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

Identification of spatial domains is a critical step in spatial transcriptomics (ST) analysis. Here, we introduce SECE, a method that incorporates global spatial proximity and local expression proximity to identify spatial domains. SECE can also obtain spatial features for low-dimensional visualization and trajectory inference. In addition, SECE can identify cell types when dealing with ST data with single-cell resolution. Compared with eight spatial domain identification and five cell type identification methods, SECE consistently achieved top performance on spatial and cellular characterization, accurately providing biological insights into ST data.

Keywords

Spatial transcriptome, Spatial domain, Spatial embedding, Cell type

Background

Spatial transcriptomics (ST) allows for the acquisition of expression profiles with spatial information, providing insight into the cellular composition and function of cells in different locations within a tissue microenvironment. This technology has led to novel discoveries in various fields such as neuroscience [1], developmental biology [2], and cancer biology [3].

A variety of ST platforms have been developed with varying throughput and resolution. *In situ* hybridization-based ST platforms, including STARmap [4], seqFISH [5, 6], seqFISH+ [7], MERFISH [8] and FISSEQ [9], provide highly accurate gene expression measurement at single cell resolution but only for a limited number of known genes [10]. On the other hand, barcode-based ST platforms, such as Spatial transcriptomics [11] and its commercial version 10X Genomics Visium, Slide-seqV2 [12], HDST [13], Seq-Scope [14] and Stereo-seq [15], can perform high-throughput sequencing on a genome-wide scale. With recent advancements, barcode-based ST platforms have greatly improved their resolution. The early ST method, Spatial transcriptomics, has an initial resolution of 200 μm, while Visium improves that to 55 μm. Slide-seqV2 has a resolution of 10 μm, which is close to a single-cell level. Stereo-seq, Seq-Scope, and HDST have subcellular resolutions ranging from 0.5 to 2 μm, with Seq-Scope and Stereo-seq capable of merging spots based on cellular location to achieve single-cell resolution.

In ST data, spatial domains refer to spatially consistent regions in terms of gene expression and physical location, each with specific anatomical structures [1, 16]. To address the challenge of spatial region identification, several algorithms have been developed. For instance, BayesSpace [17] employs a Bayesian model with a Markov random field and encourages neighbor spots to belong to the same cluster. SpatialPCA [18] introduces a spatially aware dimension reduction method to extract a low-dimensional representation of the ST data while preserving the spatial correlation structure. Another algorithm, BASS [19], takes a coherent approach by jointly inferring labels for cell types and spatial regions based on a Bayesian hierarchical model. Moreover, graph neural networks have been widely used since the spatial positional relationship can be naturally transformed into a graph structure. SpaGCN [20] and STAGATE [21] employ graph convolutional neural networks and graph attention neural networks to aggregate neighbor

information respectively, while GraphST [22] , SpaceFlow [23] , and conST [24] use self-supervised graph embedding learning strategies.

The way these approaches use similarity information between spots can be classified into global or local similarity. Global similarity information measures pairwise spot similarity, yielding high similarity for close spots and low similarity for distant ones, including SpaGCN and SpatialPCA. Local similarity information refers to selecting neighboring spots based on algorithms like K-nearest neighbors (KNN) and ensuring high consistency among them, including STAGATE, BayesSpace, BASS, GraphST, SpaceFlow, and conST. However, the simultaneous capture of local and global information which takes both advantages is lacking. In addition, some existing methods have limited effectiveness in extracting gene expression features. They 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. However, these features cannot handle expression with high noise, which is very common in barcode-based sequencing ST methods, particularly in high-resolution techniques such as Stereo-seq [20].

To this end, we developed SECE, an accurate spatial domain identification method for ST data. In contrast to the existing approaches, SECE incorporates global spatial proximity and local expression proximity of data to derive spatial domains. The spatial embedding (SE) obtained by SECE enables downstream analysis including low-dimensional visualization and trajectory inference. To infer trajectories, we utilizeed Partition-based Graph Abstraction (PAGA) [25] at the domain level and Monocle3 [26] at the single-cell level. Moreover, when applied to ST data with single-cell resolution, SECE can accurately assign cell type labels by clustering cell type-related embedding (CE). SECE was versatile on datasets from different ST platforms, for both high-resolution platforms, such as STARmap, Slide-seqV2 and Stereo-seq, and relatively low-resolution ones such as Visium (Additional file 1: Table S1). For ease of use, SECE was implemented and integrated into the workflow of ST analysis using the Anndata data structure, which can be used in the SCANPY package [27] for visualization and downstream analyses. The capability of highly accurate identification of spatial domains and cell types enables SECE to facilitate the discovery of biological insights from ST data.

Results

Overview of SECE

SECE can model ST data at various levels of resolutions, including subcellular, single-cell, near-single-cell, and multicellular (Fig. 1A). SECE obtains the result of spatial region division by clustering spatial domain-related embedding (SE). Specifically, SECE performs a two-stage procedure, consisting of an autoencoder (AE) module and a graph attention network (GAT) module. (Fig. 1B). The AE module is employed to accurately capture the CE of each spot by modeling the sequencing noise of expression profile (Methods). Then SE is achieved in GAT module by incorporating local expression proximity of neighbor spots and global spatial proximity of all spots. Expression proximity, which is based on CE, is measured using an attention mechanism, while spatial proximity, which is based on physical distance, is measured using a spatial similarity matrix (SSM). Finally, we obtain spatial domain, low-dimensional visualization, and spatial trajectory inference by running clustering, dimensionality reduction and trajectory inference algorithms on SE respectively (Fig. 1C). In addition, clustering of CE results in the identification of cell types.

To assess the contributions of CE learning and global spatial proximity in learning SE, we tested three alternatives to SECE. "SECE-PCA" eliminated the influence of CE by substituting it with PCA of equal dimension. "SECE-NG" removed the impact of SSM by setting the weight of the global spatial similarity constraint λ_{global} to 0. "SECE-PCA-NG" used PCA and set λ_{global} to 0. We tested these methods on a mouse hippocampus dataset generated by Slide-seqV2 [12], a technology that provided transcriptome-wide RNA detection with a near-cellular resolution of 10 μm. This dataset included 53,208 cells and 23,264 genes from the hippocampus, cortex, and thalamus, with a high sparsity of 98.19%.

Compared to SECE, three SECE-derived methods had reduced capability in spatial domain identification or trajectory inference. SECE-PCA-NG could not detect differences in the cortical region, while SECE-PCA detected differences but did not correctly identify near-parallel structures (Fig. 1D, left). This hinted at the importance of accurately modeling gene expression using CE. SECE-NG achieved the most similar results to SECE but it could not divide the Cornu Ammonis (CA) correctly, ignoring the region CA2. In addition, it divided Layer 5 into two sub-

layers with less smoothness. Although SECE-NG recognized more of the yellow area Dentate gyrus (DG)-molecular layer (mo) around DG compared to SECE at n=14, we observed that as the number of clusters increased, SECE was also able to identify the subtle CA1-(stratum oriens) so/(stratum radiatum) sr in the CA1 region. For instance, at n=15, SECE detected a purple region, and at n=16, it detected a brown region (Additional file 1: Fig. S1, 2). In contrast, increasing the number of categories did not lead to improved recognition of SECE-NG in the CA area or Layer 5. Spatial trajectory inference of cortex using SE generated from SECE-derived methods also showed that a spuriously high pseudo-time of deep spots in SECE-PCA-NG (Fig. 1D, right). These results showed that both CE and SSM are important to learn accurate SE.

Application to STARmap data

We tested SECE on the mouse visual cortex data generated by STARmap [4]. STARmap is a *in situ* sequencing-based ST method with single-molecule resolution. Despite its low gene throughput (160 to 1020 genes), it offers high sensitivity at single cell resolution with high efficiency and reproducibility [4]. This dataset has a total of 1,207 cells, 1,020 genes, and a sparsity of 76.88%. The cells were annotated into distinct layers, including L1, L2/3, L4, L5, L6, as well as the corpus callosum (CC) and hippocampus (HPC) of the visual cortex (Fig. 2A).

First, we compared the spatial region identification results of SECE with eight existing methods, including BASS, SpaceFlow, GraphST, STAGATE, SpatialPCA, SpaGCN, BayesSpace, and conST (Fig. 2B left, C). We assessed the consistency of each algorithm's spatial regions with the ground truth layers using Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Kuhn-Munkres [28] based Accuracy (ACC). We found that SECE achieved the highest agreement with the ground truth (Fig. 2D). It was closely followed by BASS. However, BASS incorrectly combined HPC and L5. The accuracy of other algorithms was relatively low, and they exhibited weak spatial aggregation, as indicated by the Local inverse Simpson's index (LISI) (Additional file 1: Fig. S3A). Additionally, these methods mistakenly classified aggregated endothelial cells (Endo) in L1 and L2/3 into the same region, such as region 2 in SpaceFlow, region 3 in STAGATE and BayesSpace, region 5 in GraphST and SpatialPCA, and region 6 in SpaGCN.

138 Next, we compared SE of SECE with SpaceFlow, GraphST, STAGATE, and SpatialPCA.
139 Due to the difficulty of conST in generating clear spatial regions, its SE learning comparison was
140 excluded. Our results demonstrated that SE of SECE explained the ground truth layers most
141 effectively (ASW=0.16), followed by SpaceFlow (ASW=0.12), GraphST (ASW=0.07),
142 STAGATE (ASW=0.07), and SpatialPCA (ASW=0.05) (Fig. 2E). SECE had a clearer and
143 continuous pattern in UMAP visualization compared to the other methods (Fig. 2F). When
144 comparing trajectory inference results, we selected the cortex part in ground truth, namely L1,
145 L2/3, L4, L5 and L6. The PAGA analysis based on SECE revealed a linear and continuous
146 relationship between these layers (Fig. 2G). Conversely, SpaceFlow mistakenly made the
147 connection between L1 versus L4 and L5, while the patterns in the other three methods were more
148 unclear. For individual cells, SECE exhibited a sequential increase of pseudo-time from L6 to L1
149 (Additional file 1: Fig. S3B, C), and the similar trend was also observed in SpaceFlow, while
150 GraphST, STAGATE, and SpatialPCA ordered L1 and L2/3 incorrectly (Additional file 1: Fig.
151 S3D, E).
152 Finally, we examined the results of cell type identification. We compared SECE with Seurat
153 [29], SC3 [30], CIDR [31], BASS, and SpiceMix [32]. First, we compared the identified cell types
154 with the annotations provided in the original research [4]. BASS (ARI=0.51), SECE (ARI=0.49)
155 and SC3 (ARI=0.49) had the top three performances (Additional file 1: Fig. S4A), followed by
156 SpiceMix (ARI=0.36) and Seurat (ARI=0.41), while CIDR was lowest (ARI=0.09). Specifically,
157 SECE identified seven excitatory neuronal (EX-Neuron) subtypes, two inhibitory neuronal (IN-
158 Neuron) subtypes, and seven non-neuronal cell types (Fig. 2B, right; Additional file 1: Fig. S4B,
159 S5). Among them, the seven EX-Neuron subtypes identified were all distributed in the
160 corresponding cortex (Additional file 1: Fig. S6A). Interestingly, both astrocytes (Astro) and
161 oligodendrocytes (Oligo) were divided into two subtypes, one concentrated in the CC (Astro-1,
162 Oligo-2) and one diffused in the HPC and cortical layers (Astro-2, Oligo-1). Additionally, SECE
163 was the only method that could identify subtypes of Endo with one subtype (Endo-1) concentrated
164 between L1 and L2/3 and the other (Endo-2) scattered elsewhere. Endo-1 and Endo-2 had
165 different highly expressed genes, suggesting that they might also have different functions
166 (Additional file 1: Fig. S6B). In contrast, SpiceMix failed to recognize EX L5, Seurat recognized a
167 mixed neuronal cell class, while SpiceMix, Seurat, and SC3 all failed to recognize subtypes of IN-

168 Neuron (Additional file 1: Fig. S4B, C; Fig. S5). BASS could identify the IN-Neuron subtype VIP
169 not detected by other methods. However, it assigned Astro and EX L4 into the same cluster, and it
170 detected a class (NA) with no marker genes. CIDR could only clearly identify microglia (Micro),
171 SST, Smc and two Oligo subtypes.
172 To sum, using a dataset generated by STARmap, we showed that SECE accurately identified
173 spatial domains and cell types. SECE also improved the performance in SE visualization and
174 spatial trajectory inference compared to the previous methods.
175

176 **Application to Slide-seqV2 data**

177 Next, we tested SECE on the hippocampal dataset generated by Slide-seqV2 platform. SECE
178 identified 14 distinct regions, and we annotated them according to the known structure of Allen
179 Brain Atlas (ABA) (Fig. 3A). These regions included the hippocampus (CA1, CA2, CA3, DG, and
180 CA slm/so/sr), cortex (Layers 4, 5a, 5b, 6), third ventricle, CC, and three subregions of the
181 thalamus (Fig. 3B, left). The identified regions were supported by their marker genes. For
182 example, the four important components of the hippocampal region, CA1, CA2, CA3, and DG,
183 were clearly demarcated, as evidenced by the high expression of their known markers *Wsf1*,
184 *Rgs14*, *Nptxr* and *C1ql2* [33] (Fig. 3C, Additional file 1: Fig. S7A). Layer 5 in cortical regions
185 was identified as two sublayers, Layer 5a and 5b, with different gene expression levels (Additional
186 file 1: Fig. S7). For comparison, we evaluated the performance of eight existing methods for
187 spatial region identification (Fig. 3D). Except for GraphST, none of these methods could identify
188 the CA2 region. Furthermore, for cortical areas, BASS failed to distinguish between Layer 4 and
189 Layer 5, SpaceFlow mixed Layer 5 and Layer 6, SpatialPCA exhibited suboptimal division
190 smoothness, and the remaining algorithms were unable to generate clear cortical regions.
191 Additionally, SECE exhibited the highest spatial aggregation performance as it had the smallest
192 LISI values (Additional file 1: Fig. S8A).
193 Then, we compared the performance of low-dimensional embeddings. UMAP generated from
194 SECE clearly displayed clustering patterns for the hippocampus, cortex, thalamus, third ventricle,
195 and their subregions (Fig. 3B, right). In addition, the sublayers of them like Layer4 and Layer6b
196 were arranged in a sequence, while UMAP of STAGATE and GraphST missed them (Additional

197 file 1: Fig. S8B). Moreover, we conducted spatial trajectory inference for the cortex due to its
198 continuous relationship between subregions. We selected the regions corresponding to the cortex
199 in each method, starting with the deepest clusters, such as Layer 6 for SECE, domain 1 for
200 SpaceFlow, domain 2 for GraphST, and domain 3 for SpatialPCA and STAGATE (Fig. 3E, F).
201 For SECE, the pseudo-time consistently increased with decreasing cortical depth, clearly
202 capturing the pseudo-time relationship between different layers. In contrast, SpaceFlow exhibited
203 less distinguishable differences between the learned layers. SpatialPCA reversed the relationship
204 between Layer 5b and Layer 6, while GraphST and STAGATE failed to identify the relationship
205 within the cortical region.

206 Next, we tested cell type identification and CIDR was excluded from the test due to its huge
207 memory consumption. SECE identified eleven subtypes of EX-Neuron, one IN-Neuron, and eight
208 subtypes of non-neuronal cells (Additional file 1: Fig. S9 A to C). Each cell type exhibited a
209 unique set of differentially expressed genes (Additional file 1: Fig. S10). Specifically, the EX-
210 Neurons in CA comprised of EX CA1, EX CA2, and EX CA3, while the non-neuronal subtypes
211 included Astro, Micro, two types of Oligos, polydendrocyte (Poly), mural cells (Mural),
212 ependymal cells (Epen), and choroid plexus cells (Chor). In contrast, none of the existing methods
213 were able to identify EX CA2, Poly, or Micro. Furthermore, BASS assigned Astro and the EX-
214 Neuron in the same class and recognized a cluster labeled NA without specific marker genes.
215 Seurat could not identify IN Neuron and misclassified EX-Neuron in CA into two subtypes with
216 less distinct, namely cluster 12 and 15. SC3 and SpiceMix failed to detect differences among EX-
217 Neuron in the hippocampal region, assigning both EX CA and EX DG to a single broad class,
218 namely cluster16.

219 Finally, our study revealed a compelling concordance between the spatial regions and cell
220 types identified by SECE (Additional file 1: Fig. S9D, E; Fig. S11). Each subregion of the
221 hippocampus, cortex, and thalamus corresponded to specific EX-Neuron subtypes. For instance,
222 EX CA1, EX CA2, EX CA3, and DG were predominantly localized within their respective
223 regions of CA1, CA2, CA3, and DG. Moreover, we observed that Layer 5a and Layer 5b had
224 differential distributions of EX-Neuron subtypes, and Layer 5b exhibited a higher proportion of IN
225 neuron, which further confirmed the existence of Layer 5 sublayer. Additionally, distinct spatial
226 distributions were evident among subtypes of cell types. For instance, Oligo-1 exhibited a

227 predominant presence in the CC, whereas Oligo-2 was primarily concentrated in the thalamus. In
228 contrast, the subtypes of Oligo identified by BASS did not have this spatial pattern (Additional file
229 1: Fig. S9F).
230 In conclusion, using a sparse dataset generated by Slide-seqV2, SECE outperformed the other
231 methods in various tasks. We identified functional subregions of the hippocampus, cerebral
232 cortex, and thalamus, particularly the CA2 region, and correctly reproduced the developmental
233 trajectory of the cerebral cortex. The cohesive correlation between spatial regions and cell types
234 further supported the discovery of cell subtypes and spatial subregions.
235

236 **Application to Stereo-seq data**

237 Stereo-seq is an emerging technique for ST with genome-wide throughput and subcellular
238 resolution [15]. It obtains the expression profile and spatial coordination of each DNA nanoball
239 (DNB), followed by image-based cell segmentation to identify DNBs belong to the same cell.
240 Then, an expression matrix with single-cell resolution is obtained by summing the expression
241 values within the same cell. However, the high noise levels associate with high resolution prevent
242 accurate identification of cell types and spatial domains at the single-cell level. Here, we analyzed
243 the mouse olfactory bulb data generated by Stereo-seq [34], which contained 19,527 cells and
244 27,106 genes, with 98.69% of the values being zero.
245 First, we assessed the effectiveness of cell type recognition. SECE identified 6 types of
246 neurons and 4 types of non-neurons (Fig. 4A; Additional file 1: Fig. S12; Fig. S13). Neurons were
247 three types of Granule cells (GC), two types of Mitral and tufted cells (M/TC), and Periglomerular
248 cells (PGC), while non-neurons included Astro, Oligo, and two types of Olfactory ensheathing
249 cells (OEC). In contrast, BASS produced an NA category without any marker genes. SpiceMix
250 and Seurat did not identify Astro and Oligo, respectively, while SC3 failed to obtain Astro, Oligo
251 and PGC.
252 Next, we segmented this data into spatial regions using SECE and annotated the results based
253 on the known olfactory bulb layers (Fig. 4B, left). They were Rostral migratory stream (RMS),
254 Granule cell layer (GCL), Inner plexiform layer (IPL), Mitral cell layer (MCL), External
255 plexiform layer (EPL), Glomerular layer (GL) and Olfactory nerve layer (ONL) from inside to

256 outside (Fig. 4B, right; Additional file 1: Fig. S14A). Our findings were validated using known
257 markers [33] for each layer (Additional file 1: Fig. S14B). Notably, besides the known seven
258 layers, we made a finer division of GCL and ONL, and the sublayers were named GCL-Inner,
259 GCL-Outer, ONL-Inner and ONL-Outer respectively. Several evidences confirmed these
260 sublayers. First, we found that marker gene expression differed between them, where the GCL-
261 inner highly expressed *Nrgn* and the GCL-outer highly expressed *Pcp4*. The expression level of
262 markers was also higher in the ONL-outer than in the ONL-inner. Then, we calculated the
263 relationship between layers and the corresponding cell types (Additional file 1: Fig. S14C). We
264 found that RMS contained Oligo and a small amount of GC; EPL mainly contained M/TC-2 and
265 Astro; MCL and GL mainly contained M/TC-1 and PGC, respectively, which were consistent with
266 known biological knowledge [35, 36]. Most of all, GCL-Outer was almost composed of GC, while
267 GCL-Inner contained a certain amount of Oligo, and the GC subtypes in GCL-Inner and GCL-
268 Outer also differed. Moreover, the ONL-Outer contained almost exclusively OEC cells, while the
269 ONL-Inner contained a fraction of OEC and more Astro. This supported previous findings that the
270 ONL, as a part of the olfactory bulb blood-brain barrier, had fine internal and external subregions
271 with different cell types [37] . Our study provided further support for this finding at the spatially
272 resolved single-cell level.

273 We further compared the performance of SECE to the existing methods (Fig. 4C). BASS
274 failed to separate the RMS from the GCL layer and could not distinguish GL from EPL. It also
275 identified some irrelevant regions with few cells, such as domain5, 8, and 9. STAGATE and
276 SpaGCN mixed GL and EPL. SpaceFlow, GraphST, and SpatialPCA encountered challenges in
277 accurately dividing the GCL layer. BayesSpace and conST did not yield a clear domain
278 identification. Additionally, BASS had the most spatial aggregation performance with lowest LISI
279 value, while SECE ranked second (Additional file 1: Fig. S14D). Moreover, SECE showed
280 robustness when tested with different number of clusters (7, 8, and 10), consistently providing
281 well-bounded ring stratification (Additional file 1: Fig. S15).

282 Finally, we assessed the SE learning capabilities of SECE, SpaceFlow, GraphST, STAGATE,
283 and SpatialPCA. The low-dimensional UMAP visualization based on SECE exhibited a
284 continuous low-dimensional pattern with nine layers arranged in order of spatial position from
285 inside to outside (Fig. 4D). Moreover, the trajectory inference for these layers demonstrated

286 approximately linear relationship (Fig. 4E). SpatialPCA also displayed linear patterns, except for
287 regions 6 and 7. However, the trajectory inference results of SpaceFlow, GraphST, and
288 STAGATE were challenging to interpret.
289 Collectively, SECE accurately identified cell types and spatial domains in the mouse
290 olfactory bulb. Furthermore, SECE provides a higher resolution in spatial stratification compared
291 to the existing methods.
292

293 **Application to complex hemibrain Stereo-seq data**

294 We further tested SECE to a dataset with a more complex anatomic structure generated by Stereo-
295 seq [15]. This dataset was a coronal slice of the mouse hemibrain, containing 50,140 cells and
296 25,879 genes, of which 96.94% had a value of 0. The original study annotated cell types and
297 anatomic regions using single cell and bin50 data, respectively (Fig. 2 in [15]). Bin50 data binned
298 the neighboring 50×50 DNBs with fixed area, but it did not match the cell type label. Here, single-
299 cell data was utilized for both cell type and spatial region identification, although its higher
300 sparsity increased the difficulty.
301 Clustering of SE resulted in spatial domain annotations that were in highly consistent with
302 ABA anatomy (Fig. 5A, B, left), with clear separations between regions including cortex,
303 hippocampus, midbrain, thalamus, and fiber tracts (FT). The cortex included five layers: L1, L2/3,
304 L4, L5, L6, LVC and CAA. In contrast, other methods failed to accurately identify these cortical
305 layers (Fig. 5C). Specifically, all of them encountered difficulties in identifying Cortex L5 with
306 high smoothness. In addition, LISI showed that spatial domains identified by SECE had the
307 strongest spatial pattern (Fig. 5D). Furthermore, we performed a low-dimensional visualization of
308 SE using t-SNE [38], which showed the effect of aggregation in the same region and separation in
309 different regions (Fig. 5B, right; Additional file 1: Fig. S16).
310 Then, we tested the ability of cell type recognition. Using the SECE, we identified 15 types
311 of EX-Neuron, two types of IN-Neuron, dopaminergic neuron (DA-Neuron), and granule cell of
312 dentate gyrus (GN-DG). Among the non-neuronal cell types, we distinguished four subtypes of
313 Astro, three subtypes of Oligo, Micro, erythrocytes (Ery), and meninges (Fig. 5E). However, other
314 algorithms yielded mixed results (Additional file 1: Fig. S17, 18). BASS, for instance, identified

315 two hybrid classes that grouped Astro and Ery together instead of identifying them separately. It
316 also failed to delineate meninges and EX-Thalamus, while Seurat and SpiceMix mixed meninges
317 with Ery and Astro, respectively. Seurat and SpiceMix also identified a mixture of two EX CA1
318 and EX CA3 classes, rather than detecting them separately. Additionally, Seurat encountered
319 challenges in accurately separating neurons in specific regions of the cortex, leading to the
320 identification of three EX L5/6 classes and two EX L2/3/4 classes. SpiceMix did not identify IN-
321 Neurons in the telencephalon but mixed them with three EX-Neuron subtypes. Moreover, it did
322 not recognize the DA-Neuron. The performance of the SC3 algorithm was relatively weaker,
323 possibly due to the large volume and high sparsity of the data. Classes 16 and 27 contained more
324 than one-third of the cells, but no distinct pattern of high expression was observed. Furthermore,
325 none of these algorithms were able to identify Micro accurately. These findings were further
326 supported by analyzing differentially expressed genes and assessing spatial distribution
327 (Additional file 1: Fig. S19 to S24).
328 Finally, we verified that the spatial domains and cell types identified by SECE were
329 accurately matched (Additional file 1: Fig. S25, 26). The GN-DG, EX CA1, EX CA3, and EX
330 Subiculum cell types primarily constituted the DG, CA1, CA3, and Subiculum regions,
331 respectively. Cortex L2 to L6 mainly consisted of their corresponding EX-Neuron subtypes, while
332 the Cortex L1 contained a significant number of Astro compared to other cortical layers.
333 Additionally, the substantia nigra/ventral tegmental area (SN/VTA) region was predominantly
334 populated by DA-Neuron. Furthermore, we observed distinct spatial distribution differences
335 among the four Astro subtypes and three Oligo subtypes. Astro-1 was predominantly distributed in
336 the midbrain, while Astro-2 and Astro-4 were primarily found in the telencephalon region. Astro-3
337 exhibited a concentrated presence in Cortex L1, adjacent to meningeal cells. As for Oligo, Oligo-1
338 and Oligo-2 were distributed in FT and midbrain, respectively, while Oligo-3 was found in both
339 regions.
340 In conclusion, the Stereo-seq dataset with complex spatial structure and high sparsity was
341 effectively annotated by SECE, which accurately matched spatial domains and cell types for each
342 cell and provided robust support for cell subtypes from spatial distribution disparities.
343

Application to Visium data

Finally, we tested the applicability of SECE on Visium, which had a relatively low resolution. Visium sequences a fixed number of 55 μm spots, with 1-10 cells per spot, depending on the tissue type [39]. We first applied SECE to a breast cancer dataset, containing 3,798 spots and 24,923 genes, of which only 77.44% are zero values. Based on pathological images, the dataset covered four phenotypic regions, including 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 [34] (Fig. 6A; Additional file 1: Fig. S27A).

First, we segmented 20 regions using each algorithm and performed phenotype annotations, which were then compared with image-based manual annotations (Fig. 6B; Additional file 1: Fig. S27B, C). Notably, in the context of a dataset characterized by a low missing rate, all algorithms demonstrated commendable performance. Most of the algorithms had ARI values ranging from 0.55 to 0.62, NMI values ranging from 0.5 to 0.6, and ACC values around 0.8 (Additional file 1: Fig. S27D). Among these nine algorithms, SECE's ARI ranked fifth, NMI and ACC ranked third (ARI=0.60, NMI=0.55, ACC=0.82). BayesSpace and BASS, which shared the top performance, attained ARI values of 0.62, NMI values of 0.56, and ACC values of 0.83.

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; Additional file 1: Fig. S27A). 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 further 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 of these regions by integrating the annotated breast cancer scRNA-seq data [40] and deconvoluting each spot using cell2location [41] (Additional file 1: Fig. S28, S29). Our findings showed that 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.

373 We further explored the characteristics of these subregions (-edge, -outer and -inner) of
374 DCIS/LCIS and IDC (Fig. 6E). Differential expression analysis showed that DCIS/LCIS-edge had
375 a high level of humoral immunity, which could be confirmed by the enrichment of B cell receptor
376 signaling pathway related genes (*IGLC2*, *IGLC3*, *IGHG1*, *IGHG3*, *IGHA1*). DCIS/LCIS-outer
377 overexpressed *KRT14* of the keratin family, which was a key regulator of metastasis, suggesting
378 invasive potential [42, 43]. In contrast, DCIS/LCIS-inner showed non-metastatic traits but also
379 had tumor-promoting abilities, indicated by a high expression of *LDHA*. *LDHA*, linked to
380 metastasis-free survival, catalyzes the conversion of pyruvate to lactate, creating a hypoxic
381 environment in the tumor core, and inducing metabolic changes in both tumor cells and
382 fibroblasts, promoting tumor progression [44].
383 For the IDC subregions, IDC-edge showed elevated expression of biomarkers linked to tumor
384 proliferation, invasion, and migration including tumor-associated macrophage (*APOE*),
385 complement components (*C1qA*, *C1qB*), cathepsin (*CTSD*), and apolipoprotein (*APOC1*) [45-49].
386 IDC-outer had a high level of immunity and some transferability, which was derived from the high
387 expression of MHC class I-related genes (*HLA-A*, *HLA-B*, *B2M*) [50, 51]. Besides, there were
388 increased expression of genes such as *MGST1* and *MRPS30-DT*, which were known to promote
389 breast carcinoma cell growth and metastasis [52, 53]. In IDC-inner, there were higher levels of
390 tumor activity and lower levels of immune response. *LINC00052*, known to promote breast cancer
391 cell proliferation by increasing signals of epidermal growth factor receptor (EGFR) such as *HER3*
392 [54-56], was overexpressed. Upregulation of *COX6C* and *FAM234B* implied higher levels of
393 cellular respiration [57] and lower immune response function [58], respectively.
394 Next, we inferred developmental trajectories for the IDC and DCIS/LCIS regions. The
395 starting points for calculating pseudo-time in IDC and DCIS/LCIS were the interior of the tumor
396 regions (Fig. 6F). There were increased pseudo-time values as the region of the tumor moves
397 outward, effectively mimicking the gradual progression of tumor development.
398 Moreover, to further evaluate the power of SECE in spatial domain identification, we tested
399 12 dorsolateral prefrontal cortex (DLPFC) data generated from the Visium platform. The original
400 study [1] had manually annotated the spatial regions of these datasets, encompassing white matter
401 and six cortical layers (Additional file 1: Fig S30, S31). Among the slices, eight contained all
402 seven regions, while the remaining four slices contained five of these regions. In our evaluation,

403 we ensured that the number of regions per slice remained consistent with the annotated regions.
404 We observed that SECE exhibited the highest agreement with the original annotations. The
405 median ARI for SECE was 0.58, surpassing the second and third ranked algorithms, STAGATE
406 and SpatialPCA, which achieved median ARI values of 0.55 and 0.54, respectively. These
407 findings highlighted the superior performance of SECE in accurately delineating the spatial
408 regions within the low-resolution data.
409 In summary, SECE divided tumors into finer regions compared to manual annotation based
410 on pathological images in the Visium breast cancer dataset, providing a deeper understanding of
411 the heterogeneity within a single tumor. Meanwhile, SE was able to reproduce the dynamic
412 process of tumor development.
413

414 Discussion

415 SECE was designed to accurately identify spatial domains of ST data by effectively learning low-
416 dimensional spatial features. It only uses ST data as input without additional single cell or
417 histology data and achieved superior performances on ST data generated from different platforms.
418 It effectively balanced expression and spatial approximation to learn a highly representative SE,
419 and models noise explicitly to obtain an accurate CE. SECE had excellent performance for spatial
420 domain identification, low-dimensional visualization, and trajectory inference. Its CE component
421 could accurately label cell types, even in sparse Stereo-seq datasets. To further illustrate the ability
422 of SECE to learn cell type features, we compared the feature learning capabilities of the AE
423 module in SECE with scVI on real and simulated single-cell data (details in the Additional file 1:
424 Supplementary note1, Fig S32, S33), showing the effectiveness of SECE for single-cell datasets.
425 SECE was primarily designed for high-throughput and high-resolution ST data. When
426 dealing with higher resolution datasets, such as Slide-seqV2 and Stereo-seq, which often
427 contained tens of thousands of cells with higher sparsity and noise, the identification of spatial
428 regions became more challenging (Additional file 1: Fig S34A). In such cases, SECE provided
429 more additional improvements that significantly enhance the analysis of these datasets. SECE also
430 achieved excellent performance on low-throughput data, such as STARmap, and coarse-resolution
431 data, such as Visium.

With the increasing captured area of ST data and advancements in resolution, there was a growing demand for computational methods to exhibit higher efficiency and scalability. We recorded the runtime and GPU memory consumption for each dataset (Additional file 1: Fig S34B, C). 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.

For these emerging high-resolution sequencing methods such as Stereo-seq, our exploration of spatial identification and cell type is just beginning. More algorithms are needed to uncover expression patterns as well as cell-cell communication. Additionally, methods for the integration of ST and the other types of datasets are also necessary. Ultimately, continuous improvements in ST resolution and analysis tools will provide deep biological insights and facilitate both basic and medical research.

Conclusion

In conclusion, our results demonstrated that SECE achieved great spatial domain identification, particularly for ST datasets with high resolution. It outputs spatial embedding which can be used for low dimensional visualization and other downstream analysis such as trajectory inference. Additionally, for data with single-cell resolution, accurate cell types can also be identified.

Methods

Data preprocessing

Assuming that the spatial transcriptome data expression profile contains N spots (It means that there are N cells for data with single-cell resolution, here we collectively refer to spots for the sake of convenience.) and G genes, the dimensions of spatial gene expression X and spatial coordinates Y are $N \times G$ and $N \times 2$ respectively. The raw gene expression counts X are normalized based on library size and then transformed to a natural logarithmic scale to obtain the normalized expression matrix $\tilde{X}$, which was used as input of SECE together with the coordinate matrix Y.

Learning CE with AE module

We learn cell type-related embedding (CE) based on gene expression data. We assume that each element in expression count matrix follows a zero-inflated negative binomial (ZINB) distribution [59, 60]. For x_{ng}, that is the expression count of gene g at spot n, we have

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

where δ_0 is Dirac delta function, π_{ng} is the probability that the true gene expression value is observed to be 0, (r_{ng}, p_g) are the standard parametrization for the negative binomial distribution, x_{ng} is the number of successes of independently repeated Bernoulli trials before reaching r_{ng} failures, and p_g is the probability of failure for each Bernoulli trial.

We use the auto-encoder in deep learning to learn CE and distribution parameters at the same time. The ZINB parameters are the $N \times G$-dimensional parameter matrix Π, R and the G-dimensional parameter vector P,which consist of the elements π_{ng}, r_{ng} and p_g separately.

The relationships between auto-encoder and ZINB parameters are $r_{ng} = s_n' r_{ng}'$ and $p_g = \frac{\exp(l_g)}{1+\exp(l_g)}$, where s_n' is the scaling factor for spot n, l_g is the logits of p_g. Here, we draw the assumption in SCT [61], that is, there is a linear relationship between the logarithm of gene expression level and the logarithm of the base 10 of sequencing depth, $s_n' = \exp(log_{10}s_n)$, where s_n is the sum of UMI in cell n, so the reconstructed expression is $\hat{x}_{ng} = r_{ng} * \exp(l_g)$.

The formulation of the autoencoder is given below:

$$Z = f_1(\tilde{X})$$

$$Z' = f_2(Z)$$

$$R' = f_3(Z')$$

$$\Pi = f_4(Z'),$$

f_1 is a two-layer neural network with nonlinear activation function ReLU, it takes $\tilde{X}$ as input, and outputs Z, which has dimension $N \times m_1$ and is used as CE. f_2 is a one-layer neural network with nonlinear activation function ReLU. f_3 is a one-layer neural network with exponential functions, it outputs R' consisting of elements r_{ng}'. f_4 is a one-layer neural network with activation function sigmoid. It outputs Π, a $N \times G$- dimensional matrix consists of π_{ng},

which is the zero-inflated probability of gene g at spot n. In addition, the parameter vector L consisting of elements l_g is obtained from a learnable vector with exponential transformation.

The loss function of this module is the negative likelihood of the ZINB distribution:

$$Loss_1 = NLL_{ZINB}(\mathrm{X}; \Pi, \mathrm{R}, \mathrm{P}).$$

We note that zero-inflation term in the ZINB model allows for better modeling of dropout event. For datasets with lower sparsity, such as STARmap and Visium, negative binomial (NB) is sufficient for modeling. Both models are implemented in SECE. We used ZINB for Stereo-seq and Slide-seq which have high degree of sparseness and NB for STARmap and Visium whose proportion of 0 values are relatively low.

Learning SE with GAT module

After obtaining CE of individual spots using the expression information, we can fuse the spatial coordinate information to learn the SE. We construct the ADM and SSM from spatial location, followed by learning SE with two Graph Attention (GAT) layers.

Construction of ADM

To learn a smooth and reliable SE, we need to construct an adjacency matrix A based on the coordinate position Y of spots, so that each spot can aggregate the features of neighbor spots. A is an $n \times n$-dimensional symmetric matrix with elements taking the value 0 or 1. Let the set of spots written by $V = \{v_1, v_2, \dots, v_N\}$, then for each element a_{ij} of the adjacency matrix A,

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

$\mathcal{N}(v_j)$ is the set of v_j-neighbor spots, which can be calculated by KNN or distance cutoff. We use KNN to calculate neighbors by default, and the default number of neighbors is 6 for Visium and 10 for other data.

Construction of SSM

We hypothesize that the correlation of SE between cells decreases with increasing distance and tends to decay exponentially. So ideally Σ is the $n \times n$-dimensional spatial correlation matrix based on a Gaussian kernel,

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

where the bandwidth parameter γ controls the degree of influence of spatial location on features, with larger γ having a greater influence. We set γ as the quantile value of the distance matrix, the default is 0.05 quantile.

GAT Layer

Let the input feature matrix of one GAT layer is $H = (h_1, h_2, \dots, h_N)$, which has $N \times m$ dimensions with N samples and m features. The output of GAT layer is $H' = (h'_1, h'_2, \dots, h'_N)$ with $N \times m'$ dimensions. h'_j can be formulated as

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

where W is the weight matrix with $m' \times m$ dimensions, $\mathcal{N}(v_j)$ is the set of v_j-neighbor spots, α_{ij} is the normalized attention coefficient matrix using the SoftMax function:

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

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

GAT module

GAT module consists of two GAT layers. It takes CE features Z and spatial coordinates Y as input, and outputs SE features U, which is a $n \times m_2$-dimensional matrix.

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

The goal of GAT module is to make SE at different locations satisfy the constraints of spatial approximation while preserving as much information as possible in CE, that is, the correlation of each SE at n positions UU^T is close to Σ and U is close to Z. Thus, the objective function can be written as:

$$Loss_2 = \lambda_{global} * L_{global} + \lambda_{local} * L_{local},$$

where $L_{global} = MSE(UU^T, \Sigma)$, and $L_{local} = MSE(U, Z)$.

Architecture of SECE

For the AE module, the dimension of hidden layer is 128, and that of bottle layer is 32. We use Adam optimizer to minimize $Loss_1$ with learning rate of 1e-3 and dropout rate of 0.1. For the GAT module, the dimension of two GAT layers is consistent with bottle layer. We use Adam optimizer to minimize $Loss_2$ with learning rate of 1e-2 and dropout rate of 0.2. The numbers of iterations for the two modules are set as 40 and 50 by default respectively.

λ_{local} and λ_{global} are set as 1.0 and 0.08 by default. To be more specific, for data with single-cell resolution or close to single-cell resolution (e.g., Stereo-seq and Slide-seqV2), we employed hyperparameters 1 and 0.08. For data without single-cell resolution, such as Visium, we employed hyperparameters 1 and 0.3. For STARmap data, which provided single-molecule resolution, we used hyperparameters 1 and 2. When working with spatial transcriptome data from other platforms, we recommended adjusting the value of λ_{global} based on the number of spots. This was because λ_{global} governed the role of spatial global information. Generally, as the number of spots increased, a smaller value of λ_{global} tended to be more appropriate.

Preprocessing of datasets

For Stereo-seq datasets, cells with an expression level below 200 were removed according to Stereo-seq studies [15], then we further screened genes expressed in less than 20 cells. For other datasets, we screened spots with an expression level below 20 and genes that expressed less than 20 spots. The filtered expression matrix and its corresponding coordinates were input into SECE for analysis.

Clustering

We use Leiden [62] and mClust [63] to cluster CE and SE, respectively. We cluster the CE based on Leiden and then assigned the cell type labels. We use the Leiden algorithm implemented in SCANPY package by calling scanpy.tl.leiden. We cluster the SE based on mClust and then

572 annotate the spatial domains. Clustering was performed based on SE by rmclust in mclust
573 package.
574
575 **Spatial trajectory inference**
576 We perform the trajectory inference using SE generated from SECE and STAGATE to verify their
577 reliability. We use PAGA [25] implemented in the SCANPY [27] package to perform the
578 trajectory inference between spatial domains. The PAGA graphs are calculated by scanpy.tl.paga
579 and visualized by scanpy.pl.paga. We use Monocle3 [26] to perform the trajectory inference for
580 individual spots. UMAP embeddings generated from SE are used to calculate the pseudo times by
581 learn_graph and order_cells in Monocle3 package.
582
583 **Deconvolution**
584 For Visium dataset without single-cell resolution, we perform deconvolution to analyze the cell
585 type composition using software Cell2location and annotated scRNA-seq dataset [40].
586 Cell2location is run according to the tutorial and default parameters.
587
588 **Differential expression analysis**
589 Differential expression analysis was used to identify cluster-specific marker genes where all the
590 clusters are pairwise compared using the Wilcoxon method. Each identified marker gene was
591 expressed in a minimum of 25% of cells and at a minimum log fold change threshold of 0.25.
592
593 **Comparison of spatial domain identification approaches**
594 We compared the existing spatial domain identification methods.
595
596 *BayesSpace.* BayesSpace was downloaded from https://github.com/edward130603/BayesSpace.
597 It was run using default parameters specified as top 2,000 HVGs performing PCA and top 15 PCs
598 to infer spatial domains.
599
600 *SpaGCN.* SpaGCN was downloaded from https://github.com/jianhuupenn/SpaGCN. SpaGCN was
601 run using default training parameters specified as 0.05 for learning rate and 200 for max epochs.

For the Visium data we set histology=True and for the other data we set histology=False. In addition, we set s=1 for STARmap data as recommended in the article, and the rest of the data follow the default parameters.

STAGATE. STAGATE was downloaded from https://github.com/QIFEIDKN/STAGATE. We run the default parameters except for rad_cutoff. We followed the tutorial of STAGATE and set rad_cutoff=50 for Stereo-seq and Slide-seqV2 datasets, rad_cutoff=150 for Visium datasets, and rad_cutoff=400 for STARmap datasets.

BASS. We downloaded BASS from https://github.com/zhengli09/BASS, and ran BASS following the vignettes https://zhengli09.github.io/BASS-Analysis.

SpaceFlow. We downloaded SpaceFlow from https://github.com/hongleir/SpaceFlow, and ran SpaceFlow following https://github.com/hongleir/SpaceFlow/blob/master/tutorials.

GraphST. We downloaded GraphST from https://github.com/JinmiaoChenLab/GraphST, and ran GraphST following https://deepst-tutorials.readthedocs.io/en/latest/index.html.

SpatialPCA. We downloaded the R package from https://github.com/shangll123/SpatialPCA, and ran SpatialPCA following http://lulushang.org/SpatialPCA_Tutorial/.

conST. We referred to https://github.com/ys-zong/conST to ran conST.

Compared methods for cell type identification

We compared the existing cell type identification methods.

SpiceMix. We referred to https://github.com/ma-compbio/SpiceMix to ran SpiceMix.

Seurat. We used the R package of Seurat v 4.2.1.

SC3. We downloaded SC3 from https://github.com/hemberg-lab/SC3, and ran SC3 following http://www.bioconductor.org/packages/release/bioc/vignettes/SC3/inst/doc/SC3.html.

CIDR. We referred to https://github.com/VCCRI/CIDR to ran SpiceMix..

Evaluation

Local inverse Simpson's index (LISI). LISI [64] is used to measure the degree of local mixing, and we use it to evaluate the level of spatial pattern. For every sample i, $LISI$ can be formulated as

$$LISI(i) = \frac{1}{\sum_{l \in L} p_i(l)},$$

where $p(l)$ is the probability that the spatial domain cluster label l exists in the local neighborhood of sample i, and L is the set of spatial domains.

In this paper, local neighborhoods are generated by spatial location, labels are spatial domain labels predicted by each algorithm. The value of LISI is $[1, \infty)$, and smaller LISI indicates better spatial patterns.

Adjusted Rand index (ARI). ARI evaluates the degree of overlap between the two divisions. In this paper, we use ARI to evaluate the consistency of the spatial domains divided by various methods with the ground truth domains. ARI can be written as

$$\mathrm{ARI} = \frac{\sum_{\mathrm{ij}} \binom{\mathrm{n_{ij}}}{2} - [\sum_{\mathrm{i}} \binom{\mathrm{a_i}}{2} \sum_{\mathrm{j}} \binom{\mathrm{b_j}}{2}] / \binom{\mathrm{n}}{2}}{\frac{1}{2}[\sum_{\mathrm{i}} \binom{\mathrm{a_i}}{2} + \sum_{\mathrm{j}} \binom{\mathrm{b_j}}{2}] - [\sum_{\mathrm{i}} \binom{\mathrm{a_i}}{2} \sum_{\mathrm{j}} \binom{\mathrm{b_j}}{2}] / \binom{\mathrm{n}}{2}},$$

where n_{ij}, a_i and b_i are values from the contingency table. Here, the ground truth labels are the manual annotation labels of domains, and the labels to be evaluated are the results of various spatial domain identification methods. The value of ARI is $[-1,1]$, and greater ARI indicates better agreements with true label.

Normalized mutual information (NMI). NMI evaluates the similarity between the predicted and true labels. In this paper, we use NMI to evaluate the consistency of the spatial domains divided by various methods with the ground truth domains. NMI can be written as

$$NMI(P,T) = \frac{2MI(P,T)}{H(P)+H(T)},$$

where MI is Mutual Information, and $H(P)$ is Shannon entropy. The value of NMI is [0,1], and greater NMI indicates higher similarity with true label.

Accuracy (ACC). ACC is calculated as

$$ACC = \frac{\sum_{i=1}^{N} \delta(r_i, o(s_i))}{N},$$

where N is the number of spots, r_i and s_i are the real and predicted spatial domain label of spot i, respectively. δ is a function that can be defined as

$$\delta(x,y) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}$$

o is a mapping function that takes the real label r_i as the reference label and then rearranges s_i in the same arrangement, which is implemented using the classical Kuhn-Munkres algorithm [28].

Average Silhouette width (ASW). ASW is used to describe the sharpness for each category after clustering, and we use ASW to evaluate how well the SE obtained by various methods explain the known spatial layers. For every sample i, Silhouette width $S(i)$ is calculated as

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}},$$

where $a(i)$ is the average distance from i to all data points in the same cluster C_i, and $b(i)$ is calculated similarly to $a(i)$, just need to get multiple average distance from all clusters except C_i, then choose the smallest value from it as the final $b(i)$.

In this paper, distance is calculated by SE of various methods, and cluster is the layer annotated by original author. The value of ASW is $[-1,1]$, and greater ASW indicates better embeddings.

Adjusted Rand index of cell types (ARI-Cell Type). For data with original cell type annotations, we assessed the consistency of various cell type identification algorithms with the original annotations using ARI. The calculation formula of ARI is in the Adjusted Rand index (ARI) section. In order to distinguish it from the evaluation index of the spatial domain identification, we named this ARI as ARI-Cell Type. Here, the ground truth labels are the annotation labels of the

original articles, and the labels to be evaluated are the results of various cell type identification algorithms.

Declarations

Ethics approval and consent to participate

No ethnical approval was required for this study.

Consent for publication

Not applicable.

Availability of data and materials

Slide-seq datasets are available at https://portals.broadinstitute.org/single_cell/study/slide-seq-study. Stereo-seq mouse hemibrain dataset is available at https://db.cngb.org/stomics/mosta/. Stereo-seq olfactory bulb dataset is available at https://github.com/JinmiaoChenLab/SEDR_analyses. Slide-seqV2 hippocampus datasets are available at https://singlecell.broadinstitute.org/. STARmap mouse visual cortex dataset is available at http://clarityresourcecenter.org/. Visium human breast cancer dataset is available at https://www.10xgenomics.com/resources/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0. Visium DLPFC dataset is available within the spatialLIBD package (http:://spatial.libd.org/spatialLIBD). We also organized these raw data, including expression counts and coordinates, as well as H&E images of Visium, into a format that is easy to read by SCANPY, which were available at https://drive.google.com/drive/my-drive. The SECE algorithm is implemented and provided as a pip installable Python package which is available on Github https://github.com/yuyuanyuana/SECE. The source code and datasets are available at Zenodo https://zenodo.org/record/8130682.

Competing interests

The authors declare that they have no competing interests.

Funding

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

Authors' contributions

Z.X. conceived and supervised the study. YY.Y. and Z.X. designed the study. YY.Y. analyzed the data. All the authors interpreted the data. YY.Y. and Z.X. wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

We would like to thank all the data and software contributors who make this research possible. We also thank the Center for Precision Medicine at Sun Yat-sen University for the long-term support.

Authors' information

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

Yuanyuan Yu, Yao He & Zhi Xie

References

1. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, 2nd, Tran MN, Besich Z, Tippani M, et al: **Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex.** *Nat Neurosci* 2021, **24:**425-436.

741 2. Wang M, Hu Q, Lv T, Wang Y, Lan Q, Xiang R, Tu Z, Wei Y, Han K, Shi C, et al: **High-**
742 **resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and**
743 **larvae.** *Dev Cell* 2022, **57:**1271-1283 e1274.
744 3. Hunter MV, Moncada R, Weiss JM, Yanai I, White RM: **Spatially resolved transcriptomics**
745 **reveals the architecture of the tumor-microenvironment interface.** *Nat Commun* 2021,
746 **12:**6278.
747 4. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C,
748 Ramakrishnan C, Liu J, et al: **Three-dimensional intact-tissue sequencing of single-cell**
749 **transcriptional states.** *Science* 2018, **361**.
750 5. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L: **Single-cell in situ RNA profiling by**
751 **sequential hybridization.** *Nat Methods* 2014, **11:**360-361.
752 6. Shah S, Lubeck E, Zhou W, Cai L: **In Situ Transcription Profiling of Single Cells Reveals**
753 **Spatial Organization of Cells in the Mouse Hippocampus.** *Neuron* 2016, **92:**342-357.
754 7. Eng CL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC,
755 Cai L: **Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH.** *Nature* 2019,
756 **568:**235-239.
757 8. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND,
758 Hao J, Regev A, Dulac C, Zhuang X: **Molecular, spatial, and functional single-cell profiling of**
759 **the hypothalamic preoptic region.** *Science* 2018, **362**.
760 9. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C,
761 Amamoto R, et al: **Highly multiplexed subcellular RNA sequencing in situ.** *Science* 2014,
762 **343:**1360-1363.
763 10. Moses L, Pachter L: **Museum of spatial transcriptomics.** *Nat Methods* 2022, **19:**534-546.
764 11. Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M,
765 Westholm JO, Huss M, et al: **Visualization and analysis of gene expression in tissue sections**
766 **by spatial transcriptomics.** *Science* 2016, **353:**78-82.
767 12. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, Chen F:
768 **Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2.** *Nat*
769 *Biotechnol* 2021, **39:**313-319.

770 13. Vickovic S, Eraslan G, Salmen F, Klughammer J, Stenbeck L, Schapiro D, Aijo T, Bonneau R,
771 Bergenstrahle L, Navarro JF, et al: **High-definition spatial transcriptomics for in situ tissue**
772 **profiling.** *Nat Methods* 2019, **16:**987-990.
773 14. Cho CS, Xi J, Si Y, Park SR, Hsu JE, Kim M, Jun G, Kang HM, Lee JH: **Microscopic**
774 **examination of spatial transcriptome using Seq-Scope.** *Cell* 2021, **184:**3559-3572 e3522.
775 15. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, Qiu X, Yang J, Xu J, Hao S, et al:
776 **Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned**
777 **arrays.** *Cell* 2022, **185:**1777-1792 e1721.
778 16. Ortiz C, Navarro JF, Jurek A, Martin A, Lundeberg J, Meletis K: **Molecular atlas of the adult**
779 **mouse brain.** *Sci Adv* 2020, **6:**eabb3446.
780 17. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR,
781 Taylor SEB, Nghiem P, et al: **Spatial transcriptomics at subspot resolution with BayesSpace.**
782 *Nat Biotechnol* 2021, **39:**1375-1384.
783 18. Shang L, Zhou X: **Spatially aware dimension reduction for spatial transcriptomics.** *Nat*
784 *Commun* 2022, **13:**7203.
785 19. Li Z, Zhou X: **BASS: multi-scale and multi-sample analysis enables accurate cell type**
786 **clustering and spatial domain detection in spatial transcriptomic studies.** *Genome Biol* 2022,
787 **23:**168.
788 20. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M: **SpaGCN:**
789 **Integrating gene expression, spatial location and histology to identify spatial domains and**
790 **spatially variable genes by graph convolutional network.** *Nat Methods* 2021, **18:**1342-1351.
791 21. Dong K, Zhang S: **Deciphering spatial domains from spatially resolved transcriptomics with**
792 **an adaptive graph attention auto-encoder.** *Nat Commun* 2022, **13:**1739.
793 22. Long Y, Ang KS, Li M, Chong KLK, Sethi R, Zhong C, Xu H, Ong Z, Sachaphibulkij K, Chen A,
794 et al: **Spatially informed clustering, integration, and deconvolution of spatial transcriptomics**
795 **with GraphST.** *Nat Commun* 2023, **14:**1155.
796 23. Ren H, Walker BL, Cang Z, Nie Q: **Identifying multicellular spatiotemporal organization of**
797 **cells with SpaceFlow.** *Nat Commun* 2022, **13:**4076.

24. Zong Y, Yu T, Wang X, Wang Y, Hu Z, Li Y: **conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics.** *bioRxiv* 2022**:**2022.2001.2014.476408.

25. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Gottgens B, Rajewsky N, Simon L, Theis FJ: **PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells.** *Genome Biol* 2019, **20:**59.

26. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, et al: **The single-cell transcriptional landscape of mammalian organogenesis.** *Nature* 2019, **566:**496-502.

27. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis.** *Genome Biol* 2018, **19:**15.

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

29. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al: **Integrated analysis of multimodal single-cell data.** *Cell* 2021, **184:**3573-3587 e3529.

30. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M: **SC3: consensus clustering of single-cell RNA-seq data.** *Nat Methods* 2017, **14:**483-486.

31. Lin P, Troup M, Ho JW: **CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data.** *Genome Biol* 2017, **18:**59.

32. Chidester B, Zhou T, Alam S, Ma J: **SPICEMIX enables integrative single-cell spatial modeling of cell identity.** *Nat Genet* 2023, **55:**78-88.

33. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, et al: **Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain.** *Cell* 2018, **174:**1015-1030.e1016.

34. Fu H, Xu H, Chong K, Li M, Ang KS, Lee HK, Ling J, Chen A, Shao L, Liu L, Chen J: **Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics.** *bioRxiv* 2021.

827 35. Su X, Vasilkovska T, Frohlich N, Garaschuk O: **Characterization of cell type-specific S100B**
828 **expression in the mouse olfactory bulb.** *Cell Calcium* 2021, **94:**102334.
829 36. Nagayama S, Homma R, Imamura F: **Neuronal organization of olfactory bulb circuits.** *Front*
830 *Neural Circuits* 2014, **8:**98.
831 37. Beiersdorfer A, Wolburg H, Grawe J, Scheller A, Kirchhoff F, Lohr C: **Sublamina-specific**
832 **organization of the blood brain barrier in the mouse olfactory nerve layer.** *Glia* 2020,
833 **68:**631-645.
834 38. Van der Maaten L, Hinton G: **Visualizing data using t-SNE.** *Journal of machine learning*
835 *research* 2008, **9**.
836 39. Larsson L, Frisen J, Lundeberg J: **Spatially resolved transcriptomics adds a new dimension to**
837 **genomics.** *Nat Methods* 2021, **18:**15-18.
838 40. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C,
839 Torpy JR, Bartonicek N, et al: **A single-cell and spatially resolved atlas of human breast**
840 **cancers.** *Nat Genet* 2021, **53:**1334-1347.
841 41. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Elmentaite R, Lomakin A,
842 Kedlian V, Gayoso A, et al: **Cell2location maps fine-grained cell types in spatial**
843 **transcriptomics.** *Nat Biotechnol* 2022, **40:**661-671.
844 42. Bilandzic M, Rainczuk A, Green E, Fairweather N, Jobling TW, Plebanski M, Stephens AN:
845 **Keratin-14 (KRT14) Positive Leader Cells Mediate Mesothelial Clearance and Invasion by**
846 **Ovarian Cancer Cells.** *Cancers (Basel)* 2019, **11**.
847 43. Cheung KJ, Padmanaban V, Silvestri V, Schipper K, Cohen JD, Fairchild AN, Gorin MA,
848 Verdone JE, Pienta KJ, Bader JS, Ewald AJ: **Polyclonal breast cancer metastases arise from**
849 **collective dissemination of keratin 14-expressing tumor cell clusters.** *Proc Natl Acad Sci U S A*
850 2016, **113:**E854-863.
851 44. Martinez-Ordonez A, Seoane S, Avila L, Eiro N, Macia M, Arias E, Pereira F, Garcia-Caballero
852 T, Gomez-Lado N, Aguiar P, et al: **POU1F1 transcription factor induces metabolic**
853 **reprogramming and breast cancer progression via LDHA regulation.** *Oncogene* 2021,
854 **40:**2725-2740.
855 45. Nalio Ramos R, Missolo-Koussou Y, Gerber-Ferder Y, Bromley CP, Bugatti M, Nunez NG,
856 Tosello Boari J, Richer W, Menger L, Denizeau J, et al: **Tissue-resident FOLR2(+)**

857 **macrophages associate with CD8(+) T cell infiltration in human breast cancer.** *Cell* 2022,
858 **185:**1189-1207 e1125.
859 46. Zhang H, Wang Y, Liu C, Li W, Zhou F, Wang X, Zheng J: **The Apolipoprotein C1 is involved**
860 **in breast cancer progression via EMT and MAPK/JNK pathway.** *Pathol Res Pract* 2022,
861 **229:**153746.
862 47. Seo SU, Woo SM, Im SS, Jang Y, Han E, Kim SH, Lee H, Lee HS, Nam JO, Gabrielson E, et al:
863 **Cathepsin D as a potential therapeutic target to enhance anticancer drug-induced apoptosis**
864 **via RNF183-mediated destabilization of Bcl-xL in cancer cells.** *Cell Death Dis* 2022, **13:**115.
865 48. Zhang C, Zhang M, Song S: **Cathepsin D enhances breast cancer invasion and metastasis**
866 **through promoting hepsin ubiquitin-proteasome degradation.** *Cancer Lett* 2018, **438:**105-
867 115.
868 49. Revel M, Sautes-Fridman C, Fridman WH, Roumenina LT: **C1q+ macrophages: passengers or**
869 **drivers of cancer progression.** *Trends Cancer* 2022, **8:**517-526.
870 50. Noblejas-Lopez MDM, Nieto-Jimenez C, Morcillo Garcia S, Perez-Pena J, Nuncia-Cantarero M,
871 Andres-Pretel F, Galan-Moya EM, Amir E, Pandiella A, Gyorffy B, Ocana A: **Expression of**
872 **MHC class I, HLA-A and HLA-B identifies immune-activated breast tumors with favorable**
873 **outcome.** *Oncoimmunology* 2019, **8:**e1629780.
874 51. Nomura T, Huang W-C, E Zhau H, Josson S, Mimata H, WK Chung L: **β2-Microglobulin-**
875 **mediated signaling as a target for cancer therapy.** *Anti-Cancer Agents in Medicinal Chemistry*
876 *(Formerly Current Medicinal Chemistry-Anti-Cancer Agents)* 2014, **14:**343-352.
877 52. Wu B, Pan Y, Liu G, Yang T, Jin Y, Zhou F, Wei Y: **MRPS30-DT Knockdown Inhibits Breast**
878 **Cancer Progression by Targeting Jab1/Cops5.** *Front Oncol* 2019, **9:**1170.
879 53. Zeng B, Ge C, Li R, Zhang Z, Fu Q, Li Z, Lin Z, Liu L, Xue Y, Xu Y, et al: **Knockdown of**
880 **microsomal glutathione S-transferase 1 inhibits lung adenocarcinoma cell proliferation and**
881 **induces apoptosis.** *Biomed Pharmacother* 2020, **121:**109562.
882 54. Huang X, Yu J, Lai S, Li Z, Qu F, Fu X, Li Q, Zhong X, Zhang D, Li H: **Long Non-Coding RNA**
883 **LINC00052 Targets miR-548p/Notch2/Pyk2 to Modulate Tumor Budding and Metastasis of**
884 **Human Breast Cancer.** *Biochem Genet* 2022.
885 55. Xiong D, Wang D, Chen Y: **Role of the long non-coding RNA LINC00052 in tumors.** *Oncol*
886 *Lett* 2021, **21:**316.

887 56. Salameh A, Fan X, Choi BK, Zhang S, Zhang N, An Z: **HER3 and LINC00052 interplay**
888 **promotes tumor growth in breast cancer.** *Oncotarget* 2017, **8:**6526-6539.
889 57. Grzybowska-Szatkowska L, Slaska B: **Mitochondrial NADH dehydrogenase polymorphisms**
890 **are associated with breast cancer in Poland.** *J Appl Genet* 2014, **55:**173-181.
891 58. Lyu L, Wang M, Zheng Y, Tian T, Deng Y, Xu P, Lin S, Yang S, Zhou L, Hao Q, et al:
892 **Overexpression of FAM234B Predicts Poor Prognosis in Patients with Luminal Breast**
893 **Cancer.** *Cancer Manag Res* 2020, **12:**12457-12471.
894 59. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ: **Single-cell RNA-seq denoising using a**
895 **deep count autoencoder.** *Nat Commun* 2019, **10:**390.
896 60. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N: **Deep generative modeling for single-cell**
897 **transcriptomics.** *Nat Methods* 2018, **15:**1053-1058.
898 61. Hafemeister C, Satija R: **Normalization and variance stabilization of single-cell RNA-seq data**
899 **using regularized negative binomial regression.** *Genome Biol* 2019, **20:**296.
900 62. Traag VA, Waltman L, van Eck NJ: **From Louvain to Leiden: guaranteeing well-connected**
901 **communities.** *Sci Rep* 2019, **9:**5233.
902 63. Scrucca L, Fop M, Murphy TB, Raftery AE: **mclust 5: Clustering, Classification and Density**
903 **Estimation Using Gaussian Finite Mixture Models.** *R J* 2016, **8:**289-317.
904 64. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh
905 PR, Raychaudhuri S: **Fast, sensitive and accurate integration of single-cell data with**
906 **Harmony.** *Nat Methods* 2019, **16:**1289-1296.
907

Figures

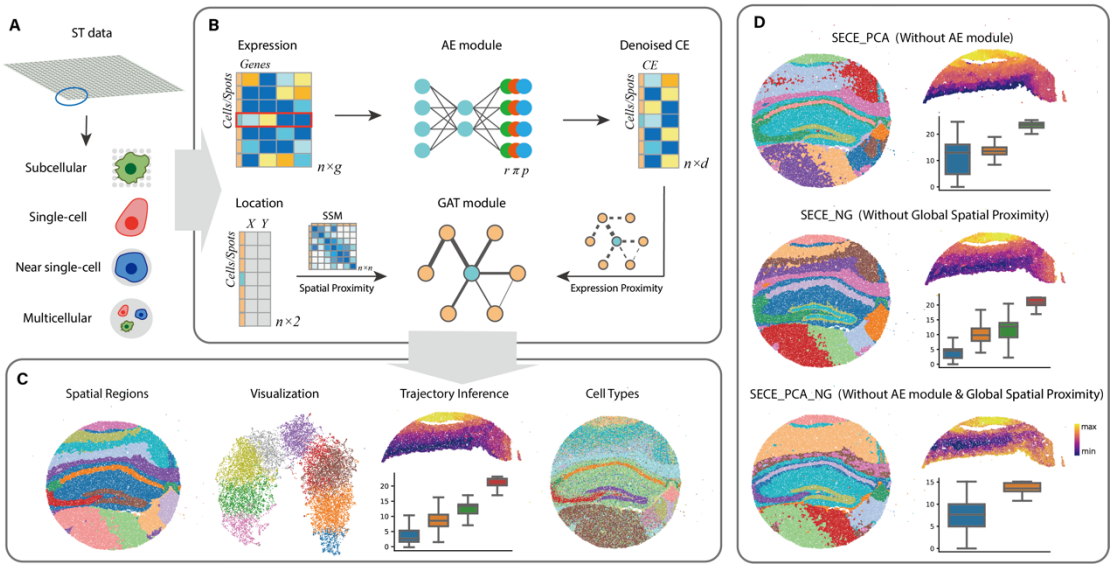


Fig.1. Overview of SECE. (A) SECE is applicable to a wide range of spatial transcriptome data with different resolution, including subcellular, single-cell, near-single-cell, and multicellular resolutions. **(B)** SECE takes gene expression profile and spatial coordinates as input. SECE begins with AE module, which learns CE under the count distribution. Then, GAT module learns the SE by balancing the expressive proximity and the spatial proximity, which are measured by the distance of the CE and the distance of the coordinate, respectively. **(C)** The main output includes spatial region and cell type labels, which are obtained by clustering CE and SE separately. SE can also perform downstream tasks such as low dimensional visualization and trajectory inference. **(D)** Spatial domain identification and trajectory inference of SEAL_PCA (top), SEAL_NG (middle) and SEAL_PCA_NG (bottom).

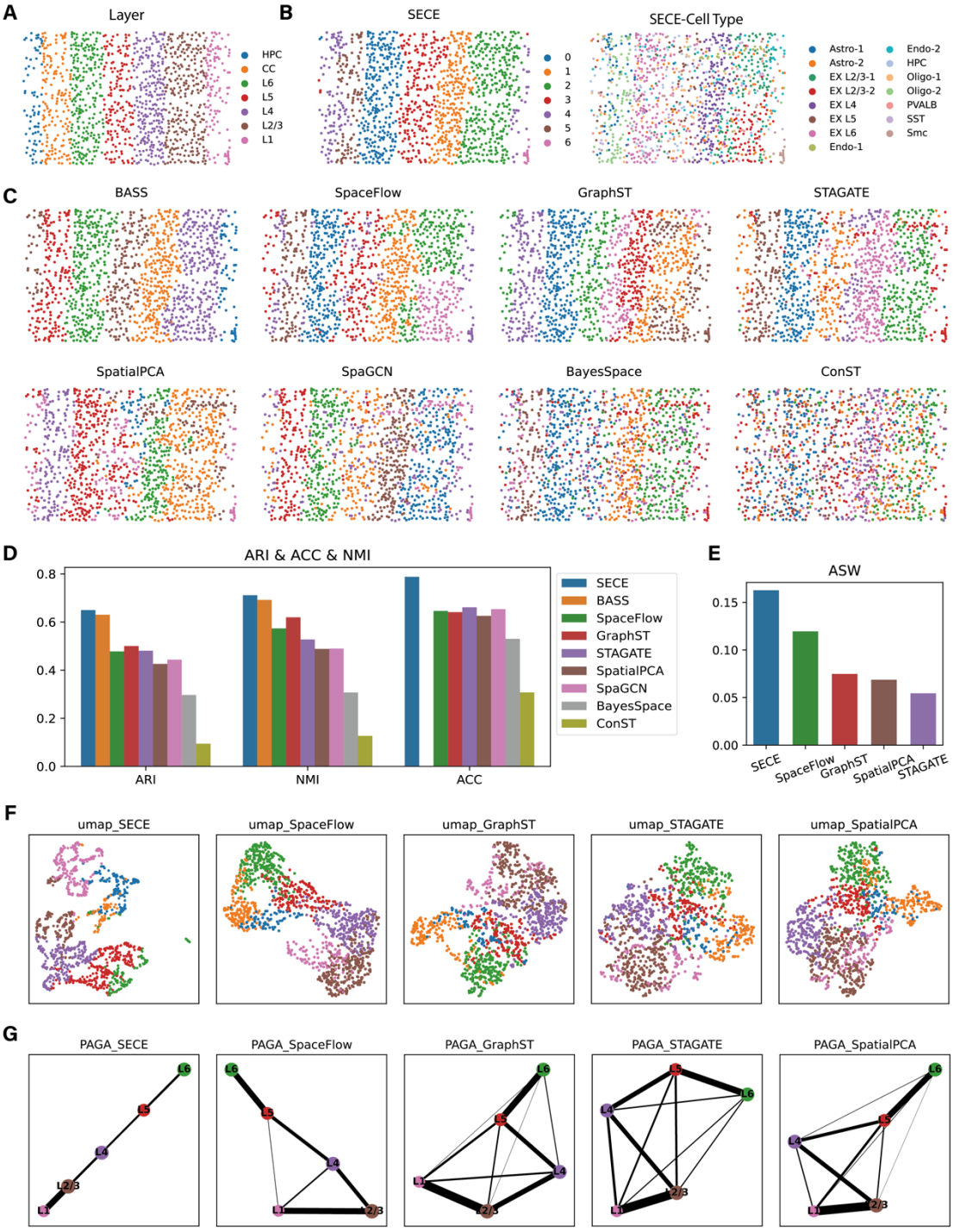
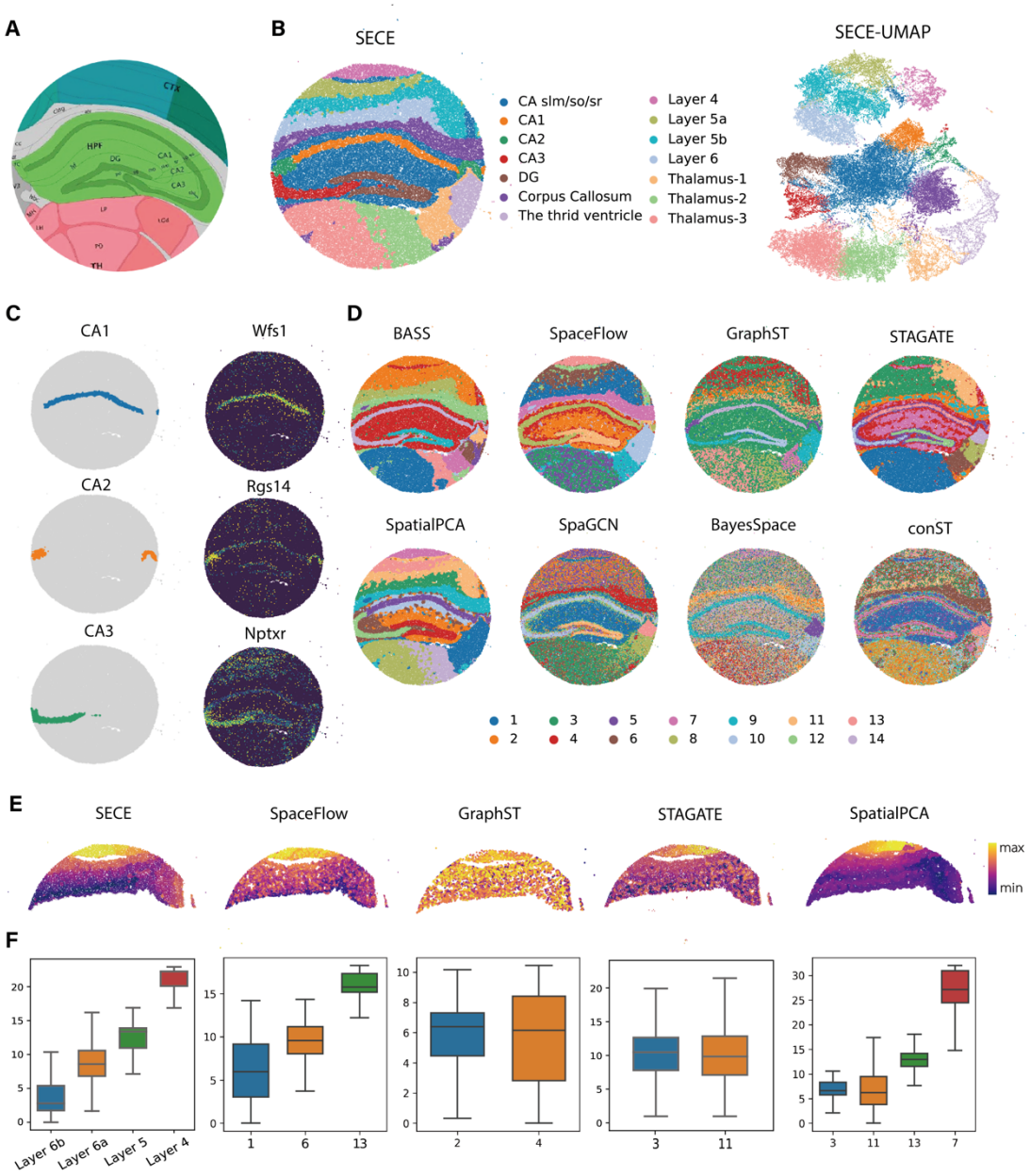


Fig. 2. Application of SECE to mouse cortex STARmap data. (A) Layer structure of the tissue section from the original study. **(B)** Left: Spatial domains identified by SECE. Right: Cell types identified by SECE. Astro, astrocytes; Oligo, oligodendrocytes; EX, excitatory neurons; IN Neuron, inhibitory neurons; Endo, endothelial cells; SMC, smooth muscle cells. **(C)** Spatial domains identified by BASS, SpaceFlow, GraphST, STAGATE, SpatialPCA, SpaGCN, BayesSpace, and conST. **(D)** Accuracy of spatial domain identification for different methods

measured by ARI, NMI, and ACC. (**E)** Barplot of ASW shows the discrimination capability in ground truth layers of SEs learned by SECE, SpaceFlow, GraphST, STAGATE, and SpatialPCA. (**F)** UMAP visualization generated by SECE, SpaceFlow, GraphST, STAGATE, and SpatialPCA, colored by annotated layers. (**G)** PAGA graph generated by SECE, SpaceFlow, GraphST, STAGATE, and SpatialPCA.



934

935 **Fig. 3. Application of SECE to mouse hippocampus Slide-seqV2 data. (A)** The annotation of
936 hippocampus structures from the Allen Reference Atlas of an adult mouse brain. **(B)** Left: Spatial
937 visualization of domains identified by SECE. Right: UMAP visualization of domains identified by
938 SECE. Spatial domains were annotated by the structure annotation showed in the Allen Reference
939 Atlas. CA, Ammon's horn; DG, Dentate gyrus. **(C)** Spatial visualization of CA1, CA2, and CA3
940 domains identified by SECE (Left) and the corresponding marker genes *Wsf1*, *Rgs14*, and *Nptxr*
941 (right). **(D)** 14 spatial regions identified by BASS, SpaceFlow, GraphST, STAGATE, SpatialPCA,
942 SpaGCN, BayesSpace, and conST. **(E)** Pseudo-time of each cell calculated by Monocle3 based on
943 embedding of SECE, SpaceFlow, GraphST, STAGATE, and SpatialPCA. **(F)** Pseudo-time of the

944 cells in each layer of the isocortex based on SECE, SpaceFlow, GraphST, STAGATE, and

945 SpatialPCA.

946

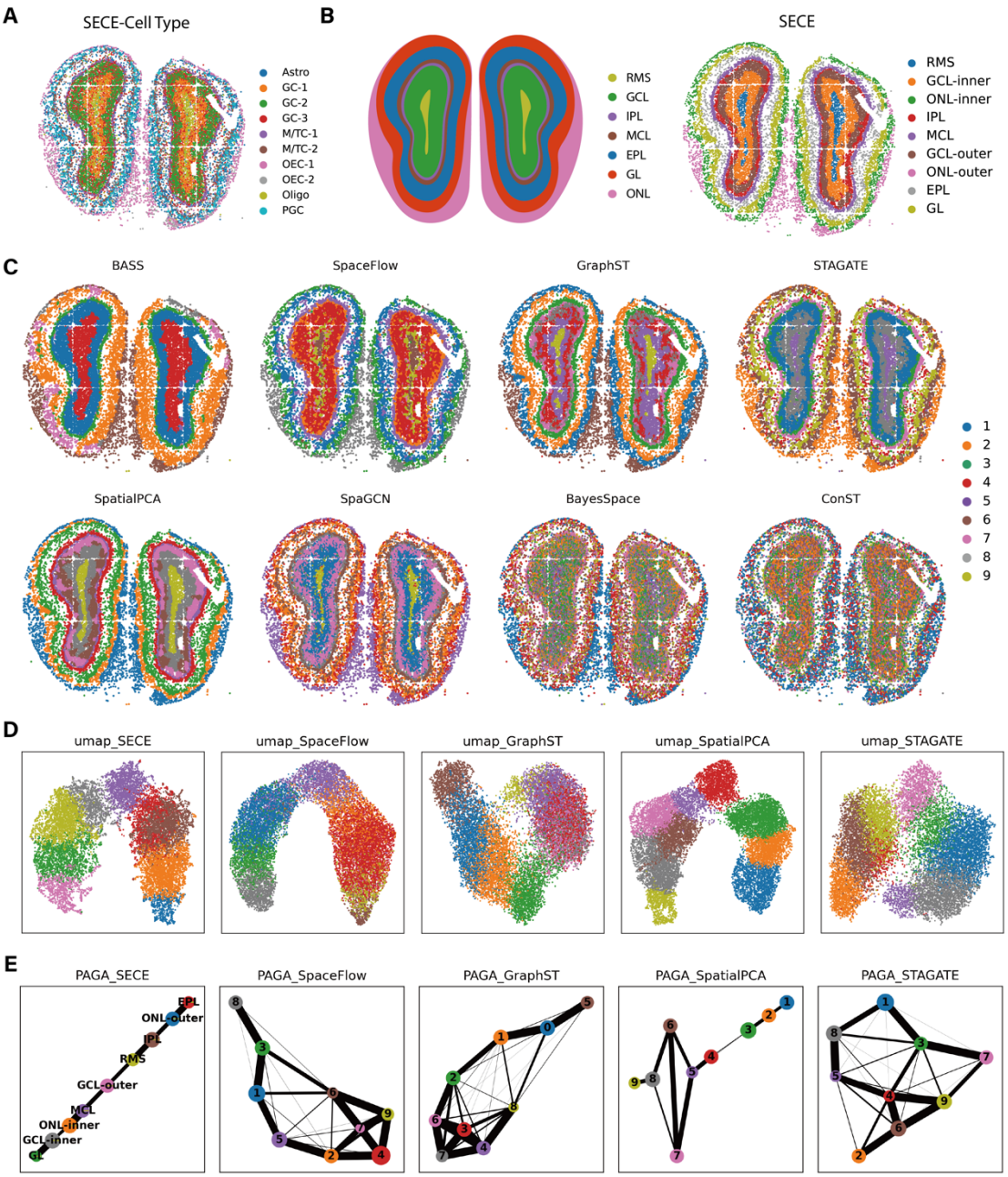


Fig. 4. Application of SECE to mouse olfactory bulb Stereo-seq data. (A) Cell types identified by SECE. GC, Granule cells; M/TC, Mitral and tufted cells; PGC, Periglomerular cells; OEC, Olfactory ensheathing cells. **(B)** Left: The annotation of mouse olfactory bulb structures from the Allen Reference Atlas. Right: Spatial visualization of domains identified by SECE. RMS, Rostral migratory stream; GCL, Granule cell layer; IPL, Inner plexiform layer; MCL, Mitral cell layer; EPL, External plexiform layer; GL, Glomerular layer; ONL, Olfactory nerve layer. **(C)** 9 spatial regions identified by BASS, SpaceFlow, GraphST, STAGATE, SpatialPCA, SpaGCN, BayesSpace, and conST. **(D)** UMAP visualization generated by SECE, SpaceFlow, GraphST,

STAGATE, and SpatialPCA, colored by annotated layers. (**E)** PAGA graph generated by SECE, SpaceFlow, GraphST, STAGATE, and SpatialPCA.

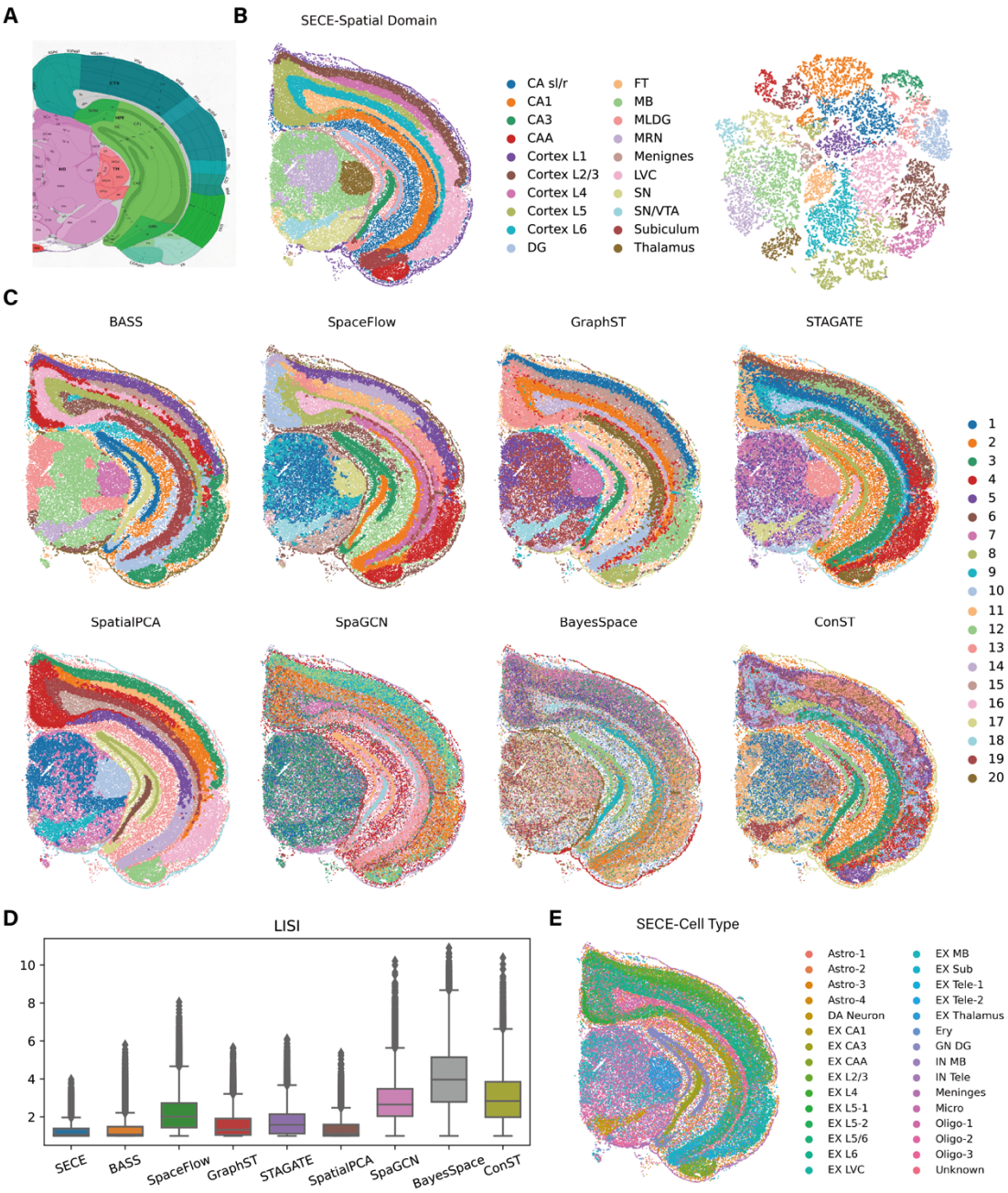


Fig. 5. Application of SECE to mouse brain Stereo-seq data. (A) The annotation of mouse olfactory bulb structures from the Allen Reference Atlas. **(B)** Left: Spatial visualization of domains identified by SECE. Right: UMAP visualization of domains identified by SECE. Spatial domains were annotated by the structure annotation showed in the Allen Reference Atlas. sl/r, stratum lacunosum / raditum cornu ammonis; DG, dentate gyrus; FT, fiber tract; MLDG, molecular layer of dentate gyrus; MRN, midbrain reticular nucleus; SN, substantia nigra; VTA, ventral tegmental area. **(C)** 20 spatial regions identified by BASS, SpaceFlow, GraphST, STAGATE, SpatialPCA, SpaGCN, BayesSpace, and conST. **(D)** Boxplot of LISI measuring the

968 degree of spatial aggregation of domains identified by different methods. (**E)** Cell types identified
969 by SECE. Astro, astrocytes; DA Neuron: dopaminergic neuron; EX, excitatory neurons; CA,
970 Cornu Ammonis; CAA, cortical amygdalar area; LVC, lateral-ventral cortex; MB, midbrain; Ery:
971 erythrocyte; GN DG: dentate gyrus of granule neurons; IN, inhibitory neurons; Tele:
972 telencephalon; Oligo, oligodendrocytes; Micro, microglia.

973

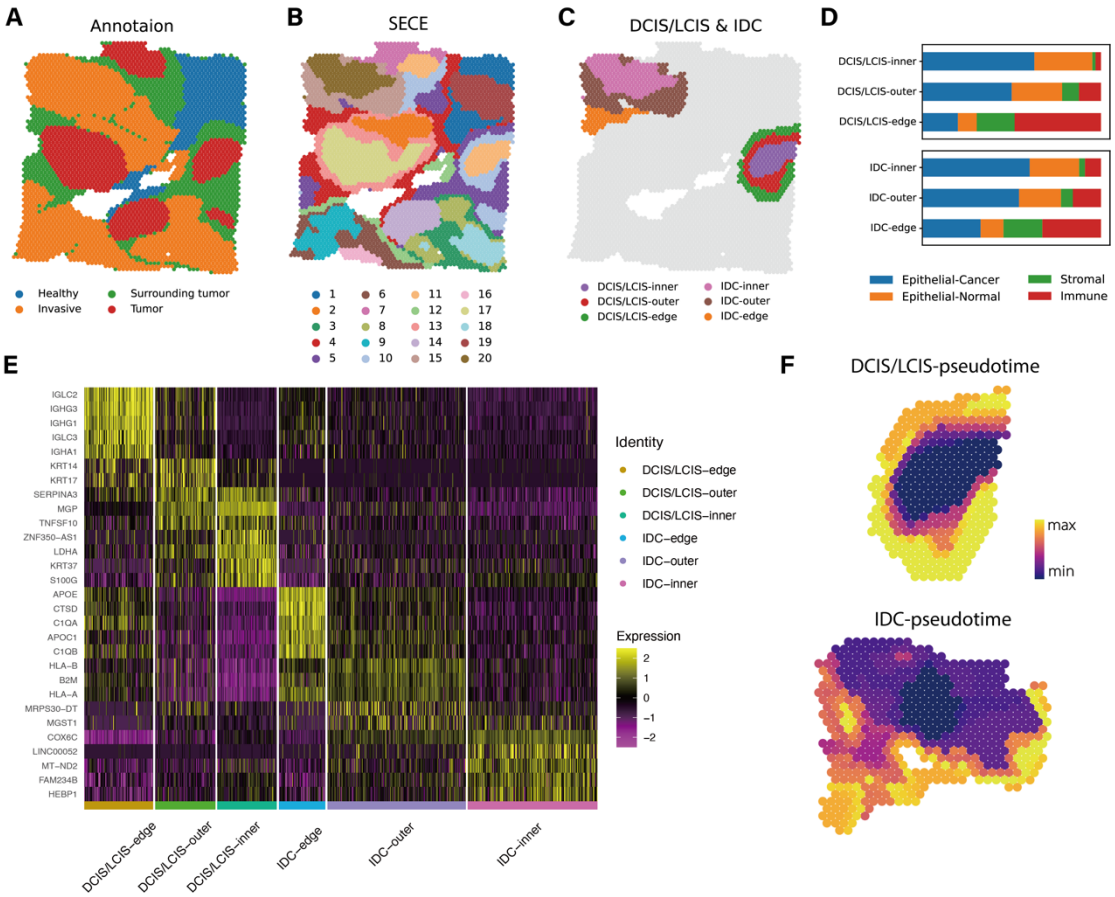


Fig. 6 Application of SECE to human breast cancer Visium data. (A) Pathology annotation of the tissue section from the original study. **(B)** Spatial regions identified by SECE. **(C)** Fine regions identified by SECE, that is, DCIS/LCIS-inner, DCIS/LCIS-outer, DCIS/LCIS-edge, IDC-inner, IDC-outer, and IDC-edge. **(D)** Cell type compositions of 6 fine regions. **(E)** Heatmaps of normalized expression of signature genes identified in the DE analysis based on 6 fine regions. **(F)** Pseudo-times of spots calculated by Monocle3 in DCIS/LCIS (top) and IDC (bottom).