Position-informed contrastive learning for spatially resolved omics deciphers hierarchical tissue structure at both cellular and niche levels

Meng Yang (yangmeng1@mgi-tech.com)
MGI, BGI-Shenzhen
Yue Wang
MGI, BGI-Shenzhen
Xingmin Liu
MGI, BGI-Shenzhen
Haiping Huang
MGI, BGI-Shenzhen
Linqi Lan
MGI, BGI-Shenzhen
Ming Ni
MGI-QingDao, BGI-Shenzhen
Yuchen Han
Department of Pathology, Chest Hospital Affiliated to Shanghai Jiao Tong University
Huanming Yang
BGI-Shenzhen
Feng Mu
MGI, BGI-Shenzhen

Article

Keywords:

Posted Date: January 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1067780/v1

Position-informed contrastive learning for spatially resolved omics deciphers hierarchical tissue structure at both cellular and niche levels

Meng Yang[1,2*], Yue Wang[1], Xingmin Liu[1], Haiping Huang[1], Linqi Lan[1], Ming Ni[4], Yuchen Han[6], Huanming Yang[3,5], Feng Mu[1*]

[1]MGI, BGI-Shenzhen, Shenzhen 518083, China.
[2]Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark.
[3]BGI-Shenzhen, Shenzhen 518083, China.
[4]MGI-QingDao, BGI-Shenzhen, Qingdao 266000, China
[5]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, 518120, China.
[6]Department of Pathology, Chest Hospital Affiliated to Shanghai Jiao Tong University, Shanghai 200030, China
*Correspondence to: yangmeng1@mgi-tech.com, mufeng@mgi-tech.com

Spatial omics and multiplexed imaging technologies provide unprecedented spatial context to characterize molecular variation in complex tissues. Developing unsupervised computational pipeline in a discovery mode is of vital importance to interpret spatially resolved data and derive novel biological insights. We develop Stereo, a unified framework leveraging contrastive learning to jointly model molecular and spatial information in a self-supervised manner. Stereo consists of two modules, StereoCell for neighbor-aware cell identity inference and clustering on spatial transcriptomics data, and StereoNiche for biologically-relevant niche identification on spatial proteomics data. StereoCell uses graph attention to aggregate immediate neighbors' content for each cell while StereoNiche exploits vision Transformer to explicitly relate position coordinates with marker readouts within certain receptive field. We validate StereoCell's superior utility on both simulated and real transcriptomics datasets of brain tissues acquired by various technologies with different resolutions. StereoCell can operate on all genes, extract layer-specific signature genes at single-cell resolution and identify clear layer-wise structure with state-of-the-art performance. Applied on tumor tissue sections, StereoCell enables detailed characterization of intratumoral heterogeneity and reveals invasive or metastasis behaviors concordant with expert annotations. Evaluated on a multiplexed fluorescence imaging dataset for colorectal cancer (CRC) patients, StereoNiche refines niche assignments, recapitulates classical follicle structure, and validates enriched T cells-macrophages interdigitation. Via self-distillation operation, StereoNiche also enables extracting bounded layout of tertiary lymphoid structure with no supervision, representing classical Crohn's-like reaction behavior. Local explanations are derived for a tree-based survival model to validate the clinical relevance of identified niches. For mass spectrometry profiled triple-negative breast cancer (TNBC) samples, StereoNiche recognizes differential tumor-immune interactions to separate compartmentalized from mixed pattern. In brief, Stereo is applicable to a variety of spatial technologies, serving as a powerful method to interrogate context-aware biological process and facilitate hierarchical signature discovery in neuroscience and computational pathology.

Tissue and organ systems consist of distinct cell subpopulations which are structurally organized and coordinately perform certain functions. Cellular states are mediated by both intrinsic factors and neighboring environment. Over the past decade, single-cell RNA sequencing (scRNA-seq) has unraveled heterogeneous composition of abundant cell types within tissues and build atlas-level annotations by marker genes. However, assays on isolated cells after tissue dissociation destroy spatial information, unable to capture localized effects of proximal environment on cellular variation. Besides, induced ectopic gene expression during dissociation might introduce artifacts and lead to mischaracterization of certain cell subpopulations[1]. Recently, a series of spatially resolved technologies[2,3] have progressed to profile cells within native tissue context. At transcriptional level, imaging-based approach, either using single-molecule fluorescence in situ hybridization (FISH) or in-situ sequencing, have overcome the challenges of spectral overlap and achieved up to 10,000 RNA detections to characterize organizational principles of brain tissues in the field of neuroscience, including multiplexed error-robust fluorescence in situ hybridization (MERFISH)[4], sequential fluorescence in situ hybridization (seqFISH+)[5], spatially resolved transcript amplicon readout mapping (STARmap)[6] and fluorescent in situ sequencing (FISSEQ)[7]. Another capturing-based roadmap leverages barcoded probes to record position coordinate and achieves unbiased

characterization of the whole transcriptome, such as commercialized 10X Genomics Visium solution with spot-level resolution[8] and recent large-field solution, Stereo-seq[9], which is based on MGI sequencing chemistry with subcellular nanoscale resolution. At protein level, recent technologies have enabled assessment of dozens of markers using antibodies staining in conjunction with various downstream signal detection methods, including cyclic immunofluorescence (cycIF)[10], co-detection by indexing (CODEX)[11] using single-step staining by DNA-conjugated antibodies, mass spectrometry-based imaging mass cytometry (IMC)[13] and multiplexed ion beam imaging by time-of-flight (MIBI-TOF)[14]. These spatial proteomics protocols have been applied to interrogate structured tumor microenvironment, derive novel hypothesis of orchestrated tumor-immune interactions[12,15], and offer insights into translational or clinical usage. With the expeditious advances of spatial profiling technologies, new computational approaches are imperative to extract unique properties from spatial-resolved dataset in a discovery mode. We consider an integral framework should contain following modules, 1) cell type inference using neighboring information, 2) niche segmentation under structured tissue context, 3) linking hierarchical spatial components with anatomical structures, biological processes, or clinical features to derive novel insights.

Firstly, for cell type assignment, many existing pipelines rely on integrating prior knowledge from single-cell reference. SPOTlight[16], cell2location[17] and RCTD[18] enable transferring annotations from dissociated reference atlas to spatial transcriptomics data. However, several cell types are hard to dissociate or lost during processing, leading to incomplete coverage of single-cell reference. Dissociation workflow might also introduce stress response and ensuing altered distribution of certain sensitive gene expression program[19]. The key drawback of purely relying on single-cell reference is ignoring spatial context to define cellular identity and underexploit spatial profiling data. With the accumulation of spatial omics, it is imperative to simultaneously model both intrinsic and adjacent factors to extract discriminative features for a localized cell. Zhu[20] proposed a hidden Markov random field (HMRF) method to identify spatial domains with coherent gene expression patterns and further improved it into Giotto[21] package. BayesSpace[22] introduces neighboring spot structure into the prior and update HMRF model with a Bayesian approach to assign clusters. LEAPH[23] uses linear factorization-based soft clustering with spatial regularization to define cell phenotypes. SpiceMix[24] introduces non-negative matrix factorization (NMF) into HMRF framework to jointly model spatial dependency and latent factors of cell identity, leading to novel cell subtypes discovery from mouse visual cortex data. All above HMRF-based methods require pre-defining number of clusters or cell types as model initialization. SpaGCN[25] leverages histological features and location to construct weighted graph for spots, which is used as adjacent matrix for graph convolutional network (GCN) to aggregate neighboring spots' gene expression. SEDR[26] is recently proposed to learn latent representation for each cell by combining deep autoencoder for gene expression reconstruction with variational graph autoencoder network for spatial information embedding. Representation learning decouples feature learning with clustering rather than integrating them as for HMRF-based approach, which is expected to be more flexible for diverse downstream tasks. Due to the intractable nature of probabilistic model to handle large feature space, BayesSpace only operates on 15 principal components derived from principal component analysis (PCA) and SpiceMix relies on similar number of meta-genes after matrix decomposition. SEDR is also applied on the first 200 principal components rather than full gene count matrix. We suggest that operating on all genes is crucial to retain subtle biological nuance.

We propose StereoCell as a cell-type inference tool for spatial transcriptomics dataset, discarding PCA-like dimension reduction to maintain original input to the greatest extent. Spatially variable genes (SVGs) are defined as differentially expressed genes in a spatially dependent fashion. Trendsceek[27], SpatialDE[28] and SPARK[29] draw lessons from geo-statistics and test each gene independently to screen candidate SVGs. These statistical methods ignore underlying tissue taxonomy and thus cannot ensure enough specificity. Implementing downstream gene-level analysis on clustering results by HMRF or unsupervised representation learning can potentially overcome these limitations via offering interpretation in alignment with layer-wise anatomical structure.

Secondly, cellular neighborhoods (CNs) are defined by Schürch as regions with a characteristic local stoichiometry of cell types. Similar concepts are proposed by other groups, including spatial domains[20], spatial communities[30], microdomains[23,31] or tissue domains[32], tissue niche[33]. There is no uniform definition recognized by the community to define CNs. It is worth mentioning that imaging based spatial proteomics methods preselect dozens of biomarkers to separate different cell types and thus maximize between cell type variance, while spatial transcriptomics protocols with more abundant features (>100 gene transcripts) can better capture within-type variation to derive novel cell subtypes[34]. In this article, we benchmark StereoCell against abovementioned competing methods for spatial transcriptomics dataset to learn discriminative representation for each cell, which can be further used for follow-up analysis. On the other hand, incorporating position-aware tissue structure is necessary to identify prototypical niches consisting of higher-order interactions between one or more cellular phenotypes. HistoCAT[35] uses statistical permutation test to identify patterns of pairwise cellular colocalization across a tissue and has been widely adopted in research communities. LEAPH[23] introduces pointwise mutual information as pairwise association statistics to discover microdomains. Both HistoCAT and LEAPH can only detect pairwise co-occurrence at a time and rely on sequential statistics analysis instead of a learning-based approach. Schürch[12] uses simple K-means clustering of cell-type frequencies within local-window defined by neighbor quantity. This strategy results in undetermined size of receptive field and cannot model higher-order intercellular communications in a position-aware manner. Jackson[30] employs community detection methods on spatially constrained adjacent matrix to identify highly interconnected spatial subunits as communities within tissue graph and then recognize recurring pattern via clustering on cell-type composition within each community. These two methods reply on clustering on frequency information of pre-defined cell types, either in a fixed window size or a connected graph. These models disentangle molecular information with cells' position coordinate and original marker intensity is only used at prerequisite cell-type assignment stage. We suggest that jointly modelling spatial and molecular information to relate protein markers with structure tissue context is necessary for niche discovery[36]. We develop StereoNiche, specifically designed for niche segmentation on spatial proteomics datasets and accounts absolute position information within a pre-determined receptive field to study highly complex tumor microenvironment.

Thirdly, biological insights hinge on how to link identified cell types and niches to certain study objectives, especially in stratifying cancer patients into relevant risks groups based on TME configurations to drive diagnostic or prognosis value relevant to clinical outcome. Linear survival model cannot satisfy the complex non-linear interactions of diverse spatial elements and enumerating all combinations to test associations is computationally intractable[37]. Machine learning

based survival analysis[38] has emerged to better tackle the non-linearity. Further developing an automatic, interpretable, quantitative approach for plausible black-box models is of vital importance to prioritize attributed features and detect outcome-specific biomarkers in the field of computational pathology. Structurally organized at different levels, certain cell phenotypes in specific niches and niche interactions, all contribute to complex host-tumor interface with huge potential as novel prognostic spatial biomarkers to answer clinical questions. Therefore, interpretability for machine learning based survival models is of vital importance for translational or clinical research.

To fill abovementioned gaps, we present Stereo, an integrative computational framework composed of StereoCell and StereoNiche. StereoCell is designed for cell identity inference and spatial clustering for spatial transcriptomics datasets with >100 RNA features, which emphasizes on capturing subtle within-cell type variance. StereoNiche is designed for niche discovery for spatial proteomics datasets with <100 protein markers, which focuses more on spatial structure modeling. StereoCell aims at incorporating immediate neighbors' information to refine cell assignment while StereoNiche leverages position-aware tissue organization to identify functional niches. Based on this distinction, StereoCell models relative position information for each cell while StereoNiche requires absolute position coordinates as input. StereoCell upgrades the self-distillation contrastive learning framework via adding another neighbor module to jointly embed molecular and spatial content into unified cell representations in a self-supervised manner. Neighbor module uses graph attention mechanism[39] to aggregate information from immediate neighbors. StereoCell is tested on simulated datasets with demonstrated utility and then applied on mouse primary visual cortex data acquired by STARmap[6], human dorsolateral prefrontal cortex (DLPFC) data[40] and human breast cancer data acquired by 10X Visium solution[3], mouse olfactory bulb data acquired by high-resolution Stereo-seq[9]. All benchmark experiments show StereoCell's superior performance over other competing methods. On the other hand, StereoNiche treats niche identification as an unsupervised semantic segmentation task and leverages self-attention based vision Transformer architecture with learnable positional embeddings to generate contextualized patch embeddings in a contrastive manner, which is expected to better model position information within a pre-defined receptive field. StereoNiche is applied on CRC[12] and TNBC[15] datasets acquired by CODEX and MIBI-TOF, respectively and successfully detect several known and novel TME signals. We further utilize an interpretable tree-based survival model[38] to attribute hierarchical tissue components with differential patient outcome[37]. StereoCell and StereoNiche both utilize dropout-layer as minimal data augmentation to generate positive pairs for contrastive learning. In a short, Stereo serves as a robust, accurate and comprehensive framework to derive novel insights from spatial-resolved omics in a fully data-driven manner.

Results

Overview of Stereo architecture

StereoCell (Figure 1-a) exploits a self-distillation contrastive learning framework[41] with an asymmetric teacher-student architecture and further introduces a neighbor module to incorporate spatial context. StereoCell begins with a quality-controlled cell by gene count matrix with 2D

coordinate. Student module accepts discrete gene counts while teacher module leverages attention operation to aggregate distributional gene embeddings scaled by count value into cell representations. Neighbor module uses graph attention operation to relate each cell with its local neighbors, which are determined by calculating pairwise Euclidean distance between spatial coordinates at certain adjustable threshold. Random masks of dropout layer are used as model-level data augmentation to generate positive pairs[42]. The pre-text task is defined as maximizing agreement between each cell's paired embeddings from teacher-student view formulated as intrinsic contrastive loss and maximizing similarities of each cell's representations with its surrounding neighbors as neighboring contrastive loss. The relative weights of two loss components can be tuned with adjustable parameters, indicating contribution at different levels from spatial context. StereoCell then learns unified cell representations by jointly modelling intrinsic and environmental factors in an adaptive manner. The former loss maintains biological invariance and the latter imposes spatial coherence. The output embedding from neighbor network is extracted as final representation for each cell. Then unsupervised clustering is implemented to blindly infer either discrete or continuous cell subpopulations, which can be mapped back onto original coordinates.

StereoNiche (Figure 1-b) treats CNs identification as an unsupervised semantic segmentation task considering both local marker count and position-aware tissue context. In contrast to StereoCell, which only incorporating immediate neighbors defined by certain distance threshold, StereoNiche using self-attention based vision Transformer[43] to directly model x-axis and y-axis coordinate together with pixel-level marker value as analogous to computer vision problems. StereoNiche first segments a whole tissue section into medium patches and further partitions them into small patches containing zero, one or two cells. Small patch embeddings are defined as average pooling of all contained cells' embeddings. Medium patch embedding are aggregated from all contained small patches through a Transformer architecture into a [CLS] token embedding. The spatial arrangement within a medium patch is modelled as learnable 1D position embeddings for flattened small patches. Random dropout masks are sampled to synthesize data augmentations. Medium and small patch embeddings are learned by pulling together token embeddings for each small patch or medium patch with corresponding augmented views and pushing apart representations from different patches via a contrastive loss. It is noted that StereoNiche uses momentum encoder[44] and maintains a negative first-in first-out queue to compute contrastive loss more efficiently, as opposed to batch-wise contrastive loss in StereoCell. The hierarchical learning objective equips StereoNiche to relate local molecular readout with spatial tissue structure though pre-defined cell-type information is never used in the learning process. Then small patch embeddings are clustered to identify spatially configured niches. After identifying both cellular and niche level TMEs, we further leverage a tree-based survival model[38] to derive explanations[37] for candidate spatial biomarkers (Figure 1-c).

StereoCell achieves superior performance of cell type inference on simulated FISH-based transcriptomics data

We first use 8 simulated spatial transcriptomics datasets provided by SpiceMix to mimic observed pattern of real primary mouse visual cortex. Three major cell types or eight subtypes, including four excitatory neurons, two inhibitory neurons and two glial cells, are distributed across four-layers in the cortex (Depicted in Figure 2-a). The simulated datasets are devised to enforce excitatory neurons

manifest layer-wise specificity, inhibitory neurons diffuse sparsely and while glial cells exhibit mixed patterns. We assess StereoCell's performance of cell type inference against classical single-cell toolkits Seurat[45] and SCANPY[46], standard HMRF[20,21], matrix factorization-based SpiceMix[24] and a degenerated version of StereoCell after removal of neighbor module. Among which, SCANPY, Seurat and StereoCell (no spatial context) all preclude spatial information, HMRF and SpiceMix are based on probabilistic modelling with pre-defined number of clusters as initialization. Solving SpiceMix requires iteratively estimating hidden states and model parameters, which incurs huge computational burden and relies on dedicated optimization software package[24] (Gurobi) to scale up for larger datasets. We set the weight of intrinsic contrastive loss at 0.1, neighboring contrastive loss at 0.9 and implement StereoCell to learn representation for each cell. Then K-means (k=8, matching the simulated ground truth) clustering is applied to StereoCell embeddings to group cells into clusters. Visualized by assigning all excitatory neurons at spatial coordinate (Figure 2-b), Seurat fails to identify layer-wise distribution of four excitatory neuron subtypes at four layers (eL1-4), while HMRF and StereoCell obtains stronger layer specificity. HMRF clusters show some inconsistency with ground truth, especially for layer 2 (L4) and layer 4 (L4). The simulated cortex structure is best recapitulated by StereoCell with clear boundary between 4 layers. The utility of learned latent space is further demonstrated by uniform manifold approximation and projection (UMAP) plot (Figure 2-c, d, e). StereoCell embeddings well separate four layers of excitatory neurons while removing neighborhood module results in intermixed pattern, revealing the importance of introducing spatial context to refine cell type assignment. Adjusted Rand index (ARI) against ground truth cell types is used as quantitative metric to compare all methods (Figure 2-f). StereoCell obtains the highest performance (mean ARI= 0.81 for 8 samples) and performs more robust than SpiceMix (ARI=0.57-0.95, mean=0.77 on average). Both HMRF (ARI=0.17-0.61, mean=0.45) and SpiceMix fluctuates more violently than other methods, whose instability might be attributed to probabilistic model's sensitivity to initialization and optimization technique. It is worth mentioning that StereoCell outperforms Seurat (mean ARI= 0.44) and SCANPY (mean ARI= 0.29) even after removing the neighbor module (StereoCell, no spatial context, mean ARI= 0.51), which again demonstrates the advantage of cell discrimination task implemented by self-distillation contrastive learning. StereoCell is trained using Intel Xeon CPU E5 node with 1 GPU (Nvidia Titan Xp). Peak memory usage is set to be 5 Gigabytes. StereoCell embeddings are extracted after 10 epochs (around 30 minutes in total). As a learning-based approach, StereoCell obtains better performance at the cost of comparable runtime with the least memory consumption to those statistical or probabilistic methods (Supplementary Figure 15). We choose Seurat and SCANPY to represent statistical methods, Giotto, BayesSpace and SpiceMix to represent probabilistic methods, SpaGCN and SEDR to represent learning-based approaches. Detailed comparison results are listed in Supplementary Table 4 and 5.

StereoCell refines cell subpopulations and reveals functionally consistent subtypes on mouse primary visual cortex data acquired by STARmap

We apply StereoCell to a real spatial transcriptomics dataset of mouse cortex acquired by STARmap (one slice containing 970 cells with measurement of 1020 genes). Clustering of StereoCell embeddings uncovers 17 subpopulations under spatial context, including 7 excitatory neural subtypes, 4 inhibitory neural subtypes and 6 non-neuron subtypes. We select k=17 to obtain the best

ARI against a recently released Mouse Whole Cortex and Hippocampus (MWCH) single-cell atlas with annotations[47]. StereoCell reveals layer-wise structure of excitatory neurons and classifies those into 7 subtypes: eL2/3，eL4-1，eL4-2，eL5，eL6-1，eL6-2 and hippocampal pyramidal cell (HPC). In comparison to STARmap assignment, StereoCell obtains clearer layer boundary with higher spatial consistency (Figure 2-g). Marker expression pattern for 17 StereoCell clusters are shown in Figure 2-h and all annotations are supported by calculating Pearson correlation coefficients of all expressed genes averaging within each cluster against MWCH[47](Supplementary Figure 1). StereoCell corrects 19 cells erroneously assigned as eL2/3 into eL4 cells, which are further validated by their specific expression of canonical eL4 markers, EGR1 gene and ZMAT4 gene (p-value=2e-07, 0.0001 respectively, Wilcoxon rank-sum test, eL4 versus eL2/3, StereoCell assignment). These cells are significant correlated with annotated eL4 annotations from MWCH reference (Pearson r: 0.41). Using confusion matrix to visualize the agreement of StereoCell assignment with original STARmap definition (Figure 2-i, Supplementary Table 2), 22 Astrocyte (Astro-1, Astro-2) subtypes assigned by STARmap are labelled as eL5 neurons by StereoCell. These cells express excitatory neuron marker genes, such as FEZF2 gene and LAMP5 gene, at higher level than original Astro-2 assignments (p-value= 2e-05, 0.048 respectively, Wilcoxon rank-sum test, eL5 versus Astro-1/2, StereoCell assignment) while expressing much lower level of intrinsic astrocyte markers APOE and SPARCL1[47] (p-value=8.3 e-12, 0.0031 respectively, Wilcoxon rank-sum test, eL5 versus Astro-1/2, StereoCell assignment), resembling excitatory neurons more than astrocytes. This misassignment of astrocytes is also validated by SpiceMix and both methods verify the diffusion pattern of astrocytes instead of illusive localized signal by STARmap. StereoCell also distinguishes an Astro/Oligo subtype (33 cells) from other Oligo class. These cells locate in deeper cortex layer and express both astrocytes marker genes APOE and GFAP (p-value=8e-07, 1e-06 respectively, Wilcoxon rank-sum test, Astro/Oligo versus all others except astrocytes, StereoCell assignment) and oligodendrocytes marker genes MBP and PLP1[47] (p-value=3e-18，2e-18 respectively, Wilcoxon rank-sum test, Astro/Oligo versus all others except oligodendrocytes, StereoCell assignment). These signals validate Astro/Oligo's intermediate state. The confusion matrix also depicts that these cells align with both Astro-2 (n=6) and Oligo (n=13) assignment by STARmap. We further conduct Gene Ontology (GO) analysis of these cells and find that gliogenesis, myelination, axon ensheathment and glial differentiation related processes are enriched (Supplementary Figure 2), manifesting behaviors of continuous myelination among oligodendrocytes[48]. This differentiation-like process is also validated by SpiceMix. In addition, StereoCell refines original eL4 assignment by STARmap and partitions these neurons into eL4-1 and eL4-2[47] (differential maker genes: GLUl (p-value=1e-7), SPARCL1 (p-value=0.0015), DUSP1 (p-value=0.048), Wilcoxon rank-sum test). To demonstrate StereoCell can identify spatial expression pattern with biological interpretability, we further compare whether differential expressed genes of StereoCell's clusters are on par with SVGs detections by other geostatistical methods, including SPARK and SpatialDE. In total, StereoCell detects 328 SVGs (False discovery rate (FDR)-adjusted p-value< 0.05, Wilcoxon rank-sum test, one versus all) and 401 SVGs are detected by SPARK (FDR-adjusted p-value< 0.05). However, SpatialDE identifies 944 genes with q-value less than 0.05, whose skewed distribution towards zero leads to large amount of false positive signals. The Venn diagram (Figure 2-k) shows that 247 SVGs are detected by all three methods with average FDR-adjusted p-value equals to 2e-4 evaluated by SPARK. Among which, we plot expression heatmap with spatial coordinate for excitatory neuron marker gene SLC17A7 (p-value=2.6e-53, Wilcoxon rank-sum test, excitatory neurons versus all

others), oligodendrocyte marker gene MBP (p-value=1.6e-64,Wilcoxon rank-sum test, Oligos versus all others), inhibitory neuron marker gene GAD2 (p-value=8.2e-8,Wilcoxon rank-sum test, inhibitory neurons versus all others), eL2/3 marker gene CUX2 (p-value=4.6e-15 ,Wilcoxon rank-sum test, eL2/3 versus all other excitatory neurons). We show that SLC17A7 manifests layer-wise expression pattern, MBP is enriched in a deeper tissue layer (L6), GAD2 diffuses across tissue and CUX2 is layer-specific (Figure 2-l). We further calculate the Moran's *I* statistic, which quantifies the spatial autocorrelation of detected SVGs (Figure 2-j). The results show that Moran's *I* of StereoCell and SPARK are comparable (median value = 0.177 for StereoCell, 0.173 for SPARK) and both significantly surpass SpatialDE (median value = 0.120 for SpatialDE) (p-value < 0.001). These results suggest that StereoCell's effectiveness as spatially-informed cell-type inference tool via combing intrinsic expression value with neighboring information. StereoCell is flexible to adjust relative contributions from either intrinsic or environmental factors simply by tuning the ratio of two contrastive loss terms. Collectively, StereoCell has demonstrated its unique advantage to accurately identify coherent spatial expression pattern of brain cell subtypes.

StereoCell achieves superior performance on spatial transcriptomics data with different resolutions to identify known layer-wise structure in brain tissues.

Beyond FISH-based methods, we further evaluate StereoCell's capability to tackle with higher-dimensional sparse spatial transcriptomics dataset acquired by barcoded capturing technology. We choose a 10X Visium spatial expression profiles of 12 human dorsolateral prefrontal cortex (DLPFC) slices with manual annotation for 6 cortical layers and white matter as a classical benchmark dataset. Here, we compare the performance of StereoCell against other spatial clustering methods including HMRF (as implemented in the Giotto package), BayesSpace, SpaGCN and a recent autoencoder-based representation learning approach, SEDR. Layers (n=7) separated by clear boundaries defined by Maynard are considered as ground truth (slice 151673 annotation is shown in Figure 3-b) and ARI for all 12 samples is used as metric to quantify the clustering consistency. It is noted that HMRF and BayesSpace explicitly model number of clusters in their algorithms, SpaGCN incorporates histological features to aggregate neighbors' information before executing iterative clustering, while SEDR and StereoCell's main objective is to learn cell embeddings under spatial context with subsequent clustering step on learned representations in a decoupled manner. We anticipate that decoupling feature learning with clustering is crucial to maintain original data attributes to the greatest extent without constrained by pre-determined discrete prior and incorporating visual histological features is unnecessary to interpret high-dimensional spatial omics data. K-means is used as clustering method for SEDR and StereoCell. Setting intrinsic contrastive loss weight at 0.1, neighboring contrastive loss weight at 0.9, StereoCell (mean ARI=0.61) achieves the highest overall performance against current best pipeline BayesSpace (mean ARI=0.46), SEDR (mean ARI=0.43), SpaGCN (mean ARI=0.40) and Giotto (mean ARI=0.36) (Figure 3-a). Detailed comparison results for 12 slices are listed in Supplementary Table 3. StereoCell effectively incorporate spatial information to improve feature learning and obtains distinct clusters with expected layer-wise pattern. BayesSpace operates on 15 PCs computed from the top 2,000 HVGs while StereoCell is applied on full gene count matrix with better accuracy. The nature of BayesSpace's probabilistic model makes it intractable to handle large feature space while StereoCell's contrastive learning objective can utilize complete information as a unique advantage. To further demonstrate the

biological meaning of learned representations by StereoCell, we implement UMAP visualization of both raw expression data and Stereo embeddings for DLPFC 151673 section (Figure 3-c). Compared to the unstructured UMAP layout of raw data, the trajectory in the StereoCell latent space is well-organized from cortical layer 1 to layer 6 and white matter in a sequential order, showing consistent spatial arrangement with known cortical developmental chronology. Cluster assignments by five methods are further visualized via mapping backed to original layout (DLPFC 151673 section) and shown in Figure 3-e. HMRF and SpaGCN obtain jagged boundaries between layers and SEDR results in extra layers. BayesSpace presents uniform axial distance distribution among layers, which are inconsistent with ground truth annotation. StereoCell almost perfectly reproduces thin and thick characteristics from Layer 1 to Layer 4 with certain unbalanced occupation of Layer 5 and Layer 6. StereoCell's teacher module and neighbor module both utilize attention to aggregate gene embeddings, and we investigate whether those attention weights can provide intrinsic interpretability via reproducing marker signals established for layer-wise pattern. The heatmap of normalized attention weights from neighbor module depicts the relative contributions of characteristic features to define specific layer, which successfully recovers canonical layer-specific molecular signature (Figure 3-d). Particularly, MYL9, GNAL, CARTPT, VAMP1, PCP4, KRT17 and PMP22 are selected as the most representative markers from Layer 1 to Layer 6 and white matter, consistent with Maynard's description[40]. However, only referring to teacher module with no spatial context cannot identify meaningful features (Supplementary Figure 3). By this means, layer-specific signatures are automatically extracted at single-cell resolution without the needs of clustering-based differential test and the results coincide with expert annotations. The only learning signal is each cell itself and its neighboring cells. We propose this novel clustering-free layer-wise marker identification approach without using any supervision signal, potentially serving as an unbiased knowledge discovery protocol.

Then we evaluate StereoCell's utility for Stereo-seq, a recent nanoscale spatial transcriptomics technology at subcellular resolution. We apply StereoCell on mouse olfactory bulb dataset with binned cellular resolution and conduct K-means (k=9) to identify known layers. The results identify distinct clusters with remarkably clear boundary and delineate laminar pattern consistent with known anatomical layers of the coronal section (Allen Brain Reference Atlas annotation[49] shown in Figure 3-h). Particularly, glomerular layer (GL) and olfactory nerve layer (ONL) are well distinguished, narrow mitral cell layer (MCL) are well separated with internal plexiform layer (IPL). Compared to SEDR, StereoCell obtains visually more obvious separation among ONL, GL and external plexiform layer (EPL) than SEDR (depicted in Figure 3-g). We use k-nearest neighbor similarity Test (kSIM, k=25) and average silhouette width (ASW) as quantitative metrics for spatial separability. StereoCell obtains larger kSIM and ASW (kSIM=0.90, ASW=0.28 for StereoCell 10,000 HVGs, kSIM=0.65, ASW=0.13 for SEDR). We also test both 2000 HVGs and 10,000 HVG scenarios for StereoCell. Visually, the latter reconstructs much clearer spatial identity than the former (kSIM=0.80, ASW=0.14 for 2000 HVGs), which again validates the utility of operating on all genes instead of subjectively selected features (Figure 3-f, g). To the best of our knowledge, StereoCell is the first method to directly learn from > 10,000 genes for high-resolution spatial transcriptomics data with proven necessity provided model capacity supports. Again, we use runtime and memory consumption to evaluate StereoCell's computational efficiency against other 8 methods on DLPFC (151673) dataset (Supplementary Figure 15). It is noted that StereoCell

removes PCA operation and is trained on 10,000 HVGs, achieving the best ARI within 10 epochs. Though all other methods are implemented on PCs as a much lower-dimensional input, StereoCell still consumes the second least memory within comparable runtime with probabilistic Giotto and BayesSpace. It is noted that SpiceMix fails to accomplish this task for high-dimensional spatial transcriptomics dataset. Detailed measurements are listed in Supplementary Table 4 and 5. This improvement significantly reforms current well-recognized practice and is expected to set an example for the community how to take full advantage of high-dimensional spatial transcriptomics.

StereoCell enables discovering heterogenous invasive tumor regions and subtle pathway signals of in situ carcinoma on a 10X Visium profiled breast cancer sample

Beyond brain tissues with established layer-wise morphological structure, we further validate whether StereoCell can characterize more heterogenous tumor samples. We apply StereoCell on a Visium spatial transcriptomics of a human breast cancer sample (10X BC, 2 slices). Hematoxylin & Eosin (H&E)-stained images are manually segmented into three major areas (Figure 4-a), including ductal carcinoma in situ (DCIS), invasive carcinoma (IC) and healthy stroma (Stroma)[50]. We also refer to another annotation by expert pathologists[3] to account for more subtle signals. Several infiltrating lymphocytes appear near or inside tumor regions (Figure 4-a). Inflammatory fibroadipose and invasive foci are distributed sporadically in the upper right corner of the slice. We implement StereoCell, SEDR, BayesSpace and Giotto on 2 slices. 20 clusters are applied for all methods as fair comparison and spatial layouts are visualized in Figure 4-b. To characterize 20 clusters, we harness recently reported 29 functional gene expression signatures (FGSE)[51] as conserved pan-cancer TME pattern and conduct Gene Set Variation Analysis (GSVA) on StereoCell clusters (Figure 4-d). Cluster 3,8,17,18,19,20 exhibit stronger Proliferation_rate signals than Cluster12. Cluster 11 and 13 show more significant Antitumor_cytokines signatures than Cluster12. Cluster 12 is depleted of cancer-associated fibroblast (CAF) signature (p-value=7.1e-52, Cluster 12 versus others). Therefore, we assign Cluster 3,8,11,13,17,18,19,20 as IC regions and Cluster 12 as DCIS, coinciding with H&E annotations. StereoCell can discriminate three DCIS regions (StereoCell cluster 12) and classify them into the same class, indicating DCIS's homogenous characteristics as also validated by BayesSpace and SEDR. However, HMRF-based Giotto fails to identify these phenotypically similar DCIS areas. Interestingly, StereoCell identifies similar outer ring-like tumor boundary (StereoCell cluster 10) for all three DCIS. We conduct Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for cluster 12 and 10 (Figure 4-f). The peripheral areas exhibit localized enrichment of focal adhesion and interleukin response, probably as early malignant progression signals of DCIS lesions indicated by surrounding areas' dysregulation of cell adhesion and extracellular matrix (ECM) pathways[53] and constrained immunoreactive effects therein. The inner regions are enriched with hypoxia and related glycolysis metabolism signals, implying DCIS's tumor-specific metabolic pathways under nutrient starvation. BayesSpace only identifies two of these 3 ring structures. On the other hand, 8 IC clusters (3,8,11,13,17,18,19,20) manifest notable heterogeneities. We use Seurat FindAllmarker() to identify cluster-specific marker genes (Wilcoxon rank-sum test, all p-values after Bonferroni correction). Cytochrome C Oxidase Subunit 6C (COX6C), related to cellular oxidative phosphorylation pathway of mitochondria, as a cataloged upregulated signature for breast cancer[56], is overly expressed in all IC except cluster 18 (log2(fold change) = -2.4, p-value = 3.7e-71, cluster 18 versus all other IC

clusters). From the visual pattern of clustering result, StereoCell cluster 20 shows similar matrix outer ring (cluster 17) features as for cluster 18 (19 as its outer ring), which is also reported by the morphological similarity pattern observed by H&E. Cysteine-rich secretory protein 3 (CRISP3) is considered as a pro-tumorigenic marker to promote cancer migration and invasion while CXCL14 as a chemokine with diversified immunological functions. Despite the controversy, CXCL14 is reported to exert anticancer role in breast cancer[57] and differential expressed in cluster 18 (log2(fold change) = 4.3, p-value = 8.5e-72, cluster 18 versus 20) while invasive marker, CRISP3, is more enriched in cluster 20 (log2(fold change) = 6.0, p-value = 8.5e-72, cluster 20 versus 18). Cluster 18/19 also shows depleted Treg_traffic signature (log2(fold change) = -0.43, p-value = 3.6e-51, cluster 18/19 versus other IC clusters). The enriched CXCL14 expression and depleted Treg trafficking behavior of Cluster 18 might imply its divergent immune reaction mechanisms. Both inner regions (17 & 5) show even higher expression of corresponding CXCL14 and CRISP3 marker than outer regions (p-value = 3.1e-17, cluster 18 versus 19 for CXCL14; p-value =1.5e-13, cluster 20 versus 17). Invasive marker, carboxypeptidase B1 (CPB1) is differential expressed in cluster 13,11 (log2(fold change) = 3.6, p-value = 2.2e-138, cluster 11,13 versus other ICs), and slightly weaker proliferative status of cluster 11,13 are also illustrated in in-situ FGSE signature heatmap in Figure 4-c, potentially revealing distinctive pro-metastasis mechanism of these two IC clusters. Another H&E annotation also reported disparity of these two regions[50]. At the upper right corner, StereoCell obtains inhomogeneous pattern similar with BayesSpace and HMRF rather than a uniform cluster by SEDR. The expert annotations show there exist fibroadipose tissue with inflammatory cells and invasive foci in this region while SEDR claims non-malignancy. We identify long noncoding RNA, MALAT1, as region specific marker (log2(fold change) = 2.8, p-value = 2.6e-152 cluster 2,6 versus others), which has controversial role of promoting or suppressing metastasis[58]. We further uncover enrichment signal of focal adhesion in cluster 6 and upregulated immune response in cluster 2 (Figure 4-e). Cluster 6 manifests invasive matrix degradation at focal adhesions as an indicator of tumor metastatic process[59]. Furthermore, CAF signature is also enriched in cluster 6 (Figure 4-c), indicating its proximity with invasive foci and involvement in ECM remodeling[60]. These signals recapitulate expert annotations about emerging invasion. Clustering visualizations for another slice are shown in Supplementary Figure 4 and StereoCell's spatial heatmaps for other 25 FGSEs are shown in Supplementary Figure 5. In summary, StereoCell can identify three DCIS regions (Giotto fails and BayesSpace only detects two of them), invasive foci (SEDR fails) and heterogenous IC regions, serving as a powerful discovery tool to decipher complex TME.

StereoNiche identifies outcome-relevant niches in an unsupervised manner on spatial proteomics datasets to dissect tumor heterogeneity of CRC and TNBC patients

We propose a different approach to process spatial proteomics dataset as dozens of markers are expected to maximize between cell-type variance to separate different cells. StereoNiche is devised to model higher-order intercellular effects as a de novo spatial pattern discovery tool, thus requiring directly representing positional information and relating protein markers with surrounding tissue context. This positioning is different from StereoCell's objective to incorporate immediate neighbors' molecular variation to capture cell-type nuance. We draw lessons from recent computer vision progress and jointly represent marker readouts at absolute x-axis and y-axis coordinate information as dozens of channels at each pixel. We consider cell niche as independent functional

module whose identity can retain even though its composition changes, i.e., pattern unaltered. We demonstrate how StereoNiche can facilitate niche identification in the field of computational pathology. We first evaluate StereoNiche on FFPE-CODEX dataset of 56 markers from 35 CRC patients and compare StereoNiche's segmentation result with 9 cellular neighborhoods (CNs) identified by window-level cell-type frequency clustering as in original article[12], where only cell neighbors' assignments rather than actual locations are directly used. Schürch identifies 28 cell types (CT) annotated by either supervised manual gating or unsupervised clustering[61]. However, StereoNiche treats niche identification as an independent task and directly learn niche-level representations without mediating through cell type information annotated by the author. Unlike Schürch's local window approach of fixed 10 nearest cell neighbors, StereoNiche leverages grid-like regular patches (containing 1-31 cells) to define tissue context under coordinate-aware spatial constraint and aim for better representation of TME organization. All 35 patients are classified into 17 CLR (''Crohn's-like reaction'') and 18 DII (diffuse inflammatory infiltration) patients according to presence or absence of tertiary lymphoid structure (TLS). We use 46 channels (excluding functional markers) suggested by Schürch to train StereoNiche for all issue sections across patients. After obtaining niche-level embeddings, we implement K-means clustering and set k=10 for the best ARI with Schürch's CN assignment. We then use cell type composition to annotate those niches since cell type identities have been validated by manual gating. Again, it is noted that StereoNiche is positioned as a de novo discovery method where cell-type information is never used in the learning process. We follow similar naming method via calculating enrichment score of certain cell types within a niche to define its identity, i.e., Smooth muscle, Tumor boundary, Lymphatics enriched, Vasculature enriched, Bulk tumor, Granulocytes enriched, Follicle, Macrophages enriched, T cells enriched, and Plasma/NK cells enriched (abbreviated by N1 to N10 in order). In general, StereoNiche's segmentation is consistent with Schürch's CNs assignment. StereoNiche identifies a novel N3 (lymphatics enriched) niche with highly expressed Podoplanin marker, while Schürch approach fails to distinguish lymphatics signal from neither vascularized smooth muscle (CN-7) nor smooth muscle (CN-8). StereoNiche also identifies N10 (Plasma and NK cells), representing non T-cell mediated immune response. Immune-infiltrated stroma (CN-3) annotated by Schürch contains significant amount of vasculature, T cells and NK cells, containing confused clues of angiogenesis process with immune infiltration, which might incur certain conflict with CN's definition as local stoichiometry. StereoNiche successfully corrects this ambiguous assignment and disentangles mixed signals into two separate niches (N4 and N10). Besides, CD3+ T cells only occur in N9 and N7 according to StereoNiche's segmentation results, showing more distinguishable and consistent pattern. These results demonstrate StereoNiche's utility to separate biologically relevant substructures either at finer granularity or with better biological consistency, probably attributed to absolute coordinate encoding and reduced interference from local density or geometry imbalance of certain cell-types. Enrichment heatmap of certain cell type in all niches is shown in Figure 5-a.

Then we assess the frequencies of niches differed between two patient groups (Supplementary Figure 6). CLR patients should have higher frequencies of B cells as defined. We recapitulate enriched N7 (Follicle enriched) in CLR patients, (p-value= 4.6e-5, U-test). Visual comparison of follicle niche identified by StereoNiche and Schürch (CN-5) for patient 11 (CLR) is shown in Figure 5-b and StereoNiche obtains TLS with more detailed substructure. N2 (Tumor boundary) appears surrounding follicle and N10 (Plasma and NK cell enriched) are scattered outside tumor boundary.

As an effective indicator of TLS, plasma cells are terminally differentiated effector B-cells and exert antibody-mediated anti-tumor immunity[62]. For Schürch's result, the sporadic tumor boundary and oversimplified T cells enriched region cannot fully explain the complexity of TLS. All Voronoi visualization results can be found in Supplementary Figure 7. TLS is a representative structure for CLR patients. StereoNiche's self-supervised configuration is expected to discover CLR's specific TLS feature and derive biologically-relevant object detection in an unsupervised manner. To test this hypothesis, we investigate the self-attention maps of [CLS] tokens from the last layer of ViT backbone (Methods). It is noted that StereoNiche is trained with no supervision and [CLS] token is not attached to any prior human knowledge. We plot normalized attention weights as well as their filtered values at median threshold for each cell. Interestingly, boundary and layout of all TLSs emerge automatically from one head of the last layer, coinciding with clustering results annotated by maker knowledge. Two representative visualizations (Patient 11 and Patient 17) are shown in Figure 5-c and all others are shown in Supplementary Figure 8. The self-attention maps automatically extract class-specific features for CLR patients, though no labels or supervision are used in the training process at all. By this means, StereoNiche effectively distills knowledge from position-aware self-attention module within ViT. We demonstrate the power of self-supervised contrastive learning as a data-driven method to discover advanced TME pattern, in particular, class-specific object detection and segmentation.

Schürch also observed overrepresented macrophages in DII patients over CLR patients yet reported non-significant CN-level difference (CN-4, macrophage enriched), exhibiting certain inconsistency. However, StereoNiche identifies higher frequencies of N8 (Macrophage enriched) in DII patients, (p-value=0.039, U-test). The better self-consistency of StereoNiche might result from less bias introduced through absolute position encoding scheme under structured context when aggregating patches related to macrophages. StereoNiche also recognizes more appearance of N5 in DII patients (p-value=0.001, U-test), probably representing diffusive tumor growth. Other niches exhibit conserved composition between CLR and DII. We further evaluate the differential enrichment of niche interactions across two patient groups. Interaction z-score is calculated to measuring spatial proximity between pairs of niches via permutation test. We first count the number of niche A located near niche B and define Z-score representing spatial enrichment of niche A close or far from niche B with background set by repeatedly randomizing locations of one niche to generate a null-distribution of A-B interactions. We apply this protocol across all patients and CLR/DII group separately (Supplementary Figure 9), observing two representative signals (Figure 5-d). N8 and N9 interacts with each other more frequently in DII patients (p-value=0.00025, Wilcoxon rank-sum test), coinciding with Schürch's conclusion that these two niches share immunological recruitment behaviors. Voronoi diagram for patient 15 illustrates frequent proximal co-occurrence of N9 (T cells enriched) and N8 (macrophages enriched) (Figure 5-b). However, Schürch fails to recapitulate this phenomenon visually, even though their tensor decomposition implementation uncovers more interdigitated pattern of corresponding CNs. Again, StereoNiche discards fixed number of neighbors, incorporates coordinate information directly and thereby avoids over-smoothing effects. In addition, N9-N5 pair shows significantly higher spatial contacts in CLR patients (p-value=7.2e-5, Wilcoxon rank-sum test), suggesting stronger lymphocytes-relevant immune infiltration and response. In short, we recommend decoupling cell type information and directly exploiting position information when evaluating niche behaviors.

We then investigate clinical utility of inferred TME elements at different levels, including their magnitude and impact direction on patient survival risk. Instead of using classical Cox proportional hazards (Cox-PH) model for survival analysis, we use a tree-based nonlinear accelerated failure time (AFT) model implemented in XGBoost (XGBoost-AFT) to account for distinct levels of censored events in CRC dataset. XGBoost-AFT allows absolute survival estimation with censored labels and its tree structure better matches complex non-linear pattern of TME than linear models. XGBoost-AFT maximizes log likelihood of input training data and the loss function contains two parts, covering both uncensored and censored patients (Methods). To equip XGBoost-AFT with interpretability, we apply game-theoretic Shapley additive explanation (SHAP)-based TreeExplainer to rank each input feature by its impact on model prediction of individual patient's survival time. It is noted that SHAP explainer leverages path coverage information from intrinsic tree structure of XGBoost and ensures that every single feature's contribution is quantified only after considering all possible ordered combinations of other features. SHAP implementation enables tractable runtime to derive accurate and consistent local explanations for tree-based models. 5-fold cross-validation is used for training and validation. SHAP-styled attribution decision plots are used to succinctly display a feature's effect, i.e., per-patient local explanation is colored by blue or red to indicate low or high feature value, respectively; location to the left (shorter survival) or right (longer survival) of zero indicates impact direction and magnitude refers to feature importance for that individual patient. Patients in the best validation splits are plotted to generate SHAP summary plots. Global feature importance is prioritized by mean absolute SHAP value of each input feature. Via calculating the directionality, feature value and attribution magnitude across all patients, we expect to prioritize several biologically meaningful TME elements as candidate prognostic biomarkers to stratify tumor patients. First, we use patient-level niche frequencies as model input. N7 (follicle) ranks first as the most attributable factor (Figure 5-e), which again validates TLS's representative role of driving effective anti-tumoral immune response and attendant better survival. Kaplan-Meier (K-M) curve (Figure 5-e) shows that follicle abundance significantly stratifies all patients (p-value=0.0064, LogRank test), recapitulating that CLR patients with TLSs have much longer overall survival than TLS-absent DII patients. We then focus on seeking effective prognostic markers to further stratify DII subgroup. Niche-level SHAP summary plot for DII patients can be found in Supplementary Figure 10. In particular, depicted by SHAP dependence plot, N3 (lymphatics enriched) shows its negative impact on survival prediction among DII patients (Figure 5-f), even though K-M analysis does not reach statistical significance (Supplementary Figure 11, p-value=0.38, LogRank test). Possible reasons might be the insufficient samples size (n=18) or distinct indicators (hazard in Cox-PH versus absolute survival in XGBoost-AFT). It is worthy of expanding DII cohort and investigating prognostic value of lymphangiogenesis, deciphering its potential role of driving lymphatic metastasis[63]. N8-N9 interaction is the key characteristic of DII patients[12]. SHAP dependence plot manifests a significant drop of survival when abundance of N8 (macrophages enriched niche) reaches high magnitude even though frequency of N9 (T cells enriched niche) is considerable (Figure 5-f). This nonmonotonicity of niche interaction effects represents complex TME behavior related to patient outcome. We then utilize the pairwise z-score as computed above to implement K-M analysis, observing that N8-N9 interaction strength can be recognized as an advanced biomarker to stratify DII patients (Figure 5-g, p-value=0.014, LogRank test). Besides niche-level biomarkers, we further evaluate whether there exist more specific combinatorial

candidates. We use certain cell type frequencies in all niches (N7 excluded) as model input for XGBoost-AFT. For all immune subtypes interacting with N9, dependence plot shows a step signal of CD68+ CD163+ macrophages in N9 (Figure 5-g), again demonstrating the immunosuppressive effects of M2-polarized macrophages interacting with T cells, leading to poorer survival[64]. SHAP summary plots of all combinations for DII patients can be found in Supplementary Figure 12.

To demonstrate the versatility of StereoNiche, we further apply it on a mass spectrometry-based spatial proteomics dataset acquired by MIBI-TOF[15]. 27 proteins (functional markers excluded) with corresponding coordinates from 41 TNBC patients are used as model input. StereoNiche obtains 8 niches annotated by canonical markers, i.e., B cells enriched, Macrophages enriched, Immune-infiltrated stroma, Bulk tumor, P53+ tumor enriched, Tumor boundary, Smooth muscle and T cells enriched (abbreviated by N1 to N8 in order). Among which, N6 (Tumor boundary) contains blended immune and tumor cells. To recapitulate global organization pattern of tumor-immune interactions, we calculate aforementioned spatial enrichment z-score between N6 and N4 (bulk tumor) to gauge the degree of mixing across all patients. Unlike Keren's approach to enumerate all paired markers to assess mixing score (defined as the number of immune-tumor interactions divided by the number of immune interactions), we probe interactions of two representative niches as a more concise metric (enrichment scores of all niche-niche interactions are shown in Supplementary Figure 14). As shown in Figure 5-i, mixed subtypes (n=19) have more frequent interactions between N6 and N4 than compartmentalized group (n=16) (p-value=0.031, Wilcoxon Rank sum test). Voronoi visualization of patient 12 and 4 (Figure 5-i) delineates interlaced boundary piece in mixed subtype while clear belt-like border structure forms in compartmentalized patients (Voronoi diagrams for all patients are shown in Supplementary Figure 13). StereoNiche also quantitatively assesses elevated proximity of N5 (P53+ tumor enriched) with N1 (B cells enriched) or N2 (Macrophages enriched) in contrast to other bulk tumor regions (N4) (p-value= 0.00064 and 3.1e-9, respectively, Wilcoxon Rank sum test) (Figure 5-j), highlighting p53+ enriched tumor sites recruit specific immune cells while non-p53 tumor regions are depleted of immune response[65].

Discussion

Emerging spatial resolved technologies has opened the door to profile hierarchical tissue architecture at omics level. Developing unsupervised computational methods to learn *in situ* from spatial signals is expected to derive novel mechanistic insights. Several current protocols underexploit spatial information. We propose a position-informed contrastive learning framework composed of two modules to fulfill distinct investigation goals for different spatial technologies, i.e., StereoCell for processing spatial transcriptomics (normally >100 features) to refine cell type assignment under adjacent spatial context and StereoNiche for analyzing spatial proteomics (with < 100 markers) to identify cell niches within organized tissue structure. Taking together of these conceptual and methodological contributions, Stereo offers general applicability for diverse spatial profiling technologies focusing on different investigation purposes.

In StereoCell, each cell learns both from itself and its immediate neighbors, which discriminates itself with all other cells not in spatial proximity. The final cell embeddings are learned by

maximizing agreement between each cell with both its augmented and neighboring view via a batch-wise contrastive loss in the latent space. In this way, each cell distills its context-aware identity from both its intrinsic information and surrounding environment which is aggregated by graph attention. StereoCell can operate on all genes to preserve genuine biological signal to the greatest extent, achieving the best quantitative performance to uncover clear layer-wise pattern for human DLPFC slice and mouse olfactory bulb. Through interrogating attention weights for each gene, StereoCell can extract layer-specific signatures at single-cell resolution in a clustering-free manner. Applied on a breast cancer sample, StereoCell recapitulates specific signals associated with invasive loci and in situ carcinoma in concordance with expert H&E annotations. StereoCell discards the needs of PCA or VAE-like operation and demonstrates dropout layer can effectively serve as a minimal augmentation strategy for omics data. StereoCell's contrastive configuration fits well with the high dimensional feature of spatial transcriptomic dataset, easily scalable to more than 10,000 features and decouples clustering operation. Current Bayesian or probabilistic models often rely on feature selection and require a pre-specified cluster number, which limits their compatibility and utility. In StereoNiche, we make an analogy between niche identification with semantic object segmentation in computer vision and investigate how different cells structurally organized to form distinct functional niches. StereoNiche learns directly from data with no cell type annotation as *a prior*. Unlike StereoCell's neighboring setting, incorporating absolute position information is important for niche discovery, considering the irregular geometry, different sizes and complex intercellular action distance among cells within certain spatial context. Therefore, StereoNiche is designed to directly model X-Y coordinates rather than simplified neighbor attribution information as in StereoCell or other popular methods. We harness Vision Transformer to extract learnable features from paired molecular and positional information in a joint manner, rendering flexibility to model both proximal and distant cell-cell interactions. Those position-aware attention operations automatically relate cell location with strength of intercellular interactions. In addition, StereoNiche utilizes a negative queue composed of abundant negative samples to enable more efficient contrastive learning unrestricted by batch-size selection. StereoNiche successfully identifies differential abundant macrophage niches and enriched T-macrophage interactions for DII patients, which otherwise missed by local-window neighborhood clustering. Without using any supervision signals, StereoNiche enables extracting TLS presence with clear boundary as a representative characteristic for CLR patients. We expect StereoNiche can set an example how to leverage self-supervision techniques to automatically extract class-specific properties in a self-distillation manner and facilitate discovering novel higher-order TME elements with minimum prior assumptions. Hierarchical interactions at cellular or niche-level shape the heterogeneous pattern of tumor microenvironment of cancer patients. We also demonstrate how to interpret clinical utility of these elements via a SHAP explainer for a tree-based survival prediction. Again, those outcome-relevant and clinically attributable features further validate the quality of segmented niches by StereoNiche.

We acknowledge several limitations. StereoCell and StereoNiche do not incorporate learnable cell-cell communication signals and intra-cellular pathways in an explicit fashion. Deconvolving hierarchical intercellular communications[34] in a context-aware manner is critical to define cell states under spatial constraint, extract state-dependent ligand-receptors' interactions within tissue niches and decipher key signaling pathways. Empowered by these intermediate features, we can expect better characterization of perturbational effects under pathological conditions. In the future, we aim

to upgrade Stereo with learnable intermediate factors to obtain biologically meaningful embeddings and automatically extract context-specific interactions as testable hypothesis. We foresee powerful computational methods developed for spatial omics technologies will facilitate discovery of advanced spatial biomarkers for better patient grouping and ultimately benefit clinical translation.

Reference

1. Longo, S. K., Guo, M. G., Ji, A. L., & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, **22**, 627–644 (2021).
2. Marx, V. Method of the Year: Spatially resolved transcriptomics. *Nature Methods*, **18**, 9–14 (2021).
3. Lewis, S. M. *et al.* Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods*, **18**, 997–1012 (2021).
4. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, **362**, eaau5324 (2018).
5. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, **568**, 235–239 (2019).
6. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, **361**, eaat5691 (2018).
7. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, **10**, 442–458 (2015).
8. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82 (2016).
9. Chen, A. *et al.* Large field of view-spatially resolved transcriptomics at nanoscale resolution. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.17.427004v2 (2021).
10. Lin, J.-R., Fallahi-Sichani, M., & Sorger, P. K. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature Communications*, **6**, 8390 (2015).
11. Goltsev, Y. *et al*. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*, **174**, 968-981.e15 (2018).
12. Schürch, C. M. *et al.* Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*, **182**, 1341-1359.e19 (2020).
13. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, **11**, 417–422 (2014).
14. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, **20**, 436–442 (2014).
15. Keren, L. *et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, **174**, 1373-1387.e19 (2018).
16. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., & Heyn, H. SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, **49**, e50 (2021).
17. Kleshchevnikov, V. *et al.* Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. Preprint at https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1 (2020).
18. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 1–10 (2021).
19. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, **14**, 935–936 (2017).

20. Zhu, Q., Shah, S., Dries, R., Cai, L., & Yuan, G.-C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature Biotechnology*, **36**, 1183–1190 (2018).
21. Dries, R. *et al.* Giotto: A toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, **22**, 78 (2021).
22. Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology* 1–10 (2021).
23. Furman, S. A. *et al.* Unsupervised cellular phenotypic hierarchy enables spatial intratumor heterogeneity characterization, recurrence-associated microdomains discovery, and harnesses network biology from hyperplexed in-situ fluorescence images of colorectal carcinoma. Preprint at https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3 (2020).
24. Chidester, B., Zhou, T., & Ma, J. SPICEMIX: Integrative single-cell spatial modeling for inferring cell identity. Preprint at https://www.biorxiv.org/content/10.1101/2020.11.29.383067 v2 (2020).
25. 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* **18,** 1342–1351 (2021).
26. Fu, H. *et al.* Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics. Preprint at https://www.biorxiv.org/content/10.1101/2021.06.15.448542v2 (2021).
27. Edsgärd, D., Johnsson, P., & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nature Methods*, **15**, 339–342 (2018).
28. Svensson, V., Teichmann, S. A., & Stegle, O. SpatialDE: Identification of spatially variable genes. *Nature Methods*, **15**, 343–346 (2018).
29. Sun, S., Zhu, J., & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, **17**, 193–200 (2020).
30. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature*, **578**, 615–620 (2020).
31. Uttam, S. *et al.* Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks. *Nature Communications*, **11**, 1-14 (2020).
32. Park, J. *et al.* Cell segmentation-free inference of cell types from in situ transcriptomics data. *Nature Communications*, **12**, 3545 (2021).
33. Browaeys, R., Saelens, W., & Saeys, Y. NicheNet: Modeling intercellular communication by linking ligands to target genes. *Nature Methods*, **17**, 159–162 (2020).
34. David S. Fischer, Anna C. Schaar, Fabian J. Theis, Learning cell communication from spatial graphs of cells. Preprint at https://www.biorxiv.org/content/10.1101/2021.07.11.451750v1.full (2021).
35. Schapiro, D. *et al.* histoCAT: Analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, **14**, 873–876 (2017).
36. Moehlin, J., Mollet, B., Colombo, B. M., & Mendoza-Parra, M. A. Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer. *Cell Systems*, **12**, 694-705.e3 (2021).
37. Lundberg, S.M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).
38. Barnwal, A., Cho, H., & Hocking, T. D. Survival regression with accelerated failure time model

in XGBoost. Preprint at https://arxiv.org/abs/2006.04920 (2021).
39. Veličković P. *et al.* Graph Attention Networks. In *International Conference on Learning Representations*, (2018).
40. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, **24**, 425–436 (2021).
41. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning,* 1597–1607 (2020).
42. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15**, 1929-1958 (2014).
43. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://arxiv.org/abs/2010.11929 (2021).
44. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729-9738 (2020).
45. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902.e21 (2019).
46. Wolf, F. A., Angerer, P., & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, **19**, 15 (2018).
47. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, **184**, 3222-3241.e26 (2021).
48. Camargo, N. *et al.* Oligodendroglia myelination requires astrocyte-derived lipids. *PLoS Biology,* **15**, e1002605 (2017).
49. Sunkin, S. M. *et al*. Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, **41**, D996–D1008 (2013).
50. Alessandri, L. *et al.* Sparsely-connected autoencoder (SCA) for single cell RNAseq data mining. *Npj Systems Biology and Applications,* **7**, 1–10 (2021).
51. Bagaev, A. *et al.* Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell*, **39**, 845-865.e7 (2021).
52. Giussani, M., Merlino, G., Cappelletti, V., Tagliabue, E., & Daidone, M. G. Tumor-extracellular matrix interactions: Identification of tools associated with breast cancer progression. *Semin Cancer Biol*, **35,** 3-10 (2015).
53. Kai F, Drain AP, Weaver VM. The Extracellular Matrix Modulates the Metastatic Journey. *Dev Cell*. **49**, 332-346 (2019).
54. Espina, V., Liotta, L. What is the malignant nature of human ductal carcinoma in situ. *Nat Rev Cancer* **11**, 68–75 (2011).
55. Ayuso, J. M. *et al.* Organotypic microfluidic breast cancer model reveals starvation-induced spatial-temporal metabolic adaptations. *EBioMedicine.* **37**, 144-157 (2018).
56. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science.* **357**, eaan2507 (2017).
57. Sjöberg, E., Augsten, M., Bergh, J., Jirström, K., & Östman, A. Expression of the chemokine CXCL14 in the tumour stroma is an independent marker of survival in breast cancer. *Br J Cancer.* **114,** 1117-24 (2016).

58. Kim, J. *et al*. Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nat Genet* **50**, 1705–1715 (2018).
59. Hoshino, D. *et al*. Network analysis of the focal adhesion to invadopodia transition identifies a PI3K-PKCα invasive signaling axis. *Sci Signal.* **5**, ra66 (2012).
60. Pereira, B. A. *et al*. CAF subpopulations: a new reservoir of stromal targets in pancreatic cancer. *Trends in cancer*, **5**, 724-741 (2019).
61. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L., & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nature Methods*, **13**, 493–496 (2016).
62. Germain, C., Gnjatic, S., & Dieu-Nosjean, M. C. Tertiary lymphoid structure-associated B cells are key players in anti-tumor immunity. *Frontiers in immunology*, **6**, 67 (2015).
63. Sundlisæter, E. *et al*. Lymphangiogenesis in colorectal cancer—prognostic and therapeutic aspects. *International journal of cancer*, **121,** 1401-1409 (2007).
64. Mantovani, A., Marchesi, F., Malesci, A. et al. Tumor-associated macrophages as treatment targets in oncology. *Nat Rev Clin Oncol*, **14**, 399–416 (2017).
65. Di Minin, G. *et al*. Mutant p53 reprograms TNF signaling in cancer cells through interaction with the tumor suppressor DAB2IP. *Molecular cell*, **56**, 617-629 (2014).

Figure Legends

Figure 1 Overview of Stereo architecture. (a) StereoCell accepts gene-count matrix with X/Y coordinate as input. StereoCell contains three modules, i.e., student, teacher, and neighbor module. Student module uses discrete input while teacher module aggregates gene embeddings into cell representations. Teacher module incorporates adjacent cell neighbors through graph attention networks to learn spatial context representations. Contrastive learning is implemented via pulling together student-teacher view and teacher-neighbor view for each cell with adjustable ratio in loss function. Neighbor embeddings are used as final output for downstream usage. **(b)** StereoNiche leverages 100×100 medium patches, each consisting of 10×10 small patches, to slide across the whole tissue section and feeds small patches with positional encodings into Vision Transformer network to extract patch-level embeddings. [CLS] tokens represent aggregated embeddings of medium patches. Contrastive learning is implemented via pulling together augmented views for each patch. Then learned embeddings can be clustered to identify cell niches. **(c)** XGBoost-AFT model is trained to relate niche-level composition or cell-type frequency in certain niche with patient survival. SHAP-based TreeExplainer offers local explanations via feature attributions. Advanced spatial biomarkers can be prioritized to stratify patients. SHAP, Shapley additive explanations.

Figure 2 StereoCell achieves excellent performance for cell type inference on both synthetic data and real transcriptomics dataset acquired by STARmap. (a) Simulated FISH-based RNA profiles for primary mouse visual cortex, containing 3 major cell types, subdivided into 8 subtypes at 4 layers. E, excitatory neurons; Inh, inhibitory neurons; Glial, glial cells. Colors distinguish different subtypes. **(b)** In situ assignment results for excitatory neurons at 4 layers by each method. **(c)** UMAP visualization of raw expression profiles. **(d)** Learned StereoCell embeddings in the latent space (discarding neighbor module). **(e)** Learned latent embeddings by StereoCell, layer-specific excitatory neurons are clearly separated and highlighted by dashed circles. **(f)** Boxplot to compare clustering results obtained by each method, measured by adjusted Rand Index (ARI) against simulated ground truth labels. Median, upper, lower quartiles with highest and lowest values are plotted. The whiskers extend to points that lie within 1.5 Interquartile range (IQRs) of the lower and upper quartile (n=8 samples). **(g)** In situ visual comparison of original STARmap assignments (top) with annotated StereoCell clusters (bottom). Deeper tissue layers are on the right. **(i)** Confusion matrix illustrates agreement of original STARmap cell-types with StereoCell clusters. Circle size stands for matched count number. **(j)** Boxplot of Moran's *I* values for spatial variable genes (SVGs) detected by StereoCell, SpatialDE and SPARK. **(k)** Venn diagram depicts overlapping SVGs detected by three methods. **(h)** Expression heatmap for 17 StereoCell clusters. **(l)** Example spatial expression patterns of 4 signature genes (Slc17a7, Mbp, Gad2, Cux2) identified by StereoCell cluster.

Figure 3 StereoCell identifies clear layer-wise structure in brain tissues for both low and high resolution spatial transcriptomics datasets. (a) Boxplot summary plot of clustering accuracy measured by ARI for 12 DLPFC slices acquired by 10X Visium technology. ARI is calculated by comparing clustering assignments by each method against annotated layer labels. In the boxplot, the center line and box limits denote the median and upper/lower quartiles, respectively. and 1.5× interquartile range are displayed as whiskers. **(b)** Ground truth of manually annotated 6 layers and

white matter (WM) in sample 151673 from DLPFC datasets. **(c)** UMAP visualization of raw data profiles (top) and StereoCell-derived embeddings (bottom) for sample 151673. **(d)** Expression heatmap of 7 makers genes for 6 layers and WM (top) compared with normalized attention weights attracted from StereoCell's neighbor module (bottom) for slice 151673. **(e)** In situ cluster assignment generated by 5 methods for slice 151673 and corresponding ARIs. **(f)** Comparison of clustering consistency by StereoCell and SEDR for Stereo-seq profiled MOB dataset. StereoCell implementation has 2 scenarios, using 10,000 or 2,000 highly variable genes (HVG), respectively. kSIM and ASW are evaluation metrics. **(g)** In situ clustering visualization results of MOB's laminar organization, obtained by 2 StereoCell scenarios and SEDR. DLPFC, dorsolateral prefrontal cortex; MOB, mouse olfactory bulb; kSIM, k-nearest neighbor similarity; ASW, Average silhouette width.

Figure 4 StereoCell discovers detailed tumor microenvironment (TME) heterogeneity for 10X Visium spatial transcriptomics profiled breast cancer slices. (a) Manual annotations by expert pathologists (IC, invasive carcinoma; DCIS, ductal carcinoma in situ). referred to Lewis *et.al.* **(b)** Visual comparison of unsupervised grouping results by 4 methods. 20 cluster labels are applied for all methods. From left to right, StereoCell, SEDR, BayesSpace and Giotto. **(c)** Spatial patterns of gene expression of 5 markers (COX6C, CXCL4, CRISP3, CPB1, MALAT) and 3 TME functional signatures (Proliferation_rate, T_reg_trafficking and Cancer-associated fibroblasts (CAF)). **(d)** Expression heatmap of functional gene expression signatures (FGSE) for 20 StereoCell clusters. **(e)** KEGG pathway enrichment analysis comparing cluster 6 with cluster 2 (StereoCell clusters). Red circles represent pathways upregulated in cluster 2. **(f)** KEGG pathway enrichment analysis comparing cluster 12 versus cluster 10. Red circles represent pathways upregulated in cluster 10 (StereoCell clusters). (e-f) Circle size represents gene counts belonging to certain pathway.

Figure 5 StereoNiche identifies functional cell niches for CRC and TNBC patients acquired by different spatial proteomics technologies. (a) Identification of 10 distinct cell niches directly based on spatial protein readouts by StereoNiche on 35 CRC patients. Heatmap shows cell-types' enrichment score within each niche. **(b)** Comparison of Voronoi diagrams of CNs from Schürch (left) and cell niches identifies by StereoNiche (right) for representative CLR (patient 11) and DII (patient 15) patients. **(c)** Normalized attention weights (left) and filtered values (right) at median threshold from the 6^{th} head at the last layer of ViT network automatically extract CLR-specific follicle structure for 2 representative patients with no supervision (top, patient 11; bottom, patient 17). **(d)** Differential paired niche interactions between CLR and DII group. Top, N8 (macrophage enriched) interacting with N9 (T cells enriched); bottom, N9 interacting with N5 (Bulk tumor). (***$p < 0.001$, Wilcoxon rank-sum test). **(e)** Top, SHAP summary plots for individual's survival attributed to niche frequencies across all patients. Each dot represents a patient in the best validation splits of 5-fold cross validation. Color corresponds to low (blue) vs. high (red) feature value, sign of SHAP value indicates direction to how the feature impacts survival likelihood (left, poorer survival vs right, better survival) and magnitude refers to feature importance. Bottom, Kaplan-Meier (K-M) curve for N7 (Follicle) frequency to stratify all patients into low and high risk group. P-value determined by log-rank test. **(f)** Top, SHAP dependence plot of N3 (Lymphatics enriched) for DII patients. Bottom, SHAP dependence plot of N9 for DII patients under the context of different N8 frequencies. **(g)** Top, SHAP dependence plot of CD68+ CD163+ macrophages in N9 for DII patients. Bottom, K-M curve for N8-N9 interaction enrichment z-score to stratify DII patients. P-value

determined by log-rank test. **(h)** Marker expression heatmap in 10 cell niches identified by StereoNiche for 41 TNBC patients. **(i)** Voronoi diagrams for two representative patients with compartmentalized pattern (top, patient 4) and mixed pattern (bottom, patient 12). Right, differential N6-N4 interactions between two groups. (*$p < 0.01$, Wilcoxon rank-sum test). N6, tumor boundary; N4, bulk tumor. **(j)** Differential immune response behaviors between N5 (P53+ tumor enriched) against N4 illustrated by their interactions with left, N2(macrophages enriched) and right, N1 (B cells enriched) (*$p < 0.05$, Wilcoxon rank-sum test). CRC, colorectal cancer; TNBC, triple-negative breast cancer; CLR, Crohn's-like reaction; DII, diffuse inflammatory infiltration.

Figure 1

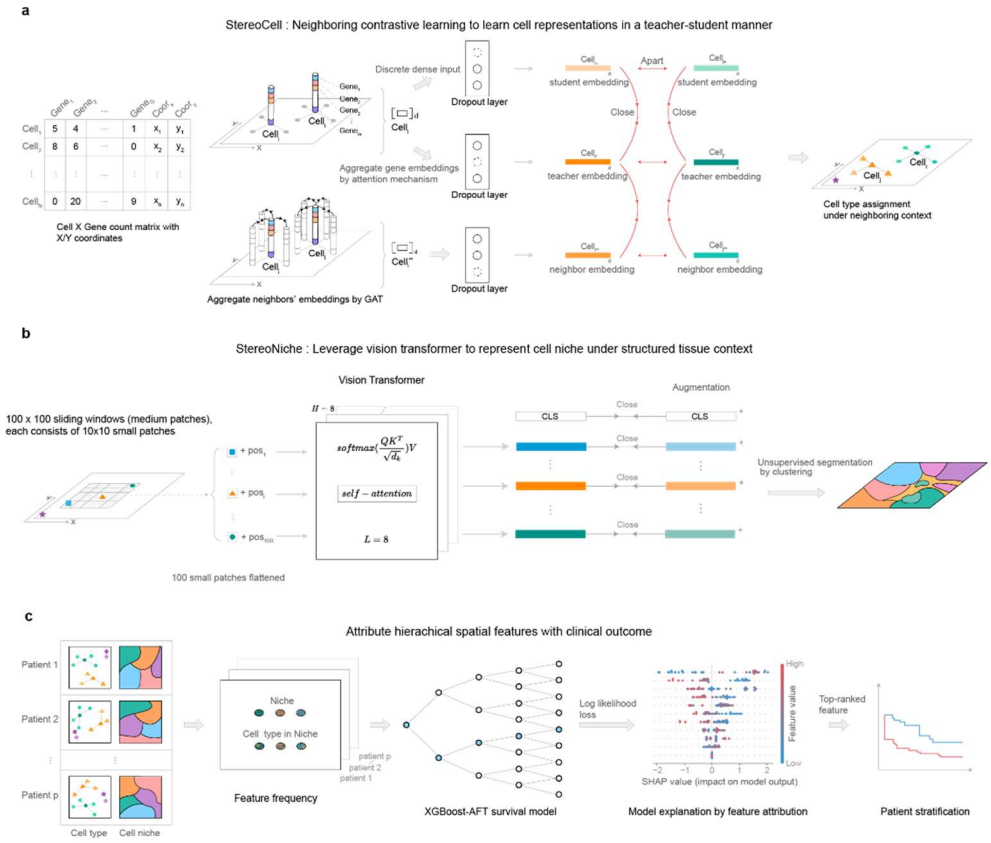

a
StereoCell : Neighboring contrastive learning to learn cell representations in a teacher-student manner
Cell X Gene count matrix with X/Y coordinates
Discrete dense input
Dropout layer
Aggregate gene embeddings by attention mechanism
Aggregate neighbors' embeddings by GAT
student embedding
teacher embedding
neighbor embedding
Apart
Close
Cell type assignment under neighboring context
b
StereoNiche : Leverage vision transformer to represent cell niche under structured tissue context
Vision Transformer
Augmentation
100 x 100 sliding windows (medium patches), each consists of 10x10 small patches
100 small patches flattened
CLS
Unsupervised segmentation by clustering
c
Attribute hierachical spatial features with clinical outcome
Patient 1
Patient 2
Patient p
Cell type
Cell niche
Niche
Cell type in Niche
Feature frequency
XGBoost-AFT survival model
Log likelihood loss
SHAP value (impact on model output)
Feature value
High
Low
Model explanation by feature attribution
Top-ranked feature
Patient stratification

Figure 2

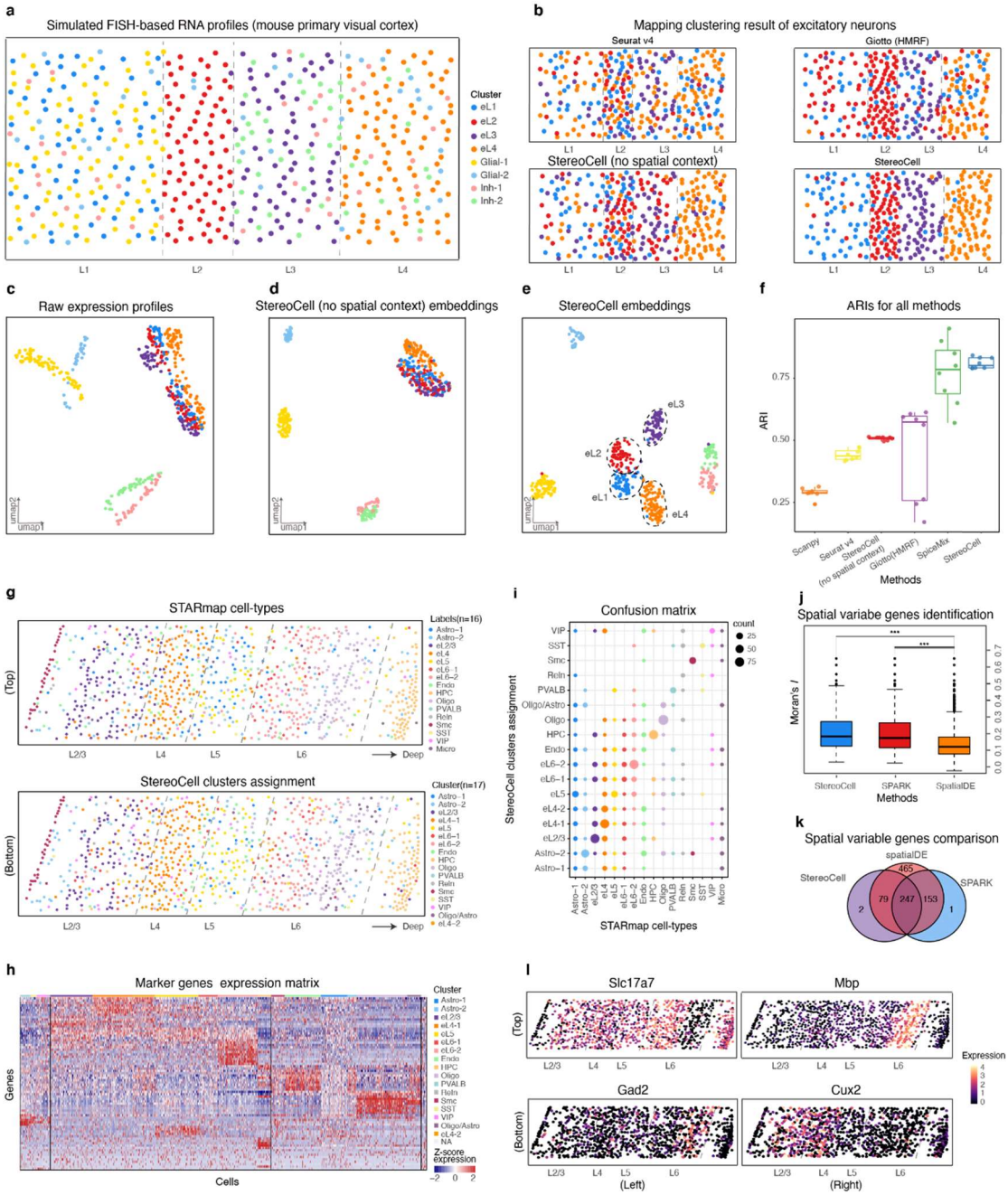

a
Simulated FISH-based RNA profiles (mouse primary visual cortex)
Cluster
eL1
eL2
eL3
eL4
Glial-1
Glial-2
Inh-1
Inh-2
L1
L2
L3
L4
b
Mapping clustering result of excitatory neurons
Seurat v4
Giotto (HMRF)
StereoCell (no spatial context)
StereoCell
c
Raw expression profiles
d
StereoCell (no spatial context) embeddings
e
StereoCell embeddings
f
ARIs for all methods
ARI
Scanpy
Seurat v4
StereoCell (no spatial context)
Giotto(HMRF)
SpiceMix
StereoCell
Methods
g
STARmap cell-types
Labels(n=16)
(Top)
Deep
StereoCell clusters assignment
Cluster(n=17)
(Bottom)
i
Confusion matrix
count
StereoCell clusters assignment
STARmap cell-types
j
Spatial variabe genes identification
Moran's I
StereoCell
SPARK
SpatialDE
Methods
k
Spatial variable genes comparison
spatialDE
StereoCell
SPARK
465
79
247
153
2
1
h
Marker genes expression matrix
Cluster
Genes
Cells
Z-score expression
l
Slc17a7
Mbp
Gad2
Cux2
(Top)
(Bottom)
Expression
(Left)
(Right)

Figure 3

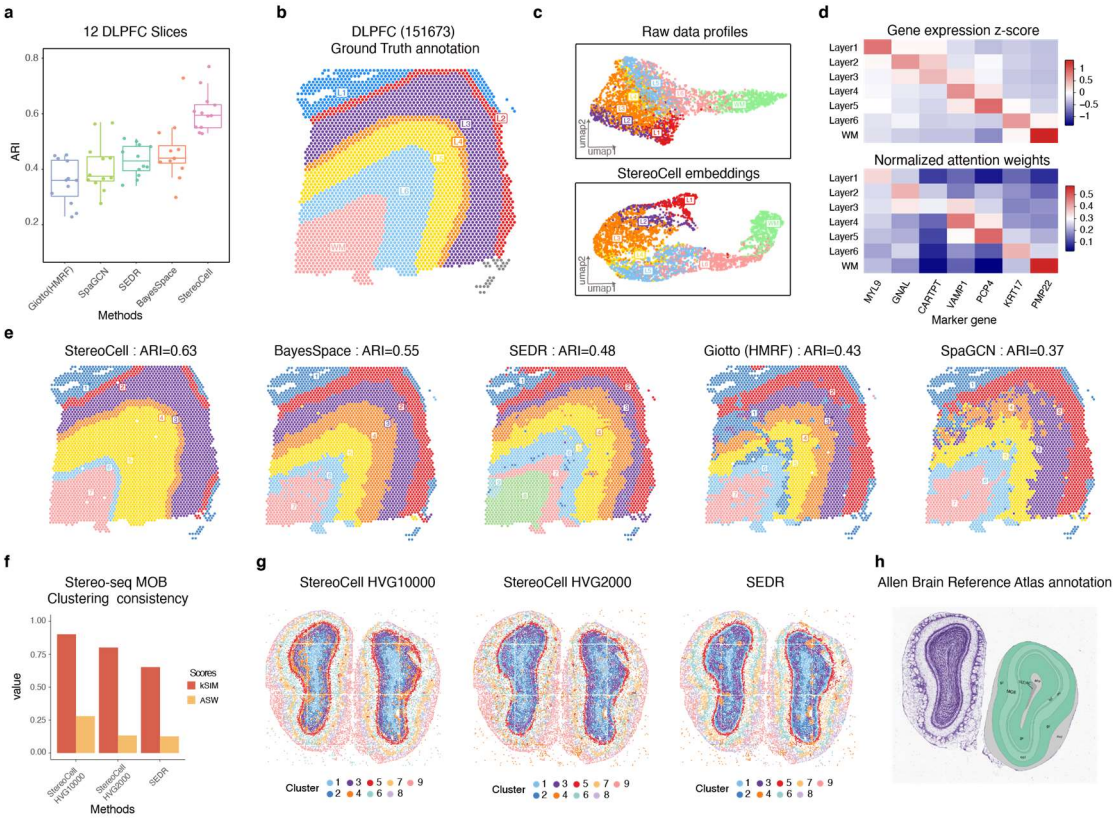

a
12 DLPFC Slices
ARI
Methods
Giotto(HMRF)
SpaGCN
SEDR
BayesSpace
StereoCell
b
DLPFC (151673)
Ground Truth annotation
c
Raw data profiles
StereoCell embeddings
umap1
umap2
d
Gene expression z-score
Normalized attention weights
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
Marker gene
MYL9
GNAL
CARTPT
VAMP1
PCP4
KRT17
PMP22
e
StereoCell : ARI=0.63
BayesSpace : ARI=0.55
SEDR : ARI=0.48
Giotto (HMRF) : ARI=0.43
SpaGCN : ARI=0.37
f
Stereo-seq MOB
Clustering consistency
value
Scores
kSIM
ASW
Methods
g
StereoCell HVG10000
StereoCell HVG2000
SEDR
Cluster
h
Allen Brain Reference Atlas annotation

Figure 4

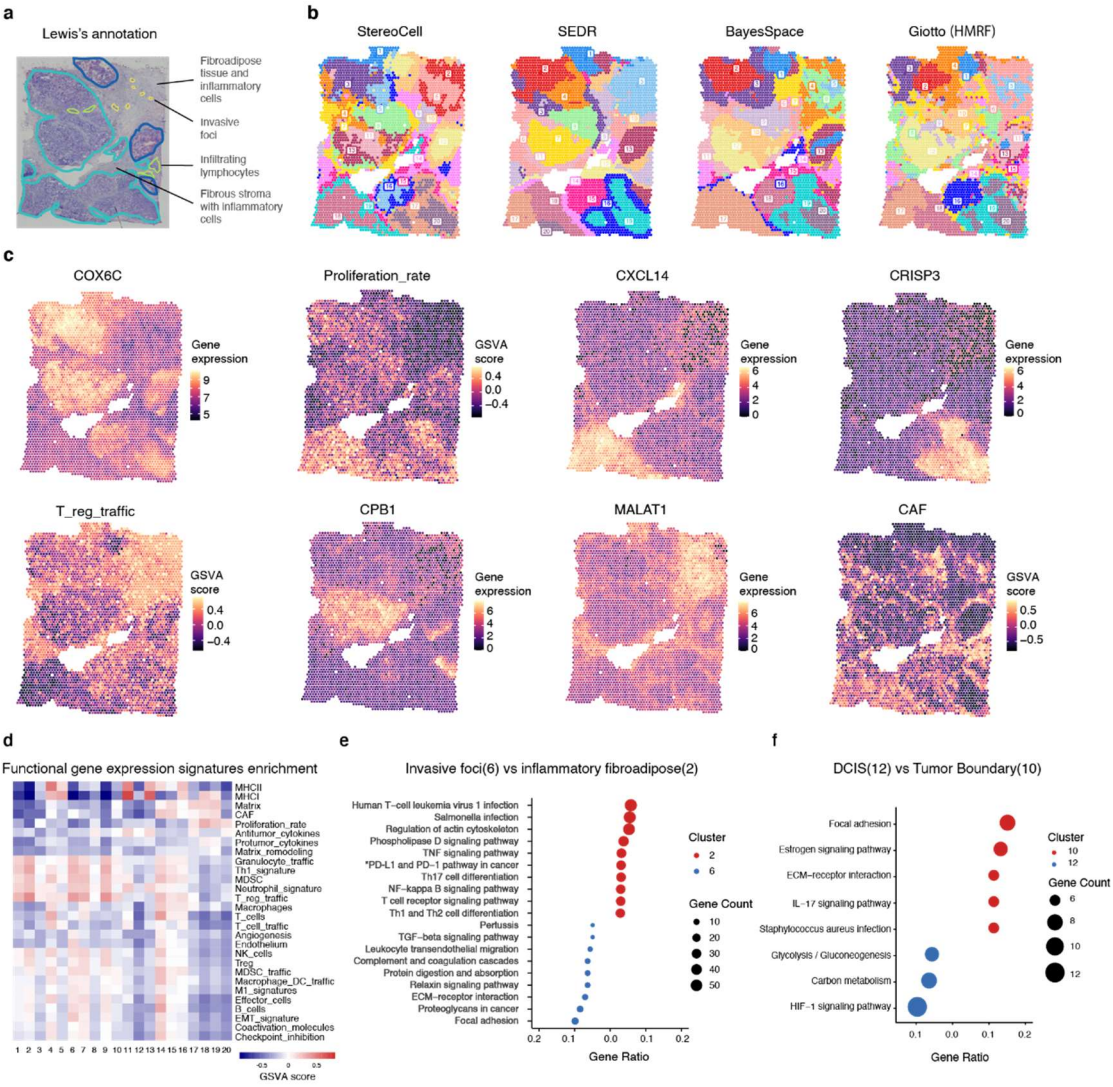
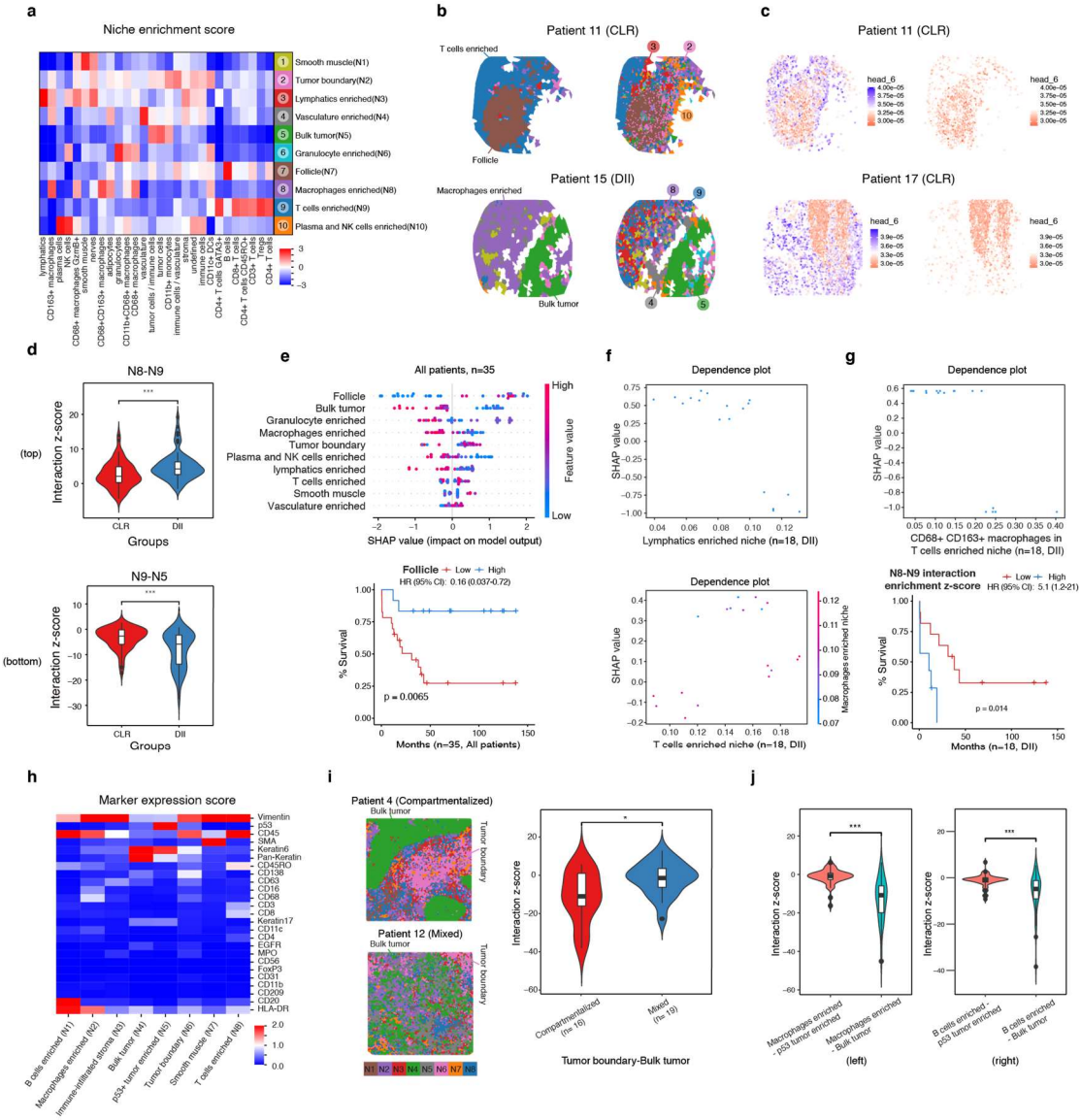

a
Lewis's annotation
Fibroadipose tissue and inflammatory cells
Invasive foci
Infiltrating lymphocytes
Fibrous stroma with inflammatory cells
b
StereoCell
SEDR
BayesSpace
Giotto (HMRF)
c
COX6C
Proliferation_rate
CXCL14
CRISP3
Gene expression
GSVA score
T_reg_traffic
CPB1
MALAT1
CAF
d
Functional gene expression signatures enrichment
MHCII
MHCI
Matrix
CAF
Proliferation_rate
Antitumor_cytokines
Protumor_cytokines
Matrix_remodeling
Granulocyte_traffic
Th1_signature
MDSC
Neutrophil_signature
T_reg_traffic
Macrophages
T_cells
T_cell_traffic
Angiogenesis
Endothelium
NK_cells
Treg
MDSC_traffic
Macrophage_DC_traffic
M1_signatures
Effector_cells
B_cells
EMT_signature
Coactivation_molecules
Checkpoint_inhibition
GSVA score
e
Invasive foci(6) vs inflammatory fibroadipose(2)
Human T-cell leukemia virus 1 infection
Salmonella infection
Regulation of actin cytoskeleton
Phospholipase D signaling pathway
TNF signaling pathway
*PD-L1 and PD-1 pathway in cancer
Th17 cell differentiation
NF-kappa B signaling pathway
T cell receptor signaling pathway
Th1 and Th2 cell differentiation
Pertussis
TGF-beta signaling pathway
Leukocyte transendothelial migration
Complement and coagulation cascades
Protein digestion and absorption
Relaxin signaling pathway
ECM-receptor interaction
Proteoglycans in cancer
Focal adhesion
Cluster
Gene Count
Gene Ratio
f
DCIS(12) vs Tumor Boundary(10)
Focal adhesion
Estrogen signaling pathway
ECM-receptor interaction
IL-17 signaling pathway
Staphylococcus aureus infection
Glycolysis / Gluconeogenesis
Carbon metabolism
HIF-1 signaling pathway
Cluster
Gene Count
Gene Ratio

Figure 5


a
Niche enrichment score
Smooth muscle(N1)
Tumor boundary(N2)
Lymphatics enriched(N3)
Vasculature enriched(N4)
Bulk tumor(N5)
Granulocyte enriched(N6)
Follicle(N7)
Macrophages enriched(N8)
T cells enriched(N9)
Plasma and NK cells enriched(N10)
b
Patient 11 (CLR)
T cells enriched
Follicle
Patient 15 (DII)
Macrophages enriched
Bulk tumor
c
Patient 11 (CLR)
Patient 17 (CLR)
head_6
d
N8-N9
Interaction z-score
(top)
CLR
DII
Groups
N9-N5
(bottom)
e
All patients, n=35
Follicle
Bulk tumor
Granulocyte enriched
Macrophages enriched
Tumor boundary
Plasma and NK cells enriched
lymphatics enriched
T cells enriched
Smooth muscle
Vasculature enriched
High
Feature value
Low
SHAP value (impact on model output)
Follicle Low High
HR (95% CI): 0.16 (0.037-0.72)
p = 0.0065
% Survival
Months (n=35, All patients)
f
Dependence plot
SHAP value
Lymphatics enriched niche (n=18, DII)
Dependence plot
SHAP value
T cells enriched niche (n=18, DII)
Macrophages enriched niche
g
Dependence plot
SHAP value
CD68+ CD163+ macrophages in T cells enriched niche (n=18, DII)
N8-N9 interaction enrichment z-score
Low High
HR (95% CI): 5.1 (1.2-21)
p = 0.014
% Survival
Months (n=18, DII)
h
Marker expression score
Vimentin
p53
CD45
SMA
Keratin6
Pan-Keratin
CD45RO
CD138
CD63
CD16
CD68
CD3
CD8
Keratin17
CD11c
CD4
EGFR
MPO
CD56
FoxP3
CD31
CD11b
CD209
CD20
HLA-DR
B cells enriched (N1)
Macrophages enriched (N2)
immune-infiltrated stroma (N3)
Bulk tumor (N4)
p53+ tumor enriched (N5)
Tumor boundary (N6)
Smooth muscle (N7)
T cells enriched (N8)
i
Patient 4 (Compartmentalized)
Bulk tumor
Tumor boundary
Patient 12 (Mixed)
Bulk tumor
Tumor boundary
N1 N2 N3 N4 N5 N6 N7 N8
Interaction z-score
Compartmentalized (n= 16)
Mixed (n= 19)
Tumor boundary-Bulk tumor
j
Interaction z-score
Macrophages enriched - p53 tumor enriched
Macrophages enriched - Bulk tumor
(left)
B cells enriched - p53 tumor enriched
B cells enriched - Bulk tumor
(right)

Methods

Stereo contains two parts, i.e., StereoCell for cell-type inference and clustering on spatial transcriptomics dataset and StereoNiche for cell niche discovery on spatial proteomics dataset. StereoCell uses both simulated and real spatial transcriptomics datasets for validation. StereoNiche is applied on published spatial proteomics datasets. Data sources with feature descriptions and corresponding download links can be found in Supplementary Table 1. Stereo is written in python package (https://github.com/melobio/Stereo). Here we describe key components of Stereo, data pre-processing as well as implementation protocols of other benchmarking tools.

Filtering, Pre-processing, and Normalization

StereoCell supports operating on all genes. For 'all genes' scenario over simulated data and STARmap dataset, StereoCell keeps the same number of input genes within a batch. For DLPFC, MOB and 10X BC dataset, StereoCell uses SCANPY () function to normalize each cell to 10,000 read counts, apply logarithmic transformation with 'log (count + 1)' operation and conduct HVGs selection at different levels for benchmark purpose.

StereoNiche is applied on CRC and TNBC datasets. Protein marker readouts ($\mathrm{protein}_1$, $\mathrm{protein}_2$, ..., $\mathrm{protein}_e$) will be standardized by z-score transformation and combined with coordinates information (x, y) to construct spatial protein expression matrix (x, y, e).

Input encoding scheme

Stereo leverages TF-record file format to encode normalized gene-count matrix. TF-record is a convenient way for sharding file in TensorFlow. A TF-record is a binary file that contains sequences of byte-strings. Data is serialized (encoded as a byte-string) before being written into a TF-record. StereoCell contains three modules, student, teacher, and neighbor module. StereoCell encapsulates 'gene index' and 'count value' into TF-record file. Student module only reads 'count value' file, teacher module accepts both 'gene index' and 'count value' as TF-record files, and neighbor module reads 'gene index' and 'neighbor count value'. StereoNiche reads TF-record file together with coordinate information and obtains tensor (d, x, y) as model input.

StereoCell Framework

StereoCell contains three modules, student, teacher, and neighbor module. Details as below.

Teacher module

Teacher module accepts $\mathrm{X}_{\mathrm{index}} \in \mathbb{R}^{\mathrm{N}\times\mathrm{G}}$ and $X_{counts} \in \mathbb{R}^{N\times G}$, where N denotes number of cells and G denotes number of genes, X_{index} represents gene index, index refers to gene ID defined by a dictionary consisted of all genes of a certain species. X_{counts} represents gene counts value. Each gene within a cell is represented by $g_i(i \in \mathbb{R}^G)$, and count of certain gene is represented by $v_i(i \in \mathbb{R}^G)$. Firstly, g_i will be embedded into a d-dimension vector space as $e_{it} \in \mathbb{R}^{G\times d}(1)$. Cross product of e_i and v_i outputs scaled hidden vector $h_{it} \in \mathbb{R}^{G\times d}$ (2).

$$e_{it} = Embedding(g_i) \qquad i \in \mathbb{R}^G \quad t \in \mathbb{R}^d \qquad (1)$$

$$h_{it} = e_{it} \times v_i \qquad h_{it} \in \mathbb{R}^{G\times d} \qquad (2)$$

StereoCell then uses attention mechanism to aggregate gene embeddings for each cell. h_{it} is firstly fed into a one Multilayer Perception (MLP) hidden layer followed by non-linear $tanh$ transformation to obtain hidden vector $u_{it} \in \mathbb{R}^{G \times d}$(3). Then a cellular context vector u_w will be dot product with u_{it} with $softmax$ operation to obtain attention weights $\alpha_i \in \mathbb{R}^G$ (4); then aggregation is implemented on all genes' h_{it} through weighted summation by attention weights α_i, obtaining aggregated vectors, $s_t \in \mathbb{R}^d$(5). MLP stands for multi-layer perceptron.

$$u_{it} = \tanh(W_1 h_{it} + b_1) \quad (3)$$

$$\alpha_i = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \quad (4)$$

$$s_t = \sum_i(\alpha_i \times h_{it}) \quad (5)$$

Attention layer output will be fed into a Dropout layer, and then Dense layer with ReLU activation, leading to final embeddings from teacher module, $z_t \in \mathbb{R}^d$ (6)

$$Z_t = \text{ReLU}(W_2 s_t + b_2) \quad (6)$$

Student module

Student module accepts only $X_{counts} \in \mathbb{R}^{N \times G}$, followed by Dropout layer, and then Dense layer with ReLU activation, leading to the output embeddings from student module, $Z_s \in \mathbb{R}^d$.

Neighbor module

StereoCell introduces neighbor module to incorporate information from adjacent cells. Based on graph attention (GAT)[39] backbone, StereoCell uses attention mechanism to extract cell neighbors' features. Neighbor definition for each cell is based on input 2D (X-Y) coordinates where Euclidean distance matrix between cell n with all other cells is calculated, then obtaining each cell's neighbor indexes (The maximum number of cell neighbors is set to 8). Each cell n will have m neighbors (m_n), these $(m_n + 1)$ cells are fed into teacher module, generating corresponding teacher embeddings, and stacked into $h_{m_n+1} \in \mathbb{R}^{(m_n+1) \times d}$ for cell n, followed by MLP linear transformation with non-linear $tanh$ activation to obtain hidden vector $u_n \in \mathbb{R}^{(m_n+1) \times d}$ (7). Attention aggregation will be applied on u_n to extract correlations between each cell with its neighbors through implementing dot product operation between spatial context vector u_c and u_n with $softmax$ operation to obtain attention weights $\alpha_c \in \mathbb{R}^{m_n+1}$ (8). α_c will be used as weights to summarize h_{m_n+1} followed by Dropout layer and Dense layer with ReLU activation to obtain final embeddings from neighbor module $Z_{neighbor} \in \mathbb{R}^d$ (9).

$$u_n = \tanh(W_3 h_{m_n+1} + b_3) \quad (7)$$

$$\alpha_c = \frac{\exp(u_n^T u_c)}{\sum_g \exp(u_n^T u_c)} \quad (8)$$

$$Z_{neighbor} = \text{ReLU}(W_4 \ (\alpha_c \times h_{m_n+1}) + b_4) \quad (9)$$

Data augmentation with dropout layer

Dropout layer operation[42] is used as a model-level minimal data augmentation strategy to obtain different views of the same cell. Via randomly masking neural units with certain probability

(*dropout rate*= 0.1 as default) before the final dense layer, each cell will obtain final embeddings from teacher-student-neighbor triple modules for subsequent contrastive learning.

Contrastive loss for StereoCell (NT-Xent loss)

Contrastive learning is conducted via comparing pairs of cell embeddings with *d* dimension in a unit hypersphere space. Specifically, StereoCell pulls together teacher-student views and teacher-neighbor views of the same cell as positive pairs while pushing apart different cells within the same batch. Assume positive pairs as $i \in Z_t$ with $j \in Z_s$; $i \in Z_t$ with $k \in Z_{neighbor}$, the distance is measured by cosine similarity of L2-normalized embeddings using dot product operation (10,13). NT-Xent loss[41] stands for normalized temperature-scaled cross entropy loss, as formalized in (11,14). τ stands for adjustable temperature coefficient, which can be used to scale the degree of pushing apart negative samples. We randomly sample a batch of N cells and compute NT-Xent loss on teacher-student and teacher-neighbor pairs, resulting in 2N data points (12,15). Given a positive pair, the other 2(N−1) examples within a batch are treated as negative examples.

$$s_{i,j} = {z_i}^T z_j / \tau \parallel z_i \parallel \parallel z_j \parallel \quad (10)$$

$$\ell_{i,j} = -log \frac{\exp(s_{i,j})}{\sum_{l=1}^{2N} \mathbb{1}_{[l \neq i]} \exp(s_{i,j})} \quad (11)$$

$$\mathcal{L}_{t-s} = \frac{1}{2N} \sum_{l=1}^{N} \left[\ell_{i,j}(2l-1, 2l) + \ell_{i,j}(2l, 2l-1) \right] \quad (12)$$

$$s_{i,k} = {z_i}^T z_k / \tau \parallel z_i \parallel \parallel z_k \parallel \quad (13)$$

$$\ell_{i,k} = -log \frac{\exp(s_{i,k})}{\sum_{l=1}^{2N} \mathbb{1}_{[l \neq i]} \exp(s_{i,k})} \quad (14)$$

$$\mathcal{L}_{t-n} = \frac{1}{2N} \sum_{l=1}^{N} \left[\ell_{i,k}(2l-1, 2l) + \ell_{i,k}(2l, 2l-1) \right] \quad (15)$$

$\mathcal{L}_{t-s}$ represents contrastive loss of teacher module and student module; $\mathcal{L}_{t-n}$ represents contrastive loss of teacher module and neighbor module. The overall loss (16) of StereoCell is obtained via summarizing these two loss terms at adjustable ratio $\theta \in [0,1]$.

$$\mathcal{L} = \theta \mathcal{L}_{t-s} + (1-\theta) \mathcal{L}_{t-n} \quad (16)$$

Attention weight extraction (DLPFC dataset)

Clustering-free signature identification for spatial transcriptomics is evaluated on attention weights α_i extracted from neighbor module. Specifically, we load the pre-training weights from the neighbor network as X_{index} and X_{counts}, and pass through a forward propagation. Then we obtain the attention weights α_i output via linear transformation and dot product with cellular context vector u_w followed by *softmax* operation. Then attention weights are logarithmic transformed followed by z-score standardization for visualization purposes. All calculations are conducted in 128-dimensional space.

StereoNiche Framework

StereoNiche leverages ViT network[43] backbone to extract niche features and implement contrastive learning to learn niche embeddings. Each section is sliced into n 100×100 pixel size of medium patches and each medium patch is then further partitioned into 100 10×10 pixel size of small patches. Thus, each medium patch is composed of 100 flattened small patches, denoted as $X_{token} \in \mathbb{R}^{100 \times 10 \times 10 \times e}$, where e is the channel depth equals to number of markers. Cell counts within each

small patch T_i is limited to 1 or 2 ($X_{cell\_count} \in (0,2]$), i.e., each small patch T_i contains 1 or 2 cells. For each small patch, we first conduct channel-wise average of contained cells (divided by X_{cell_count}) to obtain aggregated input for each small patch. Then similar encoding scheme as in StereoCell is applied to generate embeddings for small patches, i.e., scaling gene or channel embeddings by corresponding count value, followed by aggregation by attention mechanism with layer normalization to obtain final small patch embeddings $X_{input} \in \mathbb{R}^{100\times d}$, where d represents embedding dimension. Then X_{input} will be paired with corresponding positional encoding tokens and fed into ViT network in a flatten sequence as model input.

$$X_{input} = \{X_{input\_1}, X_{inpu\ \_2}, \dots, \dots, X_{input\_100}\} \quad (17)$$

Vision Transformer network (ViT)

The ViT network[43] uses the Transformer encoder with 8 layers and 8 heads, which consists of alternating layers of multiheaded self-attention and MLP blocks. Multi-head self-attention (MSA) is an extension of self-attention (18) in which k self-attention operations are implemented, called "heads", and their outputs will be concatenated. MLP contains two dense layers with non-linear Gaussian error linear units (GELU) activation. Learnable class token [CLS] is added to the sequence to represent aggregated embedding for each medium patch.

$$Attention(Q,K,V) = Softmax(\frac{Q\cdot K^T}{\sqrt{d_k}})\cdot V \quad (18)$$

Momentum Contrastive loss for StereoNiche (Moco loss)

Moco loss[44] is adopted by StereoNiche for contrastive learning, which is described as follows:

$$\ell_{i,j} = -log\frac{\exp(s_{i,j})}{\sum_{k=1\ [k\neq j]}^{K+1}\exp(s_{i,k})} \quad (19)$$

where $s_{i,j}$ stands for the similarity between positive pair (23) while $s_{i,k}$ represents the similarity between anchor with other samples as negative pairs (We regard sample i and its augmented view j as a positive pair, while sample i and other samples j $(k \neq j)$ in the queue as negative pair). Data augmentations for StereoNiche is generated via adding random noises to each medium patch.

$$s_{i,j} = z_i{}^T z_j/\tau \parallel z_i \parallel \parallel z_j \parallel \quad (20)$$

Here τ corresponds to the adjustable temperature hyper-parameter, which can be used to scale the degree of pushing apart negative samples. Contrastive loss for StereoNiche combines the [CLS] loss ℓ_{cls} for medium patch-level contrastive learning and tokens loss ℓ_{tokens} for small patch-level contrastive learning with adjustable weighting parameter θ (21). In this way, both specific marker of small patch and structured composition of medium patch are learned by StereoNiche.

$$\mathcal{L} = \theta * \ell_{cls} + (1-\theta)\ell_{tokens} \quad (21)$$

Attention weights extraction (CLR patients from CRC dataset)

We consider StereoNiche as an unsupervised object segmentation task and expect attention weights from different heads of ViT network can represent class-specific properties. Investigating the whole

slice for CLR patients, the attention weights of 8 heads using [CLS] token as a query in the last layer are exported. Note that [CLS] token is not attached to any label. By visually check the heatmap pattern of different heads, we find head 6 is consistent with the follicle-like TLS structure, then we select attention weights from head 6 to derive final visualization. As attention weights have no directions, the absolute values of z-score transformed attention weights are used for plots.

Niche frequency

Niche frequency is calculated as the number of cells assigned to certain niche divided by total number of cells from certain patient, denoted as niche frequency features f_{cn}. Patient-wise niche frequency matrix is used as survival model input, denoted as $X_{cn} \in \mathbb{R}^{P\times K}$, where P stands for patients and K stands niche types.

Cell type frequency in certain niche

For each patient, we define a combinatorial feature, cell type frequency in certain niche, denoted as f_{ct_cn}, which is calculated as the number of certain cell type assigned within a niche divided by the total cell number belonging to that niche for the patient. We can generate feature matrix of cell type frequency in certain niche, $X_{ct\_cn} \in \mathbb{R}^{P\times K\times M}$ as model input, where M stand for the number of cell types. For CRC dataset, we delete the cell type "dirt".

Niche-niche interaction enrichment score

Niche-niche interaction enrichment score is calculated to measuring spatial proximity between pairs of niches via permutation test. We first count the number of niche A located near niche B and define Z-score representing spatial enrichment of niche A close or far from niche B. The permutation test is implemented via setting the background set by repeatedly 500 times of randomizing locations for one niche to generate a null-distribution of A-B interactions.

XGBoost-AFT

We use X_{cn}, X_{ct_cn} as separate input for tree-based survival model with accelerated failure time model implemented by XGBoost. XGBoost-AFT[38] model is defined as (22), where x stands for input feature, $\mathrm{T}(x)$ represents the output from the ensembled decision trees, Y represents the survival time (model output after logarithmic operation), Z represents a random variable of a known probability distribution, and σ is a parameter that scales the size of Z. The maximum depth of the tree is set to 2, and σ is set to 0.1. 5-fold cross-validation is used for model training. Best model is selected via evaluating predicted survival time with the highest c-Index in the validation splits. We set objective parameter survival:aft and eval_metric to aft-nloglik according to the tutorials described in https://xgboost.readthedocs.io/en/latest/tutorials/aft_survival_analysis.html.

$$lnY = \mathrm{T}(x) + \sigma Z \qquad (22)$$

SHAP-based TreeExplainer

We leverage SHAP-based TreeExplainer[37] to derive explanations for XGBoost-AFT model and validate the clinical utility of identified cell niches. We use Python function shap.Treeexplainer() to calculate SHAP values, then draw SHAP summary plots to reveal the direction of attribution and its magnitude. Python function shap.dependence_plot() can plot the effect of features by showing how a feature's value (x axis) impacts the prediction (y axis) of every sample (each dot) in a dataset.

Patients in the best validation splits are used to generate SHAP summary and dependency plots. Global feature importance are prioritized by mean absolute SHAP value of each input feature.

Hyperparameters for StereoCell and StereoNiche

1) For synthesized data, the embedding dimension is set to 64, batch_size is set to 32, according to the number of simulated labels, the cluster number is set to 8 during K-means implementation.
2) For STARmap dataset, the embedding dimension is set to 64, batch_size is set to 32, cluster number is set to 15-20 and select 17 according to the best ARI against MWCH assignment.
3) For DLPFC dataset, the embedding dimension is set to 128, batch_size is set to 8. The learned cell embeddings are reduced to 10 dimensions by UMAP before K-means clustering. K is set to 8 according to the number of ground truth labels.
4) For MOB dataset, the embedding dimension is set to 128, batch_size is set to 8. The learned cell embeddings are reduced to 10 dimensions by UMAP before K-means clustering. K=9 is used to align with annotations from Allen Brain Reference Atlas.
5) For 10X BC dataset, the embedding dimension is set to 128, batch_size is set to 8. The learned cell embeddings are reduced to 10 dimensions by UMAP before K-means clustering. k=5 to 20 are implemented and 20 clusters are determined across all benchmarking methods.
6) For CRC dataset, the embedding dimension is set to 128, batch_size is set to 32. K-means is conducted on learned niche-level embeddings and k=10 is chosen according to the best ARI against Schürch assignment.
7) For TNBC dataset, the embedding dimension is set to 128, batch_size is set to 32. K-means is conducted on learned niche-level embeddings and k=8 is selected for downstream analysis.

K-means

We use Python function sklearn.cluster.KMeans() to perform K-means.

Differentially expressed analysis

Marker genes of each cluster is determined by using Seurat FindAllMarkers() function over normalized expression matrix. Significance of each maker is obtained by Wilcoxon Rank Sum test.

Gene Ontology and KEGG Pathway Enrichment Analysis

Gene symbols are mapped to Entrez Gene IDs by using R packages org.Mm.eg.db and org.Hs.eg.db for mouse and human data, respectively. Functional enrichment for each cluster is obtained by enrichGO() and enrichKEGG() functions of R package clusterProfiler[66]. The enrichment results are visualized by dotplot() function of clusterProfiler.

Gene Set Variation Analysis (GSVA)

For the 10X BC dataset, we choose Possion distribution parameter in R gsva() function to calculate the GSVA enrichment score. The R package limma is used to perform differential expression analysis between different clusters, pathways with FDR < 1% is chosen for analysis.

Analytic metrics

ARI

Adjusted Rand Index (ARI) is applied to assess clustering performance. Python library scikit-learn

is used to calculate ARI, and specifically, Python function adjusted_rand_score() is used. Before calculating ARI, we generate the contingency table (23). Given a set S of n elements, and two clusterings of these elements, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$.

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	π_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

(23)

the adjusted Rand index (ARI) can be calculated as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \quad (24)$$

where n_{ij}, a_i, b_j are values derived from the contingency table.

Average silhouette width (ASW)

We calculate ASW using Python function sklearn.metrics.cluster.silhouette_samples() from Python library scikit-learn as a reflection of clustering consistency. The Silhouette Coefficient for each cell is calculated using the mean Euclidean distance to other members in the same cluster (a) and the mean Euclidean distance to all members from the neighbor clusters (b). The Silhouette Coefficient for a cell is $(b - a) / max(a, b)$. The resulting score ranges from − 1 to 1, the best value is 1 and the worst value is -1, values near 0 indicate overlapping clusters. The average score of all cells is used to measure the overall cell-type purity according to the choice of clusters.

kSIM

We use the hierarchical navigable small world algorithm to find each cell's k-nearest neighbors (including itself) for each cell i, the number of neighbors with the same cell-type as i is denoted as n_i^k. In addition, to ensure cell *i* has a consistent neighbor, we set a minimum β fraction of the neighbors for cell i have the same cell-type as *i*. The kSIM acceptance rate is calculated as follows:

$$\text{kSIM rate} = \frac{\sum_{i=1}^{N} I\left(\frac{n_i^k}{k} \geq \beta\right)}{N} \times 100\% \quad (25)$$

where $I(x)$ is indicator function, k(default=25) and β(default=0.9) are user-specified parameters.

Moran'I

The Moran's *I* statistic is a measure of spatial autocorrelation, which can be used to quantify the degree of spatial variability for gene expression. Moran's *I* is defined as:

$$I = \frac{N}{W}\frac{\sum_i \sum_j [w_{ij}(x_i - \overline{x})(x_j - \overline{x})]}{\sum_i (x_i - \overline{x})^2} \quad (26)$$

where N is the number of spatial units indexed by i and j; x is the gene of interest; $\overline{x}$ is the mean of x; w_{ij} is a matrix of spatial weights calculated using the 2-dimensional spatial

coordinates of the spatial units; and W is the sum of all w_{ij}. Moran's I value ranges from −1 to 1, where a value significantly great than $1/(N-1)$ indicates a clear positive spatial autocorrelation, a value significantly lower than $-1/(N-1)$ indicates a negative spatial autocorrelation, and a value close to 0 indicates random spatial expression. We calculate Moran's I statistic using python function scanpy.metrics.morans_i() and evaluate SVGs identified by spatialDE, SPARK and StereoCell.

Other benchmarking tools

Seurat V4

Count matrix is firstly normalized using 'SCTransform' methods in R package Seurat V4(Seurat_4.0.0) with default parameters. PCA is then performed on the normalized data using RunPCA(). FindNeighbors() function with default parameters from Seurat is used to build shared nearest neighbor (SNN) graph. Finally, FindCluster() function is applied to assign cluster labels to each cell.

SCANPY

We run SCANPY (version 1.8.1) and follow the steps according to author's script (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html).

SpiceMix

As the SpiceMix (https://github.com/ma-compbio/SpiceMix.git) tool is still under development, we couldn't reproduce the clustering result, so the ARI score from SpiceMix paper[24] is used directly.

BayesSpace

PCA is performed over top 2,000 HVGs and the first 15 principal components are used as input to BayesSpace. For DLPFC dataset, cluster number is set to 7 to comply with manual annotation. MCMC algorithm is run for 50,000 iterations. We set the smoothing parameter gamma to 3, which is generally suggested for Visium datasets. For 10X BC dataset, cluster number is set to 20.

SpaGCN

We refer to https://github.com/jianhuupenn/SpaGCN/blob/master/tutorial/tutorial.ipynb to implement SpaGCN for DLPFC dataset.

SEDR

SEDR pipeline uses a deep autoencoder to construct a low-dimensional latent representation of the gene expression, which is then simultaneously embedded with the corresponding spatial information through a variational graph autoencoder. We use the author script to analyze spatial transcriptomics dataset. https://github.com/JinmiaoChenLab/SEDR/blob/master/run_SEDR_DLPFC_all_data.py

HMRF

Giotto implements a hidden Markov random field (HMRF) model to detect domains with coherent patterns by comparing gene expression between cells and their neighbors. HMRF is conducted following the guideline proposed by http://spatialgiotto.rc.fas.harvard.edu/giotto.visium.brain.html. To define spatial relationship between cells, we first create a spatial network using createSpatialNetwork() function. And genes of interest are listed as spatial genes returned by the spatial variable gene detection algorithm silhouetteRank() using the default parameters. Then we

choose top 500 spatial variable genes to run HMRF.

SpatialDE

SpatialDE uses Gaussian process regression to decompose expression variability into spatial and nonspatial components. Significant spatial variable genes can be detected via a likelihood ratio test and P-value is estimated by Chi-squared test. We use spatialDE() function from Giotto package to obtain candidate genes and whose q-value <0.05 will be considered as SVGs.

SPARK

SPARK is built upon a generalized linear spatial model with a variety of spatial kernels to model raw count data directly. In addition, SPARK employs ten different spatial kernels to make sure that the algorithm catches variable spatial patterns, including five periodic kernels and five Gaussian kernels. In order to combine results across multiple spatial kernels together, SPARK uses Cauchy combination approach to calculate a calibrated P-value. We use spark() function from Giotto package to obtain candidate genes and whose adjusted p-value <0.05 will be considered as SVGs.

Reference

66. Yu, G., Wang, L. G., Han, Y., & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, **16**, 284-287 (2012).

Data availability

All spatial transcriptomic and proteomic datasets in this study are published previously, and their availabilities are described in Supplementary Table 1.

Code availability

Stereo is written in Python using the TensorFlow library. The source code is available on Github at https://github.com/melobio/Stereo.

Acknowledgement

This research is supported by Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (2017B090904014), Ministry of Science and Technology of the People's Republic of China's program titled 'Science & Technology Boost Economy 2020' (SQ2020YFF0426292), and Intelligent Shanghai Program under Shanghai Health Development Planning Commission Projects (2018ZHYL0213).

Author contributions

M.Y. conceived the problem and designed all detailed studies. M.N. coordinated the resources and facilitated insightful discussions. Y.C.H. provided suggestions on pathology analysis. H.M.Y. and F.M. supervised the work. X.M.L. performed bioinformatics analysis. Y.W., H.P.H and L.Q.L. performed machine learning experiments. M.Y. wrote the manuscript.

Competing interests

F.M. declares the following competing interests: stock holdings in MGI, BGI-Shenzhen.

Additional information

Supplementary information is available for this paper in an additional Supplementary File, including 14 Supplementary Figures and 3 Supplementary Tables.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- YMengEPCflat.pdf
- StereoSupplementary20211129.pdf