SOAPy: a Python package to dissect spatial architecture, dynamics and communication

Heqi Wang[1#], Jiarong Li[1#], Siyu Jing[1], Ping Lin[1], Yu Li[2], Haibing Zhang[2],
Yujie Chen[1], Zhen Wang[1] & Hong Li[1*]

Affiliations

[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2]CAS Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

[#]These authors contributed equally

[*]Correspondence to: lihong01@sinh.ac.cn (H. L.)

Abstract

Advances in spatial omics technologies have brought opportunities to dissect tissue microenvironment, while also posing more requirements and challenges for computational methods. Here we developed a package SOAPy to systematically dissect spatial architecture, dynamics and communication from spatial omics data. Specifically, it provides analysis methods for multiple spatial-related tasks, including spatial domain, spatial expression tendency, spatiotemporal expression pattern, cellular co-localization, multi-cellular niches, and ligand-receptor-mediated and spatial-constrained cell communication. Applying SOAPy on different spatial omics technologies and diverse biological fields has demonstrated its power on elucidation of biological questions about tumors, embryonic development, and normal physiological structures. Overall, SOAPy is a universal tool for spatial omics analysis, providing a foundation for continued investigation of the microenvironment.

Keywords

spatial omics, Python package, microenvironment, expression pattern, multi-cellular niche, cell communication

Introduction

Spatially resolved transcriptomics has been crowned Method of the Year 2020 by Nature Methods[1]. Since then, more and more experimental methods for measuring expression levels of genes, proteins or metabolites in a spatial context have been developed. These technologies include barcode-based and imaging-based ones, which differ in resolution, accuracy and throughout[2,3]. The most widely used 10X Visium spatial transcriptomics measures thousands of genes in each 55μm spot that typically contains 1-10 cells[4]. And imaging-based methods reach more microscopic resolution, such as MIBI-TOF[5] and PhenoCycler-Fusion[6], both detecting dozens of proteins at subcellular resolution. Additionally, spatial multi-omics technologies that simultaneously measure multiple molecular types are emerging, e.g NanoString GeoMx DSP for 18000 RNAs and 140 proteins in the region of interest (usually >100 cells)[7].

With the development of experimental methods, corresponding analysis pipelines have been designed for pre-processing raw data from specific experimental platforms, such as Space Ranger for 10X Visium and MCMICRO for multiplexed tissue imaging[8]. Methods adapted from single-cell RNA sequencing (scRNA-seq) data analysis could be used to perform standard dimensional reduction, clustering, cell type annotation and marker selection for spatial-omics data[9] that do not require spatial information. And for low resolution spatial technologies, various deconvolution methods have been developed to impute cell-type composition from the mixture of cells.

After these pre-processing, downstream analyses are largely independent of experimental technologies, focus on the key feature of spatial omics: space. For example, identifying spatial variable genes[10–12], detecting spatial domains[13], inferring genes or cell-subtypes associated with spatial localization, and so on[3]. Earlier algorithms were often designed for one specific task, tools that fit in with various analysis tasks are becoming popular. A pioneer work Giotto not only builds a data pre-processing pipeline similar to scRNA-seq data analysis[14], but also provides modules for spatial pattern detection, cell neighborhood analysis, and interactive visualization. Squidpy provides scalable analysis framework for both spatial neighborhood graph and image, along with an interactive visualization tool[15]. stlearn is another integrated package for spatial transcriptomic analysis, which adds the functions of spatial trajectories and pseudotime analysis[16]. Investigating the spatial organization of tissue microenvironment are important applications of spatial omics, which may gain new insights in various biological fields. However, the related analysis methods are scattered or lacking, a package for integrative analysis of microenvironmental spatial organization is in an urgent need.

To address this problem, we present a package SOAPy (Spatial Omics Analysis in Python) to jointly perform multiple tasks for dissecting spatial organization, including spatial domain, spatial expression tendency, spatiotemporal expression pattern, co-localization of paired cell types, multi-cellular niches, and cell-cell communication. SOAPy improves on previous tools in three main areas (**Table S1**): (1) Providing several alternative methods for most tasks to be suitable for complex and diverse biological tissues and various analysis requirements. (2) Offering a factor decomposition strategy for high-order spatial data to discover the major modes of variations in spatial, time, sample or others. (3) Proposing a new method to combine ligand-receptor expression and spatial locations to better infer short-range and long-range cell communications. We also applied SOAPy to a wide range of public datasets to demonstrate its general applicability and interpretability. SOAPy will be one of the fundamental packages for spatial omics analysis in Python.

Results

Overview of the SOAPy package

SOAPy is composed of four modules: ***Data Preprocessing***, ***Molecular Spatial Dynamics*** containing *Spatial Tendency* and *Spatiotemporal Pattern* analysis, ***Cellular Spatial Architecture*** for analyzing *Spatial Proximity* and *Spatial Composition*, and ***Spatial Communication*** that combines spatial distance, expression level and interaction mechanism of ligand-receptors to infer cell interactions (**Figure 1**). In addition, SOAPy provides rich visualization capabilities for all of the analysis methods mentioned above.

The flexible ***Data Preprocessing*** module makes SOAPy suitable for various spatial data, fitting with different modalities and different resolutions. To demonstrate the utility of SOAPy, eight public datasets obtained from five state-of-the-art technologies were analyzed (**Table S2**). These datasets involve multiple scenarios with different molecular modalities (protein vs RNA), throughput (dozens to genome-wide), spatial resolution (0.1 ~ 55μm), and in physiological and pathological states.

Spatial domain analysis recapitulates anatomic and pathological structures

Cells are not randomly distributed in tissues. They are self-organized into specific structures to perform tissue functions. While in disease states, cells form abnormal structures. The *Spatial Domain* analysis provides unsupervised (STAGATE) and supervised (AUCell-LMI) methods to detect these structures (called spatial domains) based on gene expression profiles and spatial locations[13,17,18].

We first tested STAGATE on Slide-seq V2 data for mouse olfactory bulb and 10x Visium spatial transcriptomic data for human breast cancer[19]. Spatial domains identified by STAGATE are highly consistent with the manual-labelled structures . It successfully distinguishes truth anatomical structures (**Figure S1a**), malignant and non-malignant tissues (**Figure S1b**, ARI=0.513), and more sophisticated pathological stages (**Figure S1c**, ARI=0.580). Then we tested AUCell-LMI for finding local structures with known signature genes, such as tertiary lymphoid structure (TLS)[20]. The results showed that supervised AUCell-LMI based on known TLS signature could more accurate and more convenient identified the TLS region than unsupervised STAGATE (**Figure S1d, e**). Taken together, Spatial domain analysis in SOAPy could extract the interesting anatomic or pathological structures for downstream analysis.

Spatial tendency analysis finds genes associated with spatial structures

The aim of *Spatial Tendency* analysis is to assess whether expression features were influenced by spatial proximity to the region of interest (ROI). Expression features could be gene expression, pathway activity, cell proportion and so on. The ROI is defined by manual annotation or automatically detected by the *Spatial Domain* analysis. Two kinds of methods, statistical test and regression model, are available for tendency estimation in the *Spatial Tendency* module (Methods).

We used 10X Visium data of mouse dorsolateral prefrontal cortex (DLPFC)[21] as an example to validate the feasibility of spatial tendency estimation (**Figure 2a**). The sample is consisted of the grey matter of DLPFC (including six cortical layers) and white matter (**Figure S2a**). To find genes whose expression changes along with the distance to the white matter, three strategies were used and compared[22] (**Figure S2b, c**): 1) cortical layers were divided into two regions and applied Wilcoxon test to identify differential expressed genes; 2) cortical layers were separated to five continuous zones for Spearman correlation test; 3) a polynomial regression model was fitted between gene expressions and distances to the white matter. Some genes identified by Wilcoxon test and Spearman correlation only express in few spots, which may be the results of data sparsity instead of real biological differences (**Figure S2e**). The regression model describes the continuous spatial variation of expression, therefore it could find more complex spatial patterns than other methods[23], such as nonlinear "low-high-low" spatial pattern (**Figure S2f)**. Next, we analyzed the expression patterns of 2857 significant (FDR < 0.05, range >0.3) genes identified by polynomial regression. K-means clustering grouped them into 10 clusters (**Figure 2b**). The gene clusters were compared with previously reported cortical layer specific genes[24,25] (**Figure 2c**), showing high consistence. C3 is specifically highly expressed near white matter regions; the expression peaks of C5, C8, C2, and C7 are at layer 6, 5, 4, 2, respectively (**Figure 2d**).

Considering that there are no predetermined structures in some scenarios, we added three published methods (SpatialDE[10], SPARK[12], and SPARKX[11]) which identify spatial variable genes (SVGs) but do not need a given ROI. Comparing these SVGs methods with the above mentioned tendency estimation found shared and specific genes among methods (**Figure S2d**). SVG methods were more inclined to show the local differential expression of genes rather than the relationship with distance (**Figure S2g**). Users can select sutiable methods based on their requirements.

Tensor decomposition reveals the spatiotemporal patterns of gene expression

With advances in omics techniques, spatial-resolved and time-series molecule profilings are becoming available. One of the challenges is how to study the roles of spatial effects and temporal effects simultaneously in biological questions. The *Spatiotemporal Pattern* function in SOAPy employs tensor decomposition to extract components from the three-order expression tensor ("Time-Space-Gene"), revealing hidden patterns and reducing the complexity of data explanation.

Here, we used the mouse embryo development dataset from GeoMx Digital Spatial Profiling (DSP)[7]. Limited by the availability of expression profiles, four time points (E9, E11, E13, E15) and eight subtissues (Heart wall, Heart valve, heart trabecula, Lung epithelium, Lung mesenchyme, Midgut epithelium, Midgut mesenchyme, and Midgut neuron) from three organs were included in our analysis (**Figure 3a,b**). Canonical Polyadic (CP) decomposition[26] was used to factorize the expression tensor with 1000 high variable genes (a 4*8*1000 tensor) into seven factors, each of which is the outer product of three vectors that contain the loadings for describing the relative contribution of time, subtissues and genes (**Figure 3c**). We observed three empirical spatiotemporal patterns based on the loadings of time and subtissues: pure temporal variation (F1, F2), pure spatial variation (F3, F4), spatial and temporal variation occur together (F5, F6, F7). We also conducted functional enrichment analysis based on the loadings of genes for each factor (**Table S3**) and visualized the typical genes in images (**Figure 3d**).

Genes in F1 (e.g. *Hbb-bh1*) highly express in heart and lung sub-tissues at E9, and then gradually decrease in the later stages. Their expression pattern is consistent with the enriched function "regulation of vasculature development". F1 indicates co-development of heart and lung in the early embryo, which is consistent with previous studies[27]. The expression of F2 genes (e.g. *Epcam*) increases significantly since E11 in most sub-tissues of three organs, especially in the lungs. Expression of F3 and F4 genes is stable along the developmental time. F4 genes highly express in heart wall and heart trabecula, and their functions are enriched in cardiac cell development as expected. Both F5 and F7 genes are enriched in midgut development. F5 (e.g. *Psd*) slightly decreases from E11 to E15, while F7 (e.g. *Ndrg1*) increases obviously from E11 to E15. F6 genes are specifically highly expressed in the heart valve between E13-E15. In summary, the *Spatiotemporal Pattern* function in SOAPy could reveal spatiotemporal specificity during development and other biological processes.

Spatial proximity analysis characterizes co-localization patterns between cell types

Spatial architecture of cells is important for understanding the organization rules from single cells to tissues[28–30]. SOAPy first constructs a cell/spot network fromspatial locations; then implements two scenarios for deciphering spatial architecture: *Spatial Proximity* analysis (including neighborhood and infiltration) determines whethe two cell types or cell states within an image are significant proximal; *Spatial Composition* analysis identifies multi-cellular niches that composed by cell types with specific proportion.

We applied this analysis to a dataset of 41 triple-negative breast cancer (TNBC) patients[5], which used multiplexed ion beam imaging by time-of-flight (MIBI-TOF) to simultaneously quantify expression of 36 proteins in-situ at sub-cellular resolution. Totally 211,649 cells were annotated to eight types (epithelial cell, endothelial cell, mesenchymal cell, B, CD4 T, CD8 T, macrophage and other) based on the expression of known protein markers.

First, *Spatial Neighborhood* analysis was performed to identify significantly adjacent cell types compared to random perturbation[29]. **Figure 4a** illustrates the neighborhood score of all samples for all cell type pairs, with positive or negative scores corresponding to co-localization or avoidance. Different immune cells types such as B, CD4 T, CD8 T and macrophage have significant co-localization in many patients, which may relate with the formation of inflammatory foci (**Figure 4b**). Endothelial and mesenchymal cells also prefer to co-locate together (**Figure 4c**). Colocalization pattern of malignant epithelial cells and non-parenchymal cells were highly heterogeneous across patients. Taking malignant epithelial cells and mesenchymal cells as an example, samples with less than 200 mesenchymal cells were filtered, others are subjected to *Spatial Infiltration* analysis. Samples with higher and lower infiltration scores indicate mixed (e.g. sample 28) and compartmentalized (e.g. sample 29) patterns between malignant epithelial cells and mesenchymal cells respectively (**Figure 4 d-f**).

Spatial composition analysis discovers multi-cellular niches

For *Spatial Composition* analysis of the TNBC dataset, the cell-cell network that connected centroids of the cells within 100 pixels was built to capture the composition pattern of more surrounding cells. Niche of each cell was presented by the proportion of cell types of its surrounding cells, called I-niche. I-niches of 211,649 cells from 41 TNBC patients were clustered into 30 niche clusters, named C-niches (**Figure 5a, Figure S3a**). The major cell types of the top two C-niches (C-niche13, C-niche18) are mainly composed of malignant epithelial cells, and the percentages of other cell types are less than 15%, showing the characteristics of tumor cell aggregation (**Figure 5b**). Additionally, epithelial cells also form C-niches with other cell types. For example, C-niche25 is composed of 38% epithelial cells, 31% mesenchymal cells, and 9% macrophages; C-niche27 is composed of 23% epithelial, 28% endothelial, 10% mesenchymal cells and 10% macrophages; C-niche15 is composed of 30% epithelial, 23% CD4 T, 13% CD8 T cells and 11% macrophages, suggesting different local microenvironment exists among tumors (**Figure 5b**). We also observed four B cell dominated C-niches (C-niche10, C-niche17, C-niche28, C-niche4) that may be related to tertiary lymphoid structures. For example, sample 1 contains C-niche 10, 17, and 28 (**Figure 5c**). Around 80% of cells are B cells in C-niche10; C-niche17 majorly consists of 52% B cells, 13% CD8 T cells, 10% CD4 T cells, and 11% epithelial cells; C-niche28 majorly consists of 30% B cells, 10% CD8 T cells, and 37% epithelial cells.

In order to investigate the combinational effects of non-parenchymal cell types and niches on patient heterogeneity, the “Niche-CellType-Sample” tensor (30*7*41) was factorized to four factors (**Methods**). All samples were clustered into five groups according to the sample loadings in different factors (**Figure 5d**). Sample groups A, B, C, and E have the highest loadings in factors 3, 2, 1, and 4, respectively. By checking the loadings of cell types and niches in the major factors (**Figure S3b,c**), group B corresponds to the above mentioned B cell enriched samples; group C is characterized by niches with high proportion of mesenchymal cells; group E has niches consisted of T cells and macrophages.

Furthermore, survival analysis was performed to explore the clinical indications of niches. Eight c-niches were significantly related to survival time ($P < 0.05$, **Figure S4**). For example, patients with a higher proportion of c-niche15 had a longer survival time (**Figure 5e**). There also exists survival differences among the patient groups identified by the “Niche-CellType-Sample” tensor decomposition, such as longer survival time for group C patients that that of group D (**Figure 5f**). Taken together, spatial composition analysis could find multi-cellular niches and yield insight into how cells are organized into tissues.

Ligand-receptor-mediated and spatial-constrained cell-cell communications

The above spatial architecture analysis disregards interacting molecules and context, while expression-based methods like CellphoneDB[31] and CellChat[32] infer cell-cell communications by the expression of ligands and receptors (LRs) disregarding spatial proximity. SOAPy develops a new method that simultaneously utilizes spatial location and gene expression to calculate interaction scores (affinity and strength) and then outputs significant LR interactions (**Figure 6a, Methods)**. It can not only infers short-range cell communication that relies on contact LRs to directly deliver signaling between adjacent cells; but also infer long-range cell communication that does not require cell–cell contact, rather depending on the diffusion of signaling molecules from one cell to another after secretion[33,34].

The *Spatial Communication* module was applied to an ovarian cancer dataset generated by the MERSCOPE platform, measuring 500 genes and 71,381 cells (**Figure 6b**). Cells were classified and annotated into ten types or subtypes by Leiden clustering algorithm. The spatial locations of epithelial cells C3 are very special, which clearly separated with most of other cells. Therefore, our method did not find significant contact LRs between epithelial cells C3 and other cell types. However, CellChat, one of the most popular LR communication inference packages using scRNA-Seq data, reported many LR interactions due to lack of spatial constrain (**Tabls S4**), indicating lower false positives of our method.

We used endothelial cell as an example to present its short-range and long-range communication partners. Fibroblasts and macrophages are located closest to endothelial cell, while epithelial cell C3 and C4 are far away from endothelial cell (**Figure 6c**). Consistently, fibroblasts have the largest number of contact LRs with endothelial cells recognized by our algorithm, while there is no contact LRs for distant cell types such as epithelial cells C2, C3, C4 and C5 (**Figure 6d**). For cell types that are not spatially close to endothelial cells, *Spatial Communication* module could infer secreted LRs that mediate long-range cell communications. The average distance from epithelial cells C2 to the closet endothelial cells is significantly larger than the average distance from fibroblasts to the closet endothelial cells ($P < 3.9e\text{-}312$). There are no contact LRs between epithelial cells C2 and endothelial cells but 6 secreted LRs were identified (**Figure 6 d, e**).

Totally, we found 19 contact LRs and 66 secreted LRs that may play key roles in short-range and long-range communication between endothelial cells and others (**Figure 6f**). For example, COL1A1 (type I collagen) and its receptor ITGA1/ITGB1 (integrin α/β) highly express on spatial adjacent fibroblasts and endothelial cells, their affinity and strength scores are significantly higher than random scores (**Figure 6g**). Previous studies have reported that binding collagen to integrin may activate downstream signaling pathways contributing to cancer progression[35]. VEGFB-FLT1 is an interesting LR pair for long-range communication between epithelial and endothelial cells (**Figure 6h**). Epithelial cells C2 release ligand VEGFB, and endothelial cells high express FLT1 (also known as VEGFR1). Their interaction may promote tumor angiogenesis and are potential drug targets for anticancer therapy[36]. In summary, SOAPy provide a new way to study spatial-constrained cell-cell interactions and more accurately identify the related ligand-receptor pairs.

Discussion

Tissue microenvironment is critical for understanding homeostasis, development, regeneration and disease. Single-cell and spatial resolved omics are the most promising technologies to investigate microenvironment. Tools for systematically dissecting microenvironment and discover biologically important genes or spatial cellular architecture are still falling behind, SOAPy just fill this gap. SOAPy contains easy-to-use analysis modules for interpreting complex spatial microenvironments, such as the spatial distribution patterns of genes and cells, dynamic changes along with space and time, and cell-cell communications et al. In this article, we demonstrated all SOAPy modules with various types of spatial omics data, and provides complete tutorials to help users get started quickly.

The spatial distribution of genes or cells is associated with many elements, such as time, interaction of cells, pathological foci, sample heterogeneity and so on. In the face of these multi-dimensional data, how to extract important and meaningful features is a key task. SOAPy utilizes tensor decomposition to discover the major modes of variations from multi-dimensional data. The cases of mouse embryo development and breast cancer showed that tensor decomposition in SOAPy is powerful for interpret complex biological data. Another significant advantage of SOAPy is the innovative *Spatial Communication* module. It combines spatial distance, expression level and interaction mechanism of ligand-receptors to infer cell-cell communication. The case of ovarian cancer showed that SOAPy could markedly reduce false positives of interacting ligand-receptors compared to existing methods.

These advantages makes SOAPy differ from existing spatial data analysis tools. Future extensions of SOAPy could be the integration of multi-modal spatial data to delineate microenvironment, adaptation of methods from geoscience, network science, or artificial intelligence to better extract biological meaningful spatial patterns. We anticipate that SOAPy will be widely used by researchers to discover biological insights from spatial omics data.

Methods

Data preprocessing

Data Import

The *Data Import* function converts data from different spatial omics technology to a unified data structure that contains expression profiles of molecules (genes/proteins/metabolites) and location of cells/spots. Barcode-based data formats can be read directly by passing in tables representing expression matrix and spatial coordinate information. An image and a cell segmentation mask are provided for imaging-based data, and the representation and coordinate matrix is extracted through the tutorials on our website. We used the Scanpy toolkit[37] and generate Anndata data.

Spatial network construction

The *Spatial network* function provides four ways to build a neighborhood network of cells/spots (Figure 1a). 1) Regular network; 2) KNN network that connects each site with its K nearest neighbors; 3) Radius network that all cells/spots within the given distance are connected; 4) Neighbor network based on Voronoi Diagram.

Spatial domain identification

Unsupervised spatial domain identification: STAGATE

STAGATE is a graph attention autoencoder for spatial domain identification[13]. It firstly integrates gene expression profiles and spatial location information to learn low-dimensional latent embedding, and then assigns spatial domains by Louvain clustering.

Supervised spatial domain identification: AUCell-LMI

To detect domains whose signature genes are already known, the score of signature genes for each cell/spot is calculated by AUCell[38,39], and then local Moran index[17] (LMI) is used to estimate the degree of spatial aggregation. LMI of cell/spot i is defined as:

$$I_i = \frac{x_i - \bar{x}}{s^2} * \sum_{j \in \mathrm{n_i}} w_{ij}\left(x_j - \bar{x}\right) \#(1)$$

Where x_i is the AUCell score of cell/spot i, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, j is any neighbor cells/spots of i based on K nearest neighbors, w_{ij} is the spatial weight between i and j . The P-value is calculated by permutation test and adjusted by Benjamini-Hochberg method[40] to get the false discovery rate (FDR).

LMI of all cells/spots are illustrated by Moran scatterplot (Figure S1e). Each point represents one cell/spot, the horizontal axis shows the normalized AUCell score, and the vertical axis indicates the "spatial lag" which is calculated by spatial weighted normalized score of neighboring sites. Sites with positive AUCell scores, positive spatial lags, and low FDR were picked out as the targeted spatial domain.

Spatial tendency analysis

Definition of ROI and distance

Given a region of interest (ROI), the first step is to generate a binary mask file (**Figure S2a**). Users can manually select ROI using tools like ImageJ to generate a mask file, or get interesting cells/spots via SOAPy *Spatial domain* analysis and then use SOAPy to create a mask file: Discrete cells/spots are converted to continuously connected regions using a series of digital image processing steps in OpenCV library, such as dilation, corrosion, removal of small connected components, and removal of holes.

Next, the shortest distance from each cell/spot to the ROI boundary (contour) is calculated. When an ROI contains multiple connected components, the closest connected component is selected to calculate the distance[23].

$$d(i, C) = min_{p \in C} Enc(i, p) \# (2)$$

where i is a cell/spot, C is the boundary of ROI, and p is any pixel on the boundary. $Enc(\quad)$ is a function of Euclidean distance. Distance with positive or negative signs are used respectively to distinguish cells/spots located outside or inside the ROI boundary. Then we can study the tendency of molecule expression along with distance.

Identification of expression features with spatial tendency

SOAPy provides two statistical testing methods (**Figure S2b**): 1) wilcoxon rank sum test to compare the molecule expression of cells/spots between two regions; 2) spearman correlation between median expression and the rank of continuous zones. To resolve more complex spatial tendency (e.g., nonlinear) or analyze ROIs without prior hypothesis, SOAPy provides a parameter regression method (polynomial regression model) and a non-parametric regression method (locally weighted liner regression, LOESS).

Polynomial regression assumes that the output variable can be represented by the sum of powers of the input variable.

$$Y = a_0 + \sum_{k=1}^{n} a_k \, d^k \# (3)$$

Where d is the distance to the ROI; Y is the vector of molecule expression; n is the degree of the polynomial; a_0 is intercept; a_k are slope coefficients. P-value is calculated by F-test.

LOESS is a locally weighted polynomial regression method. Its core concept is to fit weighted linear regression models with each data point using its surrounding data points within the predefined window size and connect the centers of the regression lines. R^2 (coefficient of determination) and residual standard deviation are estimated to measure the goodness of fit.

Parameters used in both of the regression models could be customized and adjusted based on the biological scenario and goodness of fit. To summarize the spatial tendency of all molecules, the estimated expression values are fed into the K-means clustering algorithm to obtain gene clusters with similar spatial expression tendency.

Spatial architecture analysis

Spatial neighborhood analysis

For each paired cell types, a neighborhood score (NS) between cell type 1 ($ct1$) and cell type 2 ($ct2$) is calculated as follows[29]:

$$NS_{ct1,ct2} = \frac{N_{ct1,ct2}}{N_{ct1,other} + N_{ct2,other}} \#(4)$$

where $N_{ct1,ct2}$ is the number of direct connections between $ct1$ and $ct2$, $N_{ct1,other}$ is the number of direct connections between $ct1$ and all other cell types. Background distribution is generated from 1000 random permutations that fix the numbers of $ct1$ and $ct2$ and randomly change their locations. P-value is the proportion of permutations whose NS is larger or smaller than the observed one, which corresponds to either avoidance or interaction between $ct1$ and $ct2$.

Spatial infiltration analysis

An infiltration score (IS) is defined to present the degree of non-parenchymal (immune or stromal) cells infiltration into malignant tissues:

$$IS_{m,np} = \frac{N_{m,np}}{min\left(N_{m,m}, N_{np,np}\right)} \#(5)$$

where $N_{m,np}$ is the number of direct connections between malignant cells and non-parenchymal cells. Sample with too few non-parenchymal cells are regarded as cold tumor. Otherwise, larger infiltration score indicates more non-parenchymal cells are mixed into malignant tissues, while smaller infiltration score suggests non-parenchymal cells are more possible to be compartmentalized with malignant tissues.

Spatial composition analysis

Given an index cell, niche is defined as the proportion of cell types for its surrounding cells[41]. Taken all cells in one or more images, clustering algorithms like

K-means divides their niches into different clusters, called C-niches.

Spatial-constrained cell-cell communication inference

Ligand-receptor (LR) pairs were obtained from the CellChat[32] package, in which LR pairs were classified into contact and secreted based on their action mechanism. We hypothesized that the contact LR pairs mediate short-range cell communications while secreted LR pairs could mediate long-range cell communications. Therefore, SOAPy infers cell communications based on the types of LR pairs and spatial distance among cells (presented by a cell network). For short-range communication, direct neighbors on Voronoi Diagram are connected to build a cell network. For long-range communication, all cells within the given distance are connected to build a cell network. Once the cell network is built, $Affinity$ and $Strength$ scores are calculated for LRs on any two cell types. The LR pairs with $Affinity\,Pvalue <$ 0.05 and $Strength > 4.0$ are considered to be significant. Paired cell types are ranked based on the number of significant LRs.

Cell-level ligand-receptor affinity score

The interaction of LR is variable among cells/spots at different spatial locations, therefore we first define a cell-level ligand-receptor affinity score. Suppose a cell/spot i is a sender of ligand, cells/spots that have connection with i and express the matched receptor are receivers, the $Affinity\ score$ of ligand-receptor at location i is defined as:

$$Affinity\ score_{l-r,i} = \sum_{j \in n_i} \frac{l_i * r_j}{1 + d_{i,j}}\ ,\quad i\ as\ a\ ligand\ sender \#(6)$$

where j is the cell/spot that connect to i in the cell network; l and r are expression levels of the ligand and receptor; d is 0 for contact LR pairs or Euclidean distance between i and j for secreted LR pairs. Similarly, when the cell/spot i is a receptor receiver, the $Affinity\ score$ of receptor-ligand at location i is defined as:

$$Affinity\ score_{r-l,i} = \sum_{j \in n_i} \frac{r_i * l_j}{1 + d_{i,j}}\ ,\ \ i\ as\ a\ receptor\ receiver \#(7)$$

The $Affinity\ Pvalue$ is obtained by random permutation:

$$Affinity\ Pvalue = \frac{\#m\{A^{(m)} \leq A^0, m = 1,2,\cdots,M\}}{M} \#(8)$$

M is the total number of randomizations, $A^{(m)}$ is the $Affinity\ score$ under the m-th randomization. Each randomization redistributes the expression values of the LR, but keeps topology of the cell network. The affinity scores are calculated for all cells/spots, and the P-values are used to find a subset of cells/spots at which the LR exist interaction.

CellType-level communication score

Suppose $ct1$ and $ct2$ are cell types that express ligands and receptors, respectively. The $Affinity\ score$ between the ligand of $ct1$ and the receptor of $ct2$ is the sum of cell-level scores:

$$Affinity\ score_{l,r,ct1,ct2} = \sum_{i \in ct1} \sum_{j \in n_i, ct2} \frac{l_i * r_j}{1 + d_{i,j}} \#(9)$$

$Affinity\ Pvalue$ is also calculated by random permutation, which randomly assign a pseudo expression value to each cell/spot based on cell-type specific expression distribution.

$Affinity$ reflects whether spatial connected $ct1$ and $ct2$ relatively more highly express the LR genes. However, if the expression of ligand or receptor is too low in $ct1$ or $ct2$ compared to other cell types, it is difficult to say that the LR is important for cell communications; Additionally, If $ct1$ and $ct2$ are connected by too few edges in the cell network, their communication may be false positive even affinity is significant. To address these problems, another index ‘strength’ is added. $Strength_{l,r,ct1,ct2}$ consists of two components: one is the relative expression level of LR pairs on $ct1$ and $ct2$, and the other indicates the enrichment of real spatial connections between $ct1$ and $ct2$. The detailed definition is as follows:

$$Strength_{l,r,ct1,ct2} = \left(\frac{\overline{exp}_{l,ct1}}{\overline{exp}_{l,all}} * \frac{\overline{exp}_{r,ct2}}{\overline{exp}_{r,all}} \right) * \left(\frac{2E}{1+E} \right) \quad \#(10)$$

$$E = \frac{edge_{ct1,ct2}}{\overline{edge}_{ct1,ct2}} \#(11)$$

where $\overline{exp}_{l,ct1}$ and $\overline{exp}_{l,all}$ are the average expression of ligand in $ct1$ and in all cells; $edge_{ct1,ct2}$ and $\overline{edge}_{ct1,ct2}$ are the real and expected number of connections between $ct1$ and $ct2$; E is the ratio of real and expected numbers. To constrain the range of E and make the result more stable, a Hill function transforms E into a range of (0, 2) and keeps the transformed E is still 1 when the number of real and expected connections are equal.

Tensor decomposition

To discover the major modes of variation in the high-order spatial data, such as the “Time-Space-Gene” tensor or “Niche-CellType-Sample” tensor, SOAPy provides interface functions to conveniently build tensors from AnnData objects and then decomposes tensors into several latent factors or components.

SOAPy implements two tensor decomposition methods, CANDECOMP /PARAFAC (CP) and Tucker decomposition[26,42]. Moreover, SOAPy supports non-negative constraints to make the factors more interpretable. Take non-negative CP[43] as an example, an n-order tensor X is expressed as the weighted sum of R (user-defined number of factors) rank-one tensors:

$$X \approx \sum_{\mathrm{r}=1}^{\mathrm{R}} \lambda_{\mathrm{r}} \mathrm{a}_{\mathrm{r}}^{(1)} \circ \mathrm{a}_{\mathrm{r}}^{(2)} \circ \ldots \circ \mathrm{a}_{\mathrm{r}}^{(\mathrm{n})} \#(12)$$

where λ is the weight of each factor; $\mathrm{a}_{\mathrm{r}}^{(\mathrm{k})}$ is the non-negative loading values of k-th variable in the r-th factor, indicating the relative contribution of variables to factors. Each factor is the outer product of the loading vectors.

Availability

All data and code that produced the findings of the study, including all main and supplemental figures, are available at https://github.com/LiHongCSBLab/SOAPy.

Acknowledgements

We acknowledge Andrew E. Teschendorff (from Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences) for his advice on our manuscript. We thank Bihan Shen, Xi Yan (from Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences) and Biao Liu (from Center for Excellence in Molecular Cell Sciences, Chinese Academy of Sciences), for their help on programing and result interpretation.

Funding

This research was supported by National Natural Science Foundation of China (T2122018, 32170680, 32300555), National Key R&D Program of China (2021YFF1200900), CAS Youth Innovation Promotion Association (Y2022076) and Shanghai Municipal Science and Technology Major Project.

References

1. Method of the Year 2020: spatially resolved transcriptomics. *Nat. Methods* **18**, 1–1 (2021).
2. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
3. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
4. Salmén, F. *et al.* Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. *Nat. Protoc.* **13**, 2501–2534 (2018).
5. 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).
6. Keren, L. *et al.* MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci. Adv.* **5**, eaax5851 (2019).
7. Merritt, C. R. *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat. Biotechnol.* **38**, 586–599 (2020).
8. Schapiro, D. *et al.* MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nat. Methods* **19**, 311–315 (2022).

9. Cang, Z. *et al.* Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nat. Methods* **20**, 218–228 (2023).

10. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).

11. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* **22**, 184 (2021).

12. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).

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

14. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).

15. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).

16. Pham, D. *et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues.* http://biorxiv.org/lookup/doi/10.1101/2020.05.31.125658 (2020) doi:10.1101/2020.05.31.125658.

17. Anselin, L. Local Indicators of Spatial Association-LISA. *Geogr. Anal.* **27**, 93–115 (2010).

18. Jong, P., Sprenger, C. & Veen, F. On Extreme Values of Moran's I and Geary's c. *Geogr. Anal.* **16**, 17–24 (2010).

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

20. Meylan, M. *et al.* Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. *Immunity* **55**, 527-541.e5 (2022).

21. Pardo, B. *et al.* spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics* **23**, 434 (2022).

22. Bardou, P., Mariette, J., Escudié, F., Djemiel, C. & Klopp, C. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* **15**, 293 (2014).

23. Hildebrandt, F. *et al.* Spatial Transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nat. Commun.* **12**, 7046 (2021).

24. He, Z. *et al.* Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. *Nat. Neurosci.* **20**, 886–895 (2017).

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

26. Kolda, T. G. & Bader, B. W. Tensor Decompositions and Applications. *SIAM Rev.* **51**, 455–500 (2009).

27. Peng, T. *et al.* Coordination of heart and lung co-development by a multipotent cardiopulmonary progenitor. *Nature* **500**, 589–592 (2013).

28. Schapiro, D. *et al.* histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods* **14**, 873–876 (2017).

29. Bäckdahl, J. *et al.* Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell Metab.* **33**, 1869-1882.e6 (2021).

30. Yuan, Z. *et al.* SOTIP is a versatile method for microenvironment modeling with spatial omics data. *Nat. Commun.* **13**, 7330 (2022).

31. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

32. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).

33. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).

34. Cheng, J., Yan, L., Nie, Q. & Sun, X. *Modeling and inference of spatial intercellular communications and multilayer signaling regulations using stMLnet*. http://biorxiv.org/lookup/doi/10.1101/2022.06.27.497696 (2022) doi:10.1101/2022.06.27.497696.

35. Xu, S. *et al.* The role of collagen in cancer: from bench to bedside. *J. Transl. Med.* **17**, 309 (2019).

36. Fischer, C., Mazzone, M., Jonckx, B. & Carmeliet, P. FLT1 and its ligands VEGFB and PlGF: drug targets for anti-angiogenic therapy? *Nat. Rev. Cancer* **8**, 942–956 (2008).

37. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

38. Van De Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).

39. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).

40. Haynes, W. Benjamini–Hochberg Method. in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 78–78 (Springer New York, 2013). doi:10.1007/978-1-4419-9863-7_1215.

41. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981.e15 (2018).

42. Zhou, G., Cichocki, A., Zhao, Q. & Xie, S. Efficient Nonnegative Tucker Decompositions: Algorithms and Uniqueness. *IEEE Trans. Image Process.* **24**, 4990–5003 (2015).

43. Shashua, A. & Hazan, T. Non-negative tensor factorization with applications to statistics and computer vision. in *Proceedings of the 22nd international conference on Machine learning - ICML '05* 792–799 (ACM Press, 2005). doi:10.1145/1102351.1102451.

Figures

Figure 1. Schematic diagram of SOAPy. **a,** "Data Preprocessing" module that imports data, generates cell network and identifies spatial domains. Data from different spatial omics technologies are converted to a unified data structure. Cell network can be built by any of the four methods. Spatial domains are inferred by unsupervised learning from expression and morphological data, or supervised classification based on the expression of signature genes. **b,** "Molecular Spatial Dynamics" module. Spatial tendency analysis finds genes or cells whose expression change with spatial distance to the given region. **c**,Spatiotemporal Pattern analysis performs a tensor decomposition to discover the major modes of variation in space and time. **d**, "Cellular Spatial Architecture" module. Neighborhood and infiltration analysis find spatial proximal cell types. Spatial composition reveals conserved niches in which surrounding cells of the index cell are consisted of specific proportion of cell types. **e**, Innovative "Spatial Communication" module that combine spatial distance, expression level and action mechanism of ligand-receptors (LRs) to infer cell interactions. The contact and secreted LRs are considered for short-range and long-range cell communications, respectively. Results at cell/spot level indicate the heterogeneous interaction among different spatial locations, they are further integrated to cell type-level to report significant LRs for any two cell types.

Figure 2. Spatial tendency analysis finds genes associated with spatial structures. **a**, HE image of a human dorsolateral prefrontal cortex (DLPFC) sample. Regions of white matter (WM) and six neuronal layers (L6 to L1) are labeled on the image. **b**, Regression curves between gene expression and the distance to WM. Polynomial regression models were fitted to identify genes whose expression varied along with the distance to WM boundary. These genes were grouped into 10 clusters by K-means clustering algorithm. Each curve present a cluster of genes with similar spatial expression tendency. Zero at the horizontal axis indicates the outer boundary of WM. **c,** Association between gene clusters and previously reported layer specific genes. Each row corresponds to a prior gene-list that specifically expresses in one neuronal layer[24]. Each red unit indicates the cluster of genes (column) is enriched in the prior

gene-list (row). **d**, Spatial distributions and fitted curves of the representative genes.

Figure 3. Tensor decomposition reveals the spatiotemporal patterns of gene expression during mouse embryo development. **a,** The spatiotemporal dataset of mouse development is represented by a three-order tensor (4 time points * 8 sub-tissues * 1000 highly variable genes), and then it's decomposed into seven latent factors. **b,** Representative spatial locations of sub-tissues at four time points. Each spot in the subtissues represents an ROI. **c,** Loading vectors of space and time for each factor obtained by tensor decomposition. Higher loading values indicates larger contribution of sub-tissues or time points to the expression variation of this factor. **d,** Spatial expression of example genes for each factor. The contours of heart, lung and midgut are colored by red, blue and green curves. ROIs of gene expression are presented by cyan points. The darker the cyan color, the higher the gene expression level.

Figure 4. Spatial proximity analysis characterizes cellular co-localization patterns. The triple negative breast cancer (TNBC) dataset contains 41 samples and 7 cell types. **a,** Heatmap showing the neighborhood scores of any two cell types in all TNBC samples. **b,** A representative sample with strong co-localization among immune cells. **c.** A representative sample with strong co-localization between endothelial and mesenchymal cells**. d,** The red bars show the number of mesenchymal cells and the blue bars show the infiltration score of mesenchymal cells into malignant epithelial cells. **e**, A representative sample with low infiltration score, suggesting compartmentalization between mesenchymal cells and tumor tissues. **f**, A representative sample with high infiltration score, suggesting mixture of mesenchymal cells into malignant epithelial cells.

Figure 5. Spatial composition analysis discovers multi-cellular niches in TNBC samples. **a.** Heatmap on the left shows the composition of neighbor cells in each C-niche. The right barplot shows the number of cells belonging to each C-niche. **b**, Representative samples of different C-niches, characterizing tumor cell aggregation

and different local microenvironment of tumors. **c**, The left image shows an example sample that has B cell dominated C-niches (the region of red box). Cells are colored by C-niches. ‘other’ are low-frequent c-niches whose proportion is less than 2%. Right images are amplified views of three representative C-niches. Black or gray cell contours indicate cells belonging to or not belonging to the C niche. The fill colors of cells represent cell types involved in the definition of the C-niche. **d**, Heatmap showing the loading values and clusters of samples. The three-order ‘Niche-CellType-Sample’ tensor was decomposed to four latent factors (**Figure S3b, c**). Samples are clustered into five groups according to their loading vectors. **e,** Survival curves stratified by the proportion of C-niche-15. **f**, Comparison of survival curves between two groups of patients.

Figure 6. Ligand-receptor-mediated and spatial-constrained cell-cell communications. **a**, The brief flow chart of our method. Short-range interaction is mediated by contact LRs on neighbor cells, long-range interaction is mediated by secreted LRs on cells within the given radius. Two new metrics, affinity and strength, are defined to estimate the probability of LR interactions in any two cell types. Only when both metrics are high, the LR is significant to mediate the interactions of these two cell types. **b**, MERSCOPE data from an ovarian cancer sample. **c,** Barplot showing the shortest distance from other cell to the closest endothelial cell. **d, e**. Short-range and long-range cell communication networks between endothelial cells and other cell types. Edges in d and e are the number of contact and secreted LRs. Edge width is the number of significant ligand-receptor pairs (affinity P-value < 0.05, strength > 4). **f**, Dot plot with ligand-receptor interactions corresponding to d and e. Each row indicates a ligand-receptor pair, with the first and the second genes representing a ligand and a receptor, respectively. Dot size indicates P-value of affinity. Color indicates the strength score. **g**, An example of contact LR that mediates the communication between spatially colocalized fibroblast and endothelial cells. COL1A1 is the ligand on sender fibroblast cells, and ITGA1/ITGB1 is the receptor on receiver endothelial cells. Expression was scaled to the range of 0-1 by normalization. **h**, An example of secreted LR, corresponding to the communication between spatially separate epithelial and endothelial cells. VEGFB is the ligand on sender epithelial-hypoxia cells, and FLT1 is the receptor on receiver endothelial cells.

Supplementary Information

Figure S1. Spatial domain analysis recapitulates anatomic and pathological structures. **a**, Anatomical structure of mouse olfactory bulb (Slide-seq V2 data) and domains identified by STAGATE. **b-c**, Expert-annotated pathological regions of a breast cancer sample (10x Visium), and the estimated 2-class and 19-class domains based on the results of by STAGATE. **d**, Expert-annotated tertiary lymphoid structure (TLS) on a kidney cancer sample (10x Visium), and the estimated TLS by the AUCell-LMI method. **e**, Moran scatterplot. The x-axis is the Z-transformed AUC, which presents the activity for the signature genes of TLS. The y-axis is the spatial weighted normalized AUC scores of neighboring locations. Hotspot presented by red points (FDR < 0.05, x > 0, y > 0) is regarded as tertiary lymphoid structure.

Figure S2. Spatial tendency analysis. **a**, Steps of image per-processing to generate a binary mask file for the given region of interest (ROI). **b**, Illustration of three spatial tendency analysis strategies: wilcoxon test, spearman correlation, and regression. **c**, Venn diagram shows the overlap of top 1000 genes (FDR q-value < 0.05) obtained from three spatial tendency analysis strategies. There are 380 overlapped genes, 352, 209 and 227 genes uniquely identified by a method (**Figure S2c**). **d**, Intersection plot showing the agreement for seven methods. Four methods estimate the tendency of gene expressions changing with the distance to a given region: wilcoxon test, spearman correlation, polynomial regression and LOESS regression. Other three methods identify spatially variable genes (SVGs) whose expressions depend on their spatial locations: SPARK, SPARKX and SpatialDE. The top-ranked genes with equal number obtained from each method were compared. Genes of LOESS are ranked by R-square, and genes of the remaining methods are ranked by FDR values. **e-g**, Representative genes identified by different kinds of methods. **e**: MIAT that is significant by Wilcoxon test and Spearman correlation analysis but not significant by

regression methods; **f**: PVALB that is significant by regression methods; **g**: Expression of TFF1 is spatially variable but do not show tendency of change with the distance to WM.

Figure S3. **a**, Heatmap showing the proportion of niches in all TNBC samples. **b-c**, Loadings of cell types and niches obtained from the decomposition of “Niche-CellType-Sample” tensor.

Figure S4. Results of survival analysis.

Table S1. Comparison with existing tools of spatial omics data analysis.

Table S2. Examples datasets that were used in this study.

Table S3. Enriched functional terms by gene set enrichment analysis. Genes were pre-ranked based on the loading values of each factor obtained from tensor (“Time-Space-Gene”) decompositon.

Table S4. Predicted LR interactions between spatial-separated epithelial cells C3 and other cell types by CellChat and SOAPy.

Figure 1

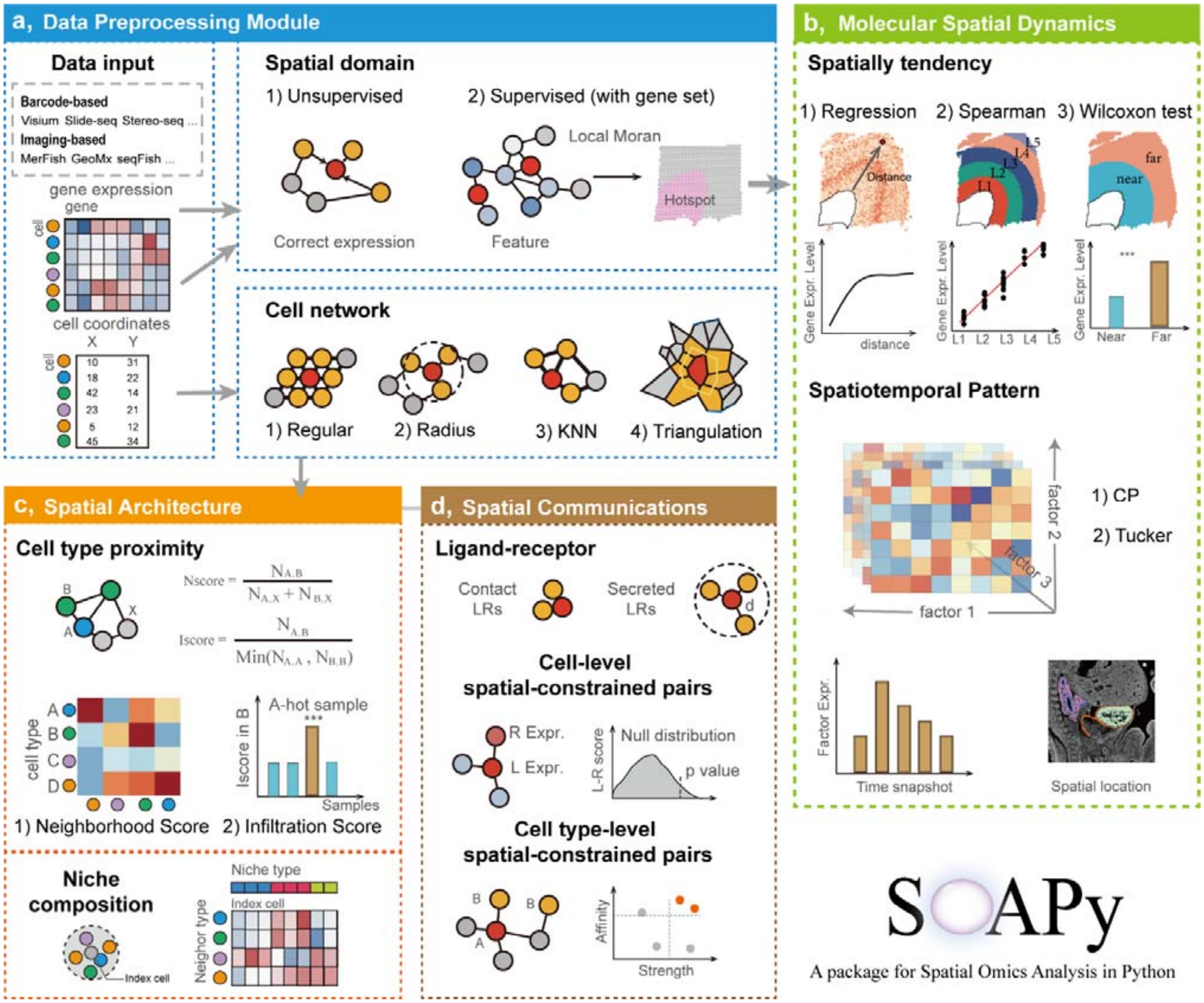
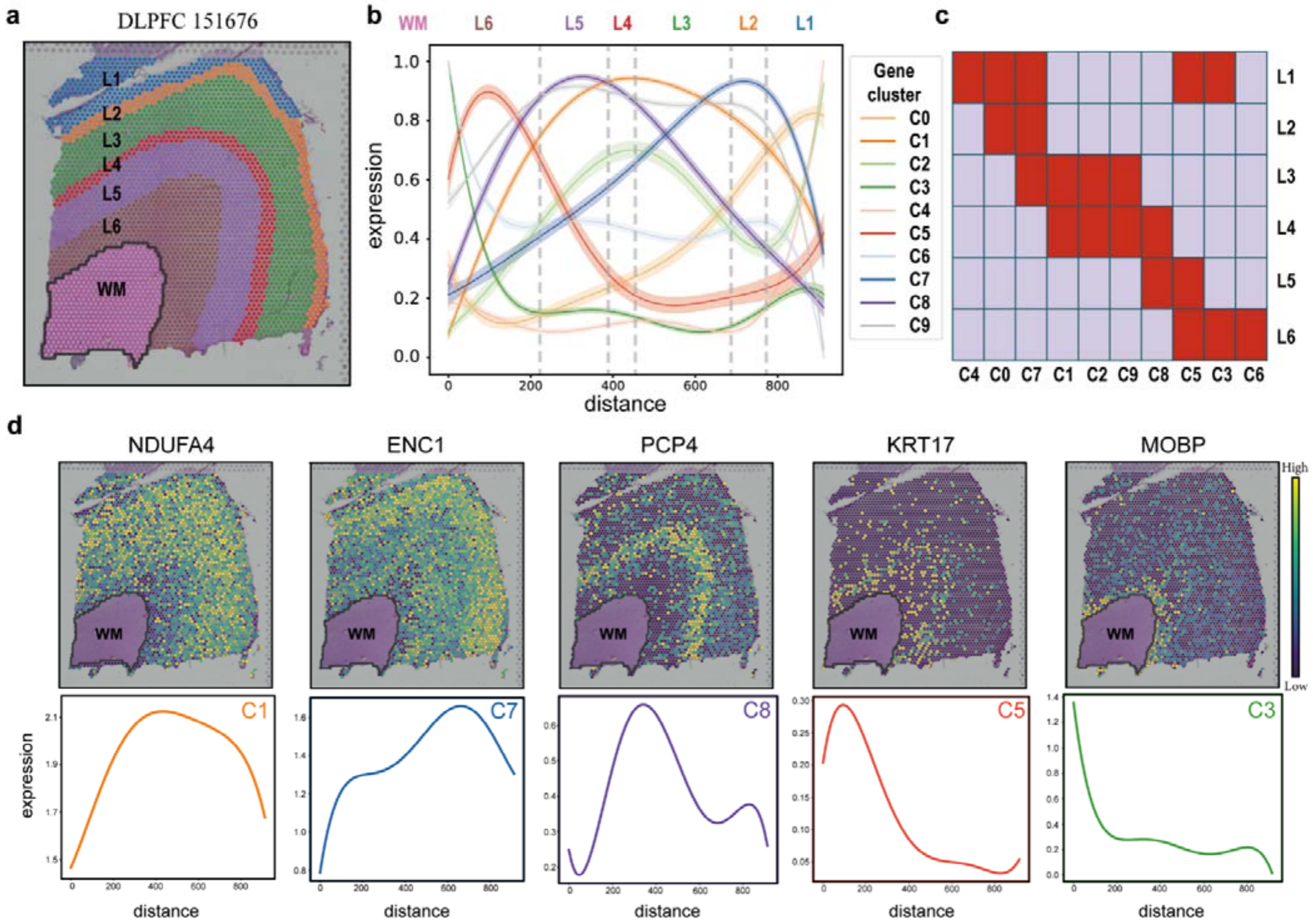

a, Data Preprocessing Module
Data input
Barcode-based
Visium Slide-seq Stereo-seq ...
Imaging-based
MerFish GeoMx seqFish ...
gene expression
gene
cell
cell coordinates
X Y
10 31
18 22
42 14
23 21
5 12
45 34
Spatial domain
1) Unsupervised
2) Supervised (with gene set)
Local Moran
Correct expression
Feature
Hotspot
Cell network
1) Regular
2) Radius
3) KNN
4) Triangulation
b, Molecular Spatial Dynamics
Spatially tendency
1) Regression
2) Spearman
3) Wilcoxon test
Distance
L1 L2 L3 L4 L5
far
near
Gene Expr. Level
distance
Near
Far
***
Spatiotemporal Pattern
factor 1
factor 2
factor 3
1) CP
2) Tucker
Factor Expr.
Time snapshot
Spatial location
c, Spatial Architecture
Cell type proximity
B
X
A
Nscore = $\frac{N_{A,B}}{N_{A,X}+N_{B,X}}$
Iscore = $\frac{N_{A,B}}{Min(N_{A,A}, N_{B,B})}$
cell type
A
B
C
D
A-hot sample
***
Iscore in B
Samples
1) Neighborhood Score
2) Infiltration Score
Niche composition
Index cell
Niche type
Neighbor type
d, Spatial Communications
Ligand-receptor
Contact LRs
Secreted LRs
d
Cell-level spatial-constrained pairs
R Expr.
L Expr.
Null distribution
L-R score
p value
Cell type-level spatial-constrained pairs
B
B
A
Affinity
Strength
SOAPy
A package for Spatial Omics Analysis in Python

Figure 2



	C4	C0	C7	C1	C2	C9	C8	C5	C3	C6
L1	🟥	🟥	🟥					🟥	🟥	
L2		🟥	🟥							
L3			🟥	🟥	🟥	🟥				
L4				🟥	🟥	🟥	🟥			
L5							🟥	🟥		
L6								🟥	🟥	🟥

Figure 3

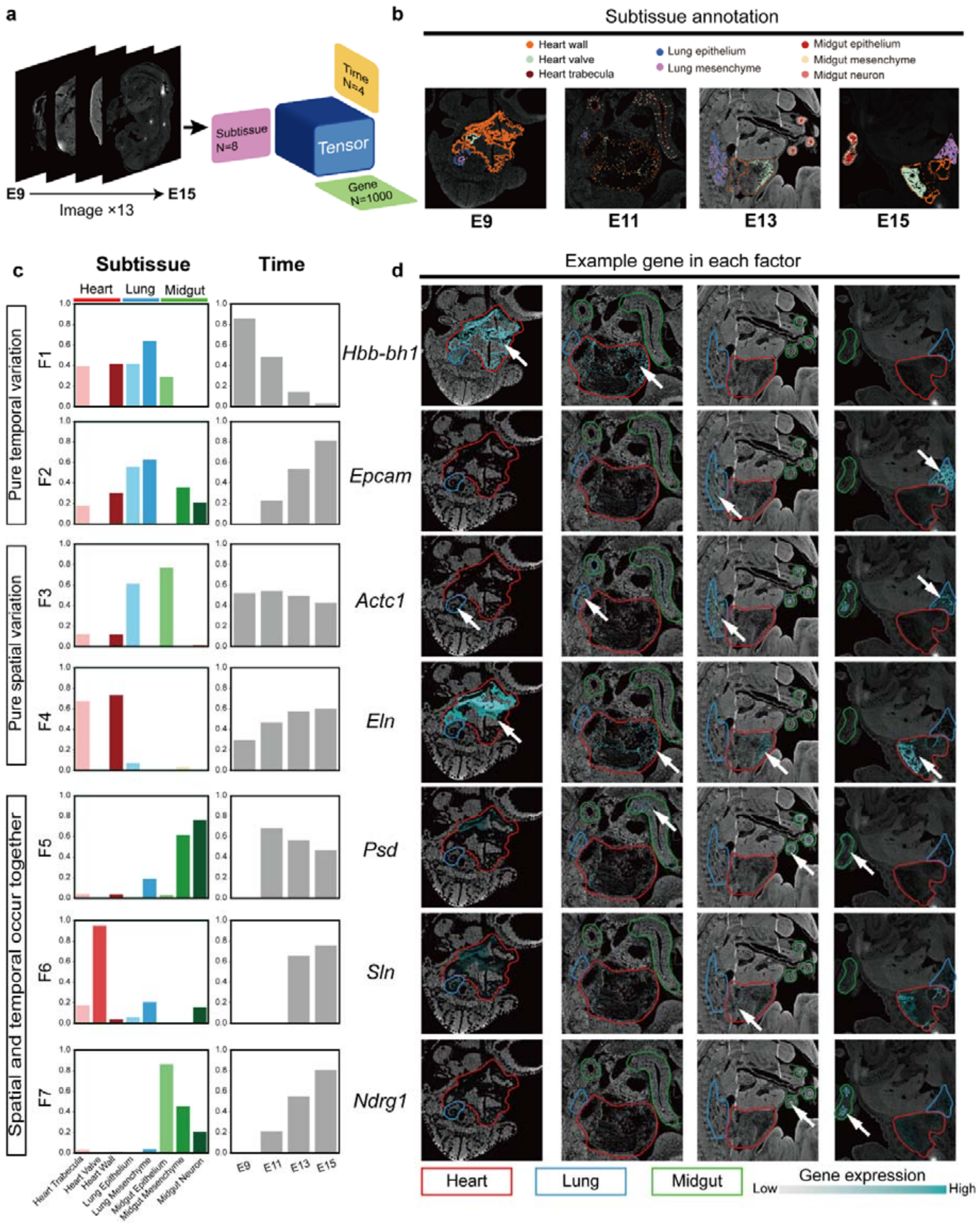

a
E9
E15
Image ×13
Subtissue N=8
Tensor
Time N=4
Gene N=1000
b
Subtissue annotation
Heart wall
Heart valve
Heart trabecula
Lung epithelium
Lung mesenchyme
Midgut epithelium
Midgut mesenchyme
Midgut neuron
E9
E11
E13
E15
c
Subtissue
Time
Heart
Lung
Midgut
Pure temporal variation
Pure spatial variation
Spatial and temporal occur together
F1
F2
F3
F4
F5
F6
F7
0.0
0.2
0.4
0.6
0.8
1.0
Heart Trabecula
Heart Valve
Heart Wall
Lung Epithelium
Lung Mesenchyme
Midgut Epithelium
Midgut Mesenchyme
Midgut Neuron
E9
E11
E13
E15
d
Example gene in each factor
Hbb-bh1
Epcam
Actc1
Eln
Psd
Sln
Ndrg1
Heart
Lung
Midgut
Gene expression
Low
High

Figure 4

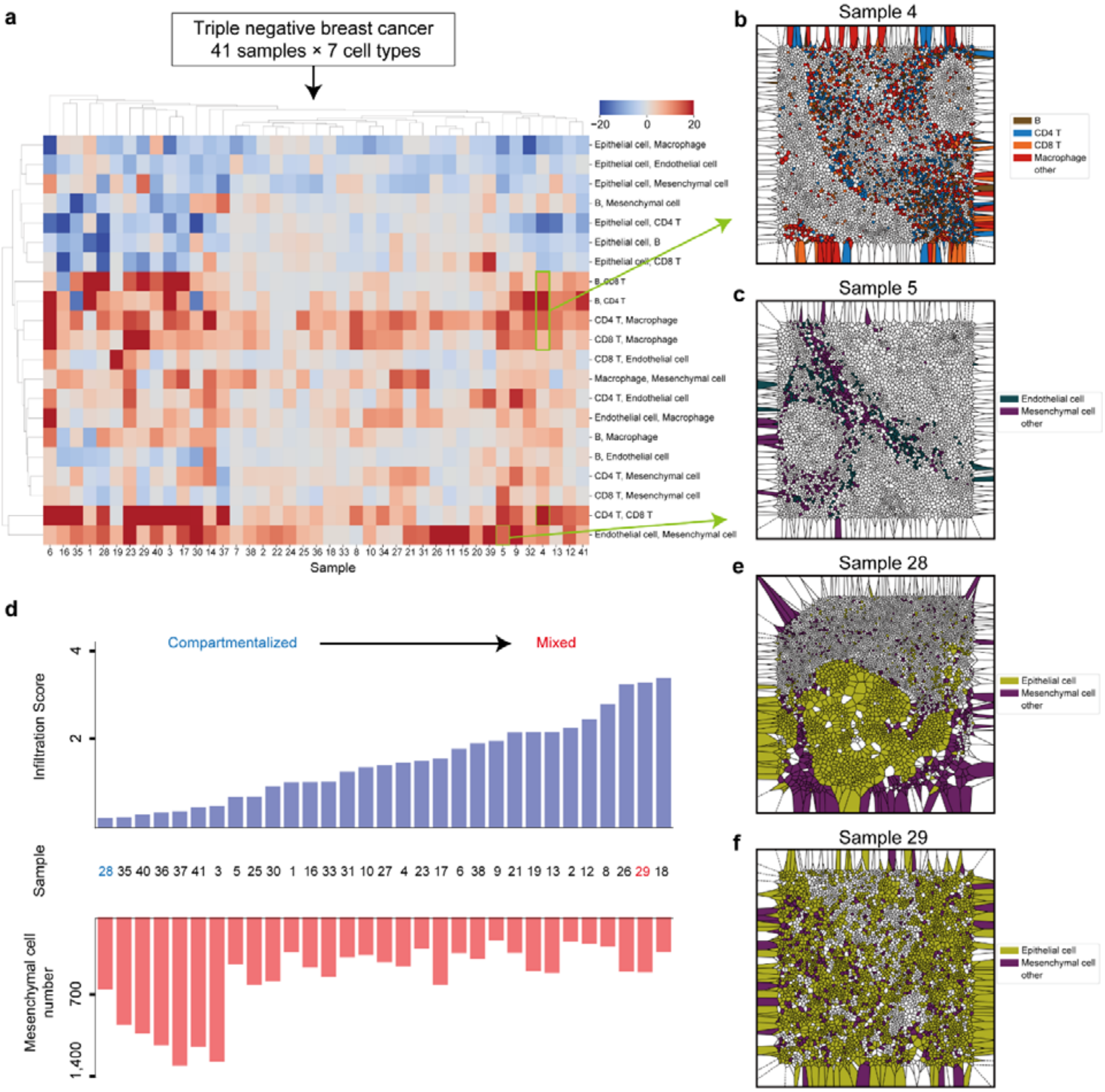

a
Triple negative breast cancer
41 samples × 7 cell types
−20
0
20
Epithelial cell, Macrophage
Epithelial cell, Endothelial cell
Epithelial cell, Mesenchymal cell
B, Mesenchymal cell
Epithelial cell, CD4 T
Epithelial cell, B
Epithelial cell, CD8 T
B, CD8 T
B, CD4 T
CD4 T, Macrophage
CD8 T, Macrophage
CD8 T, Endothelial cell
Macrophage, Mesenchymal cell
CD4 T, Endothelial cell
Endothelial cell, Macrophage
B, Macrophage
B, Endothelial cell
CD4 T, Mesenchymal cell
CD8 T, Mesenchymal cell
CD4 T, CD8 T
Endothelial cell, Mesenchymal cell
6 16 35 1 28 19 23 29 40 3 17 30 14 37 7 38 2 22 24 25 36 18 33 8 10 34 27 21 31 26 11 15 20 39 5 9 32 4 13 12 41
Sample
b
Sample 4
B
CD4 T
CD8 T
Macrophage
other
c
Sample 5
Endothelial cell
Mesenchymal cell
other
d
Compartmentalized
Mixed
Infiltration Score
4
2
Sample
28 35 40 36 37 41 3 5 25 30 1 16 33 31 10 27 4 23 17 6 38 9 21 19 13 2 12 8 26 29 18
Mesenchymal cell number
700
1,400
e
Sample 28
Epithelial cell
Mesenchymal cell
other
f
Sample 29
Epithelial cell
Mesenchymal cell
other

Figure 5

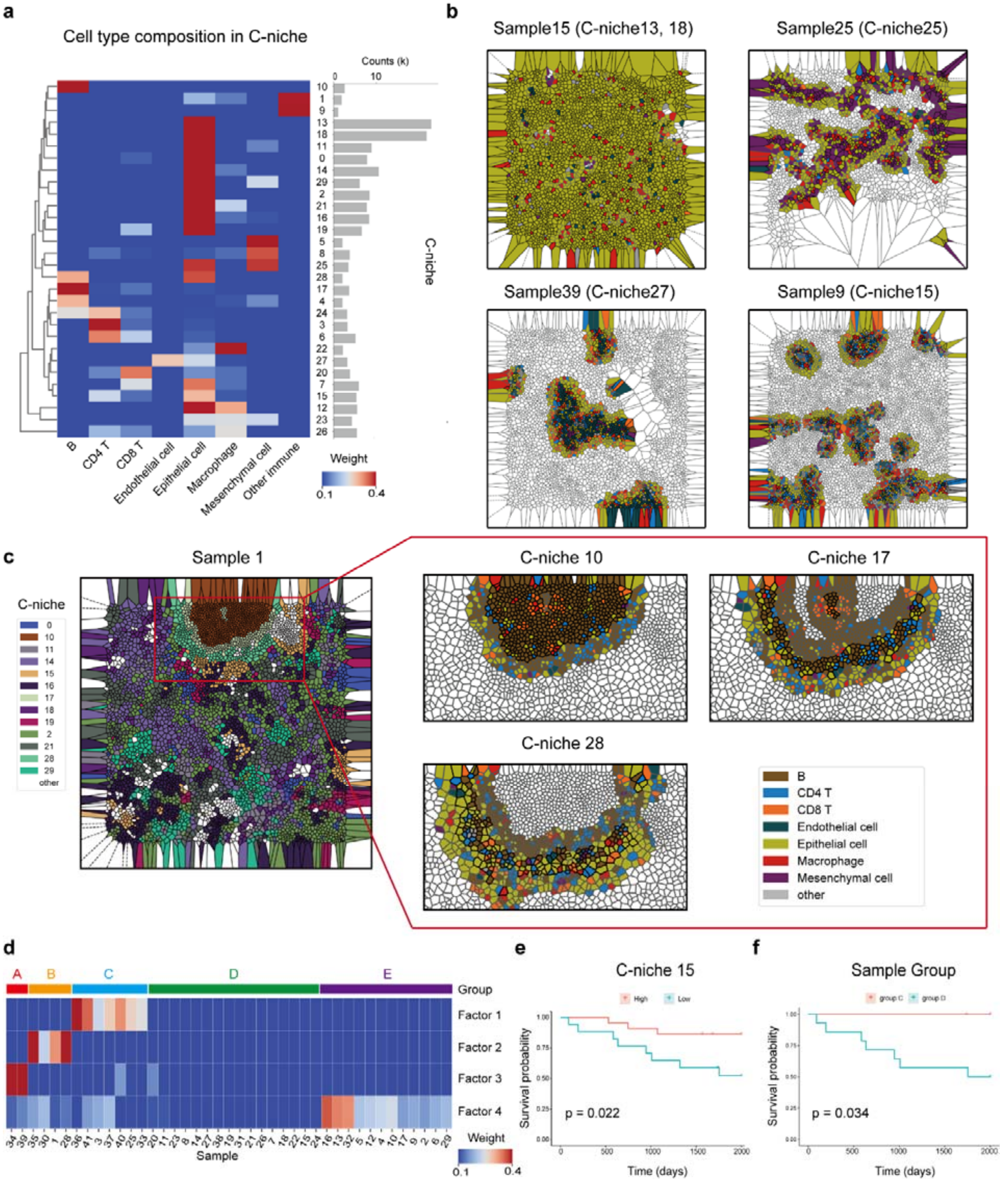
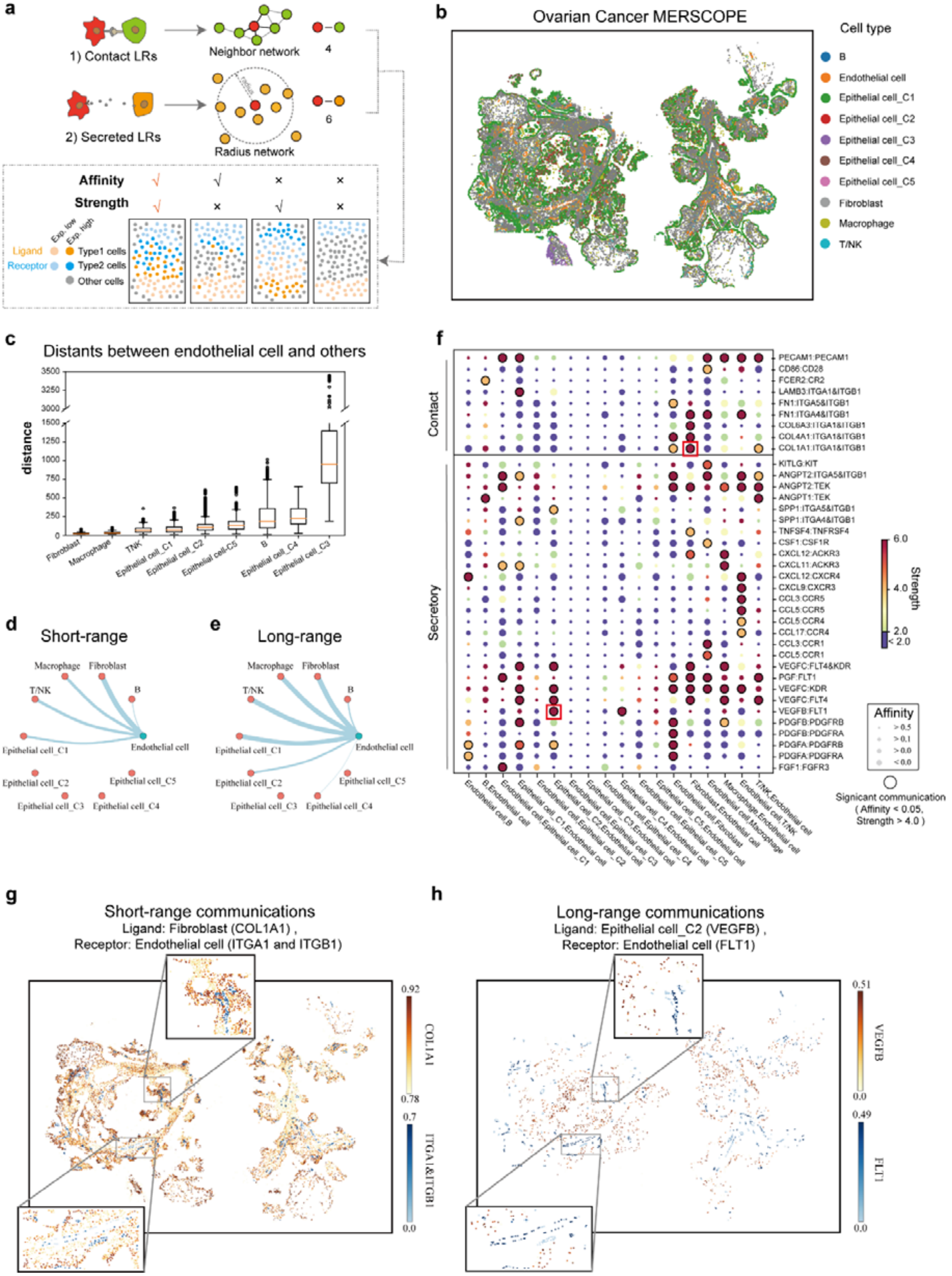

a
Cell type composition in C-niche
Counts (k)
0
10
10
1
9
13
18
11
0
14
29
2
21
16
19
5
8
25
28
17
4
24
3
6
22
27
20
7
15
12
23
26
C-niche
B
CD4 T
CD8 T
Endothelial cell
Epithelial cell
Macrophage
Mesenchymal cell
Other immune
Weight
0.1
0.4
b
Sample15 (C-niche13, 18)
Sample25 (C-niche25)
Sample39 (C-niche27)
Sample9 (C-niche15)
c
Sample 1
C-niche
0
10
11
14
15
16
17
18
19
2
21
28
29
other
C-niche 10
C-niche 17
C-niche 28
B
CD4 T
CD8 T
Endothelial cell
Epithelial cell
Macrophage
Mesenchymal cell
other
d
A
B
C
D
E
Group
Factor 1
Factor 2
Factor 3
Factor 4
34
39
35
30
1
28
36
41
3
37
40
25
33
20
11
23
8
14
27
38
19
31
21
26
7
18
22
15
24
16
13
32
5
12
4
10
17
9
2
6
29
Sample
Weight
0.1
0.4
e
C-niche 15
High
Low
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.022
0
500
1000
1500
2000
Time (days)
f
Sample Group
group C
group D
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.034
0
500
1000
1500
2000
Time (days)

Figure 6


a
1) Contact LRs
Neighbor network
4
2) Secreted LRs
Radius network
6
Affinity
Strength
√
×
Exp. low
Exp. high
Ligand
Receptor
Type1 cells
Type2 cells
Other cells
b
Ovarian Cancer MERSCOPE
Cell type
B
Endothelial cell
Epithelial cell_C1
Epithelial cell_C2
Epithelial cell_C3
Epithelial cell_C4
Epithelial cell_C5
Fibroblast
Macrophage
T/NK
c
Distants between endothelial cell and others
distance
3500
3250
3000
1500
1250
1000
750
500
250
0
Fibroblast
Macrophage
TNK
Epithelial cell_C1
Epithelial cell_C2
Epithelial cell-C5
B
Epithelial cell_C4
Epithelial cell_C3
d
Short-range
e
Long-range
Macrophage
Fibroblast
T/NK
B
Endothelial cell
Epithelial cell_C1
Epithelial cell_C2
Epithelial cell_C5
Epithelial cell_C3
Epithelial cell_C4
f
Contact
Secretory
PECAM1:PECAM1
CD86:CD28
FCER2:CR2
LAMB3:ITGA1&ITGB1
FN1:ITGA5&ITGB1
FN1:ITGA4&ITGB1
COL6A3:ITGA1&ITGB1
COL4A1:ITGA1&ITGB1
COL1A1:ITGA1&ITGB1
KITLG:KIT
ANGPT2:ITGA5&ITGB1
ANGPT2:TEK
ANGPT1:TEK
SPP1:ITGA5&ITGB1
SPP1:ITGA4&ITGB1
TNFSF4:TNFRSF4
CSF1:CSF1R
CXCL12:ACKR3
CXCL11:ACKR3
CXCL12:CXCR4
CXCL9:CXCR3
CCL3:CCR5
CCL5:CCR5
CCL5:CCR4
CCL17:CCR4
CCL3:CCR1
CCL5:CCR1
VEGFC:FLT4&KDR
PGF:FLT1
VEGFC:KDR
VEGFC:FLT4
VEGFB:FLT1
PDGFB:PDGFRB
PDGFB:PDGFRA
PDGFA:PDGFRB
PDGFA:PDGFRA
FGF1:FGFR3
Endothelial cell,B
B,Endothelial cell
Endothelial cell,Epithelial cell_C1
Epithelial cell_C1,Endothelial cell
Endothelial cell,Epithelial cell_C2
Epithelial cell_C2,Endothelial cell
Endothelial cell,Epithelial cell_C3
Epithelial cell_C3,Endothelial cell
Endothelial cell,Epithelial cell_C4
Epithelial cell_C4,Endothelial cell
Endothelial cell,Epithelial cell_C5
Epithelial cell_C5,Endothelial cell
Endothelial cell,Fibroblast
Fibroblast,Endothelial cell
Endothelial cell,Macrophage
Macrophage,Endothelial cell
Endothelial cell,T/NK
T/NK,Endothelial cell
6.0
4.0
2.0
< 2.0
Strength
Affinity
> 0.5
> 0.1
> 0.0
< 0.0
Signicant communication
(Affinity < 0.05,
Strength > 4.0)
g
Short-range communications
Ligand: Fibroblast (COL1A1) ,
Receptor: Endothelial cell (ITGA1 and ITGB1)
0.92
COL1A1
0.78
0.7
ITGA1&ITGB1
0.0
h
Long-range communications
Ligand: Epithelial cell_C2 (VEGFB) ,
Receptor: Endothelial cell (FLT1)
0.51
VEGFB
0.0
0.49
FLT1
0.0