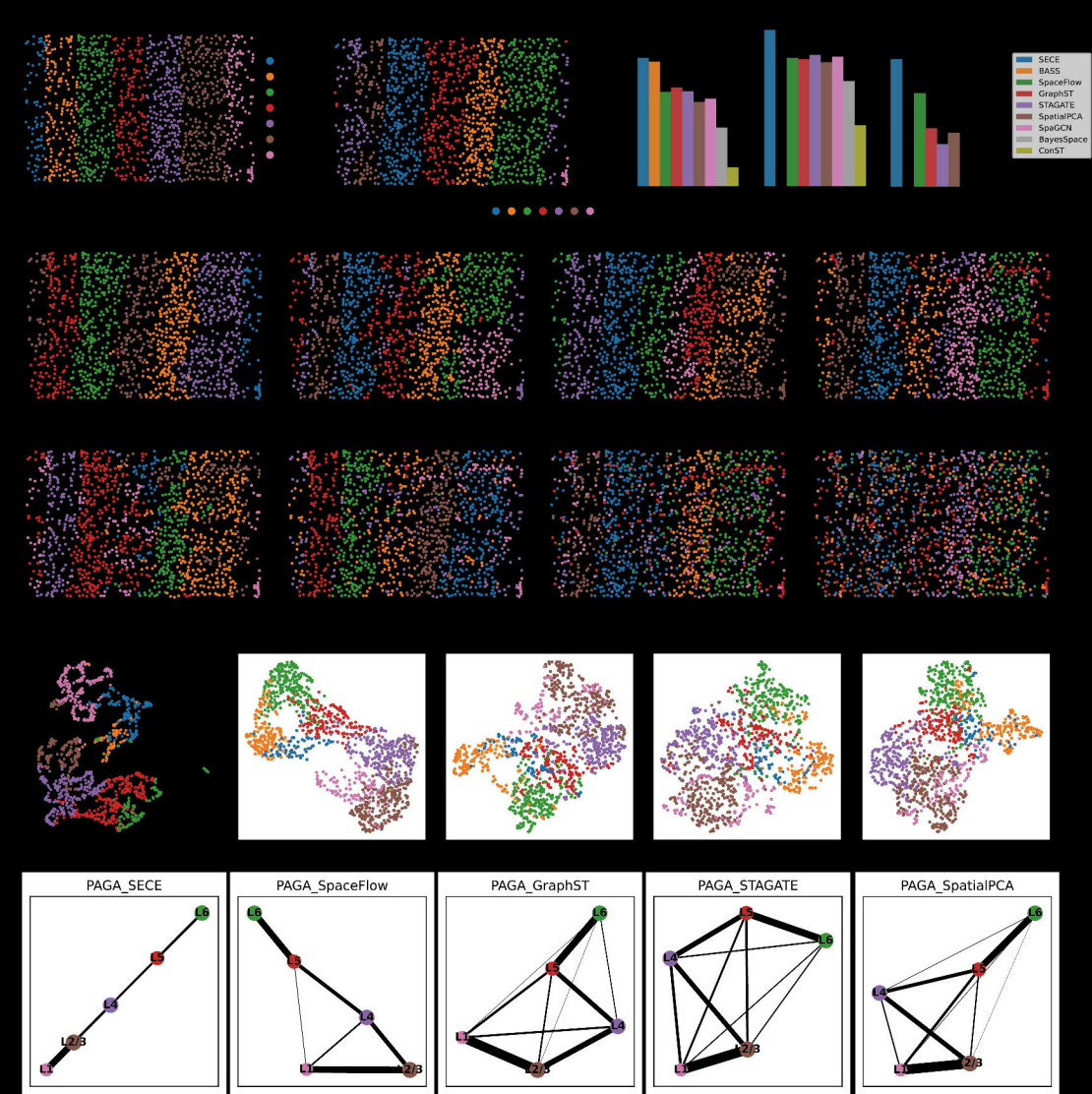
Zhang, C., Zhang, M. and Song, S. Cathepsin D enhances breast cancer invasion and metastasis through promoting hepsin ubiquitin-proteasome degradation. *Cancer Lett* 2018;438:105-115.

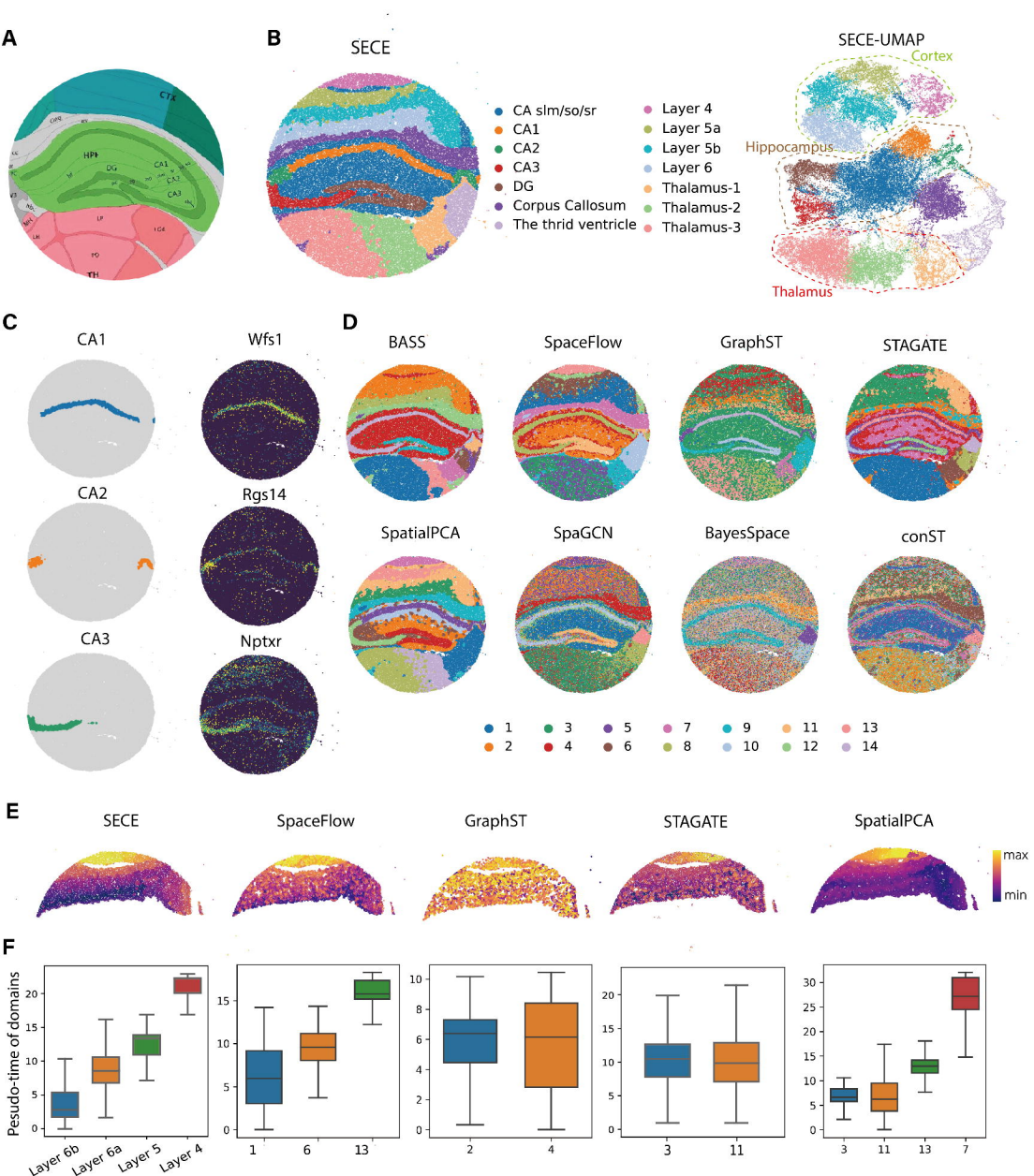
Zhang, H., *et al*. The Apolipoprotein C1 is involved in breast cancer progression via EMT and MAPK/JNK pathway. *Pathol Res Pract* 2022;229:153746.

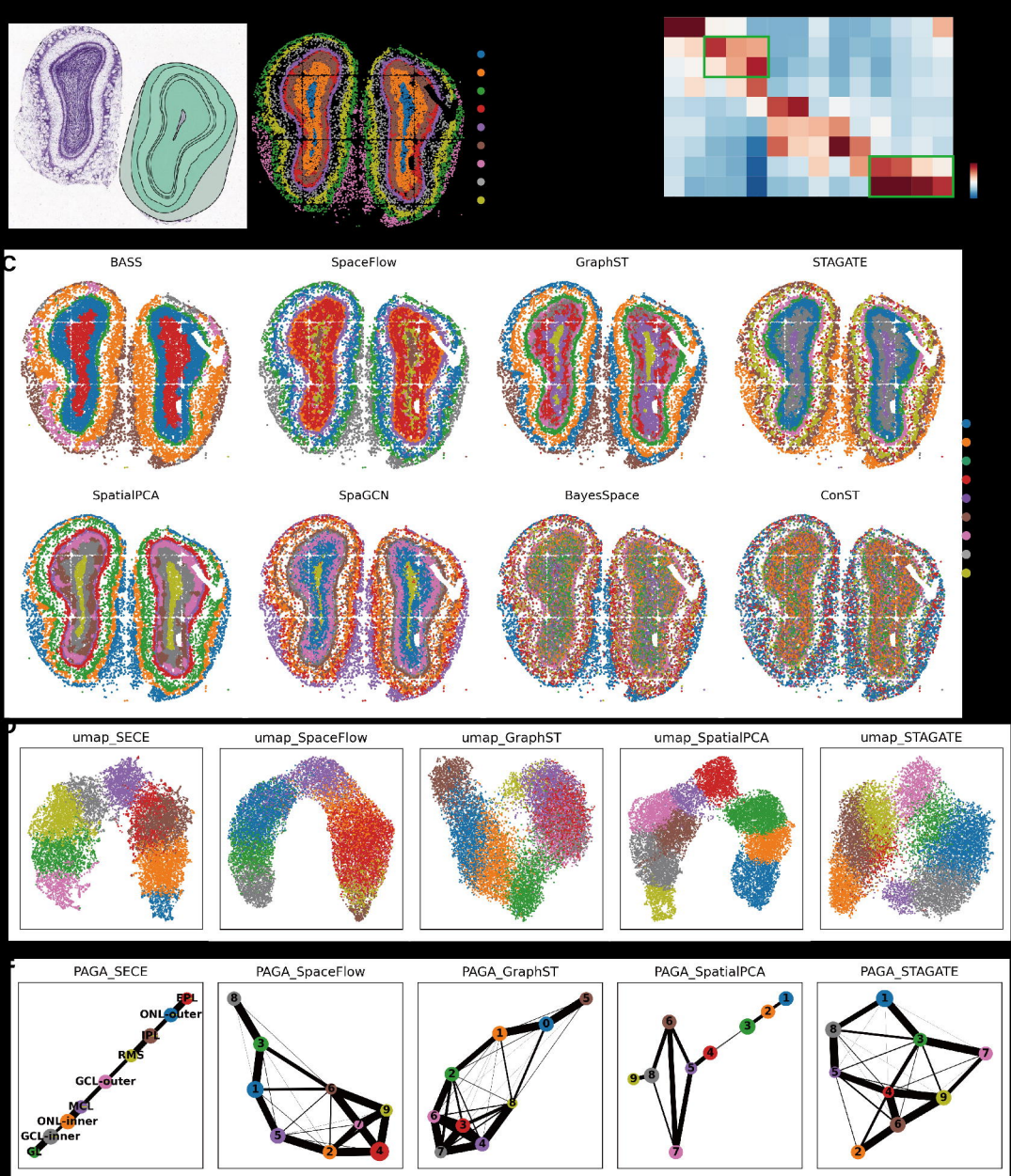
Zhao, E., *et al*. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* 2021;39(11):1375-1384.

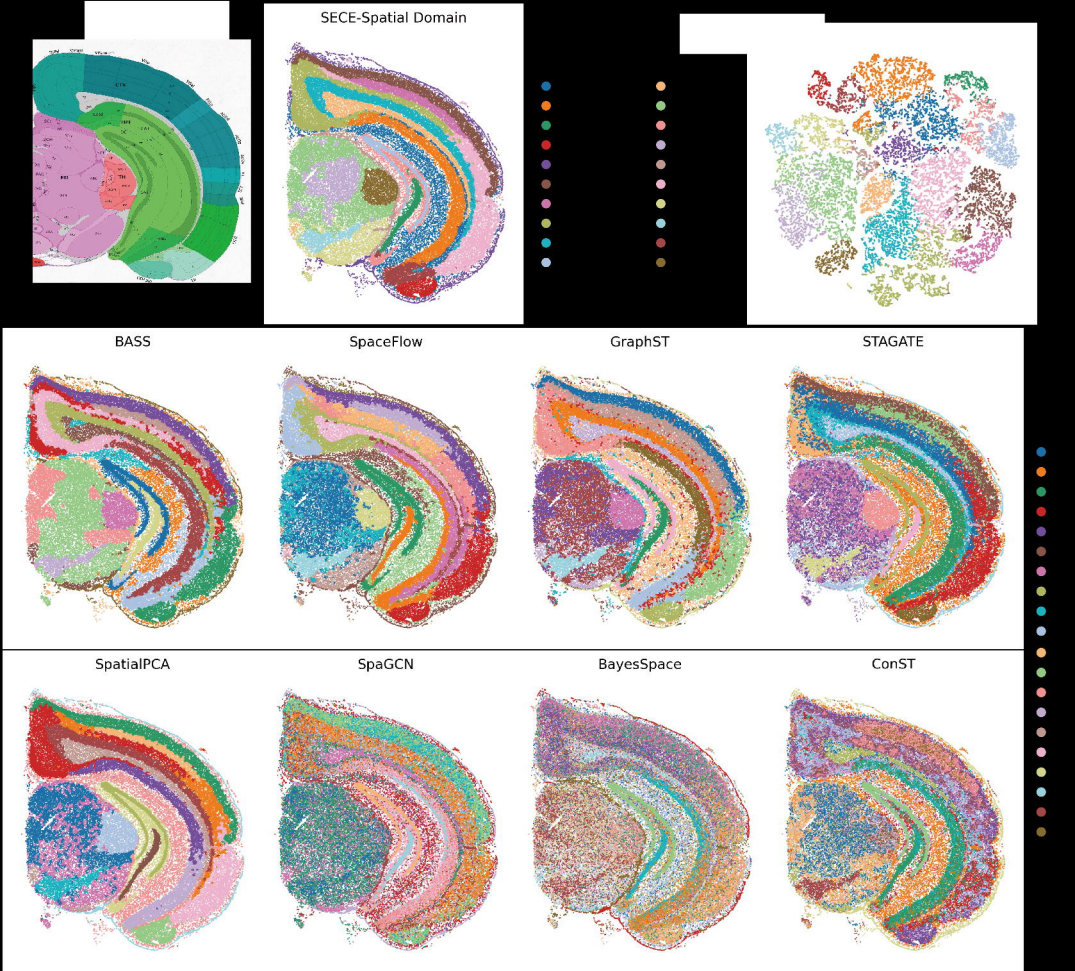
Zong, Y., *et al*. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. *bioRxiv* 2022:2022.2001.2014.476408.

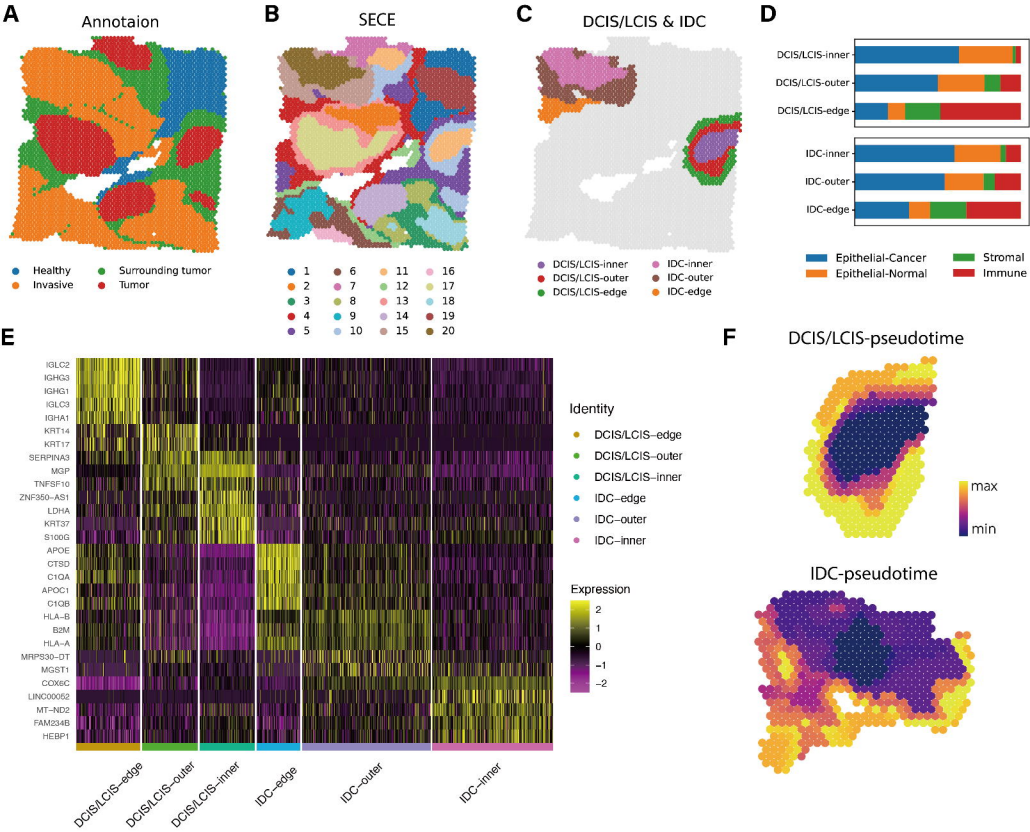

A
ST data
Subcellular
Single-cell
Near single-cell
Multicellular
B
Expression
Genes
Cells/Spots
N×G
AE module
Π R P
Denoised CE
CE
Cells/Spots
N×m
Location
y₁ y₂
Cells/Spots
N×2
SSM
N×N
Spatial Proximity
GAT module
ADM
Expression Proximity
C
Spatial Domains
Visualization
Trajectory Inference
20
15
10
5
0
L6
L5
L4
L2/3
L1


SECE
BASS
SpaceFlow
GraphST
STAGATE
SpatialPCA
SpaGCN
BayesSpace
ConST
PAGA_SECE
PAGA_SpaceFlow
PAGA_GraphST
PAGA_STAGATE
PAGA_SpatialPCA
L6
L5
L4
L2/3
L1


A
B
SECE
CA slm/so/sr
CA1
CA2
CA3
DG
Corpus Callosum
The thrid ventricle
Layer 4
Layer 5a
Layer 5b
Layer 6
Thalamus-1
Thalamus-2
Thalamus-3
SECE-UMAP
Cortex
Hippocampus
Thalamus
C
CA1
Wfs1
CA2
Rgs14
CA3
Nptxr
D
BASS
SpaceFlow
GraphST
STAGATE
SpatialPCA
SpaGCN
BayesSpace
conST
1 2 3 4 5 6 7 8 9 10 11 12 13 14
E
SECE
SpaceFlow
GraphST
STAGATE
SpatialPCA
max
min
F
Pseudo-time of domains
Layer 6b
Layer 6a
Layer 5
Layer 4
1
6
13
2
4
3
11
3
11
13
7


c
BASS
SpaceFlow
GraphST
STAGATE
SpatialPCA
SpaGCN
BayesSpace
ConST
umap_SECE
umap_SpaceFlow
umap_GraphST
umap_SpatialPCA
umap_STAGATE
PAGA_SECE
EPL
ONL-outer
IPL
RMS
GCL-outer
MCL
ONL-inner
GCL-inner
GL
PAGA_SpaceFlow
PAGA_GraphST
PAGA_SpatialPCA
PAGA_STAGATE


SECE-Spatial Domain
BASS
SpaceFlow
GraphST
STAGATE
SpatialPCA
SpaGCN
BayesSpace
ConST


A
Annotaion
Healthy
Invasive
Surrounding tumor
Tumor
B
SECE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
C
DCIS/LCIS & IDC
DCIS/LCIS-inner
DCIS/LCIS-outer
DCIS/LCIS-edge
IDC-inner
IDC-outer
IDC-edge
D
DCIS/LCIS-inner
DCIS/LCIS-outer
DCIS/LCIS-edge
IDC-inner
IDC-outer
IDC-edge
Epithelial-Cancer
Epithelial-Normal
Stromal
Immune
E
IGLC2
IGHG3
IGHG1
IGLC3
IGHA1
KRT14
KRT17
SERPINA3
MGP
TNFSF10
ZNF350-AS1
LDHA
KRT37
S100G
APOE
CTSD
C1QA
APOC1
C1QB
HLA-B
B2M
HLA-A
MRP530-DT
MGST1
COX6C
LINC00052
MT-ND2
FAM234B
HEBP1
DCIS/LCIS-edge
DCIS/LCIS-outer
DCIS/LCIS-inner
IDC-edge
IDC-outer
IDC-inner
Identity
DCIS/LCIS−edge
DCIS/LCIS−outer
DCIS/LCIS−inner
IDC−edge
IDC−outer
IDC−inner
Expression
2
1
0
−1
−2
F
DCIS/LCIS-pseudotime
max
min
IDC-pseudotime