Genome Biology

METHOD

Open Access

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

Heqi Wang[1†], Jiarong Li[1†], Siyu Jing[1], Ping Lin[1], Yiling Qiu[1], Xi Yan[1], Jiao Yuan[1], ZhiXuan Tang[1], Yu Li[2], Haibing Zhang[2], Yujie Chen[1], Zhen Wang[1] and Hong Li[1*]

†Heqi Wang and Jiarong Li contributed equally to this work.

*Correspondence: lihong01@sinh.ac.cn

[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 200031, China

Abstract

Advances in spatial omics enable deeper insights into tissue microenvironments while posing computational challenges. Therefore, we developed SOAPy, a comprehensive tool for analyzing spatial omics data, which offers methods for spatial domain identification, spatial expression tendency, spatiotemporal expression pattern, cellular co-localization, multi-cellular niches, cell–cell communication, and so on. SOAPy can be applied to diverse spatial omics technologies and multiple areas in physiological and pathological contexts, such as tumor biology and developmental biology. Its versatility and robust performance make it a universal platform for spatial omics analysis, providing diverse insights into the dynamics and architecture of tissue microenvironments.

Keywords: Spatial omics, Python package, Microenvironment, Expression pattern, Multi-cellular niche, Cell communication

Background

Spatially resolved transcriptomics was recognized as *Nature Methods* "Method of the Year" in 2020 [1]. Since then, an increasing number of experimental methods for measuring the expression levels of genes, proteins, or metabolites in a spatial context have been developed. These technologies include barcode-based and imaging-based methods, which differ in resolution, accuracy, and throughput [2, 3]. The most widely used 10X Visium spatial transcriptomics measures thousands of genes in each 55-μm spot that typically contains 1–10 cells [4]. And imaging-based methods reach to microscopic resolution, such as MIBI-TOF [5] and PhenoCycler-Fusion [6], both detecting dozens of proteins at a subcellular resolution. Additionally, spatial multi-omics technologies that simultaneously measure multiple molecular types are emerging, e.g., NanoString GeoMx DSP for 18,000 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 preprocessing 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, 10] that do not require spatial information. For low-resolution spatial technologies, various deconvolution methods have been developed to impute the cell-type composition from a mixture of cells.

After preprocessing, downstream analyses, including identifying spatially variable genes [11–13], detecting spatial domains [14], and inferring genes or cell-subtypes associated with spatial localization, are largely independent of experimental technologies, focusing on spatial context, the key feature of spatial omics. Earlier algorithms were often designed for one specific task, while tools that fit in with various analysis tasks are becoming popular. Giotto is a pioneering tool that integrates a pre-processing pipeline similar to scRNA-seq analysis and offers modules for spatial pattern detection, cell neighborhood analysis, and interactive visualization [15]. Squidpy provides a scalable framework for analyzing spatial neighborhood graphs and images, along with interactive visualization tools [16]. Spateo focuses on spatiotemporal analysis, supporting spatial data across multiple time points and continuous slices [17]. STUtility [18] and stLearn [19] are tailored for 10X Visium data. STUtility is marked by its pre-processing workflows and various automated and manual functions for spatial and image data, whereas stLearn extends functionality with spatial trajectory and pseudotime analysis. Investigating the spatial organization of tissue microenvironment is an important application of spatial omics, which may provide new insights into various biological fields. However, the related analysis methods are scattered or lacking, so a package for integrative analysis of microenvironmental spatial organization is in an urgent need.

To address this problem, we presented a package SOAPy (Spatial Omics Analysis in Python) to perform multiple tasks for dissecting spatial organization, including spatial domains, spatial expression tendency, spatiotemporal expression patterns, co-localization of paired cell types, multi-cellular niches, and cell–cell communication. SOAPy improves on previous tools in three main areas (Additional file 2: 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) Modeling ligand-receptor communication in tissue spatially considering different interaction modes distinguishing between secretory and membrane-binding ligands. We also applied SOAPy to a wide range of public datasets to demonstrate its general applicability and interpretability. SOAPy would be a powerful and promising tool for spatial-omics microenvironment analysis in Python.

Results

Overview of the SOAPy package

SOAPy integrates and develops a suite of algorithms to investigate gene expression variability and cell type distribution heterogeneity in spatial omics. It features four main modules: ***Data Preprocessing***, ***Molecular Spatial Dynamics***, ***Cellular Spatial Architecture***, and ***Spatial Communication***.

The flexible ***Data Preprocessing*** module could construct spatial networks through four methods and identify spatial domain in unsupervised or supervised ways (Fig. 1a). The ***Molecular Spatial Dynamics*** module includes *Spatial Tendency* and *Spatiotemporal Pattern*, to discover the trend of gene expression spatially or in other complex dimensions (Fig. 1b). The ***Spatial Architecture*** module includes *Spatial Proximity* and *Spatial Composition*, which unravels the colocalization patterns of cell types with several methods (Fig. 1c). In the ***Spatial Communication*** module, we constructed models for contact ligand-receptor pairs and secretory ligand-receptor pairs respectively and differentiated cell communication modes under different chemical properties (Fig. 1d). In addition, SOAPy provides rich visualization capabilities for all of the analysis methods mentioned above.

To demonstrate the utility of SOAPy, seven state-of-the-art public datasets obtained from four technologies were analyzed (Additional file 3: Table S2). These datasets involve

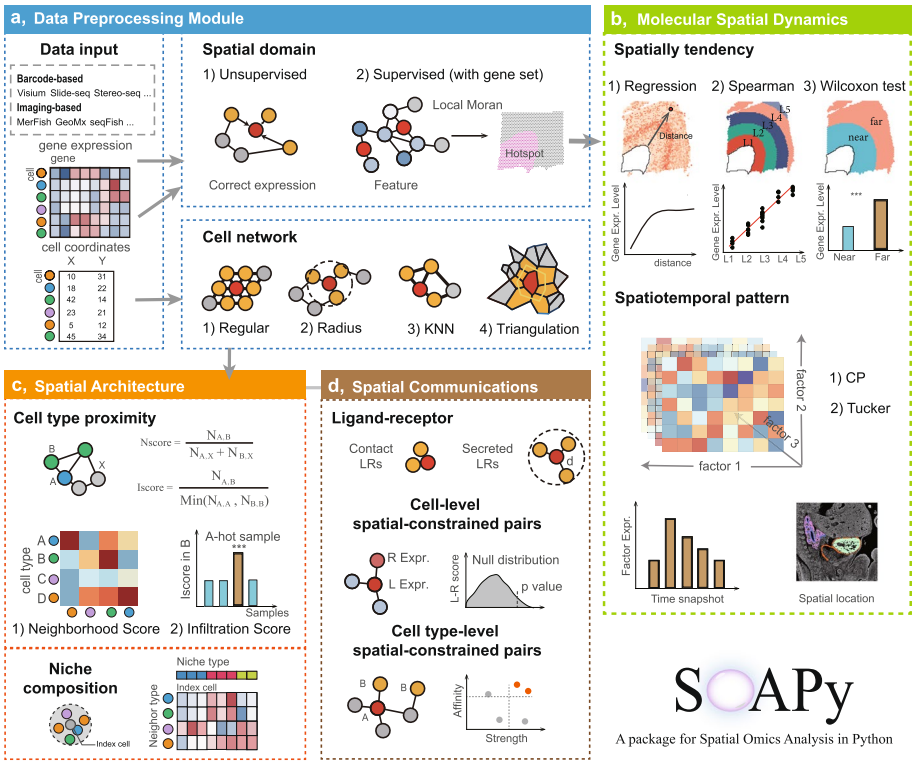

Fig. 1 Schematic diagram of SOAPy. **a Data Preprocessing** module that imports data, generates cell network, and identifies spatial domains. Data from different spatial omics technologies are converted to a unified data structure. Cell network could be built by any of the four methods. Spatial domains 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 changes with spatial distance to the given region. Spatiotemporal Pattern analysis performs a tensor decomposition to discover the major modes of variation in space and time. **c Spatial Architecture** module. Neighborhood and infiltration analysis find spatial proximal cell types. Spatial composition reveals conserved C-niches to delineate the cell type composition of the neighbors. **d Spatial Communication** module that combines 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

multiple scenarios with different molecular modalities (protein vs RNA), throughput (dozens to genome-wide), spatial resolution (0.1 ~ 55 μm), and in physiological and pathological states.

Spatial domain analysis recapitulates anatomic and pathological structures

Cells are not randomly distributed in the tissues. The self-organized structures built of various types of cells perform biological activities holistically, while these structures would become disordered in disease states. The *Spatial Domain* module provides three published unsupervised methods (STAGATE [14], GraphST [20], and SCAN-IT [21]) and one supervised method (AUCell-LMI) to detect these structures (called spatial domains) based on gene expression profiles and spatial locations [22, 23].

STAGATE, GraphST, and Scan-IT are tools that perform well in the previous benchmark [24]. We first tested these methods on 10 × Visium spatial transcriptomic data for human breast cancer [25] (Additional file 4: Table S3). In the binary classification task, STAGATE is the only method to successfully separate the malignant region from the non-malignant region (Additional file 1: Fig. S1a). In the multi-class classification task, all of the three methods have achieved good performance (Additional file 1: Fig. S1b). The results showed that supervised AUCell-LMI, based on known TLS signatures [26], was more accurate and convenient for identifying TLS regions than the unsupervised STAGATE method (Additional file 1: Fig. S1c, d). In summary, the *Spatial Domain* module in SOAPy effectively extracts physiologically or pathologically relevant structures for downstream analysis.

Spatial tendency analysis finds genes associated with spatial structures

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

We used a 10X Visium dataset of mouse dorsolateral prefrontal cortex (DLPFC) [27] as an example to validate the feasibility of spatial tendency estimation (Fig. 2a). The sample is consisted of the gray matter of DLPFC (including six cortical layers) and white matter (Additional file 1: Fig. S2a). To find genes whose expression changes along with the distance to the white matter, three strategies were used and compared [28] (Additional file 1: Fig. S2b–d): (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 (Additional file 1: Fig. S2e). The regression model describes the continuous spatial variation of expression; therefore, it could find more complex spatial patterns than other methods [29], such as nonlinear "low–high-low" spatial pattern (Additional file 1: Fig. S2f). Next, we analyzed the expression patterns of

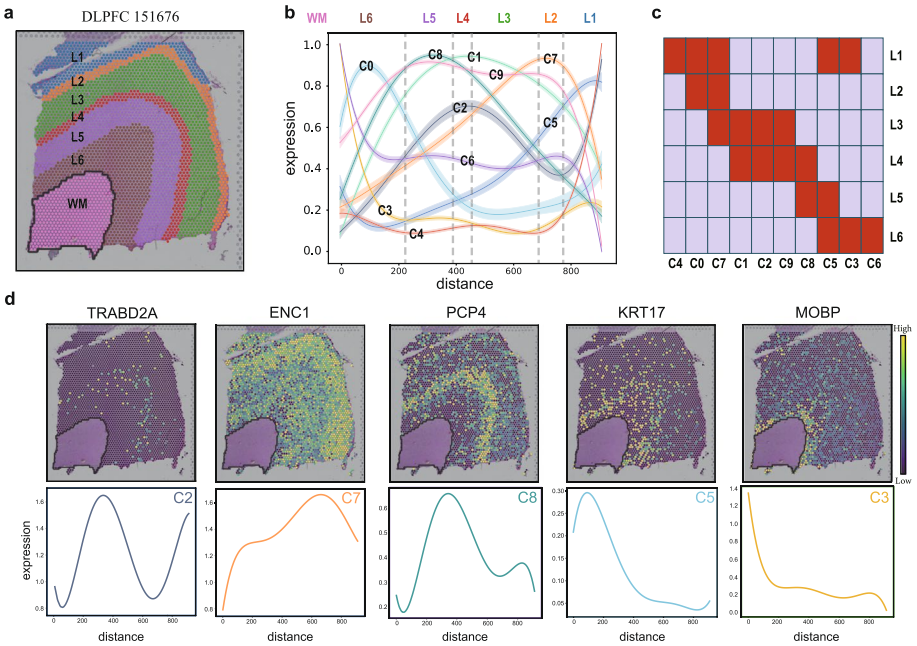

Fig. 2 Spatial tendency analysis finds genes associated with spatial structures. **a** HE image of the human dorsolateral prefrontal cortex (DLPFC) sample. Regions of white matter (WM) and six cortical layers (L6 to L1) are labeled on the image. **b** Regression curves of gene clusters between gene expression 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 presents 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 [30]. Each red unit indicates the cluster of genes (column) is enriched in the prior gene-list (row). **d** Spatial distributions (top) and fitted curves (bottom) of the representative genes

2857 significant (FDR < 0.05, range > 0.3) genes identified via polynomial regression. These genes were further grouped into 10 clusters by K-means clustering (Fig. 2b). The gene clusters were compared with previously reported cortical layer specific genes [30, 31] (Fig. 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, and 2, respectively (Fig. 2d).

Considering that there are no predetermined structures in some scenarios, we added three published methods (SpatialDE [11], SPARK [13], and SPARKX [12]) which identify spatial variable genes (SVGs) without ROI. Comparing these SVG methods with the abovementioned tendency estimation method, we found shared and specific genes among methods (Additional file 1: Fig. S2d). The SVG methods were more inclined to show the local differential expression of genes rather than the relationship with distance (Additional file 1: Fig. S2g). Users can select suitable methods based on their requirements.

Tensor decomposition reveals the spatiotemporal patterns of gene expression

With advancements in omics technologies, spatially resolved and time-series molecular profiling data are increasingly accessible. One of the challenges is how to study the roles of spatial effects and temporal effects simultaneously in biological issues. The *Spatiotemporal Pattern* function in SOAPy employs tensor decomposition to extract components from the three-order expression tensor ("Time–Space-Gene"), reducing the complexity of data explanation and revealing hidden biological patterns.

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 sub-tissues (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 (Fig. 3a, b). Canonical polyadic (CP) decomposition [32] was used to factorize the expression tensor with 1000 highly 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 contributions of time, sub-tissues, and genes (Fig. 3c). We observed three empirical spatiotemporal patterns based on the loadings of time and sub-tissues: pure temporal variation (F1, F2), pure spatial variation (F3, F4), spatial and temporal variation (F5, F6, F7). We also conducted functional enrichment analysis based on the loadings of genes for each factor (Additional file 5: Table S4) and visualized the typical genes in the images (Fig. 3d).

The genes in F1 (e.g., *Hbb-bh1*) were highly expressed in heart and lung sub-tissues at E9, and then gradually decrease in the later stages. The expression pattern of these genes is enriched in "regulation of vasculature development." F1 indicates co-development of heart and lung in the early embryo, which is consistent with previous study [33]. F2 indicates the presence of highly expressed genes at later embryonic stages (Fig. 3c), and *Epcam* (Fig. 3d) was chosen as a representative gene, as demonstrated in a previous study [34]. Expression of genes in F3 and F4 is stable during development. The genes in F4 highly express in the heart wall and heart trabecula, enriched in cardiac cell development as expected. The genes in F5 and F7 are enriched in midgut development. F5 (e.g., *Psd*) slightly decreases from E11 to E15, while F7 (e.g., *Ndrg1*) clearly increases from E11 to E15. The genes in F7 is highly expressed specifically in the midgut epithelium between E11 and E15, which has also been reported in previous study [35].

We applied this function to investigate liver regeneration across lobular zones following acute injury [36]. By averaging gene expression of the spots in each time point after injury and lobular zone to construct a 3D tensor (Additional file 1: Fig. S3a), we performed tensor decomposition to identify six distinct factors (Additional file 1: Fig. S3b, c). We calculated the contribution of each factor to gene expression. F6 indicates genes related to 48 h in pericentral vein (PV) zone, and *Sds* as a marker gene of the PV zone was highly accounted for by this factor. F1 and F5 represent the periportal vein (CV) zone for different time points. *Glul*, a marker gene of the CV zone, expresses in both F1 and F5 [29]. *Il11*, *Tagln*, and *Acta2* are mainly expressed in F5 that is dominated in the early stage after acute injury of the CV zone. This phenomenon may be related to immune modulators in the CV zone and activation of hepatic stellate cell (HSC). *Col1a1*, the gene expressed in F1, is associated with expression of extracellular matrix, in line with the required matrix buildup in the CV zone at 48 h [36].

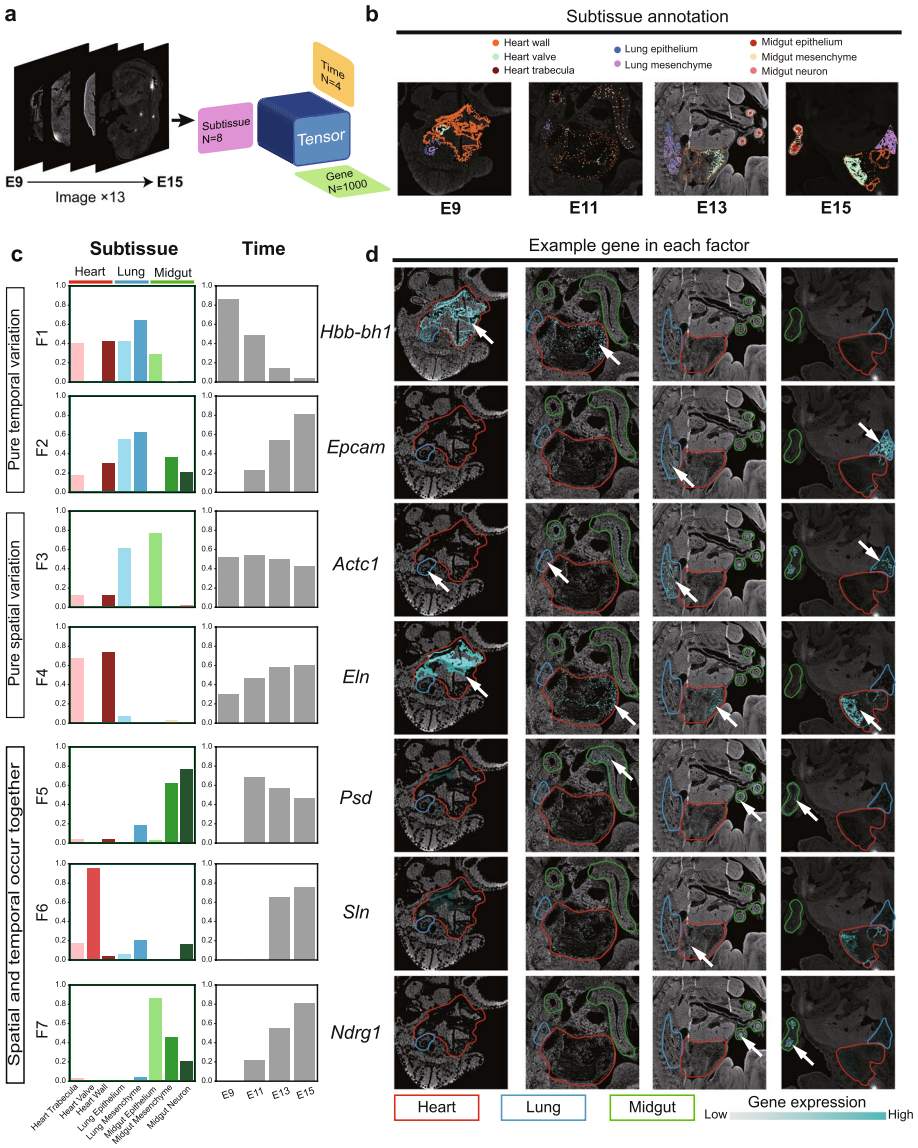

Fig. 3 Tensor decomposition reveals the spatiotemporal patterns of gene expression during mouse embryo development. **a** The schematic of spatiotemporal dataset of mouse development is represented by a three-order tensor (4 time points * 8 sub-tissues * 1000 highly variable genes), and then it is decomposed into seven latent factors. **b** Spatial locations of the sub-tissues at four time points. Each spot in the sub-tissues represents an ROI. **c** Loading vectors of space and time for each factor obtained by tensor decomposition. Higher loading values indicate larger contribution of the 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, respectively. ROIs of gene expression are presented by cyan points. The darker the cyan color, the higher the gene expression level

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

The spatial architecture of cells is important for understanding the organization patterns from single cells to tissues [37–39]. SOAPy first constructs a cell/spot network from

spatial locations and then implements two scenarios for deciphering spatial architecture: (1) *Spatial Proximity* analysis (including neighborhood and infiltration) determines whether two cell types or cell states within an image are significant proximal; (2) *Spatial Composition* analysis identifies multi-cellular niches composed of cell types with specific proportion.

We applied this analysis to a dataset of 41 triple-negative breast cancer (TNBC) patients [5] and used multiplexed ion beam imaging by time-of-flight (MIBI-TOF) to simultaneously quantify the expression of 36 proteins in situ at sub-cellular resolution. In total, 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 with random perturbation [38]. Figure 4a shows the neighborhood

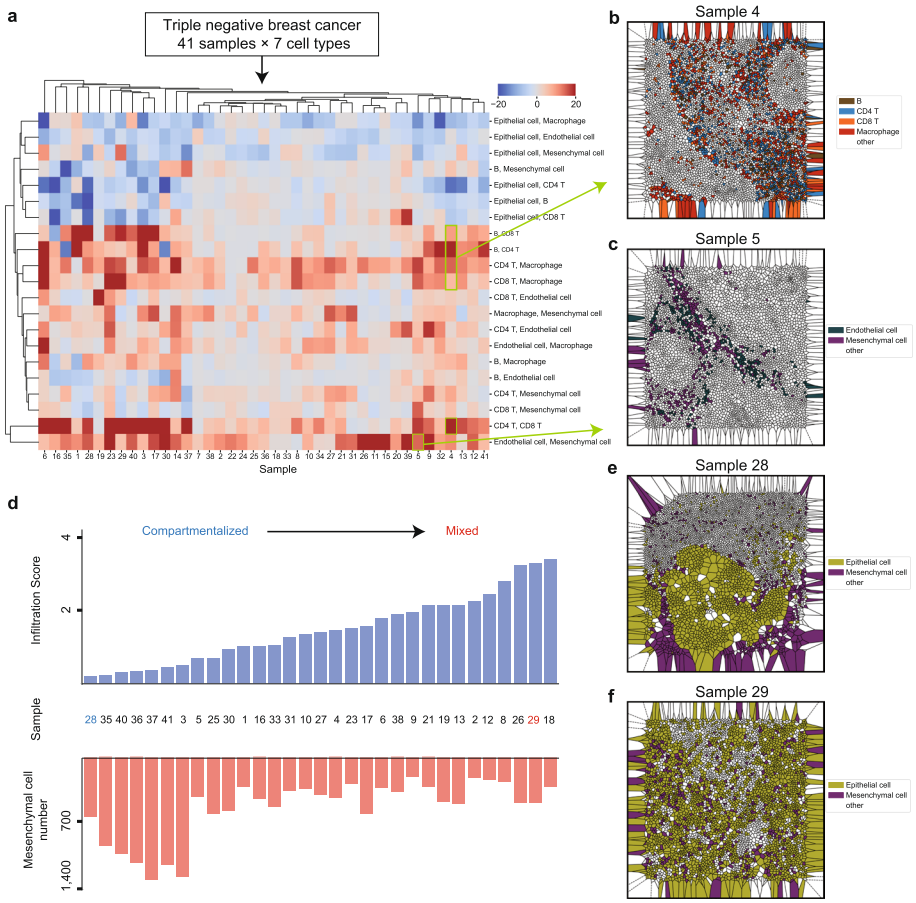

Fig. 4 Spatial proximity analysis characterizes cellular co-localization patterns. The triple-negative breast cancer (TNBC) dataset contains 41 samples and 7 cell types. **a** Heatmap showing the neighborhood scores of any two cell types in all TNBC samples. **b** A representative sample with strong 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 and malignant epithelial cells

scores of all the samples for all the cell type pairs, with positive or negative scores corresponding to co-localization or avoidance. Different immune cell types such as B cells, CD4 T cells, CD8 T cells, and macrophages are significantly colocalized in many patients, which may be related with the formation of inflammatory foci (Fig. 4b). Endothelial and mesenchymal cells also prefer to co-locate together (Fig. 4c). Colocalization pattern of malignant epithelial cells and non-parenchymal cells is highly heterogeneous across patients. Taking malignant epithelial cells and mesenchymal cells as an example, samples with fewer than 200 mesenchymal cells were filtered out, while the remaining samples were subjected to *Spatial Infiltration* analysis. Samples with higher and lower infiltration scores indicate compartmentalized (e.g., sample 28) and mixed (e.g., sample 29) patterns between malignant epithelial cells and mesenchymal cells, respectively (Fig. 4d–f).

Spatial composition analysis discovers multi-cellular niches

For *Spatial Composition* analysis of the TNBC dataset, a cell–cell network that connected the centroids of the cells within 100 pixels was built to capture the composition pattern of surrounding cells. The niche of each cell is represented by the proportion of the 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 (Fig. 5a, Additional file 1: Fig. S4a). The major cell types of the top two C-niches (C-niche13 and C-niche18) are mainly malignant epithelial cells, and the percentages of other cell types are less than 15%, indicating the characteristics of tumor cell aggregation (Fig. 5b). Additionally, epithelial cells also form C-niches with other cell types. For example, C-niche25 is composed of 38% epithelial cells, 38% mesenchymal cells, and 9% macrophages; C-niche27 is composed of 23% epithelial cells, 28% endothelial cells, 10% mesenchymal cells, and 10% macrophages; and C-niche15 is composed of 30% epithelial cells, 23% CD4 T cells, 13% CD8 T cells, and 11% macrophages, suggesting different local microenvironments exist among tumors. We also observed four B cell dominated C-niches (C-niche10, C-niche17, C-niche28, and C-niche4) that may be related to TLS (tertiary lymphoid structure). For example, sample 1 contains C-niche 10, 17, and 28 (Fig. 5c). Approximately 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 primarily consists of 31% B cells, 10% CD8 T cells, and 37% epithelial cells. Patients were divided into two groups for survival analysis on the basis of the abundance levels of these four C-niches. Those with higher levels of these C-niches presented significantly longer survival time [40].

To investigate the combined 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 (Fig. 5d). Sample groups A, B, C, and E have the highest loadings in factors 3, 2, 1, and 4, respectively. By scrutinizing the loadings of cell types and niches in the major factors (Additional file 1: Fig. S4b, c), group B corresponds to the B cell enriched samples mentioned above; group C is characterized by niches with high proportion of mesenchymal cells; and group E has niches consisting of T cells and macrophages.

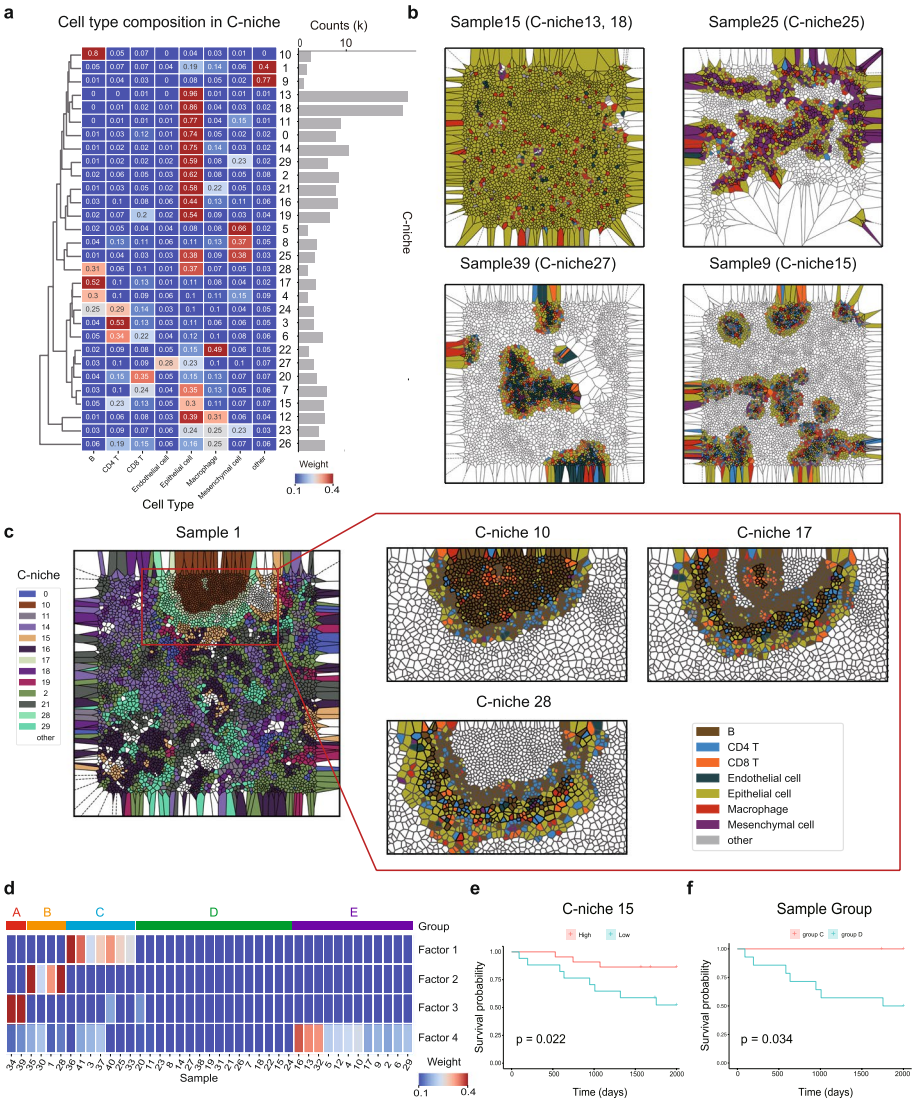

Fig. 5 Spatial composition analysis discovers multi-cellular niches in TNBC samples. **a** Heatmap on the left shows the composition of neighbor cells in each C-niche. The right bar plot shows the number of cells belonging to each C-niche. **b** Representative samples of different C-niches, characterizing tumor cell aggregation and different local microenvironment of tumors. **c** The left image shows an example that contains B cell dominated C-niches (the region of red box). Cells are colored by C-niches. "other" are infrequent C-niches with proportion less than 2%. The 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 colors of cells represent cell types involved in the definition of the C-niche. **d** Heatmap showing the loading values and the clusters of the samples. The three-order "Niche-CellType-Sample" tensor was decomposed to four latent factors (Additional file 1: Fig. S3b, c). Samples are clustered into five groups according to the loading vectors. **e** Survival curves stratified by the proportion of C-niche 15. The stratification standard is determined by "maxstat" package. **f** Comparison of survival curves between the patients from group C and group D

Furthermore, survival analysis was performed to explore the clinical indications of niches. Eight C-niches were significantly related to survival time ($P < 0.05$, Additional file 1: Fig. S5). For example, patients with a higher proportion of C-niche15 had a longer survival time (Fig. 5e). There also are survival differences among the

patient groups identified by the "Niche-CellType-Sample" tensor decomposition, such as longer survival time for group C patients than that of group D (Fig. 5f). Previous papers reported that fibroblasts may play an important role in immune regulation, which activates tumor immunity and leads to better survival [41–43]. 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 communication

Besides spatial architecture analysis of cell types, delineating the molecular communication between cells is also an important task for spatial-omics data analysis. Popular expression-based methods such as CellphoneDB [44] and CellChat [45] predict cell–cell communication by analyzing the expression of ligands and receptor (LR) pairs, but they fail to account for spatial proximity. In addition, membrane-binding ligands or the extracellular matrix could only target to the cells in direct contact, which is called juxtacrine signaling, while secretory ligands could be secreted to extracellular microenvironment and deliver signals in the local environment, which is called autocrine or paracrine signaling [46]. Different communication modes caused by ligand characteristics were not taken into consideration by existing spatial modeling tools such as Giotto, stLearn, and Squidpy. Here, SOAPy develops a new method that simultaneously utilizes spatial location and gene expression to calculate communication scores (affinity and strength) to identify significant LR communications (Fig. 6a, Methods). "Affinity" reflects the proximity of the LR signal between cells, whereas "Strength" is a comprehensive indicator related to the proximity and ligand/receptor expression level of two cell types. By leveraging the relationship between "Affinity" and "Strength," users could effectively identify LR pairs in the spatial context that are most biologically relevant for their specific context. In addition, the model is adaptable for the two communication modes mentioned above. For contact-dependent communication, we applied a short-range communication mode, which is limited

(See figure on next page.)

Fig. 6 Ligand-receptor-mediated and spatial-constrained cell–cell communications. **a** The brief flow chart of our method. Short-range communication is mediated by contact LRs on neighboring cells; long-range communication is mediated by secreted LRs on cells within a defined radius. Two new metrics, affinity and strength, are defined to estimate the probability of LR communications 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** Spatial distribution of ten cell types in an ovarian MERSCOPE sample. **c** Bar plot showing the shortest distance between cells belonging to all other cell types to their 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 indicates the number of significant ligand-receptor pairs (affinity P value < 0.05, strength > 4). **f** Dot plot with LR communications between endothelial cells and other cell types. Each row indicates one LR pair, with the first and the second genes representing ligand and 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. The expression level of the ligand, *COL1A1*, in fibroblasts is displayed by the darkness of the red color; and the expression level of the receptor, *ITGA1/ITGB1*, in endothelial cells is displayed by the darkness of the blue color. Expression levels were normalized to the range of 0–1. **h** An example of secreted LR, corresponding to the communication between slightly separated epithelial cells C2 and endothelial cells. The expression level of the ligand, *VEGFB*, in epithelial cells C2 is displayed by the darkness of the red color, and the expression level of the receptor, *FLT1*, is displayed by the darkness of the blue color. Expression levels were normalized to the range of 0–1

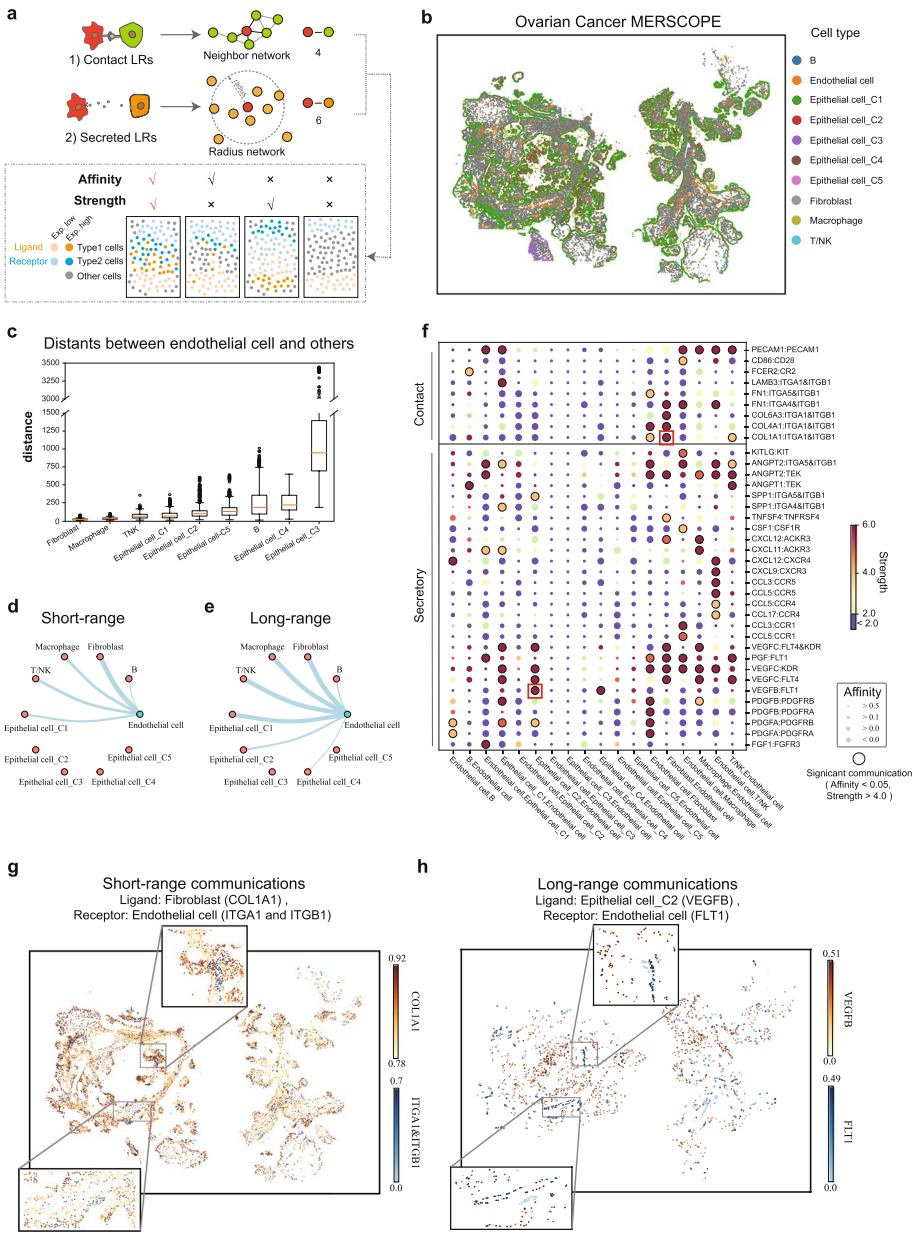

Fig. 6 (See legend on previous page.)

to a small distance. For secretory ligands, we proposed a long-range communication mode, in which the communication intensity decreases with the distance [47, 48].

The ***Spatial Communication*** module was applied to an ovarian cancer dataset generated by the MERSCOPE platform, measuring 500 genes in 71,381 cells (Fig. 6b). The cells were classified and annotated into ten types or subtypes via the Leiden clustering algorithm. The spatial locations of epithelial cells C3 are very special and clearly separated from those of most other cells. Therefore, our method did not detect significant LR pairs between epithelial cell C3 and other cell types located far away (Additional file 6: Table S5).

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

In total, we identified 19 contact LRs and 66 secreted LRs that may play key roles in short-range and long-range communication between endothelial cells and other cells (Fig. 6f). For example, *COL1A1* (type I collagen) and its receptor *ITGA1/ITGB1* (integrin α/β) are highly expressed on spatial adjacent fibroblasts and endothelial cells, and their affinity and strength scores are significantly higher than random scores (Fig. 6g). Previous studies have reported that the binding of collagen and integrin may activate downstream signaling pathways that contribute to cancer progression [49]. *VEGFB-FLT1* is an interesting LR pair for long-range communication between epithelial and endothelial cells (Fig. 6h). Epithelial cells C2 release ligand *VEGFB*, and endothelial cells highly express *FLT1* (also known as *VEGFR1*). Their communications may promote tumor angiogenesis and would be potential drug targets for anti-cancer therapy [50].

We compared the results of SOAPy with CellChat v2 [51] and Giotto [15], which are also developed based on spatial omics, on this MERSCOPE dataset. Although the false positives of the results from CellChat v2 are rather low, the identified LR pairs are still highly related to the cell types rather than the spatial location between cells. Both of CellChat v2 and Giotto detect a number of LR pairs between other cell types and epithelial cells C1 or C3, but epithelial cells C3 do not co-localize with epithelial cells C1 (Fig. 6b, Additional file 1: Fig. S6a, b). In our approach, both "Affinity" and "Strength" metrics are calculated to infer the communication between cell types. The results of SOAPy without filtering with "Strength" are similar to Giotto, but there are some differences between the contact and secretory modules (Additional file 1: Fig. S6c, d). As the filter criteria of "Strength" index is gradually raised, some ligands with significant "Affinity" score are filtered out (Additional file 1: Fig. S6e, f). For example, Giotto identified an LR pair of *COL1A1:ITGA1&ITGB1* between epithelial cells C1 and C3 (Giotto P value: 0.04, SOAPy-Affinity P value = 0.036), while it would be filtered out via setting the "Strength," consistent with the fact that epithelial cells C1 and C3 are not co-localized (Fig. 6b, Additional file 1: Fig. S7a).

The spatial patterns of the LRs can be divided into the following four conditions:

1) Significant "Affinity" and high "Strength": High expression of ligand and receptors in two cell types and predominantly localized at the junction of the two cell types, indicating strong and spatially relevant interactions (Additional file 1: Fig. S7b (2)).

2) Not significant "Affinity" but high "Strength": High expression of ligand and receptors in two cell types but not enriched in the junction of the two cell types, suggesting interactions that are less spatially constrained (Additional file 1: Fig. S7b (1)).
3) Significant "Affinity" but low "Strength": The cells at the junction have rather high expression of ligand and receptor, but the expression is low compared with the other cell types, or the adjacency of the two cell types is low (Additional file 1: Fig. S7b (4)).
4) Not significant "Affinity" and low "Strength": Ligand-receptor expression is negligible, indicating minimal or no interaction (Additional file 1: Fig. S7b (3)).

In summary, SOAPy provides a new way to study spatial-constrained cell–cell interactions and more accurately identify the related ligand-receptor pairs.

Discussion

The tissue microenvironment is critical for understanding homeostasis, development, regeneration, and disease. Single-cell and spatially resolved omics are the most promising technologies to investigate microenvironment. Tools for systematically dissecting microenvironment and discovering biologically important genes or spatial cellular architecture are in need, and SOAPy has just filled 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 communication. In this work, we demonstrated all SOAPy modules with various types of spatial omics data and provided complete tutorials to help users start quickly.

The spatial distribution of genes or cells is affected by various factors, such as time, interaction of cells, pathological foci, and sample heterogeneity. Given 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. Studies of mouse embryo development, liver regeneration, and breast cancer have shown that tensor decomposition in SOAPy is powerful for interpreting complex biological data. Another significant advantage of SOAPy is the innovative ***Spatial Communication*** module. It combines the spatial distance, expression level, and interaction mechanism of LRs to infer cell–cell communication. In the case of ovarian cancer showed that SOAPy could markedly reduce false positives of interacting ligand-receptors compared to existing methods.

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

Conclusions

SOAPy provides a powerful and flexible framework for analyzing spatial omics data. Four modules—***Data Preprocessing***, ***Molecular Spatial Dynamics***, ***Cellular Spatial Architecture***, and ***Spatial Communication***—offer alternations for dissecting multiple spatial

distribution patterns of genes and cells. Tested with several datasets from multiple spatial-omics technologies demonstrated its advantages on wide application and biological interpretation, making SOAPy a promising and competitive package for analyzing spatial omics data.

Methods

Data preprocessing

Data Import

The Data Import function converts data from different spatial omics technologys to a unified data structure that contains expression profiles of molecules (genes/proteins/metabolites) and location of cells/spots. Barcode-based data formats could be read directly by passing in tables representing expression matrix and spatial coordinate information. For imaging-based data, a multiplexed image and cell segmentation mask are needed, and the expression matrix and coordinate matrix would be extracted automatically. In this way, the raw data of spatial-omics are transferred to an Anndata object, which is adaptable for Scanpy package [10].

Spatial network construction

The Spatial network function provides four approaches to build a neighborhood network of cells/spots (Fig. 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

We have encapsulated three unsupervised methods of spatial domain in SOAPy. STAGATE is a graph attention autoencoder for spatial domain identification [14]. It integrates gene expression profiles and spatial location information to learn low-dimensional latent embedding. GraphST integrates GNNs and self-supervised comparative learning to effectively learn spot representations in spatial transcriptomics data by modeling both gene expression and spatial localization information [20]. SCAN-IT is developed by transforming the spatial domain identification problem into an image segmentation problem [21], with cells mimicking pixels and expression values of genes within a cell representing the color channels, and generates low-dimensional embeddings of the spots through the application of deep learning. After using any method to obtain the reduced features, the spatial domain is obtained by performing the clustering operation.

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 [52, 53], and then local Moran index [22] (LMI) is used to estimate the degree of spatial aggregation. LMI of cell/spot i is defined as:

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

where x_i is the AUCell score of cell/spot i, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, j is any neighbor cells/spots of i based on K nearest neighbors, and w_{ij} is the spatial weight between i and j. The P value is calculated by permutation test and adjusted by Benjamini–Hochberg method [54] to obtain the false discovery rate (FDR).

LMI of all cells/spots is illustrated by Moran scatterplot (Additional file 1: Fig. 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 are 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 (Additional file 1: Fig. S2a). Users could select ROI using tools like ImageJ to manually generate a mask file, or acquire from SOAPy *Spatial domain* analysis. Given an ROI, SOAPy creates the mask of ROI: discrete cells/spots are converted to continuously connected regions via 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 [29]:

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

where i is a cell/spot, C is the boundary of ROI, and p is any pixel on the boundary. $Enc()$ is a function of Euclidean distance. Distance with positive or negative signs is 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 (Additional file 1: Fig. S2b): (1) Wilcoxon rank sum test to compare the molecular 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 one parametric regression method (polynomial regression model) and one 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.

$$Y = a_0 + \sum_{k=1}^{n} a_k d^k \tag{3}$$

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; and 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 [38]:

$$NS_{ct1,ct2} = \frac{N_{ct1,ct2}}{N_{ct1,other}+N_{ct2,other}} \tag{4}$$

where $N_{ct1,ct2}$ is the number of direct connections between $ct1$ and $ct2$ and $N_{ct1,other}$ 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) cell infiltration into malignant tissues:

$$IS_{m,np} = \frac{N_{m,np}}{min(N_{m,m},N_{np,np})} \tag{5}$$

where $N_{m,np}$ is the number of direct connections between malignant cells and non-parenchymal cells. Samples 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, I-niche is defined as the proportion of cell types for its surrounding cells [55]. After taking all cells in one or more images, clustering algorithms like K-means divides I-niches into different clusters, called C-niches.

Spatial-constrained cell–cell communication inference

Ligand-receptor (LR) pairs were obtained from the CellChatDB [45], 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, which is (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. Paired cell types are ranked based on the number of significant LRs.

Cell-level ligand-receptor affinity score

The communication of LR is variable among cells/spots at different spatial locations; therefore, we first defined a cell-level ligand-receptor affinity score. Suppose a cell/spot i is a sender of ligand, and cells/spots that are connected to i and express the corresponding receptor are the receivers, the *Affinity score* of ligand-receptor at location i is defined as:

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

where j is the cell/spot that connects to i in the cell network; l and r are expression levels of the ligand and receptor, respectively; and 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:

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

The *Affinity Pvalue* is obtained by random permutation:

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

M is the total number of randomizations, and $A^{(m)}$ 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 pairs have communication.

CellType-level communication score

Supposing $ct1$ 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:

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

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

Affinity reflects whether spatial connected $ct1$ and $ct2$ have relatively more highly expression of 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 determine whether the LR communication occurs between ct1 and ct2. 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:

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

$$E = \frac{edge_{ct1,ct2}}{\overline{edge}_{ct1,ct2}} \tag{11}$$

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 observed and expected number of connections between $ct1$ and $ct2$; and E is the ratio of observed and expected number of the connections. To constrain the range of E and make the result more stable, a Hill function transforms E into a range of (0, 2) and ensures the transformed E is 1 when the number of the observed 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 [32, 56]. Moreover, SOAPy supports non-negative constraints to make the factors more interpretable. Taking non-negative CP decomposition [57] as an example, an n-order tensor X is expressed as the weighted sum of R (user-defined number of factors) rank-one tensors:

$$X \approx \sum_{r=1}^{R} \lambda_r a_r^{(1)} \circ a_r^{(2)} \circ \cdots \circ a_r^{(n)} \tag{12}$$

where λ is the weight of each factor; $a_r^{(k)}$ is the non-negative loading values of kth variable in the rth factor, indicating the relative contribution of variables to factors. Each factor is the outer product of the loading vectors.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03550-5.

Additional file 1: Figures S1–S7. All supplementary figures in this study.

Additional file 2: Table S1 Comparison of existing tools of spatial omics data analysis.

Additional file 3: Table S2 Datasets that were used in this study.

Additional file 4: Table S3 Confusion matrix for the three spatial domain methods.

Additional file 5: Table S4 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") decomposition.

Additional file 6: Table S5 Predicted LR interactions between spatial-separated epithelial cells C3 and other cell types by SOAPy.

Additional file 7: Review history.

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 (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.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
Conceptualization: H.L., H.W., and J.L.; methodology: H.W., J.L., and H.L.; software development: H.W. and J.L.; software test: H.W., Y.Q., and J.Y.; visualization: H.W., J.L., S.J., and P.L.; case study: H.W., J.L., P.L., and X.Y.; writing: H.W., J.L., Z.T., and H.L.; writing review: Z.T., Y.L., H.Z., Y.C., and Z.W.; supervision: H.L.; all authors read and approved the final manuscript.

Funding
This research was supported by the National Natural Science Foundation of China (T2122018, 32470707, 32300555), Shanghai Municipal Science and Technology Major Project, Shanghai Sailing Program (22YF1458000), and CAS Youth Innovation Promotion Association (Y2022076).

Data availability
All codes that produced the findings of the study, including all main and supplemental figures, are available at https://github.com/LiHongCSBLab/SOAPy [58], under the AGPL-3.0 license. The version of the code used in this study is available at Zenodo https://doi.org/10.5281/zenodo.15029958 [59]. For tutorials about this package, please visit https://soapy-st.readthedocs.io/en/latest/. Datasets used in this manuscript are all available for download from the original publication using the links provided in Additional file 3: Table S2. For the raw data, the dorsolateral prefrontal cortex 10X Visium data is available from http://spatial.libd.org/spatialLIBD/ [60], the breast cancer 10X Visium data is available from https://www.10xgenomics.com/resources/datasets [61], the kidney cancer 10X Visium data is available from GSE175540 [62], the liver following acute injury 10X Visium data is available from https://zenodo.org/records/6035873 [63], the mouse embryo GeoMx DSP data is available from https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/mouse-development/ [64], the breast cancer MIBI-TOF data is available from https://mibi-share.ionpath.com [65], and the human ovarian MERSCOPE data is available from https://info.vizgen.com/ffpe-showcase [66]. For convenience and reproducibility, we uploaded processed datasets in h5ad format in Zenodo: https://zenodo.org/records/14588408 [67].

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 4 January 2024 Accepted: 18 March 2025
Published online: 29 March 2025

References

1. Marx V. Method of the Year: spatially resolved transcriptomics. Nat Methods. 2021;18:9–14.
2. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021;596:211–20.
3. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods. 2022;19:534–46.
4. Salmén F, Ståhl PL, Mollbrink A, Navarro JF, Vickovic S, Frisén J, et al. Barcoded solid-phase RNA capture for spatial transcriptomics profiling in mammalian tissue sections. Nat Protoc. 2018;13:2501–34.
5. Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. Cell. 2018;174:1373–1387.e19.
6. Keren L, Bosse M, Thompson S, Risom T, Vijayaragavan K, McCaffrey E, et al. MIBI-TOF: a multiplexed imaging platform relates cellular phenotypes and tissue structure. Sci Adv. 2019;5: eaax5851.
7. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nat Biotechnol. 2020;38:586–99.
8. Schapiro D, Sokolov A, Yapp C, Chen Y-A, Muhlich JL, Hess J, et al. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nat Methods. 2022;19:311–5.
9. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42:293–304.
10. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
11. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15:343–6.
12. 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. 2021;22:184.
13. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.

14. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun. 2022;13:1739.
15. Dries R, Zhu Q, Dong R, Eng C-HL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. 2021;22:78.
16. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19:171–8.
17. Qiu X, Zhu DY, Lu Y, Yao J, Jing Z, Min KH, et al. Spatiotemporal modeling of molecular holograms. Cell. 2024;187:7351–73.e61.
18. Bergenstråhle J, Larsson L, Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics. 2020;21:482.
19. Pham D, Tan X, Balderson B, Xu J, Grice LF, Yoon S, et al. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. Nat Commun. 2023;14:7739.
20. Long Y, Ang KS, Li M, Chong KLK, Sethi R, Zhong C, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nat Commun. 2023;14. Available from: https://www.nature.com/articles/s41467-023-36796-3. Cited 2025 Jan 14.
21. Cang Z, Ning X, Nie A, Xu M, Zhang J. SCAN-IT: domain segmentation of spatial transcriptomics images by graph neural network. 2022.
22. Anselin L. Local indicators of spatial association-LISA. Geogr Anal. 2010;27:93–115.
23. Jong P, Sprenger C, Veen F. On extreme values of Moran's I and Geary's c. Geogr Anal. 2010;16:17–24.
24. Yuan Z, Zhao F, Lin S, Zhao Y, Yao J, Cui Y, et al. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. Nat Methods. 2024. Available from: https://www.nature.com/articles/s41592-024-02215-8. Cited 2024 Mar 16.
25. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat Biotechnol. 2021;39:313–9.
26. Meylan M, Petitprez F, Becht E, Bougoüin A, Pupier G, Calvez A, et al. Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. Immunity. 2022;55:527–541.e5.
27. Pardo B, Spangler A, Weber LM, Page SC, Hicks SC, Jaffe AE, et al. spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. BMC Genomics. 2022;23:434.
28. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15:293.
29. Hildebrandt F, Andersson A, Saarenpää S, Larsson L, Van Hul N, Kanatani S, et al. Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. Nat Commun. 2021;12:7046.
30. He Z, Han D, Efimova O, Guijarro P, Yu Q, Oleksiak A, et al. Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. Nat Neurosci. 2017;20:886–95.
31. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24:425–36.
32. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev. 2009;51:455–500.
33. Peng T, Tian Y, Boogerd CJ, Lu MM, Kadzik RS, Stewart KM, et al. Coordination of heart and lung co-development by a multipotent cardiopulmonary progenitor. Nature. 2013;500:589–92.
34. Sarrach S, Huang Y, Niedermeyer S, Hachmeister M, Fischer L, Gille S, et al. Spatiotemporal patterning of EpCAM is important for murine embryonic endo- and mesodermal differentiation. Sci Rep. 2018;8. Available from: https://www.nature.com/articles/s41598-018-20131-8. Cited 2025 Jan 14.
35. Vaes N, Schonkeren SL, Brosens E, Koch A, McCann CJ, Thapar N, et al. A combined literature and in silico analysis enlightens the role of the NDRG family in the gut. Biochim Biophys Acta BBA - Gen Subj. 2018;1862:2140–51.
36. Ben-Moshe S, Veg T, Manco R, Dan S, Papinutti D, Lifshitz A, et al. The spatiotemporal program of zonal liver regeneration following acute injury. Cell Stem Cell. 2022;29:973–989.e10.
37. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat Methods. 2017;14:873–6.
38. Bäckdahl J, Franzén L, Massier L, Li Q, Jalkanen J, Gao H, et al. Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. Cell Metab. 2021;33:1869–1882.e6.
39. Yuan Z, Li Y, Shi M, Yang F, Gao J, Yao J, et al. SOTIP is a versatile method for microenvironment modeling with spatial omics data. Nat Commun. 2022;13:7330.
40. Schumacher TN, Thommen DS. Tertiary lymphoid structures in cancer. Science. 2022;375: eabf9419.
41. Lavie D, Ben-Shmuel A, Erez N, Scherz-Shouval R. Cancer-associated fibroblasts in the single-cell era. Nat Cancer. 2022;3:793–807.
42. Friedman G, Levi-Galibov O, David E, Bornstein C, Giladi A, Dadiani M, et al. Cancer-associated fibroblast compositions change with breast cancer progression linking the ratio of S100A4+ and PDPN+ CAFs to clinical outcome. Nat Cancer. 2020;1:692–708.
43. Davidson S, Coles M, Thomas T, Kollias G, Ludewig B, Turley S, et al. Fibroblasts as immune regulators in infection, inflammation and cancer. Nat Rev Immunol. 2021;21:704–17.
44. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020;15:1484–506.
45. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, et al. Inference and analysis of cell-cell communication using Cell Chat. Nat Commun. 2021;12:1088.
46. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet. 2021;22:71–88.
47. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22:627–44.
48. Cheng J, Yan L, Nie Q, Sun X. Modeling and inference of spatial intercellular communications and multilayer signaling regulations using stMLnet. Syst Biol. 2022. Available from: http://biorxiv.org/lookup/doi10.1101/2022.06.27.497696.

49. Xu S, Xu H, Wang W, Li S, Li H, Li T, et al. The role of collagen in cancer: from bench to bedside. J Transl Med. 2019;17:309.
50. Fischer C, Mazzone M, Jonckx B, Carmeliet P. FLT1 and its ligands VEGFB and PlGF: drug targets for anti-angiogenic therapy? Nat Rev Cancer. 2008;8:942–56.
51. Jin S, Plikus MV, Nie Q. Cell Chat for systematic analysis of cell–cell communication from single-cell transcriptomics. Nat Protoc. 2025;20:180–219.
52. Van De Sande B, Flerin C, Davie K, De Waegeneer M, Hulselmans G, Aibar S, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc. 2020;15:2247–76.
53. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Lu Z, editor. Bioinformatics. 2023;39:btac757.
54. Haynes W. Benjamini–Hochberg method. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. Encycl Syst Biol. New York: Springer New York; 2013. p. 78–78. Available from: http://link.springer.com/10.1007/978-1-4419-9863-7_1215. Cited 2023 Nov 23.
55. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell. 2018;174:968–981.e15.
56. Zhou G, Cichocki A, Zhao Q, Xie S. Efficient nonnegative tucker decompositions: algorithms and uniqueness. IEEE Trans Image Process. 2015;24:4990–5003.
57. Shashua A, Hazan T. Non-negative tensor factorization with applications to statistics and computer vision. In: Proc 22nd Int Conf Mach Learn - ICML 05. Bonn, Germany: ACM Press; 2005. p. 792–9. Available from: http://portal.acm.org/citation.cfm?d=1102351.1102451. Cited 2023 Nov 23.
58. Wang H. SOAPy: a package for spatial-omics analysis in Python. Github; 2024. https://github.com/LiHongCSBLab/SOAPy.
59. Wang H. SOAPy: a Python package to dissect spatial architecture, dynamics and communication v2. Zenodo; 2024. https://doi.org/10.5281/zenodo.15029958.
60. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, 2nd, Tran MN, Besich Z, Tippani M, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Datasets; 2021. http://spatial.libd.org/spatialLIBD.
61. 10X Visium human breast cancer. Datasets; 2019. https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Breast_Cancer_Block_A_Section_1.
62. Meylan M, Petitprez F, Becht E, Bougoüin A et al. Tertiary lymphoid structures generate and propagate anti-tumor antibody-producing plasma cells in renal cell cancer. Datasets; 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175540.
63. Ben-Moshe S, Itzkovitz S. The spatio-temporal program of liver zonal regeneration. Zenodo; 2021. https://zenodo.org/records/6035873.
64. Nanostring GeoMx WTA mouse development data. Datasets; 2022. https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/mouse-development/.
65. Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. Datasets; 2018. https://mibi-share.ionpath.com/tracker/imageset.
66. Vizgen MERFISH FFPE human immuno-oncology data set. Datasets; 2022. https://info.vizgen.com/ffpe-showcase.
67. Wang H. SOAPy: a Python package to dissect spatial architecture, dynamics and communication v2. Zenodo; 2025. https://zenodo.org/records/14588408.

Publisher's Note