bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Analysis of spatial transcriptomics at varying resolution levels using the unified framework of SpaSEG

Yong Bai[1,*,†], Xiangyu Guo[1,†], Keyin Liu[1,2,†], Qiuhong Luo[1], Yingyue Wang[1,3], Jianhua Yin[1], Yuxiang Li[1], Ao Chen[1,4], Xun Xu[1,5], Xin Jin[1,*]

[1]BGI-Shenzhen, Shenzhen, 518083, Guangdong, China

[2]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, 518055, China

[3]School of Life Sciences, Southern University of Science and Technology, Shenzhen Guangdong, 518055, China

[4]Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark

[5]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen 518120, China

[*]Correspondence email: baiyong@genomics.cn, jinxin@genomics.cn

[†]These authors contributed equally to this work: Yong Bai, Xiangyu Guo, Keyin Liu.

Abstract

Recent improvements in spatial transcriptomics technologies have enabled the characterization of complex cellular mechanisms within tissue context through unbiased profiling of genome-wide transcriptomes in conjunction with spatial

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

coordinates. These technologies require a systematic analysis approach to deciphering the complex tissue architecture. Here, we develop SpaSEG, an unsupervised convolutional neural network-based method towards this end by jointly learning gene expression similarity of spots and their spatial contiguousness via adopting a loss function for spatial boundary continuity. Using several spatial transcriptomics datasets generated by different platforms with varying resolutions and assayed tissue sizes, we extensively demonstrate that not only can SpaSEG better identify spatial domains, but also be much more computationally and memory efficient than existing methods. In addition, SpaSEG is able to effectively detect genes with spatial expression patterns and infer spot-wise intercellular interactions as well as cell-type colocalization within the tissue section by utilizing the identified domains. Taken together, our results have indicated the flexibility of SpaSEG in multiple analysis tasks in spatial transcriptomics, making it as a desirable tool in facilitating the exploration of tissue architecture and the knowledge of underlying biology.

Introduction

Coordinated activities of diverse cells with the spatial context in tissues that underlie their communications with surroundings and sophisticated biological processes can be characterized by spatial gene expression patterns. Emerging spatial transcriptomics (ST) technology has allowed the unbiased profiling of genome-wide gene expressions with physical capture sites (referred to as spots), offering a quantitative and spatial snapshot of cellular heterogeneity across a tissue section [1-3]. Recent years have witnessed considerable progress in the ST experimental methods including the imaging-based *in-situ* transcriptomics approaches like MERFISH [4] and seqFISH[5], and the next-generation sequencing (NGS)-based approaches such as Slide-seqV2 [6], 10x

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genomics Visium [7], and Stereo-seq [8]. These methods have reached an astonishing resolution from multicellular to single-cell or even subcellular level with varying gene throughput, delivering unprecedented insights into tissue-specific function, development, and pathology through elucidating tissue architectures with myriad cell types [9, 10].

A pivotal task in ST data analysis is to identify spatial domains defined as regions in the tissue section with coherence in both gene expressions and spatial dependency. The identification of spatial domains serves unravelling tissue structures, facilitating the characterization of cell type composition and transcriptomic profiles in the tissue microenvironments [8, 11, 12]. Conventional approaches to arrange spots into distinct spatial domains resort to clustering methods such as Leiden [13] that only take into account gene expressions without considering localization information, usually leading to the loss of spatial contiguousness. Recently, several deep learning-based and statistical based methods have been developed to enhance spatial domain identification through incorporating gene expression with spatial information, including SEDR [14], BayesSpace [15], SpaGCN [16], stLearn [17] and Giotto [18], to name a few. Despite promising performance, these methods are only applied on the ST datasets with limited spots and thereby may be fragile for larger tissue section with higher capture resolution. Besides, these methods performing spatial clustering is absent of considering the boundary constraint of spatial domains.

Additionally, knowledge of biological functions associated with spatial domains necessitates pinpointing genes that exhibit spatial expression variations and patterns known as spatially variable genes (SVGs). A handful of methods such as trendsceek [19], SpatialDE[20] and SPARK[21] have been proposed to identify SVGs by modeling

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

spatial dependency of gene expressions and conducting correlation testing between the distribution of gene expression and spatial localization. These methods apparently neglected spatial domains and may obtain suspectable spatial gene expression patterns, failing to fully reflect tissue-specific spatial functions. Moreover, spatial variations in gene expression across spatial domains can imply spatial patterns in cell-cell interactions (CCIs) in a tissue [22]. However, the majority of existing methods do not present the automatic detection of CCIs across whole tissue section using ST data.

Here, we propose SpaSEG, a simply yet powerful unsupervised convolutional neural network (CNN)-based model to jointly identify spatial domains, SVGs and CCIs. In brief, SpaSEG first establishes an unsupervised CNN network through learning gene expression similarity in conjunction with spatial coordinates to detect spatial domains. SVGs are then detected for each spatial domain to ensure spatial gene expression patterns. CCIs are further examined by investigating enriched expression of ligand-receptor (L-R) pairs in both intra and inter spatial domains. By analyzing several datasets that encompass a wide range of ST platforms, including 10x Genomics Visium, Slide-seqV2, seqFISH, MERFISH, and Stereo-seq, we extensively demonstrate that SpaSEG exhibits superior performance on the identification of spatial domain over the existing state-of-the-art methods. We also use ST datasets of mouse embryo to examine the detection of SVGs, as well as mouse brain and human breast cancer section to delineate L-R pairs in normal and tumor microenvironment, respectively. Through comprehensive analyses, we show that SpaSEG is computational and memory efficient as well as applicable for diverse ST platforms and analysis tasks, serving as a desirable tool to explore tissue architecture and cellular characterization for different size of tissue section at varying resolutions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Overview of SpaSEG

SpaSEG starts with raw spatial transcriptomic data preprocessing through low-quality genes and poor spots removal, gene measure normalization and log-transformation, as well as principal component analysis (PCA) and z-score scaling, leading to a d-dimensional feature vector $\mathbf{s}_n \in \mathbb{R}^d$ for each spot n (Fig. 1a). SpaSEG then converts the low-dimensional representation of spots with spatial coordinates to an image-like tensor, where spots are analogous to image pixels while the corresponding d-dimensional feature vectors to image channels. Accordingly, the spatial domain identification problem can be regarded as the pixel-wise image segmentation problem in an unsupervised fashion.

SpaSEG responsible for spatial domain identification is a CNN-based network model that consists of a batch normalization layer, two stacking convolutional modules and a refinement module (Fig.1b). The network model yields a response representation $\mathbf{y}_i \in \mathbb{R}^d$ for each spot i. To initialize model parameters, we first pre-train the model using the mean squared error (MSE) loss between $\mathbf{s}_n$ and $\mathbf{y}_n$ for all spots. Then, the softmax likelihood $\mathbf{p}_n$ and the related pseudo-label c_n for each spot n can be obtained, and thus the classic cross entropy loss $\mathcal{L}_{\text{seg}}$ is applied in the subsequent training iterations. To encourage continuity of neighboring spots, we additionally calculate L1-norm of boundary gradients for each spot with its spatial location as domain boundary constraint loss $\mathcal{L}_{\text{spa}}$, inspired by the previous study [23]. To this end, we optimize the joint loss of the weighted sum of $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{spa}}$ for progressively enhancing spot classification during iterative learning while preserving spatial

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dependency and continuity (Supplementary Figure 1). Finally, spots that present similar gene expression and spatially continuous coordinates are clustered into the same domain, and the detection of SVGs and CCIs based on the identified domains are further investigated (Fig.1c). We also showcase the capability of SpaSEG to integrated analysis of multiple tissue sections, thus helping to discover spatial domains in different tissue samples.

SpaSEG improves spatial domain identification on the human dorsolateral prefrontal cortex dataset

To evaluate the performance of SpaSEG on spatial domain identification, we first downloaded the publicly available dataset of the 10x Genomics Visium human dorsolateral prefrontal cortex (DLPFC) [12] and used as a benchmark. This manually annotated dataset is composed of 12 sections that cover six neuron layers and white matter with the number of spots ranging from 3460 to 4789 (Supplementary Table 1). To compare with SpaSEG, we chose a commonly used non-spatial clustering method Leiden plus five recently published state-of-the-art methods, including stLearn, Giotto, SpaGCN, BayesSpace, and SEDR. Apart from qualitative visualization analysis, two widely used evaluation metrics of adjusted rand index (ARI) [24] and normalized mutual information (NMI) [25] were employed to quantitatively assess performance of these methods.

SpaSEG outperformed the competitive methods for the identification of spatial domains on the 12 DLPFC sections in terms of its highest values of ARI (0.532±0.058; mean ± s.d.) and NMI (0.644±0.020) (Fig. 2a and Supplementary Table 3,4) while requiring the least running time and memory usage except for Leiden partially due to its lack of leveraging spatial information during clustering (Supplementary Figure 2,

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 5, 6). SpaSEG also yielded the neatest spatial domains with clear boundaries to depict the tissue structures of all 12 sections in the comparison of other methods (Supplementary Figure 3). As a representative example of section 151673 (Fig. 2b), we observed that spatial domains unraveled by SpaSEG (ARI=0.554) had the best consistency with the manual annotations in spite of failure to detect the thinnest layer 4 separately (Fig.2c). Interestingly, this failure also took place in all other methods, plausibly because of the small number of spots in the layer 4 that may have gene expressions similar to the adjacent layer. All methods struggled discerning layers 4 and 5. Although being able to obtain comparable clustering accuracies and promising layer structures, SpaGCN (ARI=0.457) and BayesSpace (ARI=0.546) appeared to improperly separate the white matter into two domains with ragged boundaries, while SEDR (ARI=0.522) incorrectly merged the layers 4, 5 and 6 into a single layer. The spatial domains detected by Leiden, stLearn and Giotto massively mixed many unexpected outliers, leading to the rough tissue structures and the poorest clustering ARI values of 0.335, 0.306 and 0.291, respectively.

SpaSEG displays high robustness on diverse ST platform datasets and high scalability on large tissue section with high resolution

Next, we sought to test whether SpaSEG was robust to identify spatial domains on the datasets generated by different ST platforms such as Slide-seqV2, Stereo-seq, MERFISH, and seqFISH. Considering the above results regarding clustering accuracy and concordance of identified spatial domains with manual annotations, we only compared SpaSEG with SpaGCN and BayesSpace, as well as Leiden serving as a baseline method in the following experiments.

We first applied SpaSEG on the mouse hemibrain Stereo-seq data [8]. This image-based

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cell segmentation dataset has 50140 segmented cells (i.e., spots) and 25879 genes. As expected, SpaSEG can well uncover spatial regions of mouse hemibrain in comparison with Leiden and SpaGCN (Fig. 3a). BayesSpace did not successfully perform spatial clustering on this dataset because of the excessive large number of spots. Moreover, the LISI values displayed that SpaSEG reached significant more accuracy than Leiden ($p<2.2e-16$, Mann-Whitney U test) except for SpaGCN (Fig.3d). However, SpaSEG can clearer outline many cell-type spatial localizations than SpaGCN (Fig.3a), including neuroprotective astrocyte 2 types in thalamus, granule cells of dentate gyrus (GN DG), excitatory glutamatergic neurons from CA1 (EX CA1), and excitatory glutamatergic neurons from CA3 (EX CA3).

Then, we examined the scalability and efficiency of SpaSEG on large tissue section with high resolution. To achieve this, we analyzed an unreported whole adult mouse brain spatial transcriptomic data generated by Stereo-seq [8]. To facilitate our analysis at different resolution levels, we aggregated transcripts of the same gene into non-overlapping bin areas that covered corresponding DNB spots. These bins were of sizes in 10 μm diameter (bin20; 20×20 DNB sites; equivalent to ~1 medium mammal cell size), 25 μm diameter (bin50; 50×50 DNB sites), 50 μm diameter (bin100; 100×100 DNB sites), and 100 μm diameter (bin200; 200×200 DNB sites). As a result, we obtained four binned Stereo-seq ST datasets with the number of bins from 5420 to 526716 (Supplementary Table 7). The annotation of whole mouse brain from Allen Reference Atlas [26] is used as the reference and we choose resolution bin50 as a representative in our study. SpaSEG can well characterize the structures of the whole adult mouse brain such as cortex layers and hippocampus (including DG-sp, CA3sp or CA1sp subfields) at all four resolution levels (Fig.3b, Supplement Figure 4). In contrast, Leiden mixed the domains with other spots at bin20 and bin50 resolution levels, neither

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

uncovering clear cortex layers at bin100 resolution nor separating DG-sp and CA3sp at bin200. SpaGCN was not able to handle Stereo-seq data at bin20 resolution due to the substantial number of bins (526716) and running out of memory. Neither it yielded continuous and neat spatial domains nor uncover DG-sp and CA3sp separately at resolution levels of bin50, bin100, and bin200. BayesSpace cannot successfully perform spatial domain identification for Stereo-seq data at high resolution levels of bin20, bin50, and bin100 because of large number of bins. The LISI values of resolution bin 50 demonstrated that SpaSEG identifying spatial domain was significant more accurate than Leiden, SpaGCN ($p<2.2e-16$, Mann-Whitney U test; Fig.3d). Moreover, SpaSEG took 12.1 minutes to perform spatial domain identification for the bin20 data (2.5 times faster than Leiden) while 4.2 minutes with only 3.5G memory for the bin50 data, 26 times extraordinary faster and 35 times fewer memory usage than SpaGCN that suffered from considerable computational burden and took 110 minutes and 122.4G memory (Supplementary Table 7).

We next used the mouse hippocampus Slide-seqV2 data with 53208 spots and 23264 genes at 10 μm diameter resolution that can capture gene expressions at cellular level [6] (Supplementary Table 8). The annotation of hippocampus structures from the Allen Reference Atlas was employed as reference [26] (Fig. 3c). As expected, SpaSEG can better outline the topology of the tissue based on the identified neat spatial domains and sharp boundaries than that of Leiden, SpaGCN and BayesSpace (Fig.3c). For example, in addition to different cortical layers, SpaSEG was also able to clearly delineate the pyramidal layer of Ammon's horn and the granule cell layer of the dentate gyrus. More specifically, SpaSEG successfully depicted subfields of Ammon's horn such as CA1 (CA1so, CA1sp, and CA1sr) and CA3 (CA3so, CA3sp, and CA3sr), as well as subfields of dentate gyrus including DG-mo, DG-sg, and DG-po. Spatial domains for CA2 were

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not uncovered separately partly due to the few spots with gene expression similar to nearby domains. Owing to the lack of manual annotation on this dataset, we calculated the local inverse Simpson's index (LISI) to measure the clustering performance. As a result, SpaSEG reached a significantly lower LISI value than other three methods ($p<2.2e-16$, Mann-Whitney U test; Fig 2d), suggesting its highest accuracy for spatial domains detection with high resolution data.

We then utilized the mouse hypothalamic preoptic region data generated by MERFISH [4]. This annotated dataset contains 4975 single cells (i.e., spots) and 160 genes. SpaSEG achieved the ARI value of 0.46, which was higher than all other methods of Leiden (0.38), SpaGCN (0.26) and BayesSpace (0.33) (Fig. 2e. Moreover, SpaSEG also can delineate the spatial distribution of cell classes with spatial dependency such as ependymal, inhibitory, excitatory, mature OD, and mural, which were agreement to the annotations. We further employed the mouse organogenesis seqFISH data [5]. This dataset consists of 19416 single cells and 351 genes with a total of 22 cell types annotated. Compared to Leiden, SpaGCN and BayesSpace, SpaSEG yielded the highest ARI value of 0.46 (Fig. 2f). SpaSEG can better depict the spatial distribution of cell classes than other method, including three germ layers of ectoderm, mesoderm and endoderm, which were in consistent with the original study and known anatomy [27]. These results demonstrated that SpaSEG had the high accuracy for spatially clustering imaging-based *in-situ* transcriptomic data.

SpaSEG can successfully detect spatially variable genes (SVGs)

Next, we applied SpaSEG to detect SVGs for the validation of the identified spatial domain. Similar to previous study [16], we first examined the detected SVGs for each domain in the DLPFC section 151673 originally with 3639 spots and 33538 genes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SpaSEG finally detected 143 SVGs with low false discovery rate (FDR)-adjusted P values (<0.05), of which 126 genes were specifically expressed highly in domain 0, while the rest 17 genes were in the remaining domains (Supplementary Table 10). The Gene Ontology (GO) term enrichment analysis indicated the most SVGs enriched in domain 0 were significant related to white matter such as central nervous system myelination, neural myelin sheath, and structural constituent of myelin sheath (Supplementary Figure 5d). SpaSEG detected single representative genes for each of neuronal layers and white matter. For example, *PLP1, CNP, GFAP, CRYAB, TF, MOBP* gene was enriched in domain 0 (white matter), *CAMK2N1, ENC1, HPCAL1, HOPX* in domain 2(layer2, 3), *NEFL, NEFM, SNCG* in domain 3(layer 3), *PCP4, TMSB10, TUBB2A* in domain 4(layer 4, 5) and *MALAT1* was in domain 6(layer 1) (Fig.4a, b, Supplementary Figure 5a). By contrast, SpaGCN detected only 67 SVGs while SpatialDE and SPARK can totally detect 3661 and 3187 SVGs, respectively (Supplementary Figure 5c). However, SVGs detected by SpatialDE and SPARK did not necessarily show domain specificity. The Moran's I values and Geary's C values for SVGs detected by SpaSEG were significantly lower than that detected by SpatialDE ($p<2.2e-16$, Mann-Whitney U test) and SPARK ($p<2.2e-16$, Mann-Whitney U test) but showed no significant difference against that detected by SpaGCN ($p=0.07,0.01$ Supplementary Figure 5b). These results suggested that SpaSEG can detected more domain-specific SVGs than SpaGCN, SpatialDE and SPARK while maintained the similar accuracy against SpaGCN in spite of being slightly inferior to SpaGCN in terms of the Moran's I value and Geary's C value. These results demonstrated that SpaSEG outperforms SpaGCN in identifying spatial patterns for genes.

Then, we applied SpaSEG to detect SVGs on the unannotated mouse embryo Stereo-seq data with 72944 spots (bin50, 25 μm diameter per spot) and 28879 genes. Based

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on the 30 identified spatial domains (Fig.4c), SpaSEG detected a total of 490 SVGs that was more than SpaGCN (n=458) (mean of Moran's I for SpaSEG 0.361 and mean of Geary's C = 0.616, Fig.4d, e). These results demonstrated that SpaSEG outperforms SpaGCN in identifying spatial patterns for genes. Of particular interest in the following analyses were domain 1 (brain), domain 4 (epidermis), and domain 7 (cartilage primordium/bone), which were associated with 178 SVGs, 18 SVGs, and 18 SVGs respectively (Supplementary Table 11-13). These SVGs showed transcriptionally distinct patterns that distinguished the three spatial domains (Fig.4h). We further select top 5 genes that were highly expressed for each domain. For example, top 5 SVGs highly expressed in domain 1 contains brain development associated genes of *Nnat*, *Tuba1a*, *Mapt*, and brain marker genes *Stmn2*, *Tubb2a*, and top 5 SVGs in domain 4 of *Krt10*, *Krt15*, *Krt77*, *Lor*, *Krtdap,* and top 5 SVGs in domain 7 of *lbsp*, *Col1a1*, Col1a2, *Sparc*, *Serpinh1* (Fig.4f). We also depicted spatial expression for each of top 2 SVGs that demonstrate strong spatial patterns in corresponding spatial domain (Fig.4g). GO enrichment analysis of the SVGs showed that a total 457 GO terms and 40 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were enriched in cluster 1 (brain), of which many of these GO terms and KEGG paths are associated with the development of the nervous system (Fig.4i), for example, growth cone (GO 0030426), site of polarized growth (GO 0030427), distal axon (GO 0150034). In cluster 4 (epidermis), 18 SVGs are selected (q value < 0.05 and p value < 0.05) for the enrichment analysis, a total 72 GO terms and 11 KEGG pathways were enriched and most of these GO terms and KEGG pathways are related to keratin and epidermal cells, (Fig.4i) such as intermediate filament (GO 0005882), intermediate filament cytoskeleton (GO 0045111), keratin filament (GO 0045095). While in cluster 7 (cartilage primordium/bone), the most significant GO terms and KEGG paths are

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

related to collagen and cartilage or bone development (Fig.4i), collagen-containing extracellular matrix (GO 0062023), fibrillar collagen trimer (GO 0005583), banded collagen fibril (GO 0098643). In addition, SpaSEG is also capable to identified fine structure of mouse embryos and we further analysis the spatial variable genes in toes. Several representative marker genes are identified (Fig.4j), such as *Krt10* (skin tissue), *Dcn* (connective tissue), *Col2a1* (cartilage primordium). These results demonstrate that SpaSEG could effectively and accurately identify spatial variable based on our spatial segmentation results.

SpaSEG facilitates the investigation of ligand-receptor interactions

Most of cell-cell interactions and crosstalks are mediated by ligand-receptor (LR) interactions [ref]. To facilitate the exploration of putative intercellular interaction across the entire tissue section, we proposed a method to conduct LR interaction analysis by leveraging the spatial domains identified by SpaSEG and the co-expressions of the ligands and receptors. Here we applied the LR interaction analysis on the adult mouse brain Stereo-seq data at bin200 resolution with spatial domains identified by SpaSEG (Fig.5a). In order to further validate the accuracy of our clustering result, we first analyzed the SVGs in cortex (Fig.5d) and found that *Lamp5, Nrgn* are specifically enriched in cluster 5 (cortex layer 2/3), *Pvalb* gene in cluster 16 (cortex layer 4 or 5) and *Tbr1* in cluster 7 (cortex layer 5/6). Region specific SVGs are also found in hippocampus (Fig.5e), such as *Tmem54*, *Pantr1* in cluster 4 (CA1/2), *Hpca* and *Ddn* in cluster 17 (CA3) and *Wipf3* in cluster 11 (DG). A total of 267 significant ligand-receptor pairs are first identified by CellPhoneDB (Supplementary Table 14), of which a large number of LR interactions are enriched in cortex area as showed in the overall LR score

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

heatmap (Fig.5b). The cell types are annotated by cell2location according to the max confidential spot-level cell type deconvolution score and we observed that the deconvolution result could well depict the analogy of the adult mouse brain (Fig. 5c). Then we calculated the Spearman correlations between spot-wise cell type deconvolution scores with its corresponding LR scores, in which a positive correlation suggests as the colocalization of LR pairs with specific cell types. We found a majority of cell types between cortex, hippocampus and amygdala displayed highly positive correlations, including cell types of Astrocytes and Excitatory neuros that suggested their well colocalization in spatial context of the tissue (Fig. 5f). For example, the ligand Il34 (Interleukin-34) and receptor Csf1r are highly active in the cortex, hippocampus and amygdala (cluster 5,7,8,9,10,12,16), which corresponding to the conclusion that Il34 identified as a tissue-specific ligand of Csf-1 receptor (Csf1r) is mainly expressed in brain cerebral cortex (Fig. 5f, g). Our result also shows that Cholecystokinin (Cck) and its receptor Cckbr are enriched in cortex, hippocampus, amygdala and piriform cortex [28]. Meanwhile, we find that Bdnf_Sort1 and Bdnf_Ntrk2 are enriched in cortex and hippocampus [29], which may be related to increased or decreased volume of the hippocampus.

SpaSEG enables to elucidate the ligand-receptor interactions in Invasive ductal carcinoma

To further validate our LR interaction results identified by SpaSEG, we analyzed a breast cancer sample originally published in BayesSpace [15] with tumor regions being annotated by the pathologist (Fig. 6a). The sample was an estrogen receptor-positive (ER^+), progesterone receptor-negative (PR^-), human epidermal growth factor receptor (HER)2-amplified ($HER2^+$) invasive ductal carcinoma (IDC). The dataset was

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

generated by 10x Genomics Visium, leading to a total of 4727 spots in tissue and 36601 genes with a median of 2964 genes per spot. To allow cell2location for cell type mapping at spot level, we downloaded a published breast cancer scRNA-seq dataset as the reference that comprised 16 primary tumors [30] (11 ER^+ and 5 $HER2^+$) with cell types being annotated.

Spatial clusters obtained by applying SpaSEG were able to accurately distinguish regions among invasive carcinoma (cluster 0, 2,3, 8 and 9), carcinoma in situ (cluster 6), and benign hyperplasia (cluster 2) as well as non-tumor tissue (cluster 1,5 and 7), which were strong accordance with histopathological annotations (Fig. 6b). Cell types mapping using cell2location showed that, compared to other clusters, predominant proportions of cancer cells resided at invasive tumor regions (cluster 0, 2,3, 8 and 9) while non-tumor regions (cluster 1,5 and 7) were enriched for more immune-related cells than other regions such as B cells, cancers associated fibroblasts (CAFs), T cells, and plasmablasts (Fig. 6d and 6f). These findings indicated that SpaSEG can well capture the inter- and intra-tumor heterogeneities at molecular level. Spatially co-expressed LR pairs and its corresponding potential interaction cell types are showed in Figure 6e. By conducting cell-cell interaction analysis, we observed many interactions around the immune-invasive areas, especially between cluster 5 and cluster 0,4,3,9 (Fig.6c), and found colocalizations of cell types such as B cells and T cells, macrophages and T cells, CAFs and T cells (Fig. 6e) For example, ligand PTPRC secreted by T cells is an essential regulator in mediating T- and B-cell antigen processing by targeting the CD22 receptor in B cells [31, 32], playing a major role in adaptive immune response. T cells communicated with dendritic cells (DCs) through ligand PTPRC and receptor MRC1 [32, 33] (Fig. 6g, h). The mannose receptor (MRC1) expressed on DCs acts as a direct regulator of CD8+ T-cell activity by interacting with

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CD45[34, 35], which will result in the up-regulation of cytotoxic T-lymphocyte–associated Protein 4 (CTLA-4) and the induction of T-cell tolerance. The cytokine macrophage migration inhibitory factor (MIF) which constitutively found in macrophage sustains pro-inflammatory function and cell proliferation. And its receptor CD74 is also found in T-cells as previous literatures indicates and MIF_CD74 (Fig. 6g, h) shows significant high Spearman correlation with T cells in our study. In addition, we also detected that Galectin-9 (LGALS9 secreted by macrophages, monocytes) served as a ligand for immune checkpoints HAVCR2[30, 36, 37] (Fig. 6 e, g, h) (highly correlated with NKT cells, CD4+, CD8 T cells) and contributes to anti-cancer immune suppression by killing cytotoxic T lymphocytes and impairing the activity of natural killer (NK) cells[38], which is a promising target for immunotherapy. Apart from the active immune cell-cell interaction in TME, crosstalk between stromal cells and immune cells is also of great importance for angiogenesis, tumor invasion and metastasis. We detected that CAF ligand CXCL12 and its cognate T cell receptor (CXCR4/CXCR3) (Fig. 6g, h) are among the top ranked cell types for CXCL12_ CXCR4/CXCR3 pairs [39, 40], of which the CAF mainly promotes tumor growth by the secretion of SDF-1. Besides, CAF associated LR pairs like TIMP1_FGFR2, C3_C3AR1 [41, 42] (Fig. 6g, h) could also be observed. And the endothelial cells derived gene VEGFB, PDGFB, ACKR1 that could induce new blood vessel formation and stimulate cell proliferation and migration via interaction with NRP1 [43], LRP1 and the chemokine ligand CCL5.

Therefore, it is significant to understand the cell-cell interactions between macrophages and other immune cells and the factors that enhance existing anticancer treatments. These results proved that SpaSEG could be served as a handful tool for LR analysis in pathology, suggesting potential patterns of most likely cell-cell interactions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Identification of spatial domain is a significant process in spatial transcriptomic data analyses. SpaSEG harmoniously integrated gene expression information and spatial coordinates into one three-dimensional matrix as model input. Through feeding the input data to the convolutional neural network, SpaSEG learns the gene expression similarity and spatial contiguity simultaneously with the optimization of the gene expression similarity loss and the spatial continuity loss and ultimately identifies separate spatial domains. In our study, we firstly demonstrate its strong ability in differentiating distinct layers and superior performance relative to existing alternatives in DLPFC benchmarking dataset. To further validate the utility of SpaSEG in ST data of diverse resolutions, we artificially simulated four datasets representing different levels of resolution by utilizing the adult mouse brain data generated by Stereo-seq, SpaSEG could uniformly profile significant functional regions, such as Cerebral Cortex layers (CTX), Thalamus (TH), Hypothalamus (HY) and HPF (hippocampal formation areas) [44]. Although SEDR also performed clustering analysis on mouse olfactory bulb data from Stereo-seq and exhibited its efficacy to handle high resolution data, it suffers from high computational burden in constructing graphs for such high throughput and high-resolution spatial omics data [14].

In our work, we have also demonstrated the utility of SpaSEG in detecting spatial variable genes based on our clustering result. For fair and comprehensive comparison, we analyze two separate SVG detection methods: SpaGCN [16]and SpatialDE [45]. The former one takes spatial information into consideration when identifying SVGs, while the other one detects SVGs without the guidance of spatial domains. SpaSEG

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

identified 490 SVGs compared to SpaGCN (458 SVGs) in total in sample 151673. More specifically, there are about half of the SVGs enriched in white matter that are only detected by our methods while SpaGCN not, although both methods have many overlapping SVGs. Besides, there are more than 4000 SVGs identified by SpatialDE. However, many of them failed to exhibit spatial patterns. Thus, SpaSEG could serve as a potential tool for researchers to discover novel marker genes. In addition, our research has also shown the ability of SpaSEG in dissecting the spatial cell-cell interaction in adult mouse brain and IDC. The clustering results of SpaSEG in IDC and MB are concordant with the manual annotation from pathologist and the reference panel from Allen Mouse Brain Atlas respectively. More concretely, the heatmap of L-R pairs indicates that the most active regions in mouse brain are cerebral cortexes and hippocampus, which corresponds to previous studies [46]. Moreover, SpaSEG could spatially visualize the distribution of immune cell in IDC through deconvolution and further research could be undertaken to study the tumor microenvironment between the tumor cells and macrophage, T cells and B cells, etc in molecular level through L-R analysis.

Histology images are a new modality along with the ST data and the tissue morphological features have a strong connection with the corresponding gene expression around the specific spot [47, 48]. Recent studies such as SpaGCN and stLearn have made efforts to incorporate histological images into their algorithms but they are unable to achieve the finest clustering results, as demonstrated in our study. One possible reason could be that there are no obvious morphological features among the adjacent layers in DLPFC. Besides, the artifacts such as batch effects in the process

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of Hematoxylin and Eosin (HE) staining and noise(quality) control of the H&E images have a great influence on the final results. Our future research direction may lie in the development of robust algorithms that seamlessly incorporate both modalities' information together to achieve optimal performance. Another utility of H&E images is resolution enhancement in ST data. While our work mainly focused on ST data from 10X genomics and Stereo-seq, the former one usually has lower resolution and sensitivity, which limits their usefulness in studying detailed expression patterns and uncovering comprehensive tissue anatomical structure. Recent works have demonstrated the applicability of histological images in inferring accurate full-transcriptome spatial gene expression at the same resolution as the image data (XFuse [49]), which could be further used as a data enhancement approach for pre-processing step of the low-resolution ST data. Further research could be undertaken to explore how the imputed ST data would be used to train our algorithm and improve the clustering accuracy.

In conclusion, this work presents SpaSEG, an efficient and scalable unsupervised deep learning algorithm for ST data clustering. The results of our study also indicate the applicability of our algorithm in various downstream analyses, such as SVGs identification, cell-cell interaction and trajectory inference. Therefore, we believe that SpaSEG could serve as a valuable tool to benefit ST data analysis in the future.

Methods

Data preprocessing:

SpaSEG takes transcriptome-wide gene expression profile with spatial coordinates as inputs. Genes expressed in less than five spots or bins are excluded for all datasets. We

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

also eliminate poor spots with fewer than 200 expressed genes. The reserved raw gene counts per spot are further normalized using the library size, and scaled by log-transformation. Principal component analysis (PCA) is then performed on the gene expression data in an $N \times M$ matrix with N spots and M genes, and top d PCs per spot were subsequently extracted. The optimal value $d \in \{15, 50, 100\}$, varies from 15 to 100, depending on different sequencing platforms (Supplementary Table 9). These PCs are able to explain the sufficient variability in the data and mitigate the computational intension, as well as yield the best spatial clustering performance (Supplementary Figure 1). We further perform z-score normalization such that each PC has zero mean and unit variance.

Conversion of spatial gene expression data into image-like tensor

To enable SRT data analysis through SpaSEG, we convert the gene expression data with spatial information into an image-like tensor. In this tensor, a spot n in the SRT array at row i and column j was represented as a feature vector of $\mathbf{s}_{i,j}^{n} \in \mathbb{R}^{d}$, where d is the number of extracted PCs. As a result, an d-channel image-like tensor $\mathbf{X} = \{\mathbf{s}_{i,j}^{n}\}_{n=1}^{N} \in \mathbb{R}^{H \times W \times d}$ is created, where N is the number of spots, H and W are the height and width of the SRT array, $i \in \{1, 2, \ldots, H\}$, $j \in \{1,2,, \ldots, W\}$. For simplification, we denote the image-like tensor by $\mathbf{X} = \{\mathbf{s}_{n}\}_{n=1}^{N}$ unless otherwise specification.

SpaSEG model development

SpaSEG architecture. Relying on convolutional neural network (CNN) architecture, our model starts with a batch normalization layer, followed by two convolutional blocks and a refinement module consecutively (Fig. 1b). Each convolutional block is

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

composed of a 3×3 (kernel size) convolutional layer, a batch normalization layer, and a leaky ReLU activation layer (intermediate parameter $\alpha = 0.2$). The two convolution modules have u and v output channels, respectively. Finally, the refinement module consists of only a 1×1 convolutional layer and a batch normalization layer, yielding k output channels. It should be noted that the output size of each convolutional layer in SpaSEG is maintained the same as input. For simplification, we set u, v and k be equal to d in all experiments.

Formally, given the SRT data of a tissue slice that was represented by an image-like tensor $\mathbf{X} = \{\mathbf{s}_n\}_{n=1}^{N}$, the feature representation of spot n can be learned by

$$\mathbf{y}_n = f_{\boldsymbol{\Theta}}(\mathbf{s}_n) \quad (1)$$

where $f_{\boldsymbol{\Theta}}(\cdot)$ is the SpaSEG network with the trainable parameter $\boldsymbol{\Theta}$ that can be updated during an iterative training process, $\mathbf{y}_n = [y_n^1, y_n^2, \dots, y_n^k] \in \mathbb{R}^k$. Then, the pseudo-label for spot n can be given by

$$c_n = \underset{t}{\operatorname{argmax}}\, y_n^t, \ t = 1,2,\dots,k \quad (2)$$

Loss Function. We treat the spatial domain identification as spot-wise classification problem, where the class label of each spot can be viewed as a segment. To train SpaSEG, we first consider the most commonly used cross entropy loss with L2-norm regularization over pseudo-label as follows.

$$\mathcal{L}_{\text{seg}} = -\sum_{n=1}^{N}\sum_{t=1}^{k} \mathbb{I}(c_n = t)\log(p_n^t) + \lambda\|\boldsymbol{\Theta}\|_2^2 \quad (3)$$

where $\mathbb{I}(x) = \begin{cases} 1, \text{if } x \text{ is true} \\ 0, \text{otherwise} \end{cases}$, $p_n^t = e^{y_n^t} / \sum_{t=1}^{k} e^{y_n^t}$, and λ is regularization parameter that controls the L2-norm regularization penalty and we set it to 0.00001

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in our experiments.

To encourage the class label to be the same as those of spatially adjacent spots, we follow the previous work [23] to introduce a spatial-smoothness loss function that considers the horizontal and vertical differences of feature representations, which is defined as

$$\mathcal{L}_{\text{spa}} = \sum_{i=1}^{W-1} \sum_{j=1}^{H-1} \left\| \mathbf{y}_{i+1,j} - \mathbf{y}_{i,j} \right\|_1 + \left\| \mathbf{y}_{i,j+1} - \mathbf{y}_{i,j} \right\|_1$$

Consequently, the overall loss is then given by

$$\mathcal{L}_{\text{overall}} = \alpha \mathcal{L}_{\text{seg}} + \beta \mathcal{L}_{\text{spa}}$$

where α and β are weighting factors for segmentation and spatial smoothness, which are set to be 0.4 and 0.7, respectively.

SpaSEG training

Rather than randomly initializing parameters of SpaSEG that usually yields unstable results, we pre-train SpaSEG using MSE loss defined as $\mathcal{L}_{\text{pre}} = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{s}_n - \mathbf{y}_n\|_2^2$ during the first 400 training epochs to initialize the model parameters. this iterative process could stabilize the model performance and reinforce the entire algorithm to update the model in a desirable direction. In the subsequent epochs, feature representation $\mathbf{y}_n$ and the corresponding pseudo-label c_n for each spot was obtained. SpaSEG then calculates and backpropagates the overall loss $\mathcal{L}_{\text{overall}}$ to update the model parameters. This process is repeated until either the number of iterations exceeds the pre-defined maximum, or a minimum number of unique class labels is attained.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Otherwise, if the model failed to achieve the minimum number of class labels, an optional refinement process is proposed to enhance the final segmentation result from SpaSEG. In this process, the mean of the image-like tensor $\{\boldsymbol{s}_n^k\}_{n=1}^N$ for each cluster along the d-channel are calculated, where k is the cluster label assigned by SpaSEG. Then the pairwise Euclidian distances are calculated between the clusters, denoted as $D_{i,j}$ and i, j is the cluster labels. While the candidate spots that are spatially separatable among the clusters will be relabeled according to the smallest $D_{i,j}$. We employed Adam optimizer with the default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as optimization method for the backpropagation. The learning rate was set to 0.002 and the total number of epochs were set to 2100. Those optimal values for the hyper-parameters of SpaSEG were determined via a combination of grid search and manual tuning such that the best performance can be achieved.

Spatial variable gene detection

In order to identify spatially variable genes (SVGs) that have high expression in each spatial cluster, we combined the cluster results with pre-processing datasets. For each cluster, Scanpy implementation of the Wilcoxon rank-sum test was used to identify SVGs (adjusted p value < 0.05). To further confirm that the SVGs have abundant expression, we added three conditions, this refers to SpaGCN's method of identifying SVG:

1) In the target cluster, count the ratio of gene expression spots to total spot;

2) For outside the target cluster, the percentage of spots expressing genes within the target cluster and outside the clusters;

3) Expression fold change in target cluster and outside clusters.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

And then we drew the spatial pattern map of each SVG expression on the tissue, which are identified that SVGs are truly enriched in target cluster.

And further discovering the functions with SVGs, we did GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis of SVGs by clusterProfiler package in R software.

Cell type deconvolution

Each spatial spot was annotated to a specific cell type deconvoluted from spatial data and corresponding single cell reference datasets. The deconvolution of cell types was implemented using cell2location [50], which is a Bayesian model for spatial mapping of cell types. Given the complementary information from spatial resolved transcriptomic data and single-cell RNA sequence, we applied this Bayesian model to infer different cell types in different spatial locations. The training hyperparameters defined manually are selected depending on the cell number of a spot and RNA detection sensitivity. The cell type corresponding to the maximum score of each spot is regarded as the cell type of the spot in this research.

Ligand-receptor interaction

The cellular interactions mediated by protein-protein interactions are significant for understanding tissue structures and functions. Firstly, we randomly permuted the SpaSEG cluster labels of all spots to create a null distribution for each L-R pair in each

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

pairwise comparison between two clusters by using cellphoneDB. Then the significant interacting pairs were used to analysis ligand-receptor interactions between the clusters.

With the prior of significant L-R pairs, we calculated the expression of spatially co-expressed L-R pairs using:

$$\boldsymbol{e}_{i,j}^{L_k_R_k} = \sqrt{\mathbf{x}_i^{L_k} \cdot \mathbf{x}_j^{R_k}}$$

$$E_{L_k_R_k} = \sum_{i=1}^{n} \sqrt{\mathbf{x}_n^{L_k} \cdot \mathbf{x}_n^{R_k}}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ represents the gene expression vector of spot i and spot j, respectively. $E_{i,j}$ represents the value of L-R pair in all spots. If one of the ligands and receptors in a spot is zero (not expressed), the co-expression value of the spot is zero. By calculating the geometric mean of gene expression values, we know that the geometric mean of each receptor ligand is large when the expression of each receptor ligand is similar, and the geometric mean is small when the expression difference is large. We calculated Spearman correlations between cell-type score and L-R score, which represents the corresponding L-R pairs expression in related cell type.

Data description

We applied SpaSEG to ST datasets with different resolutions generated by various platforms, such as 10X genomics, Stereo-Seq, MERFISH, SeqFISH, Slide-SeqV2. The DLPFC dataset generated by Visium platform contains 12 slice and each consists of around 4000 spots [12]. These 12 sections are all manually annotated and are used to benchmark our algorithm. For the Stereo-seq data, we artificially divided the expression

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

matrix into non-overlapping bins covering an area of X×X DNB, with $X \in (20,50,100,200)$ and the transcripts of the same gene are aggregated within each bin. Specifically, the raw matrix of adult mouse coronal brain sample [8] contains 27,279 genes, which was divided into bin20(526,716 spots with 10 μm diameter), bin50(84,724 spots with 25 μm diameter), bin100(21,368 spots with 50 μm diameter) and bin200(5,420 spots with 100 μm diameter). The MERFISH sample (animal id = 1, Bregma = -0.24) was collected from a female mouse with no treatment performed in the hypothalamic preoptic region [51], this sample data contains 6412 spots and measured 161 genes expression values. The mouse hippocampus data from Slide-seqV2[6] (id: Puck_200115_08) consists of 53208 spots and measured 23264 genes. The sagittal sections of the seqFISH[27] sample was collected from 8-12 somite-stage embryos (Embryonic days (E)8.5-E8.75) and contains 19416 locations and 351 barcoded genes are measured. The mouse embryo dataset [8] with 76453 spots and 27009 genes collected from E15.5 embryos that was from pregnant C57BL/6J female mice, produced by Stereo-seq have been deposited to CNGB Nucleotide Sequence Archive. We aggregated the raw bin 1 matrix into the final bin 50 matrix and assigned coordinates for each bin. The IDC (invasive ductal carcinoma) [15] data was downloaded from 10X Genomics, which was stained for nuclei with DAPI and anti-CD3. There are totally 4,727 spots detected under tissue. The mean reads per spot is 40,795. The median genes per spots is 2,964. The IDC sample approximately include a median of 21 cells per spot.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data availability

The DLPFC dataset is publicly available at http://research.libd.org/spatialLIBD/. The MERFISH sample data could be downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248. The Slide-seqV2 data could be accessible at Single Cell Portal (broadinstitute.org). The seqFISH data can be downloaded from https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/. The IDC data is publicly available at https://www.10xgenomics.com/resources/datasets/invasive-ductal-carcinoma-stained-with-fluorescent-cd-3-antibody-1-standard-1-2-0. The mouse embryo bin50 could be down form https://db.cngb.org/search/project/CNP0001543 .

Comparison with state-of-arts methods and evaluation.

To demonstrate the superior performance in spatial transcriptomics data clustering, we chose a commonly used non-spatial clustering method Leiden plus five recently published state-of-the-art methods, including stLearn, Giotto, SpaGCN, BayesSpace, and SEDR (Supplementary Table 2). To evaluate the effectiveness of SpaSEG in integrating multiple tissue sections, two commonly used algorithms in scRNA-seq data batch correction, Harmony and LIGER are utilized to compare with SpaSEG.

Leiden. Leiden is a popular tool for single cell transcriptomics data clustering integrated in Scanpy. The data preprocessing step is the same as SpaSEG and we ran *sc.tl.leiden* in Scanpy and tune the resolution parameter to give us desirable number of clusters.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

stLearn. stLearn is the first algorithm simultaneously integrating H&E information and spatial transcriptomics data and allows various downstream analysis like cell-cell interaction and trajectory inference. We ran the data processing with filtering gene at least expressing in 1 cell and setting the number of principal components to 15, then we follow the clustering pipeline of stLearn(version: 0.3.2) with the guidance of official tutorial https://stlearn.readthedocs.io/en/latest/.

Giotto. Giotto is a toolbox for spatial data integrative analysis by utilizing hidden Markov random field (HMRF) model. We follow the online tutorial of Giotto(version: 1.0.4): https://github.com/RubD/Giotto_site and set the *expression_threshold* parameter to 1 in filterGiotto function and set *hvg* = 'yes', *perc_cells* > 3, *mean_expr_det*>0.4 in gene_metadata function when preparing data for dimensional reduction. The spatial neighborhood network is created with the default parameters and the number of ground truth clusters is employed for HMRF model clustering.

SEDR. SEDR uses a deep autoencoder to construct latent gene representation and a variational graph autoencoder to embed spatial information. We ran SEDR code indicated on the github repository: https://github.com/JinmiaoChenLab/SEDR with default parameter settings.

SpaGCN. SpaGCN utilizes a graph convolutional network that combines gene expression, spatial location and histology in spatial transcriptomics data analysis. We follow the official tutorial of SpaGCN: https://github.com/jianhuupenn/SpaGCN (version: 1.2.0) and set the *min_cell* to 3 when filtering genes, *alpha* to 1, *beta* to 49

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

when calculating adjacent matrix with histology image available. The learning rate and max training epoch were set to 0.05 and 200, respectively.

BayesSpace. BayesSpace performs spatial clustering by introducing a spatial prior and encouraging neighboring spots to belong to the same cluster. We ran BayesSpace (version:1.2.1) followed by https://github.com/edward130603/BayesSpace. Top 200 highly variable genes are selected to perform PCA and we set q to the number of PCA and d to the desirable cluster number, nrep to 50000 and gamma to 3, respectively.

Reference

1. Sonawane, A.R., et al., *Understanding Tissue-Specific Gene Regulation.* Cell Rep, 2017. **21**(4): p. 1077-1088.
2. Lee, J.H., *Quantitative approaches for investigating the spatial context of gene expression.* Wiley Interdiscip Rev Syst Biol Med, 2017. **9**(2).
3. Armingol, E., et al., *Deciphering cell-cell interactions and communication from gene expression.* Nat Rev Genet, 2021. **22**(2): p. 71-88.
4. Moffitt, J.R., et al., *High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization.* Proc Natl Acad Sci U S A, 2016. **113**(39): p. 11046-51.
5. Lubeck, E., et al., *Single-cell in situ RNA profiling by sequential hybridization.* Nat Methods, 2014. **11**(4): p. 360-1.
6. Stickels, R.R., et al., *Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2.* Nat Biotechnol, 2021. **39**(3): p. 313-319.
7. Stahl, P.L., et al., *Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.* Science, 2016. **353**(6294): p. 78-82.
8. Chen, A., et al., *Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays.* Cell, 2022. **185**(10): p. 1777-1792 e21.
9. Marx, V., *Method of the Year: spatially resolved transcriptomics.* Nat Methods, 2021. **18**(1): p. 9-14.
10. Rao, A., et al., *Exploring tissue architecture using spatial transcriptomics.* Nature, 2021. **596**(7871): p. 211-220.
11. Ji, A.L., et al., *Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma.* Cell, 2020. **182**(2): p. 497-514 e22.
12. Maynard, K.R., et al., *Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex.* Nat Neurosci, 2021. **24**(3): p. 425-436.
13. Traag, V.A., L. Waltman, and N.J. van Eck, *From Louvain to Leiden: guaranteeing well-connected communities.* Sci Rep, 2019. **9**(1): p. 5233.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14. Fu, H., et al., 2021.
15. Zhao, E., et al., *Spatial transcriptomics at subspot resolution with BayesSpace.* Nat Biotechnol, 2021. **39**(11): p. 1375-1384.
16. 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): p. 1342-1351.
17. Pham, D., et al., 2020.
18. Dries, R., et al., *Giotto: a toolbox for integrative analysis and visualization of spatial expression data.* Genome Biol, 2021. **22**(1): p. 78.
19. Edsgard, D., P. Johnsson, and R. Sandberg, *Identification of spatial expression trends in single-cell gene expression data.* Nat Methods, 2018. **15**(5): p. 339-342.
20. Svensson, V., S.A. Teichmann, and O. Stegle, *SpatialDE: identification of spatially variable genes.* Nature Methods, 2018. **15**(5): p. 343-346.
21. Sun, S., J. Zhu, and X. Zhou, *Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies.* Nature Methods, 2020. **17**(2): p. 193-200.
22. Walker, B.L., et al., *Deciphering tissue structure and function using spatial transcriptomics.* Communications Biology, 2022. **5**(1).
23. Kim, W., A. Kanezaki, and M. Tanaka, *Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering.* IEEE Transactions on Image Processing, 2020. **29**: p. 8055-8068.
24. Hubert, L. and P. Arabie, *Comparing partitions.* Journal of Classification, 1985. **2**(1): p. 193-218.
25. Fowlkes, E.B. and C.L. Mallows, *A Method for Comparing Two Hierarchical Clusterings.* Journal of the American Statistical Association, 1983. **78**(383): p. 553-569.
26. Sunkin, S.M., et al., *Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system.* Nucleic Acids Res, 2013. **41**(Database issue): p. D996-D1008.
27. Lohoff, T., et al., *Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis.* Nat Biotechnol, 2022. **40**(1): p. 74-85.
28. McCullough, K.M., et al., *Cell-type-specific interrogation of CeA Drd2 neurons to identify targets for pharmacological modulation of fear extinction.* Transl Psychiatry, 2018. **8**(1): p. 164.
29. Li, Z., et al., *The role of BDNF, NTRK2 gene and their interaction in development of treatment-resistant depression: data from multicenter, prospective, longitudinal clinic practice.* J Psychiatr Res, 2013. **47**(1): p. 8-14.
30. Wu, S.Z., et al., *A single-cell and spatially resolved atlas of human breast cancers.* Nat Genet, 2021. **53**(9): p. 1334-1347.
31. Akatsu, C., et al., *The inhibitory coreceptor CD22 restores B cell signaling by developmentally regulating Cd45(-/-) immunodeficient B cells.* Sci Signal, 2022. **15**(723): p. eabf9570.
32. Coughlin, S., et al., *An extracatalytic function of CD45 in B cells is mediated by*

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CD22. Proc Natl Acad Sci U S A, 2015. **112**(47): p. E6515-24.

33. Akatsu, C., et al., *The inhibitory coreceptor CD22 restores B cell signaling by developmentally regulating Cd45 −/− immunodeficient B cells.* Science Signaling, 2022. **15**(723).
34. Broncy, L. and P. Paterlini-Bréchot, *Cancer-associated circulating atypical cells with both epithelial and macrophage-specific markers.* Journal of Laboratory and Precision Medicine, 2018. **3**: p. 91-91.
35. Schuette, V., et al., *Mannose receptor induces T-cell tolerance via inhibition of CD45 and up-regulation of CTLA-4.* Proc Natl Acad Sci U S A, 2016. **113**(38): p. 10649-54.
36. Holderried, T.A.W., et al., *Molecular and immune correlates of TIM-3 (HAVCR2) and galectin 9 (LGALS9) mRNA expression and DNA methylation in melanoma.* Clin Epigenetics, 2019. **11**(1): p. 161.
37. Yang, R., et al., *Galectin-9 interacts with PD-1 and TIM-3 to regulate T cell death and is a target for cancer immunotherapy.* Nat Commun, 2021. **12**(1): p. 832.
38. Yasinska, I.M., et al., *The Tim-3-Galectin-9 Pathway and Its Regulatory Mechanisms in Human Breast Cancer.* Front Immunol, 2019. **10**: p. 1594.
39. Orimo, A., et al., *Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion.* Cell, 2005. **121**(3): p. 335-48.
40. Zhu, G., et al., *CXCR3 as a molecular target in breast cancer metastasis: inhibition of tumor cell migration and promotion of host anti-tumor immunity.* Oncotarget, 2015. **6**(41): p. 43408-19.
41. Jackson, H.W., et al., *TIMPs: versatile extracellular regulators in cancer.* Nat Rev Cancer, 2017. **17**(1): p. 38-53.
42. Shu, C., et al., *C3a-C3aR signaling promotes breast cancer lung metastasis via modulating carcinoma associated fibroblasts.* J Exp Clin Cancer Res, 2020. **39**(1): p. 11.
43. Ferrara, N., H.P. Gerber, and J. LeCouter, *The biology of VEGF and its receptors.* Nat Med, 2003. **9**(6): p. 669-76.
44. Lein, E.S., et al., *Genome-wide atlas of gene expression in the adult mouse brain.* Nature, 2007. **445**(7124): p. 168-76.
45. Svensson, V., S.A. Teichmann, and O. Stegle, *SpatialDE: identification of spatially variable genes.* Nat Methods, 2018. **15**(5): p. 343-346.
46. Ero, C., et al., *A Cell Atlas for the Mouse Brain.* Front Neuroinform, 2018. **12**: p. 84.
47. He, B., et al., *Integrating spatial gene expression and breast tumour morphology via deep learning.* Nat Biomed Eng, 2020. **4**(8): p. 827-834.
48. Tan, X., et al., *SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells.* Bioinformatics, 2020. **36**(7): p. 2293-2294.
49. Bergenstrahle, L., et al., *Super-resolved spatial transcriptomics by deep data fusion.* Nat Biotechnol, 2022. **40**(4): p. 476-479.
50. Kleshchevnikov, V., et al., *Cell2location maps fine-grained cell types in spatial*

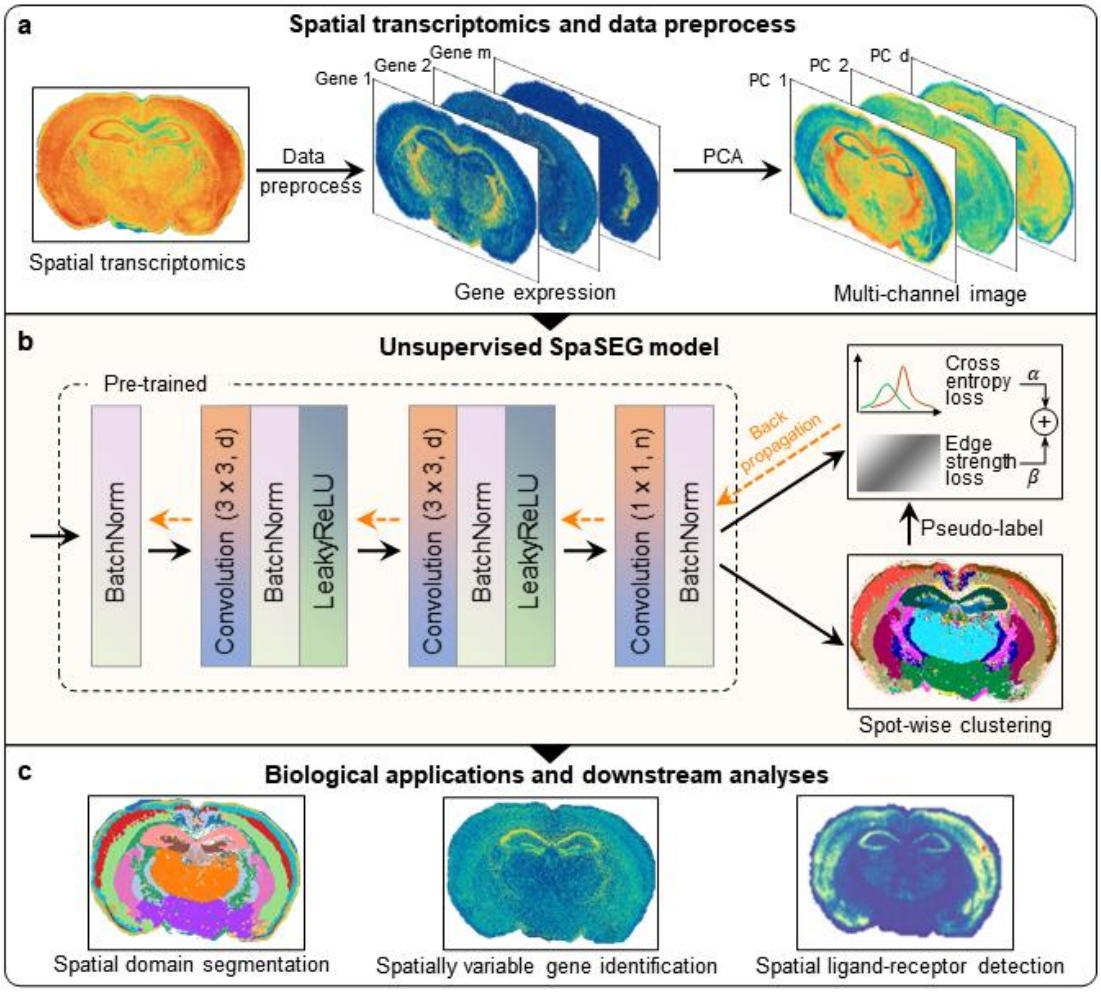
bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

transcriptomics. Nat Biotechnol, 2022. **40**(5): p. 661-671.

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

Figure legend

Figure 1. Overview of SpaSEG. a. Spatial transcriptomics data preprocessing and data preparation step for SpaSEG. **b**. SpaSEG takes the image-like low dimensional feature vector as input, then spot-wise labels are assigned through iterative unsupervised CNN model training with gene similarity loss and spatial continuity loss. **c.** Biological application and downstream analysis for SpaSEG, including spatial domain identification, spatial variable detection, spatial ligand receptor detection.



bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2. Algorithms comparison in 10X DLPFC. **a**, The average ARI, NMI score among 12 DLPFC slices between SpaSEG and the other spatial methods. **b**, Annotated ground truth for sample 151673. **c,** Spatial domain results for SpaSEG and other spatial clustering methods in sample 151673.

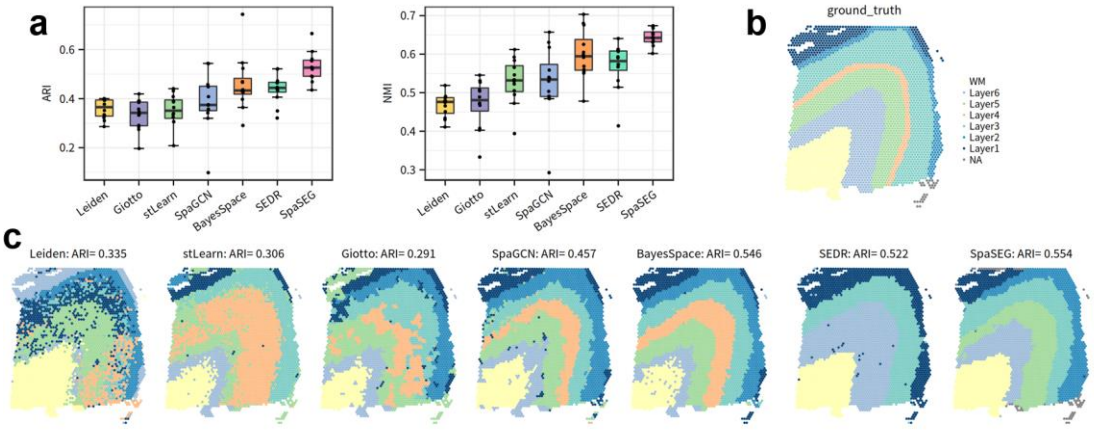
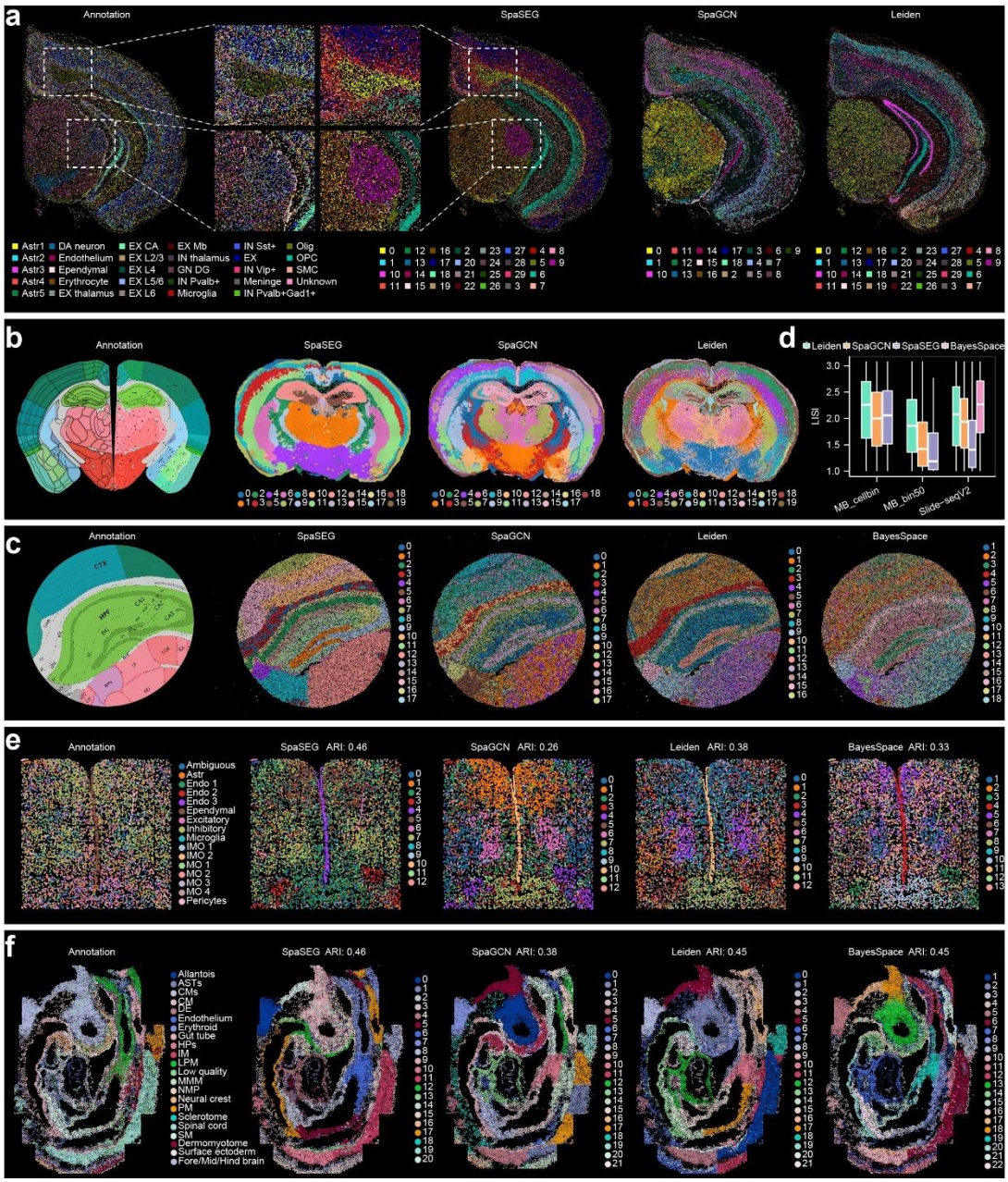


Figure 3. Robustness and scalability of SpaSEG in different spatial platforms. **a**, from left to right, the segmented Stereo-seq mouse brain cellbin annotation in the original paper. The cell composition between the annotation and SpaSEG clusters in two specific zoomed-in areas. SpaSEG cell segmentation results. SpaGCN clustering results. Leiden clustering results. **b**, from left to right. Annotated mouse brain coronal sections from Allen Brain Atlas. The clustering results of SpaSEG, SpaGCN and leiden respectively in Stereo-seq MB bin50 data. **c**, The LISI score calculated from SpaSEG, SpaGCN and Leiden spatial clustering labels in MB cellbin, MB bin50 and Slideseq2 datasets. **d**, from left to right. The annotated structure of mouse hippocampus from Allen Brain Atlas. Spatial domain results of SpaSEG, SpaGCN, Leiden and BayesSpace in SlideseqV2. **e**, from left panel to right panel. The annotated celltype spatial distribution of mouse hypothalamic preoptic region in the original paper. Spatial clustering results of SpaSEG, SpaGCN, Leiden and BayesSpace in MERFISH dataset. **f**. from left panel to right panel. The spatial map of cell composition during mouse organogenesis in seqFISH data. Spatial clustering results of SpaSEG, SpaGCN, Leiden and BayesSpace in the forementioned dataset. Astr: Astrocyte; Endo: Endothelial; IMO:

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

OD Immature; MO: OD Mature; ASTs: Anterior somitic tissues; CMs: Cardiomyocytes; CM: Cranial mesoderm; DE: Definitive endoderm; HPs: Haematoendothelial progenitors; IM: Intermediate mesoderm; LPM: Lateral plate mesoderm; MMM: Mixed mesenchymal mesoderm; PM: Presomitic mesoderm; SM: Splanchnic mesoderm.



bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4. Spatial clusters and SVGs detected in the DLPFC slices and Mouse embryo. **a**, Spatial expression pattern of the DLPFC's SVGs detected SpaSEG. **b**, Bubble map of spatially differential genes obtained by SpaSEG in DLPFC (151673). **c**, Spatial domain results for SpaSEG spatial clustering methods in mouse embryo. **d**, Moran's I and Geary's C values for SVGs detected by SpaSEG(n=490) and SVGs detected by SpaGCN (n=458). **e**, Venn diagram for SVGs detected by SpaSEG and SpaGCN in the mouse embryo data. f, Violin plot of spatially differential genes obtained by SpaSEG in mouse embryonic. **g,** spatial expression of mouse embryo's brain, epidermis, cp. **h**, Spatial expression heatmap of the SVGs in the mouse embryo's brain, epidermis and cp. **i**, GO pathway enrichment analysis of mouse embryo's brain, epidermis, cp. **j**, spatially clusters of mouse toe and spatial expression of mouse toe.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

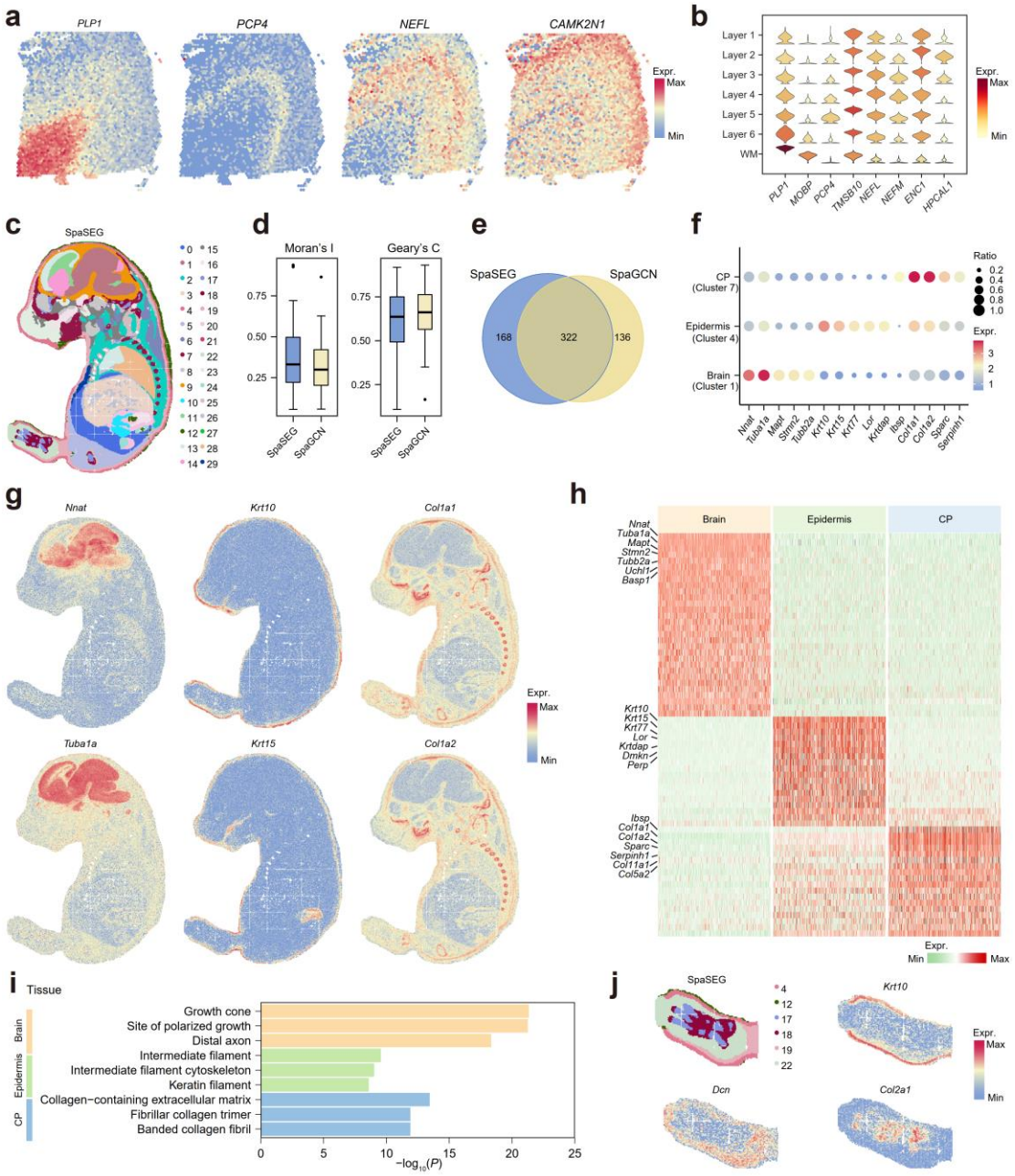


Figure 5. Cell-cell interaction analysis of the adult mouse brain. a, The workflow of cell-cell interaction. **b**, The result of SpaSEG clustering. **c**, Spatial expression of all significant L-R pairs. **d**, Spatially variable gene expression of zoomed-in cortex (cluster 5, 7, 12, 16). The top5 SVGs and SpaSEG clustering for cortex are shown. **e**, Spatially variable gene expression of hippocampus (cluster 4, 7, 11) based on SpaSEG clustering. **f**, The Spearman correlation of cell type scores and L-R expression values. The

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Spearman correlation values are represented by color. P values are represented by circle size. **g**, Significant L-R pairs of cortex area and hippocampus. **h**, A subset of cell types in the cortex and hippocampus, the proportions of which are estimated by cell2location.

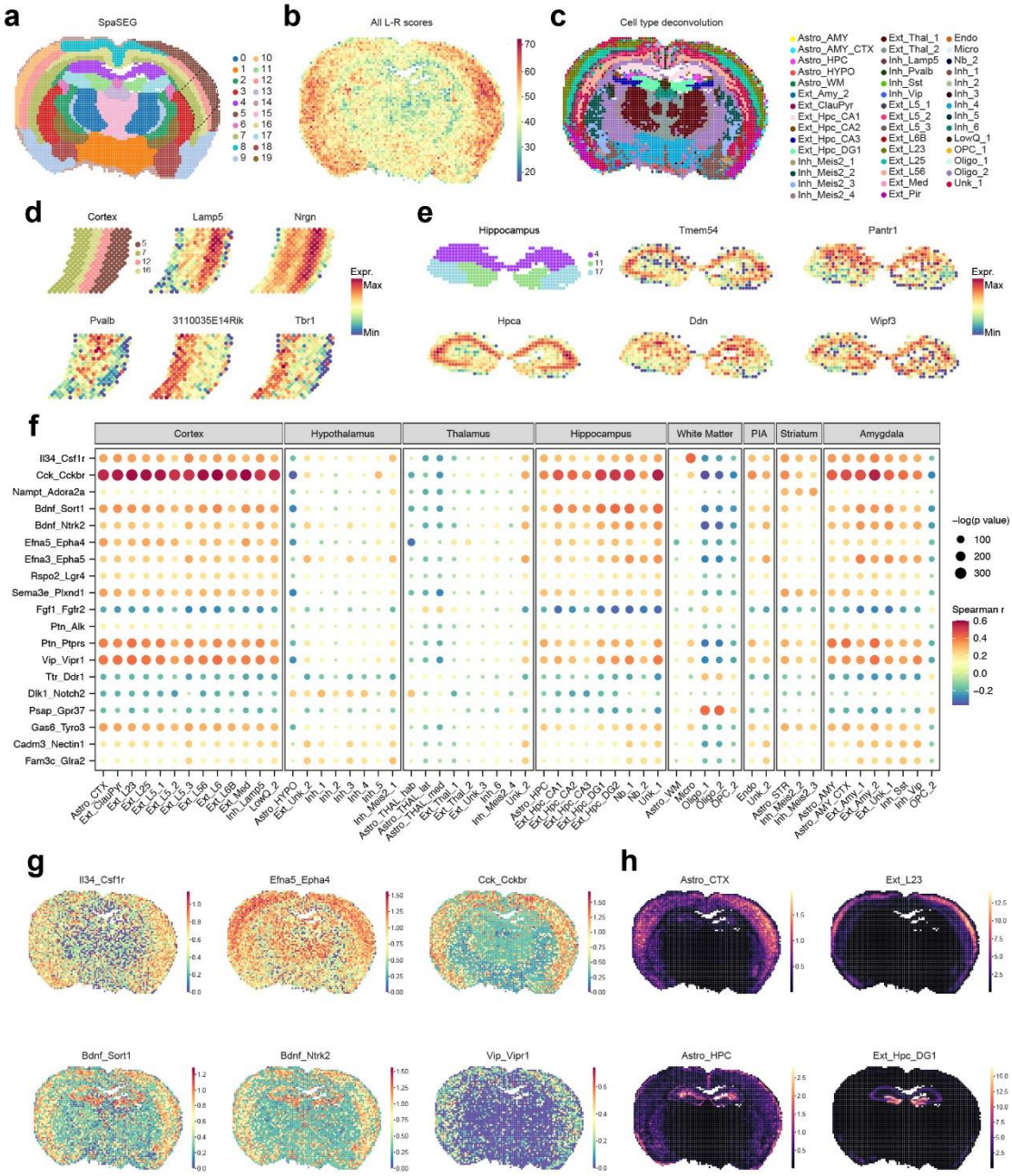
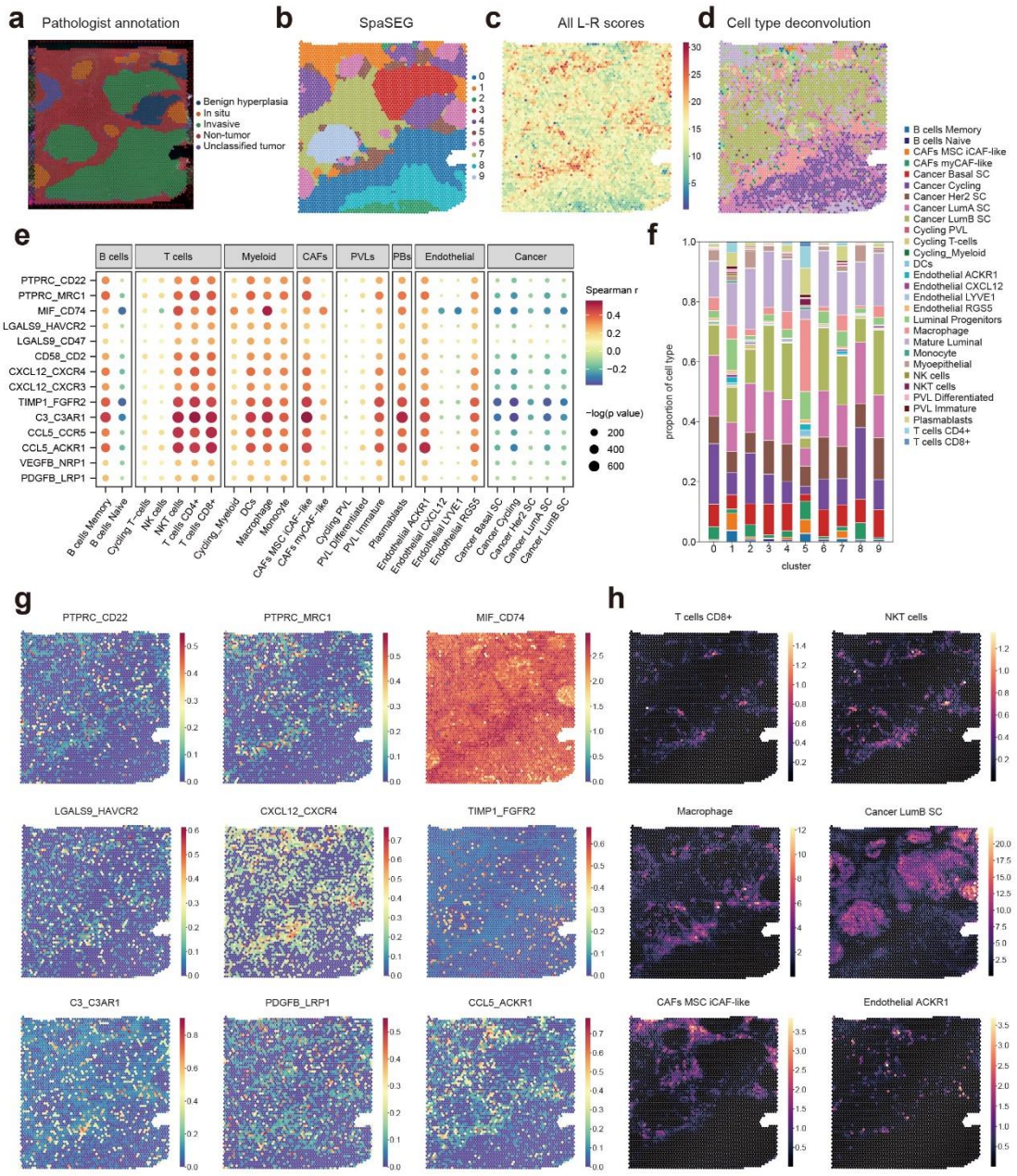


Figure 6. Cell-cell interaction analysis of IDC. a, The raw image of the IDC sample. **b**, The result of SpaSEG clustering. **c**, Spatial expression of all significant L-R pairs. **d**, The distribution of dominant cell types corresponding to the maximum score. **e**, The

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.16.516728; this version posted November 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Spearman correlation of cell type scores and L-R expression values. The Spearman correlation values are represented by color. P values are represented by circle size. **f**, Proportion of 29 cell types in each cluster. **g**. Significant L-R pairs of immune area and cancer area. **h**, Spatial distribution of a subset of immune cells (T cells CD8+, NKT cells, Macrophage and DCs) and stromal cells (CAFs MSC iCAF-like and Endothelial ACKR1).