Genome Biology

METHODOLOGY

Open Access

SpaNiche: spatial niche analysis to explore colocalization patterns and cellular interactions in spatial transcriptomics data

Siyuan Huang[1†], Qinghua Ran[1†], Junjie Tang[2,3†], Xiaochen Wang[4†], Junqing Xi[5], Shiyang Ma[6,7*] and Ruibin Xi[1,3,4*]

†Siyuan Huang, Qinghua Ran, Junjie Tang and Xiaochen Wang contributed equally to this work.

*Correspondence: mashiyang1991@sjtu.edu.cn; ruibinxi@math.pku.edu.cn

[1] Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[2] Present Address: Ray and Stephanie Lane Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[3] Center for Statistical Science, Peking University, Beijing 100871, China
[4] School of Mathematical Sciences, Peking University, Beijing 100871, China
[5] Department of Interventional Therapy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China
[6] Institute of Clinical Medicine, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
[7] School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract

We propose a computational framework for spatial niche analysis (SpaNiche) in spatial transcriptomics data to uncover colocalization patterns and infer potential ligand-receptor interactions. SpaNiche leverages graph-regularized joint non-negative matrix factorization to integrate information from cell abundance and ligand-receptor expression, identifying spatial colocalization patterns among cell types while providing insights into associated ligand-receptor interactions. Besides, SpaNiche employs consensus clustering to define "ecotypes", enhancing its utility in multi-sample spatial transcriptomics datasets. We investigate microenvironmental ecotypes in colorectal cancer, prostate cancer and cerebral cortex in early Alzheimer's disease, highlighting the broad applicability in dissecting complex colocalization patterns within spatial tissue microenvironments.

Background

Understanding how cells are organized within tissues and interact from a spatial perspective is pivotal for advancing our knowledge of developmental processes and tumor biology [1]. For instance, in the brain's intricate architecture, various types of neurons collaborate synergistically to maintain normal signal transmission [2]. Such spatial insights are not only valuable for exploring the organization of healthy tissue microenvironments but also for elucidating alterations in disease-associated microenvironments. The tumor microenvironment (TME) comprises a diverse array of cell types, including tumor cells, immune cells, and stromal cells, which interact within specific niches to form a complex and dynamic ecosystem [3]. Spatial transcriptomics (ST) has emerged as a revolutionary approach for unraveling the complexities of the TME [4]. ST enables the mapping of gene expression to specific locations within tissues, providing transformative insights into the spatial organization of biological processes and disease mechanisms.

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Recent advancements in ST technologies have significantly improved their throughput and resolution, enabling more precise spatial analyses [5, 6]. Andersson et al. [7] identified specific colocalization pattern between macrophage and T-cell subtypes during type I interferon responses in HER2-positive breast tumors. They also observed the colocalization of B cells and T cells, revealing the presence of tertiary lymphoid-like structures within the spatial landscape. Liu et al. [8] combined ST with single-cell RNA-sequencing (scRNA-seq) to identify a tumor immune barrier structure composed of SPP1 + macrophages and cancer-associated fibroblasts located near the tumor boundary in hepatocellular carcinoma microenvironment. This study highlighted the critical role of the spatial structure in influencing the efficacy of immunotherapy.

In the field of ST, numerous methods have been developed to identify spatial domains by leveraging transcriptional expression, spatial position, and graphical features [9–11]. However, spatial domains are conceptually distinct from colocalization patterns, as they do not specifically emphasize the spatial arrangement and interactions between cell types. Currently, many existing tools for analyzing ST data focus on identifying spatial domains, but there is a lack of approaches tailored to explore spatial colocalization patterns of different cell types. These current methods primarily rely on techniques such as non-negative matrix factorization (NMF) [12, 13], pairwise correlation-like metrics [14, 15], or multivariate predictive models [16]. Nevertheless, they often neglect to incorporate ligand-receptor information, which is crucial for understanding spatial colocalization patterns. Spatial proximity between cells frequently implies enhanced interaction strength mediated through ligand-receptor signaling [17]. Therefore, integrating spatial distribution features with ligand-receptor interaction information has the potential to significantly improve the identification of spatial colocalization patterns. Moreover, existing methods typically focus on colocalization within immediate "neighbors" (e.g., the spatial extent of a single spot in 10 × Visium data). However, cell–cell interactions are not confined to immediate neighbors and may extend across broader spatial ranges [18, 19]. In addition, some methods, such as correlation-based approaches, while effective in capturing the global characteristics of tissue slices, often fail to accurately detect specific local features within the tissue microenvironment.

To address these limitations, we propose SpaNiche, a computational framework for spatial niche analysis based on graph-regularized joint NMF that integrates both the spatial distribution of cell types and the strength of ligand-receptor interactions. This framework enables systematic exploration of colocalization patterns across cell types from multiple spatial perspectives, thereby deepening our understanding of spatial cellular architecture under different physiological conditions. Furthermore, SpaNiche can infer the molecular signatures of colocalization-defined local microenvironments. Beyond spatial pattern discovery, SpaNiche advances our understanding of ecotypes and enables the deconvolution of ecotype proportions in bulk transcriptomes, facilitating their potential clinical applications.

Results

Overview of SpaNiche

The SpaNiche framework comprises a core computational module and four downstream analytical modules: visualization, enrichment analysis, ligand-receptor-induced pathway

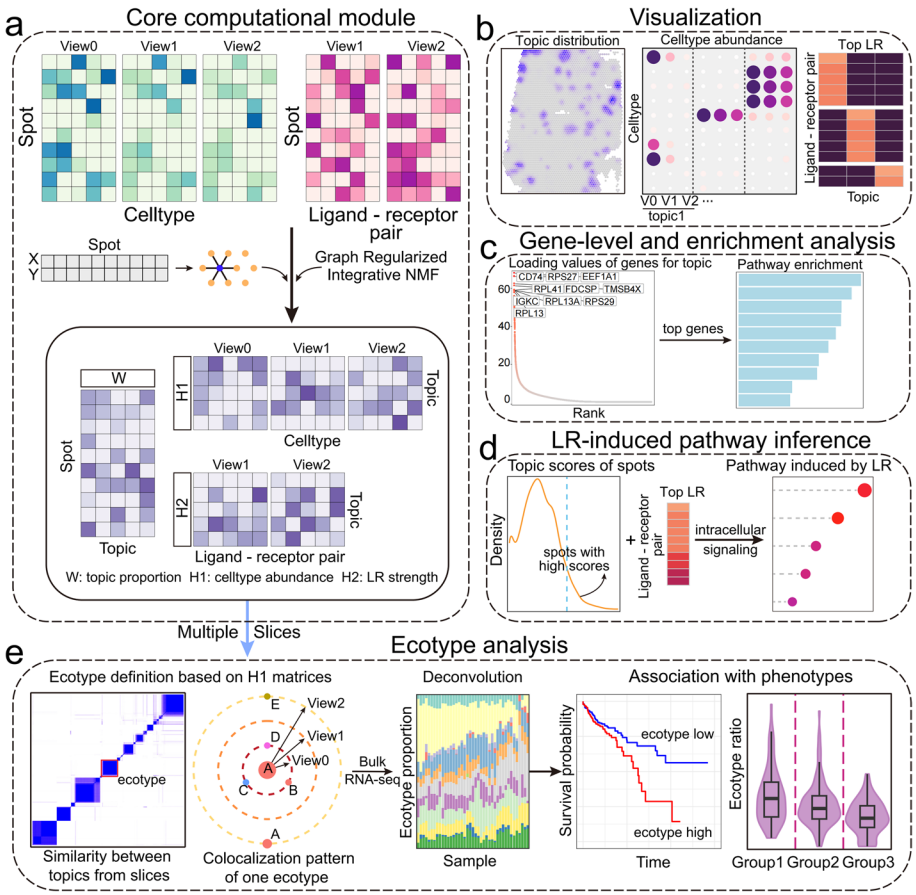

Fig. 1 Schematic overview of SpaNiche. **a** Core computational module of the SpaNiche framework. The graph-regularized joint NMF decomposes both the smoothed cell type abundance matrix and the smoothed ligand-receptor expression matrix, producing three matrices for downstream analysis: the spatial distribution matrix of distinct topics W, the topic weights for various cell types across different views H_1, and the topic weights for ligand-receptor pairs across different views H_2. Views represent the spatial ranges at which we interrogate tissue structure and function, operationalized as different spatial distances. In 10 × Visium data, the default distances for view1 and view2 correspond to 100 μm and 200 μm, respectively. **b** Visualization of the three output matrices. **c** Identification of highly expressed genes and upregulated pathways associated with each topic. **d** Identification of activated pathways potentially induced by intercellular interactions. **e** Definition of ecotypes using SpaNiche for multi-sample ST datasets, and estimation of ecotype abundance in bulk transcriptomics samples

inference, and ecotype analysis approach (Fig. 1). The computational module takes the smoothed cell type abundance matrix and the ligand-receptor expression matrix as inputs for graph-regularized joint NMF (Fig. 1a). Departing from conventional applications of NMF in single-cell transcriptomics [20], our methodology in this matrix factorization step incorporates both the spatial distribution of cell types and cell–cell interactions from diverse spatial views, further enhanced by the implementation of graph constraints to bolster the consideration of spatial location (see "Methods"). Views represent the spatial ranges at which we interrogate tissue structure and function, operationalized as different spatial distances (see the "Cellular composition" and "Cell–cell interactions" subsections in the "Methods" section; Additional file 1: Fig. S1a, b). This process produces three output matrices: the spatial distribution matrix of distinct topics W (topics can be interpreted

as fundamental units derived from the decomposition), the topic weights for various cell types across different views H_1, and the topic weights for ligand-receptor pairs across different views H_2 (Fig. 1a).

The first step in analyzing these output matrices involves visualization, as demonstrated in Fig. 1b. Matrix operations between W and the ST expression matrix, constrained by non-negativity, yield the weight distribution of each gene across topics (Additional file 1: Fig. S1c). Genes with the highest weights can be used to characterize the topics, providing potential biological interpretability. Subsequent enrichment analysis can be conducted in the context of specific research questions (Fig. 1c). After deriving the matrix W, we identify positive and negative spots for each topic, defined as those within the top 20% and bottom 20% of topic scores, respectively. Ligand-receptor pairs strongly associated with each topic are identified from the loading matrix H_2. By integrating this information with the expression profiles of topic-related positive and negative spots, we infer the biological pathways activated by high ligand-receptor pair expression (Fig. 1d). Unlike the enriched pathways presented in Fig. 1c, the pathways highlighted in Fig. 1d focus on those that are upregulated specifically under the scenario of ligand-receptor interactions. This analysis provides functional insights into cellular responses driven by intracellular signal transduction [21].

For multi-sample ST datasets, we introduce an ecotype analysis approach (Fig. 1e). Here, an "ecotype" refers to a local microenvironment comprising multiple cell types that exhibits a certain level of consistency across samples. Cells within the same ecotype are more likely to facilitate cell–cell interactions due to their spatial proximity. After running graph-regularized joint NMF on each sample, all H_1 matrices are integrated for consensus clustering [22], where topics belonging to the same module are considered part of the same ecotype. For tumor ST datasets, we extract ecotype characteristics and use weighted linear regression to estimate the proportions of different ecotypes in bulk transcriptomics data. These estimates are then correlated with clinical phenotypes, providing novel insights into tumorigenesis, progression, prognosis, and therapeutic responses.

Benchmarking SpaNiche on simulated ST data

To evaluate the performance of SpaNiche, we analyze simulated ST data and real ST data with extensive prior biological annotations. To generate simulated ST datasets, we organize ST spots in a square grid, with each spot surrounded by six neighboring spots. We utilize a scRNA-seq dataset from breast cancer as a reference to simulate cell type proportions and gene expression levels for each spot [23]. We define multiple spatial niches, where each niche was distributed across two equally sized circular regions, with spots inside each region being spatially adjacent. Spots belonging to the same niche were enriched for specific cell types ("Methods"). To construct these niches, we select specific cell types and their corresponding spots, ensuring that spots do not overlap across niches. Each niche is enriched with one or more specific cell types (Fig. 2a). For each spot, we simulate cell type proportions using a Dirichlet distribution, ensuring that the proportions within each spot sum to 1. These proportions vary depending on whether a spot is located within a niche, with additional noise introduced to simulate deconvolution errors. To simulate gene expression, we assume that gene expression follows a zero-inflated negative binomial (ZINB) distribution and

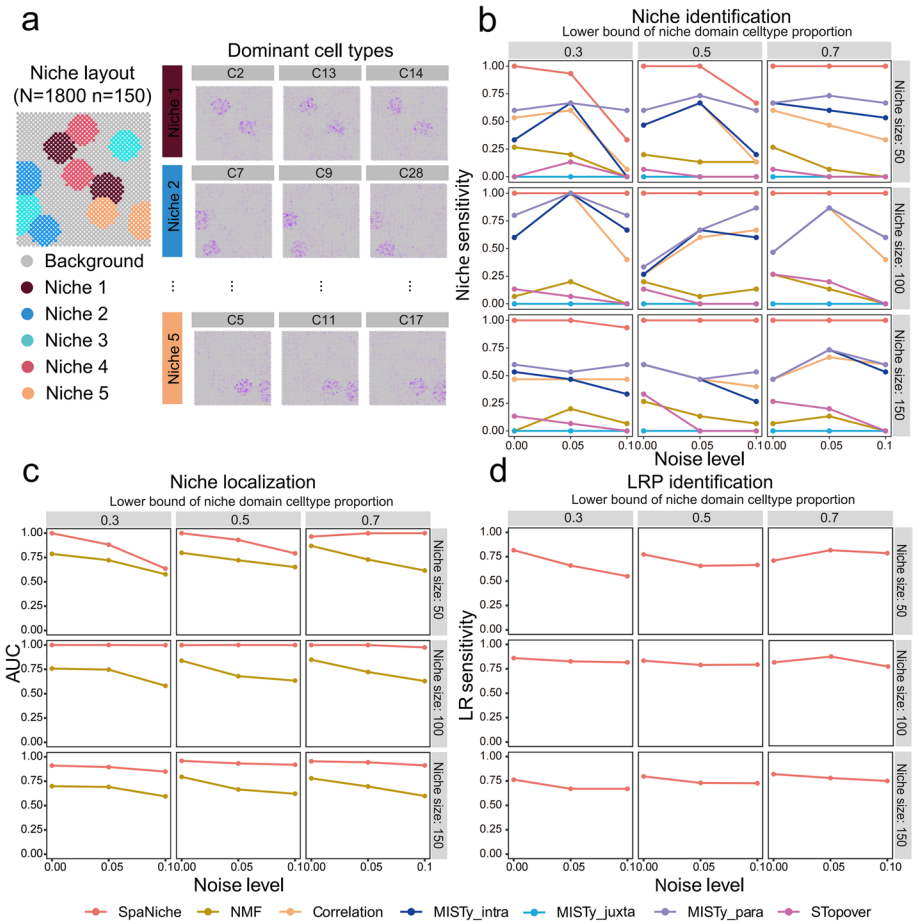

Fig. 2 Performance evaluation based on simulated ST data. **a** Examples of the simulated data generated in this study. Each niche is enriched with one or several specific cell types. **b-d** Results of three evaluation metrics calculated on the simulated data with a fixed parameter $\boldsymbol{\alpha}$=1 while varying other parameters. **b** Niche identification; **c** Niche localization; **d** LRP identification. The parameter β_1 is used to control the proportion of enriched cell types within each niche. A larger value of β_1 indicates a higher proportion of the dominant cell type. The parameter Niche size controls the number of spots in the niche region; a larger value corresponds to a larger tumor area. The parameter σ controls the noise level in the simulated data; higher noise results in smaller differences between cell types

estimate the distribution parameters for each cell type from the reference scRNA-seq dataset. Based on the simulated cell-type proportions, we then generate single-cell expression profiles for each spot and aggregate these single-cell values to obtain spot-level expression. Finally, dropout is introduced to better replicate the real ST datasets.

By varying simulation parameters, we generate 162 distinct settings to comprehensively evaluate the performance of each algorithm. In our evaluation, we assess our method's performance from three perspectives: identifying niches, localizing niche-associated spots, and detecting niche-specific ligand-receptor pairs. We use sensitivity, defined as the proportion of correctly identified niches or ligand-receptor pairs, and the area under the receiver operating characteristic curve (ROC AUC) as evaluation metrics. These metrics quantify the effectiveness of our method in replicating and elucidating complex spatial and molecular interactions in the simulated ST datasets.

Using one group of settings as an example, we present the simulation results in Fig. 2. We observe that when varying noise in the cell type proportion matrix, the minimum proportion of enriched cell types within a niche, and the number of spots within a niche, SpaNiche consistently exhibits the highest sensitivity in detecting niches. Compared with four other celltype composition-based methods (NMF, Correlation, MISTy and STopover) [12, 14–16], the results of SpaNiche remain relatively stable and do not fluctuate significantly with parameter changes (Fig. 2b). From the weight matrix W obtained after decomposition, we derive the weight of each topic in spots, enabling the identification of spots associated with specific niches. The AUC is used to measure classification accuracy. The accuracy of SpaNiche in locating niche-associated spots, based on topic weights, is significantly higher than that of NMF (Fig. 2c). Several other methods based on correlation-like metrics cannot directly infer the classification or probability of spot assignment, making the computation of AUC infeasible. Unlike other methods, SpaNiche is uniquely capable of identifying ligand-receptor pairs enriched within specific topics. Under different parameters, the sensitivity of SpaNiche in identifying ligand-receptor pairs remained consistently around 0.75 (Fig. 2d). The results under all other parameter settings are presented in Additional file 1: Fig. S2, where SpaNiche consistently outperforms the four comparison methods in terms of the ability of identifying niches, the accuracy of localizing niche-related spots, and the capability of discovering niche-specific ligand-receptor pairs. It is worth noting that the noise introduced into the cell-type proportion matrix is mainly used to mimic inaccuracies arising from deconvolution methods. In our simulations, SpaNiche remains robust across different noise levels.

Performance validation by real ST data from a human lymph node

To evaluate the performance of SpaNiche on real ST data, we analyze a public dataset of human lymph node (Fig. 3a), which has been extensively characterized and analyzed in previous studies [12, 24]. The well-established biological prior knowledge of lymph nodes makes this dataset particularly suitable for comparing computational results with real-world biological observations. Based on the functional maturation process of B cells in lymph nodes [25] and prior findings [12], the dataset is expected to contain the following local niches (Fig. 3b): naïve B cells in the B follicle region surrounding the germinal center; B cells in the dark zone of the germinal center undergoing clonal expansion and proliferation; B cells in the light zone of the germinal center differentiating into plasma cells under the influence of T follicular helper cells and follicular dendritic cells; reactivated B cells before entering the germinal center, distinct from both upstream activated B cells and downstream germinal center B cells; a vascular niche containing endothelial cells and vascular smooth muscle cells; and a perivascular region enriched with T helper cells and B memory cells.

To comprehensively compare the performance of SpaNiche with other state-of-the-art niche detection methods, we include NicheCompass [26] and EcoTyper [27] in the benchmarking analysis of real ST data. Our comparative analysis shows that all methods could capture germinal center-associated niches, except for NicheCompass and EcoTyper (Fig. 3c-e and Additional file 1: Fig. S4c-f). For instance, these include cluster1 detected by the Pearson correlation-based method, cluster2 detected by

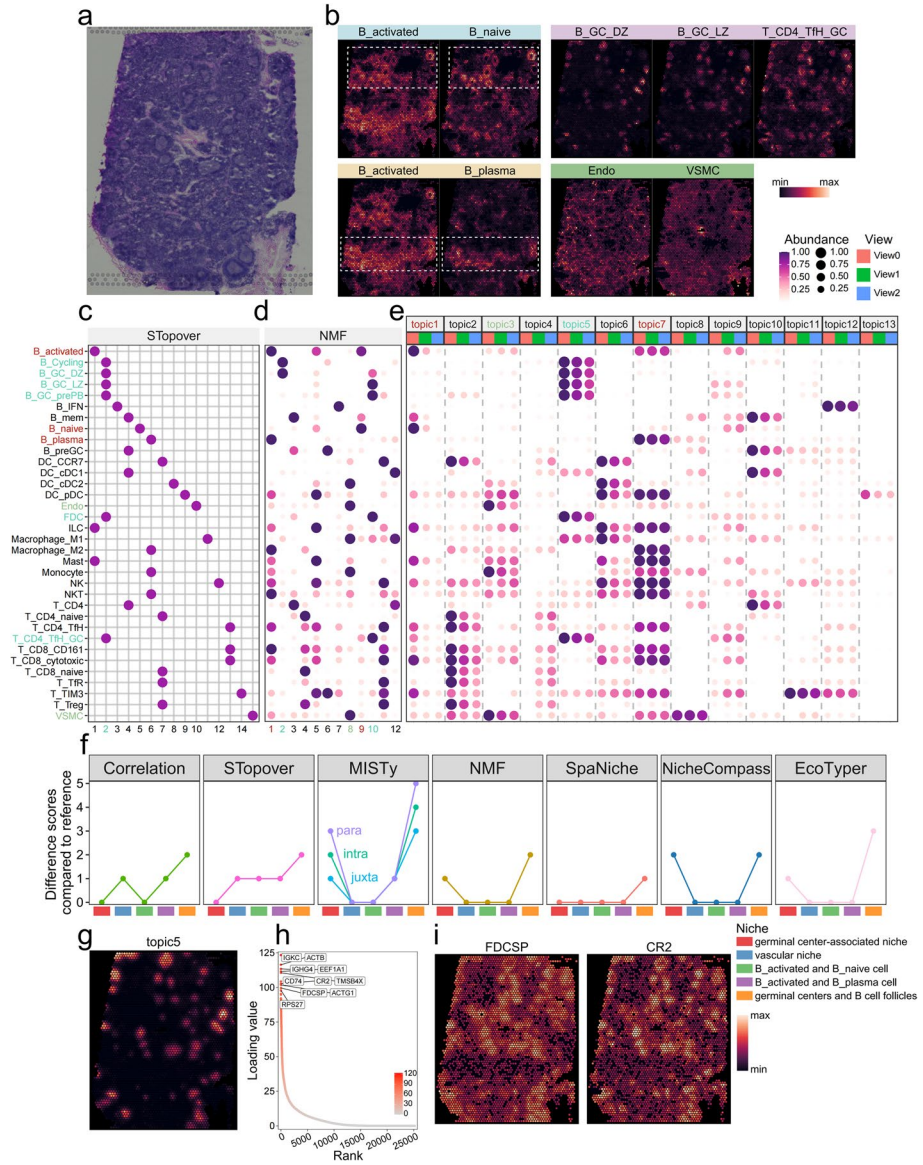

Fig. 3 Performance evaluation based on lymph node ST data. **a** HE-stained image of a lymph node ST sample from 10 × Genomics. **b** Established cell type colocalization patterns from prior knowledge and observed spatial distribution of cell types. Spot colors indicate relative abundance of cell types. **c** STopover-based colocalization detection. The x-axis denotes community clustering categories, while the y-axis represents cell types. Each colocalization pattern consists of distinct cellular compositions. **d** NMF-based colocalization detection. The x-axis represents decomposition-derived topics, and the y-axis shows cell types. Dot size and color intensity reflect cell type weight values within each topic. **e** SpaNiche-based colocalization detection. The x-axis represents decomposition-derived topics partitioned into three spatial views, and the y-axis shows cell types. Dot size and color intensity reflect cell type weight values within each topic. **f** Difference scores compared to reference, based on the results of spatial colocalization detection from different methods. **g** Spatial distribution of colocalization pattern (topic 5). **h** Gene weight values and rankings within topic 5. **i** Spatial distribution of high-weight genes (*FDCSP* and *CR2*)

STopover, cluster2 and cluster3 detected by MISTy with intra mode, factor2 and factor10 detected by NMF, and topic5 detected by SpaNiche. This is attributed to the strong signals of germinal center-related cell types, such as B_GC_LZ and B_GC_DZ, which make them relatively easy to detect. The spatial distribution of cell types

indicates that the B_activated subtype exhibits two distinct co-localization patterns: one with B_naive cells in the upper part of the section and the other with B_plasma cells in the lower part (Fig. 3b). The Pearson correlation-based method detects only one of these patterns (cluster3), whereas STopover fails to detect either. The MISTy-based method detects only one pattern, represented by cluster1, indicating the spatial relationship between B_activated and B_naive cells. NMF detects both spatial relationships, corresponding to factor1 and factor9. EcoTyper also detects both patterns, represented by E2 and E3. Similarly, SpaNiche detects both patterns, represented by topic1 and topic7. In addition, based on the SpaNiche results, we observe that the spatial niche associated with topic 7 includes not only B_activated + B_plasma cells but also other cell types such as M2 macrophages (Fig. 3e and Additional file 1: Fig. S3). This observation suggests a potential role for B cells in promoting the polarization of M2-like macrophages [28]. The vascular niche containing endothelial cells and vascular smooth muscle cells is detected by all methods except for the Pearson correlation-based method, STopover and NicheCompass (Additional file 1: Fig. S5).

In topic3 detected by SpaNiche (Fig. 3e), expanding the scope of spatial colocalization analysis from a single spot to three layers of spots reveals a gradual increase in the abundance of B_activated, B_naive, and germinal center-related cell types, accompanied by a corresponding decrease in abundance of endothelial and vascular smooth muscle cells. This pattern indicates the spatial proximity relationship between vascular and immune hotspot regions. Such multi-perspective information is not achievable using Pearson correlation, STopover, NMF, NicheCompass, or EcoTyper. Although MISTy also provides three different perspectives, these perspectives are relatively independent and do not detect similar spatial patterns in this dataset. Another similar pattern detected by SpaNiche is topic9 (Fig. 3e), where, as the perspective expands, signals of B_activated and B_naive cells appear in addition to the original germinal center signals, consistent with the spatial proximity relationship between the germinal center and B cell follicles [29]. After quantitatively comparing the differences between the results obtained by different methods and the reference cell composition within each niche, SpaNiche still demonstrates the highest accuracy (Fig. 3f).

Furthermore, SpaNiche can infer transcriptional features corresponding to spatial colocalization patterns. We find that topics corresponding to the germinal center exhibit high expression of *CR2* (complement component 3 d receptor 2, involved in B lymphocyte activation), *FDCSP* (a gene encoding a secreted protein expressed in follicular dendritic cells, which can specifically bind to activated B cells and regulate the antibody response) (Fig. 3g-i). This topic exhibits overall high expression of pathways associated with adaptive immunity, such as "immunoglobulin-mediated immune response," "antigen processing and presentation," and "B cell activation" (Additional file 1: Fig. S4h). The topic corresponds to the high interaction between ligands and receptors such as CXCL13-CXCR5 and CCL18-ACKR1 (Additional file 1: Fig. S4g). *CXCL13*, primarily produced by follicular dendritic cells, can attract B cells and T cells expressing *CXCR5* to the lymph node. These transcriptional features align with the known biological characteristics of the germinal center.

Identification of colorectal cancer ecotypes using SpaNiche

To systematically characterize the spatial colocalization-based ecotypes within the colorectal cancer (CRC) microenvironment, we analyze publicly available ST datasets of CRC, covering various developmental stages and primary locations, totaling 32 samples [30–34]. For each sample, the abundance of cell types is estimated using cell2location [12], followed by the joint factorization of the cellular composition matrix and the ligand-receptor expression matrix using SpaNiche. After quality control, 14 biologically meaningful ecotypes are identified. Based on their predominant cell types, these ecotypes are classified into five categories: normal intestinal epithelial cells, tumor cells, immune cells, stromal cells, and mixed types (Fig. 4a and Additional file 1: Fig. S6a-c). Ecotype 6 is characterized by the colocalization of pericytes, stromal cells, and endothelial cells, with high expression levels of collagen genes (*COL4A1* and *COL6A2*), Galectin 1 (*LGALS1*), and *SPARC*, which is involved in extracellular matrix synthesis (Fig. 4c). This ecotype represents the stromal subtype. Ecotype 8 indicates the spatial proximity of tumor cells, macrophages, and myofibroblasts. Previous studies have reported the colocalization of FAP + fibroblasts and SPP1 + macrophages in CRC, which shows a negative correlation with lymphocyte infiltration, predicting poor patient survival [32]. In hepatocellular carcinoma (HCC), interactions between SPP1 + macrophages and fibroblasts have also been observed, promoting the formation of tumor immune barrier and limiting immune infiltration into the tumor, thereby impacting the efficacy of immunotherapy [8]. Ecotype 11 represents an immune cell composition that forms an immunosuppressive microenvironment, comprising anti-inflammatory macrophages and Treg cells. This ecotype is primarily characterized by the expression of macrophage marker genes, including *APOE* and *C1QB*, the interactions mediated by HLA class II molecules and CD4, and the pathways including antigen processing and presentation, and phagocytosis (Additional file 1: Fig. S7a, b and Fig. 4d). Previous studies have demonstrated that Treg cells within the tumor microenvironment can suppress the ability of macrophages to produce pro-inflammatory cytokines [35]. Conversely, macrophages can promote Treg generation through reactive oxygen species (ROS) production [36]. Ecotype 12 is characterized by the spatial proximity of germinal center B cells, memory B cells, and helper T cells, which are features typically associated with lymph nodes or tertiary lymphoid structures (TLS) [37]. This spatial organization is confirmed by corresponding hematoxylin and eosin (HE)-stained images (Fig. 4b). This ecotype exhibits

(See figure on next page.)

Fig. 4 SpaNiche identifies clinical phenotype-associated ecotypes in colorectal cancer patients. **a** Four ecotypes identified using SpaNiche in CRC spatial transcriptomics data. Different concentric circles represent different spatial views, from the innermost to the outermost: view0, view1, and view2. Each dot represents a cell subtype, with different colors indicating several major cell types. Epi denotes epithelial cells, and Mye denotes myeloid cells. The size of each dot corresponds to the frequency of that subcluster across all samples for one ecotype. **b** Representative topics from ecotypes 6, 8, 11, and 12. Columns (left to right): HE-stained tissue section, spatial distribution of the topic, and spatial distribution of cell types constituting the colocalization pattern. Color intensity indicates relative abundance in spots. **c** Gene weight values and rankings within selected topics from Fig. **b**. **d** Identification of activated pathways induced by intercellular communication for the two selected topics in Fig. **b**. **e** Relative proportions of ecotypes in CRC tumor and normal tissues from TCGA-COAD cohort. **f** Significant associations of ecotypes 8 and 12 with patient survival outcomes

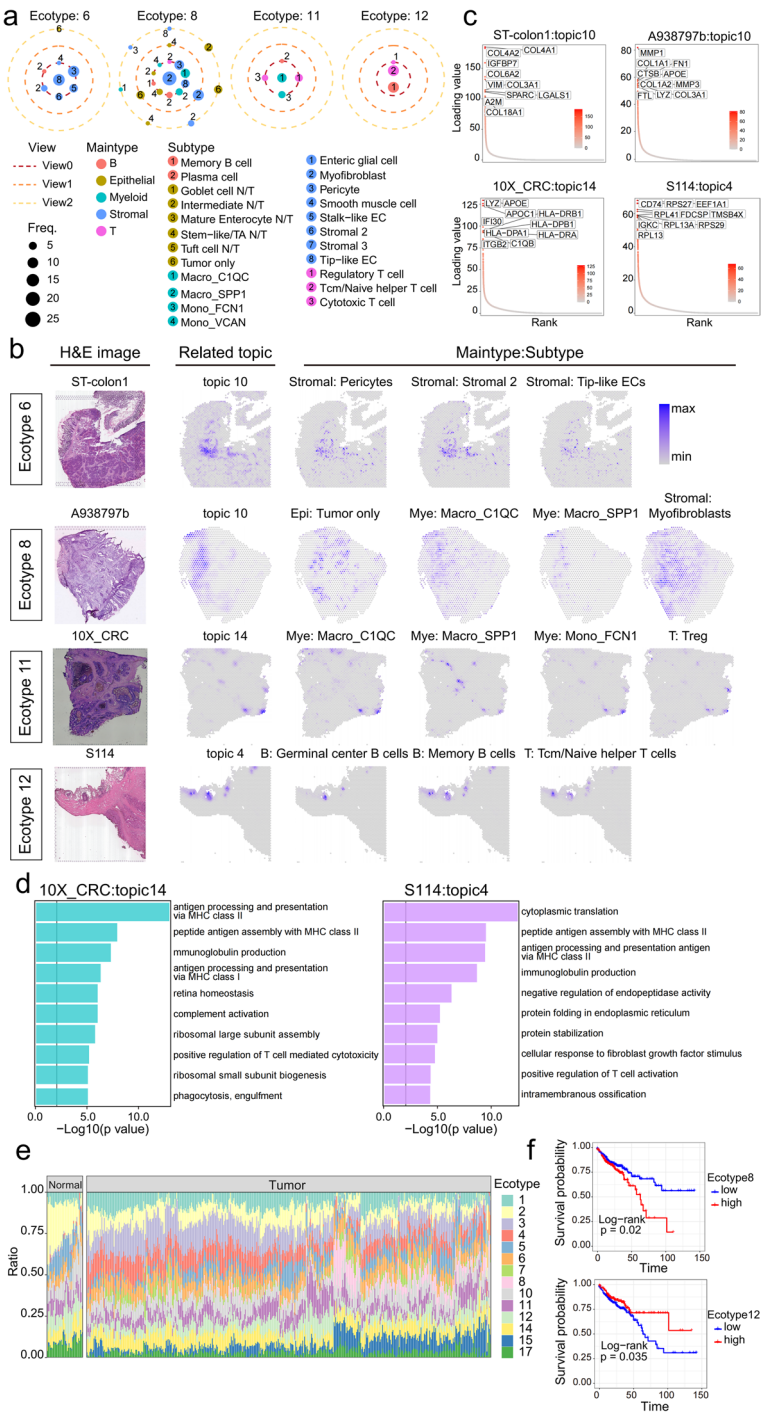

Fig. 4 (See legend on previous page.)

high expression of ligand-receptor interactions and pathways associated with antigen recognition (Additional file 1: Fig. S7a, b and Fig. 4d).

We discover that ecotypes can be used to characterize the spatial heterogeneity of tumors. Taking the CRC1 sample as an example, spots enriched with epithelial cells are categorized into three distinct niches (Additional file 1: Fig. S7c). Niche 1 is

characterized by high expression of Ecotype 2 signals and is enriched with cell types such as goblet cells and intestinal cells, located in intestinal crypts. This niche likely represents normal intestinal tissue or a pre-tumor state. Niche 2 is defined by the simultaneous high expression of Ecotype 3 and Ecotype 4 signals. Specifically, it contains stem cell-like/transitional tumor cells associated with Ecotype 3, as well as tuft cell-like/tumor tissue-specific tumor cells associated with Ecotype 4. Niche 3 exclusively exhibits Ecotype 4 signals. Further analysis reveals that spots in Niche 3 highly express *MMP11* and *POSTN* (Additional file 1: Fig. S7e), genes associated with epithelial-mesenchymal transition (EMT) in tumor cells. Additional signature scores (Additional file 1: Fig. S7f) reveal that Niche 3 exhibits elevated angiogenesis and EMT pathway activity, suggesting a more malignant tumor phenotype. In contrast, spots in Niche 2 exhibit high expression of MYC target genes and pathways associated with cell proliferation and oxidative phosphorylation metabolism (Additional file 1: Fig. S7f). These findings highlight significant tumor heterogeneity within a single ST slide and demonstrate that SpaNiche is a powerful tool for uncovering intrinsic tumor differences.

SpaNiche can map ecotypes back onto bulk transcriptome data and estimate the relative abundance of each ecotype in bulk samples. As shown in Fig. 4e, proportions of ecotypes are illustrated in TCGA normal colon and tumor samples. For example, we find that Ecotype 2, representing intestinal crypts, and Ecotype 12, representing mesenteric lymph nodes or TLS, are more abundant in normal control samples (Additional file 1: Fig. S7g). In contrast, Ecotype 6, representing stromal cells, and Ecotype 8, representing the spatial proximity of fibroblasts and macrophages, are predominantly observed in tumor samples (Additional file 1: Fig. S7g). Subsequently, we investigate the association between ecotypes and clinical phenotypes. We discover that Ecotype 12 is linked to improved survival, whereas Ecotype 8 correlates with poorer survival, underscoring the prognostic significance of ecotypes (Fig. 4f). Additionally, we find that Ecotype 3, which represents tumor cells, exhibited an increasing trend with advancing tumor stages (Additional file 1: Fig. S7h), suggesting the more malignant nature of this ecotype. Conversely, Ecotype 12, which indicates TLS-like structure and anti-tumor activity, shows a gradual decline with advancing tumor stages (Additional file 1: Fig. S7i), suggesting a reduced capacity of the immune system to eliminate tumor cells during the later stages of tumor progression.

Distinct ecotypes between colorectal cancer and liver metastases

We further utilize SpaNiche to investigate the microenvironmental characteristics of CRC liver metastases, aiming to identify distinct ecotypes between primary and metastatic sites. Ecotype analysis is conducted on 10 ST samples from liver metastases [30, 33, 38], resulting in the identification of seven ecotypes after filtering (Additional file 1: Fig. S8a). Compared to the ecotypes identified in primary tumor sites, those in metastatic sites are relatively less complex. Ecotype 1 is characterized by the spatial colocalization of fibroblasts and endothelial cells. Ecotype 2 indicates the colocalization of hepatocytes, liver sinusoidal endothelial cells, and Kupffer cells, characteristic of normal liver tissue. Ecotypes 4 and 6 are associated with tumor cells in metastatic sites. Ecotype 8 exhibits the colocalization of plasma cells and tumor cells. Ecotype 9 is characterized by macrophages. Ecotype 12 features the colocalization of effector T helper cells

and memory B cells. The spatial distribution of Ecotype 12 reveals that although T cells and B cells are co-localized, they do not exhibit spatial aggregation (Additional file 1: Fig. S8b). In contrast, spots with high Ecotype 12 scores are spatially dispersed, differing from the TLS-like structures observed in primary sites. Notably, none of the liver metastasis samples displays significant TLS-like spatial patterns. We hypothesize that the absence of ecotypes associated with TLS in metastatic sites promotes immune evasion of tumor cells, thereby facilitating their colonization in the liver.

Decoding the microenvironment of human prostate cancer using SpaNiche

SpaNiche can also be applied to ST data with single-cell resolution (Fig. 5a). We select windows with varying physical radius to calculate the abundance of different cell types and assess interaction strengths across different distance scales, followed by joint matrix decomposition using SpaNiche.

We reanalyze published Slide-seqV2 data from human prostate cancer [39], to investigate microenvironmental differences between high-grade (HG) and low-grade (LG) tumors. The cells are annotated into 18 distinct cell types (Fig. 5b). By analyzing the spatial tumor microenvironment, we discover that HG tissues are uniquely characterized by a spatial niche dominated by hillock cells and a vascular niche composed of endothelial cells and other stromal cells (Additional file 1: Fig. S9a), corresponding to the heightened demand for nutrients and oxygen supply associated with malignant tumor cell proliferation. In contrast, LG tissues exhibit a unique spatial pattern characterized by macrophage and fibroblast colocalization, alongside a well-structured parenchyma formed by epithelial and fibroblast cells (Additional file 1: Fig. S9b). We measure the spatial distances between endothelial and fibroblast cells, finding that HG samples exhibited significantly shorter distances than LG samples. Conversely, the spatial proximity between fibroblasts and macrophages is closer in LG samples than in HG samples (Fig. 5c-f). These observations further support our findings described above.

Additionally, we identify some similar spatial niches across samples with different malignancy levels, primarily based on the similarity of their cellular compositions. For instance, niches composed of pericytes and fibroblasts, club cells and fibroblasts, tumor niches densely populated by tumor cells, and tumor-adjacent niches occupied by normal epithelial cells are observed (Fig. 5g, h, and Additional file 1: Fig. S9c). The epithelial cell types enriched in the tumor-adjacent niches differ between HG and LG samples (Fig. 5i). HG samples are enriched with more basal epithelial cells, while LG samples have more luminal epithelial cells, highlighting the differences in

(See figure on next page.)

Fig. 5 SpaNiche deconstructs the tumor microenvironment across prostate cancer grades. **a** Analytical workflow of SpaNiche for single-cell resolution ST data. **b** Spatial distribution of cell types across ST samples. **c** Spatial distribution of fibroblasts and Endothelial cells-1 in ST samples. **d** Relative spatial proximity analysis between fibroblasts and Endothelial cells-1. **e** Spatial distribution of fibroblasts and macrophages in ST samples. **f** Relative spatial proximity analysis between fibroblasts and macrophages. **g** Bar plots showing cell type weight distributions across multi-view topics. Both high-grade and low-grade samples exhibit tumor-context and tumor-adjacent-context topics. **h** Feature plots displaying topic scores across distinct samples. **i** Cell type abundance in topic-enriched regions. Spots with the highest topic scores (from Fig. **h**) are defined as topic-positive regions. **j** Gene set enrichment analysis based on the differences of gene ranking related to tumor-context topics reveals pathways enriched in the HG group

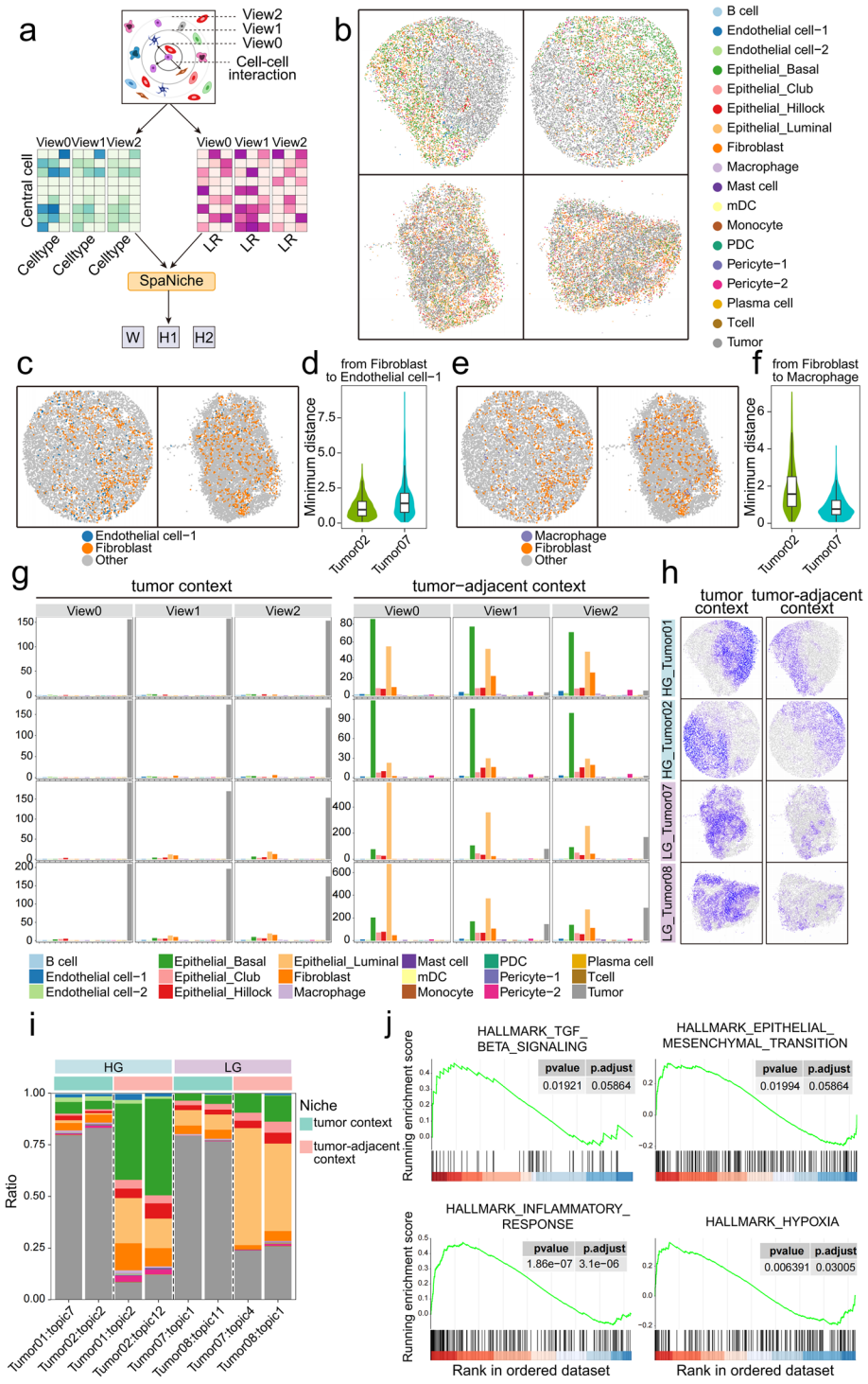

Fig. 5 (See legend on previous page.)

tumor-adjacent niches between tumors of different grades. By analyzing the high-weight genes corresponding to the tumor niche, we find that HG tumor niches exhibit high invasiveness and metastatic potential, enabling tumor cells to survive and thrive in an immune-active and hypoxic microenvironment (Fig. 5j). Moreover, we discover

that gene weight rankings provided by SpaNiche are more accurate than traditional gene expression levels for enrichment analysis, revealing more potential pathway differences (Additional file 1: Fig. S9d, e). Scoring cells with positive topic scores further confirm these pathway differences between the two groups (Additional file 1: Fig. S9f). SpaNiche also demonstrates high robustness. The cell type weights, ligand-receptor pair weights, and gene weights output by SpaNiche are highly correlated between the two tumor samples within the same group for the same niche (Additional file 1: Fig. S9g).

Identification of ecotypes linked to early Alzheimer's disease using SpaNiche

To demonstrate the broad applicability of SpaNiche in uncovering complex colocalization patterns within the tissue microenvironment, we reanalyze ST data from 18 cerebral cortex samples [40, 41] using single-nucleus RNA-seq data from 3 dorsolateral prefrontal cortex (DLPFC) samples as a reference [42] to define five major classes of ecotypes in the cortex (Fig. 6a). The first category predominantly comprises excitatory neurons. The second primarily consists of inhibitory neurons. The third is enriched with astrocytes, microglia, and oligodendrocytes. The fourth is primarily made up of astrocytes and microglia, while the fifth is enriched with oligodendrocyte progenitor cells. These ecotypes and their associated primary cell types show similar distributions across cortical regions, with some regions enriched for specific ecotypes (Fig. 6b). For example, Ecotype3 is predominantly found in white matter regions, Ecotype12 is found in layer 5 (L5), and Ecotype1 is found in layer 2 (L2).

We observe a certain level of ecotype conservation in different cortical functional regions and physiological states. For instance, Ecotype3, enriched with astrocytes, microglia, and oligodendrocytes, is predominantly located in white matter regions across DLPFC and temporal cortex samples, including those from control and Alzheimer's disease (AD) cases (Fig. 6d). Ecotype3 is linked to biological processes like ATP metabolism, axonal development, and neuronal myelination (Additional file 1: Fig. S10a). These functions reflect the characteristics of white matter, which is rich in axons and myelin, enabling rapid neural signal transmission and interregional functional coordination.

The pathogenesis of AD involves the accumulation of several toxic proteins, among which β-amyloid and Tau proteins are the most prominent [43]. Spots from AD-related ST samples are categorized into four groups based on pathological annotations: those containing only β-amyloid, only Tau protein, both β-amyloid and Tau proteins, or neither β-amyloid nor Tau proteins (Additional file 1: Fig. S10c). Spots enriched with Tau protein, whether containing only Tau or both Tau and β-amyloid, exhibit higher scores for Ecotype12 or Ecotype1 (Fig. 6e). Pathways enriched in Ecotype12 and Ecotype1 reveal transcriptional characteristics associated with synaptic signaling and vesicle-mediated transport (Fig. 6c and Additional file 1: Fig. S10b). Ecotype12 and Ecotype1 are predominantly enriched in excitatory neurons, indicating a specific association between Tau protein accumulation and excitatory neurons [44]. These findings suggest that Tau protein-mediated pathological mechanisms may differentially affect excitatory and inhibitory synaptic functions, resulting in abnormal excitatory neuronal activity. This dysregulation likely contributes to neural network dysfunction observed in AD [44, 45].

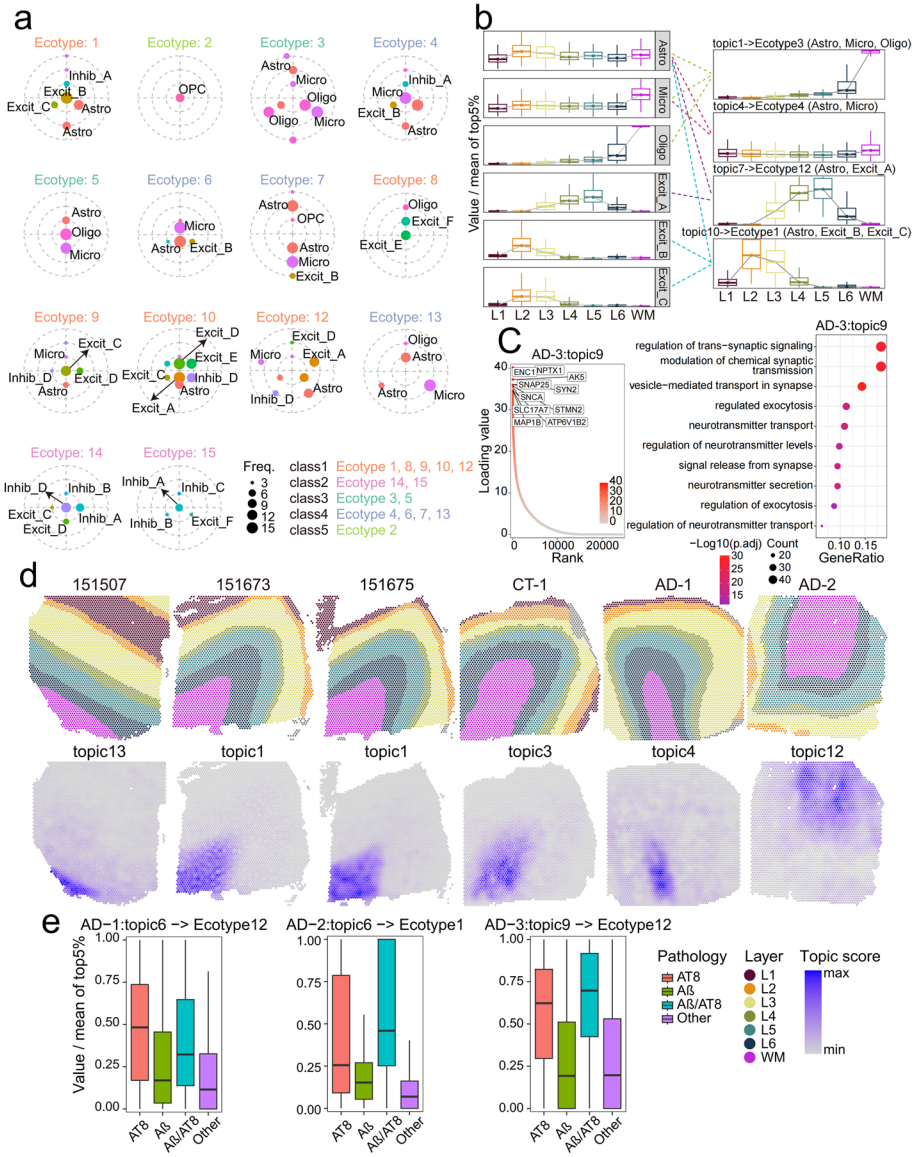

Fig. 6 SpaNiche deconstructs cortical and Alzheimer's disease-associated ecotypes. **a** Identification of 14 ecotypes in cortical ST data using SpaNiche. **b** Distribution of ecotypes and their cellular compositions across cortical subregions. **c** Gene weight rankings for ecotype 12-associated topics (left panel). Pathway enrichment analysis of high-weight genes (right panel). **d** Spatial mapping of cortical subregions across samples (top panel). Spatial distribution of ecotype 3-associated topic scores (bottom panel). **e** Box plots comparing normalized topic scores across pathological groups

Discussion

We present SpaNiche, a novel framework incorporating several algorithmic innovations to address the challenge of identifying spatial colocalization patterns and cellular interactions in the tissue microenvironment using ST data. We demonstrate the superior performance of SpaNiche compared with other commonly used methods, using both simulated and real ST datasets. SpaNiche identifies colocalization patterns among diverse cell types across distinct spatial perspectives and infers potential ligand-receptor (LR) interactions using a graph-regularized joint NMF algorithm.

Following the topic decomposition by SpaNiche, we introduce the concept of "ecotype" for spatial colocalization, which is commonly used in oncology to describe localized microenvironments composed of diverse cell types [27]. Specifically, we define ecotypes as spatial colocalization patterns that are consistently observed across multiple samples and involve potential intercellular interactions. This concept is further extended beyond oncology, broadening its applicability to multi-sample ST datasets. Additionally, we develop a built-in module to estimate ecotype proportions in bulk transcriptomics data, enabling the integration of spatial colocalization patterns with clinical phenotypes. SpaNiche is a novel tool that provides such a comprehensive range of insights simultaneously. Using SpaNiche, we explore ecotypes within the tissue microenvironment across varying cellular resolutions and pathological conditions, demonstrating its broad applicability in uncovering complex spatial colocalization patterns.

NMF has been widely applied in inferring cell-state co-occurrence patterns in bulk transcriptomics [27], as well as single-cell transcriptomics clustering [46, 47], data integration [48], and ST deconvolution [49]. However, its application in detecting spatial colocalization patterns remains underexplored. EcoTyper, an NMF-based method, was originally designed to investigate multicellular communities from conventional bulk transcriptomic data [27]. However, due to the lack of refined cell-type annotation and the absence of essential spatial information, its applicability in analyzing spatial colocalization patterns is limited. Our approach, based on graph-regularized joint NMF, can effectively integrate multimodal information and spatial localization, enabling the discovery of complex spatial colocalization patterns. In addition, correlation-based methods in their original form are limited to identifying colocalization between two cell types, as exemplified by STopover [15]. Although some studies have employed STopover to detect potential colocalization between two cell types, this approach risks oversimplifying the intricate complexity of the tissue microenvironment [50]. SpaNiche overcomes this limitation by offering a more comprehensive and nuanced characterization of spatial colocalization patterns.

As single-cell resolution ST technologies continue to advance, SpaNiche can be seamlessly adapted to high-resolution datasets, thereby eliminating the need for deconvolution analysis. In this context, the deconvoluted spot-cell type matrix can be replaced by the actual cell type abundance matrix derived through binning. This highlights the adaptability of SpaNiche in spatial research across varying data resolutions.

A limitation of SpaNiche should be noted. For multi-cell resolution ST data, SpaNiche depends on external algorithms to construct the cell type abundance matrix, which directly affects the accuracy of spatial colocalization patterns and related outcomes. We therefore advise users to approach the deconvolution step with caution, carefully selecting both the reference single-cell dataset and the deconvolution method that are most appropriate for their research context.

In addition, SpaNiche assumes a shared latent space for cell type abundance and ligand-receptor information to characterize spatial niches defined by both cellular composition and intercellular communication. This assumption may be limited in settings where the spatial distributions of cell abundance and LR activity are misaligned, such as tumor–immune interfaces, leading to increased complexity in interpretation. Exploring

modality-specific latent spaces with appropriate alignment strategies is a promising direction for future work.

Conclusions

SpaNiche is a computational framework designed to investigate spatial colocalization patterns and cellular interactions within the tissue microenvironment. In benchmark analyses of simulated and real datasets, SpaNiche demonstrates superior overall performance compared to other methods. Additionally, SpaNiche includes a built-in module for estimating ecotype proportions in bulk transcriptomics data, providing a valuable link between spatial colocalization patterns and clinical phenotypes. As ST data from clinical research continues to grow, we anticipate that SpaNiche would uncover mechanistic insights into diverse biological processes and play a pivotal role in addressing key challenges in therapeutic strategies. SpaNiche has been implemented in a computationally efficient package and can be applied widely to ST data.

Methods

Data preprocessing

SpaNiche takes the raw expression matrix of ST data $X \in \mathbb{R}^{n \times g}$ and its spatial coordinates $S = \{(x_i, y_i)\}_{i=1}^{n}$ as input, where n and g represent the number of spots and genes, respectively. To characterize the microenvironment around each spot, we aggregate information from its neighboring spots across multiple spatial views and describe them based on two complementary perspectives: cellular composition and cell–cell interactions. Specifically, views correspond to distinct spatial scales and are operationalized as different spatial distance thresholds.

Cellular composition

Given that multiple cell types may coexist within a single spot, we employ a deconvolution method, cell2location [12], to estimate the abundance of cell types. In addition to the raw expression matrix X, we provide cell2location with a reference scRNA-seq dataset annotated with t cell types. When running this method, we set the training iterations to 20,000, the number of cells per spot to 15, and the parameter "*detection_alpha*" to 20. This process yields a spot-by-cell-type matrix $M \in \mathbb{R}^{n \times t}$, which is used for subsequent analyses. It is generally advisable to evaluate multiple deconvolution methods in advance to identify the one most appropriate for the biological context. In our analysis, we also tested RCTD [51], which similarly produced highly reliable estimates of cellular composition (Additional file 1: Fig. S11).

The cellular composition within an individual spot is insufficient to fully represent the entire microenvironment. Therefore, we further calculate the average cell type abundance of neighboring spots to characterize the cell composition patterns across the microenvironment (Additional file 1: Fig. S1a). We first determine the pairwise distances between spots from the spatial coordinates S as:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

In SpaNiche, spots within a distance smaller than a threshold are defined as first-order neighbors, whereas spots within a distance smaller than a larger threshold are defined as second-order neighbors. Specifically, we construct a 0–1 adjacency matrix $A^{(1)} \in \mathbb{R}^{n\times n}$, where $A_{ij}^{(1)} = 1$ if $d_{ij} \le c_1$, and another matrix $A^{(2)}$ is defined as:

$$A_{ij}^{(2)} = \begin{cases} 1, d_{ij} \le c_2 \\ 0, otherwise \end{cases}.$$

Where, $c_1 < c_2$ are two distance thresholds. These adjacency matrices represent different levels of neighboring relationships. From each $A^{(k)}$, we identify the neighbors for each spot and compute the average cell type abundance:

$$M^{(k)} = \left(D^{(k)}\right)^{-1} A^{(k)} M \in \mathbb{R}^{n\times t}, k = 1,2.$$

where $D^{(k)}$ is the diagonal degree matrix of $A^{(k)}$. By concatenating these matrices, we obtain $X_1 = \left[M, M^{(1)}, M^{(2)}\right] \in \mathbb{R}^{n\times 3t}$. Through this multi-level aggregation of neighboring spots, X_1 provides a comprehensive representation of cell type composition within each microenvironment.

Cell–cell interactions

We obtain curated ligand-receptor pairs from a public ligand-receptor database [52]. Based on the ST expression profiles, we extract the ligand expression matrix $L \in \mathbb{R}^{n\times p}$ and receptor expression matrix $R \in \mathbb{R}^{n\times p}$. Here, p is the number of ligand-receptor pairs (LRPs), L_{ij} and R_{ij} correspond to the ligand expression and receptor expression of the j-th LRP in the i-th spot. If a ligand or receptor consists of multiple subunits, their average expression is used. For each spot and its neighbors, the interaction strength for each LRP is measured by the product of ligand and receptor expression.

Similarly to the previous step, we account for the potential interactions across varying distance ranges. Using thresholds c_1 and c_2, we construct two adjacencies $\widetilde{A}^{(1)}$ and $\widetilde{A}^{(2)}$ as following:

$$\widetilde{A}_{ij}^{(1)} = \begin{cases} 1, d_{ij} \le c_1 \\ 0, otherwise \end{cases}, \widetilde{A}_{ij}^{(2)} = \begin{cases} 1, c_1 < d_{ij} \le c_2 \\ 0, otherwise \end{cases}.$$

Then, we calculate the average ligand and receptor expression for neighboring spots:

$$L^{(k)} = \left(\widetilde{D}^{(k)}\right)^{-1} \widetilde{A}^{(k)} L \in \mathbb{R}^{n\times p}, R^{(k)} = \left(\widetilde{D}^{(k)}\right)^{-1} \widetilde{A}^{(k)} R \in \mathbb{R}^{n\times p}, k = 1,2.$$

where $\widetilde{D}^{(k)}$ is the diagonal degree matrix of $\widetilde{A}^{(k)}$. The average interaction strength at different distances is then computed through element-wise multiplication:

$$P_L{}^{(k)} = R^{(k)} \odot L, P_R{}^{(k)} = L^{(k)} \odot R, k = 1,2.$$

Here, $P_L{}^{(k)}$ and $P_R{}^{(k)}$ represent the spot's role as a signal sender (ligand) and receiver (receptor), respectively. By merging these matrices, we obtain $X_2 = \left[P_L{}^{(1)}, P_L{}^{(2)}, P_R{}^{(1)}, P_R{}^{(2)}\right] \in \mathbb{R}^{n\times 4p}$. This representation captures the cell–cell interaction patterns centered at each spot across different spatial ranges. (Additional file 1: Fig. S1b).

Joint NMF with graph regularization

After obtaining the abundance matrix X_1 and the ligand-receptor matrix X_2, SpaNiche employs a multimodal joint NMF algorithm to derive a low-dimensional embedding that integrates information from both matrices. Joint NMF has been widely applied in cancer-related bulk multi-omics research [53], due to its ability to incorporate data from multiple modalities simultaneously and uncover biologically meaningful patterns. In the context of SpaNiche, this approach facilitates the integration of cell type distributions and cell–cell interaction patterns, providing a comprehensive and unified representation of the microenvironment.

In the framework of joint NMF, we assume that the non-negative matrix factorizations of the two modalities share the same factor matrix. Specifically, we aim for:

$$X_1 \approx WH_1, X_2 \approx WH_2, W \in \mathbb{R}^{n\times m} \geq 0, H_1 \in \mathbb{R}^{m\times 3t} \geq 0, H_2 \in \mathbb{R}^{m\times 4p} \geq 0.$$

Each row of the loading matrix H_i is referred to as a "topic", representing a tissue microenvironment characterized by specific cell types and cell–cell interactions. The hyperparameter m defines the dimensionality of NMF, which corresponds to the number of topics and needs to be specified in advance.

Then joint NMF estimates W and H_i by solving the following optimization problem:

$$\min_{W\in\mathbb{R}^{n\times m}, H_1\in\mathbb{R}^{m\times 3t}, H_2\in\mathbb{R}^{m\times 4p}} \sum_{i=1}^{2} \omega_i ||X_i - WH_i||_F^2, \text{s.t.} H_i, W \geq 0.$$

Here, $\|\cdot\|_{\mathrm{F}}^2$ denotes the Frobenius norm, and ω_i are the weights of the two components. To incorporate the similarity between neighbors, we further introduce a graph regularization term to the objective function. This term encourages adjacent spots to share similar weights, which promotes spatial smoothness. The modified optimization problem becomes:

$$\min_{W,H_1,H_2} \sum_{i=1}^{2} \omega_i ||X_i - WH_i||_F^2 + \lambda \mathrm{Tr}(W^T \Lambda W), \text{s.t.} H_i, W \geq 0.$$

In the setup of graph NMF, Λ is the graph Laplacian derived from a weighted adjacency matrix:

$$A = \left[a_{ij}\right] \in \mathbb{R}^{n\times n}, \textit{where } a_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right).$$

Here, σ represents the width of the Gaussian kernel, which determines how the similarity between spots decreases with increasing spatial distance. Then Λ is defined as:

$$\Lambda = \mathrm{diag}\left(\sum\nolimits_j a_{ij}\right) - A.$$

To solve the above optimization problem, we first ignore the graph regularization term and provide the initial values for W and H_i based on the joint NMF setting. To better balance the contributions of the two modalities, we predefine the default values of ω_i based on the scale and dimensionality of the two components:

$$\theta_i^0 = \left\{ \frac{\max\left(\frac{\|X_1\|_F}{N_1}, \frac{\|X_2\|_F}{N_2}\right)}{||X_i||_F} \right\} \times \left\{ \frac{\max(N_1, N_2)}{N_i} \right\},$$

where $N_i = \#(X_i)$ is the number of elements in matrix X_i. Then, SpaNiche employs the nmf.mnnals() function from R package "IntNMF" [53] to derive the initial values for W and H_i, denoted as W^0 and H_i^0, respectively.

Next, SpaNiche takes these values as the starting point and iteratively minimizes the objective function with the graph regularization term. Here, we additionally require the user to input the expected weights for the two modalities, ω_1^0 and ω_2^0, to priorly control the importance of both. The iteration process is summarized in Algorithm 1.

Algorithm 1 Joint NMF

Require: $X_1,\ X_2,\ \boldsymbol{\omega^0},\ A,\ D,\ \Lambda,\ \epsilon,\ T$
Initialize: $W \leftarrow W^0,\ H_1 \leftarrow H_1^0,\ H_2 \leftarrow H_2^0,\ iter \leftarrow 0.$
Calculate:

$$\mathcal{L}_0 \leftarrow \sum_{i=1}^{2} \omega_i ||X_i - WH_i||_F^2 + \lambda \mathrm{Tr}(W^T \Lambda W),$$

$$\omega_i \leftarrow \frac{\omega_i^0}{||X_i - WH_i||_F^2}.$$

While $\mathcal{L}_n - \mathcal{L}_{n-1} > \epsilon$ or iter $< T$, **do**

$$W \leftarrow W \odot \frac{\sum_{i=1}^{2} \omega_i X_i H_i^T + \lambda DW}{\sum_{i=1}^{2} \omega_i W H_i H_i^T + \lambda \Lambda W},$$

$$H_i \leftarrow H_i \odot \frac{W^T X_i}{W^T W H_i},$$

$$\text{iter} \leftarrow \text{iter} + 1,$$

$$\mathcal{L}_n \leftarrow \sum_{i=1}^{2} \omega_i ||X_i - WH_i||_F^2 + \lambda \mathrm{Tr}(W^T \Lambda W).$$

Inference of pathways regulated by cell–cell interactions

In the fields of single-cell and ST, most methods for inferring cell–cell interactions primarily focus on the interplay between cells, often characterized by the expression levels of receptors and ligands. However, very few approaches account for both intercellular and intracellular signaling, where the latter refers to downstream gene expression changes that occurs within cells engaged in interactions. In this section, SpaNiche incorporates the R package "scSeqComm" [21], a tool specifically designed to identify and quantify both intercellular and intracellular signaling from single-cell transcriptomics

data while also providing functional characterizations of cellular communication. To quantify the evidence of ongoing intracellular signaling influenced by cellular communication, "scSeqComm" integrates static prior knowledge of signaling pathways and transcription factor (TF)-target gene regulatory networks from public databases, along with dynamic information from the analyzed single-cell transcriptomics dataset.

The methodology employed by "scSeqComm" for quantifying intracellular signaling can be summarized as follows: First, "scSeqComm" translates known signaling pathways into directed graphs and calculates the prior association scores between receptors and TFs by leveraging the topological structure of the graph through the Personalized PageRank (PPR) algorithm [54]. Subsequently, for each cell subpopulation and signaling graph, "scSeqComm" assesses the activity of TFs by measuring the changes in expression levels of their target genes. Finally, for each combination of receptors, cell subpopulations, and signaling pathways, "scSeqComm" integrates the receptor-TF association scores with TF activity scores to determine the strength of receptor involvement in ongoing intracellular communication. To functionally characterize the effects of cell–cell communication, "scSeqComm" performs Gene Ontology (GO) enrichment analysis on the target genes implicated in the detected intracellular signaling.

After obtaining the shared factor matrix W, for each topic, we identify the spots that are positive and negative for the topic, defined as those in the top 20% and bottom 20% of topic scores, respectively. These spots are used to construct a refined expression matrix annotated with positive and negative labels. Subsequently, from the loading matrix H_2, ligand-receptor pairs most strongly associated with each topic are selected. Finally, by utilizing the "scSeqComm_analyze" and "scSeqComm_GO_analysis" functions, we can derive functional insights of cellular responses driven by intracellular signal transduction.

Construction of multi-sample analysis modules

Applying joint NMF to a single ST sample allows us to uncover the relationship between spatial spots and latent topics. However, this approach has two key limitations: first, it is unclear whether these topics are universally applicable across multiple samples, and second, there is no direct linkage between the topics and sample phenotypes. To address these limitations, we extend the downstream analysis capabilities of SpaNiche. Specifically, we introduce the concept of "ecotypes" to represent spatial co-localization patterns that are consistently observed across diverse samples. Furthermore, to integrate phenotypic information from publicly available databases, we develop a novel deconvolution approach to estimate the abundance of ecotypes in bulk samples, thereby establishing a link between ecotypes and phenotypes.

Definition of ecotype

For ST datasets with multiple samples, joint NMF is first applied to each individual sample. The resulting H_1 matrices from all samples are then combined into a single matrix after assessing the distributions of topic weights and applying normalization when

appropriate. Consistency clustering is subsequently performed on the combined matrix [22]. To determine the optimal number of clusters, we utilize the proportion of ambiguously clustered pairs (PAC) method [55]. In this context, each cluster represents a group of topics with highly similar cell distribution patterns, which we define as an "ecotype".

As described in the joint NMF section, the rows of the loading matrix H_i represent the importance of cell types ($i = 1$) and LRPs ($i = 2$) within each topic. We also quantify the characteristics of ecotypes from these two perspectives. Considering an ecotype comprising n_e topics, we extract the cell type importance vectors corresponding to these topics, denoted as $h_1^1, h_1^2, \ldots h_1^{n_e} \in \mathbb{R}^{3t}$. For each vector, we select the top 20% largest elements and map their positions to the corresponding cell types (e.g., elements at positions i, $i + t$, and $i + 2t$ correspond to cell type i). We then calculate the frequency of each cell type appearing among these top-ranked cell types and denote the results as a vector $h_1^{ecotype} \in \mathbb{R}^t$. Similarly, we extract the LRP importance vectors $h_2^1, h_2^2, \ldots h_2^{n_e} \in \mathbb{R}^{4p}$ and obtain $h_2^{ecotype} \in \mathbb{R}^p$.

Ecotype-based bulk transcriptome data deconvolution

Our goal is to estimate the presence of each ecotype in a sample from bulk RNA data. We denote the bulk gene expression as $y_1 \in \mathbb{R}^g$. From y_1, we further calculate the LRP strength from the expression of ligands and receptors, denoted as $y_2 \in \mathbb{R}^p$. Suppose e ecotypes are identified from ST data. Based on the definitions above, we derive feature vectors for each ecotype at the cell type and LRP levels and concatenate them into matrices $H_1^{ecotype} \in \mathbb{R}^{t \times e}$ and $H_2^{ecotype} \in \mathbb{R}^{p \times e}$, respectively.

We use the vector $\beta \in \mathbb{R}^e$ to represent the ecotype abundance to be estimated in the bulk data. The cell type proportion of bulk data can be given by $H_1^{ecotype}\beta$. Using the annotated scRNA-seq data in the deconvolution step, we calculate the average expression of each cell type, represented as a matrix $E \in \mathbb{R}^{g \times t}$. This allows us to establish a relationship between ecotypes and the bulk gene expression at the cell type level:

$$y_1 \approx EH_1^{ecotype}\beta.$$

At the LRP level, we directly have:

$$y_2 \approx H_2^{ecotype}\beta.$$

Let $Z_1 = (EH_1^{ecotype})^T$ and $Z_2 = (H_2^{ecotype})^T$. We estimate β using the least square method:

$$\min_{\beta} \|y_1 - Z_1^T\beta\|_{\mathrm{F}}^2 + \lambda\|y_2 - Z_2^T\beta\|_{\mathrm{F}}^2.$$

Here, the parameter λ is used to balance the two components of the objective function. During the optimization process, we iteratively adjust λ to ensure that the ratio of the two loss functions approaches a predefined expected value, r_0.

Algorithm 2 Estimation of ecotype proportions

Require: $y_1,\ y_2,\ Z_1,\ Z_2,\ r_0,\ \epsilon,\ T$

Initialize:

$$\beta \leftarrow (Z_1 Z_1^T + Z_2 Z_2^T)^{-1}(Z_1 y_1 + Z_2 y_2),$$
$$r \leftarrow \frac{\|y_1 - Z_1^T \beta\|_F^2}{\|y_2 - Z_2^T \beta\|_F^2},$$
$$\text{iter} \leftarrow 0.$$

While $|r - r_0| > \epsilon$ and $\text{iter} < T$, **do**

$$\lambda \leftarrow \frac{r}{r_0},$$
$$\beta \leftarrow (Z_1 Z_1^T + \lambda Z_2 Z_2^T)^{-1}(Z_1 y_1 + \lambda Z_2 y_2),$$
$$r \leftarrow \frac{\|y_1 - Z_1^T \beta\|_F^2}{\|y_2 - Z_2^T \beta\|_F^2},$$
$$\text{iter} \leftarrow \text{iter} + 1.$$

The correlation between ecotypes and clinical phenotypes

After obtaining the proportion of each ecotype in bulk transcriptome samples, SpaNiche investigates the relationship between ecotypes and clinical phenotypes, such as survival outcomes and tumor staging. For this study, we retrieve phenotypic information for colon cancer samples from the TCGA database. Then we group them according to tumor staging and analyze the distribution of each ecotype across these groups. Additionally, based on the relative abundance of each ecotype in individual tumor samples, we stratify the samples into high- and low-content groups. Survival analyses are then performed to assess the prognostic significance of ecotypes in guiding clinical outcomes.

Benchmarking with simulated datasets

To comprehensively evaluate the statistical performance of SpaNiche, we generate a series of simulated ST datasets with manually defined niche types. We then assess the ability of our method to identify niches from multiple perspectives and compare it with other methods. Here, a niche type is characterized by a set of colocalized cell types within a cluster of spots, as well as their ligand-receptor interactions.

Generation of simulated datasets

For simplicity, we assume that the ST spots are arranged in a square layout with a side length of 60, where each spot is surrounded by 6 neighbors. Such a simulated dataset contains $N = 1800$ spots, denoted as $S = \{s_1, \ldots, s_n\}$. To simulate the cell proportions and gene expression levels of each spot, we use a scRNA-seq dataset from breast cancer as reference [23]. This reference dataset has been annotated into $t = 29$ cell types, including various tumor and immune subtypes, denoted as $C = \{c_1, \ldots, c_t\}$.

In general, a niche should contain several cell types with relatively high abundance, and these cell types form spatial clusters in certain spots. We assume that there are $N_{\text{niche}} = 5$ niches in total, each containing k cell types, and these cell types are enriched in n spots. We use $C_l \subseteq C$ and $S_l \subseteq S$ to represent the two sets corresponding to the l-th

niche. We randomly sample k cell types from C to form C_l, then randomly sample two center spots and select the nearest $n/2$ spots around each center to form S_l. We sequentially generate C_l and S_l, ensuring that all S_l are non-overlapping to guarantee that each spot belongs to at most one niche.

We next simulate the cell type proportions for each spot. Specifically, we aim to generate a non-negative matrix $X \in \mathbb{R}^{N\times t}$, where the sum of each row equals 1. The element at the i-th row and j-th column, X_{ij}, represents the proportion of cell type c_j in spot s_i. If spot i is not assigned to any niche, we assume that all cell types have equal occurrence probabilities and generate them using the following Dirichlet distribution:

$$\boldsymbol{x}_i = (X_{i1}, \ldots, X_{it}) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_t), \alpha_1 = \cdots = \alpha_t = 1.$$

If a spot i is located within niche l, we naturally assume that the cell types in C_l have higher proportions, while still allowing other cell types to appear. In this case, we first sample the total proportion of C_l from a uniform distribution $U(\beta_1, \beta_2)$, then use Dirichlet distributions to separately generate the proportions for cell types in C_l and for the remaining cell types:

$$p_i \sim U(\beta_1, \beta_2),$$

$$\left\{X_{ij} : c_j \in C_l\right\} \sim p_i * \text{Dirichlet}\left(\alpha_1^l, \ldots, \alpha_{|C_l|}^l\right), \alpha_1^l = \cdots = \alpha_{|C_l|}^l = \alpha,$$

$$\left\{X_{ij} : c_j \notin C_l\right\} \sim (1 - p_i) * \text{Dirichlet}\left(\alpha_1, \ldots, \alpha_{t-|C_l|}\right), \alpha_1 = \cdots = \alpha_{t-|C_l|} = 1.$$

Here, α is an adjustable parameter that controls the relative differences among the cell types. When α is large, the proportions of different cell types have smaller differences. When α is small, a spot is usually dominated by a single cell type. Since the true proportions cannot be directly observed in real data, and can only be estimated using deconvolution algorithms, we further add noise to X to simulate the errors in deconvolution:

$$X_{ij}^{\text{obs}} = \max(X_{ij} + \varepsilon_{ij}, 0), \varepsilon_{ij} \sim N\left(0, \sigma^2\right).$$

Then we use X^{obs} as the input for each method in our evaluation.

After generating the cell-type proportion matrix $X \in \mathbb{R}^{N\times t}$, we construct a simulated spatial expression matrix that retains realistic statistical properties. Specifically, we use a breast cancer scRNA-seq dataset [8] as the reference and compute the mean and variance of each gene per cell type, denoted as $U \in \mathbb{R}^{t\times g}$ and $V \in \mathbb{R}^{t\times g}$, where U_{kg} and V_{kg} represent the mean and variance of gene g within cell type c_k, respectively. To map continuous proportions to discrete cell counts, we assume a fixed total number of cells per spot M, and compute the number of cells belonging to type c_k in spot s_i as $n_{ik} = \lfloor M \cdot X_{ik} \rceil$, forming the spot-cell type count matrix $N_{\text{cell}} \in \mathbb{N}^{N\times t}$.

Considering the potential cell–cell interactions within the niche, we also simulate the expression patterns of niche-specific ligand-receptor pairs. Given p candidate LR pairs, we compute LR scores across spots using the reference data and select the top $10 \times N_{\text{niche}}$ highest-expressing pairs. These are randomly partitioned into N_{niche} disjoint subsets, each containing 10 LR pairs and corresponding to one spatial niche (denoted as

M_l^{LR} for niche l). For all cells in spots belonging to niche l (spot set S_l), the mean expression of ligand and receptor genes in M_l^{LR} is increased by a factor of $\sqrt{1.5}$, such that for all $g \in \mathcal{G}(M_l^{LR})$,

$$U_{kg}^{(l)} = \sqrt{1.5} \cdot U_{kg}.$$

While non-LR genes remain unchanged, i.e., $U_{kg}^{(l)} = U_{kg}$, where $g \in \mathcal{G}(M_l^{LR})$ denotes the set of all ligand and receptor genes belonging to the LR pairs assigned to niche l. For spots not assigned to any niche, the unmodified mean matrix U is used. Under this construction, the expected LR product signal is strengthened by approximately 1.5-fold within each niche, introducing spatially localized ligand-receptor signaling "hotspots".

Based on the niche-adjusted mean matrix $U_{kg}^{(\ell(i))}$, we next simulate gene expression at the single-cell level using a ZINB distribution to capture both over-dispersion and dropout. Specifically, for the c-th cell type c_k in spot s_i, expression of gene g is sampled as:

$$Y_{i,c,g} \sim \text{ZINB}(\mu = U_{kg}^{(\ell(i))}, \phi = \phi_{kg}, \theta = 0.3),$$

where θ represents the dropout probability, ϕ_{kg} is derived from the cell-type-specific variance V_{kg}, and $\ell(i)$ denotes the niche assignment of spot s_i (defaulting to U_{kg} if the spot is not part of a niche). Under this model, expression equals zero with probability θ. Otherwise, it follows a negative binomial distribution with mean μ and dispersion ϕ, introducing realistic noise structure and sparsity into the simulation.

Finally, simulated single-cell expression values are aggregated into spot-level profiles,

$$Y_{ig}^{\text{spot}} = \sum_{k=1}^{t} \sum_{c=1}^{n_{ik}} Y_{i,c,g}^{(k)}.$$

The number of cell types within a single niche, k, is selected from {3, 4}, and the number of spots, n, is chosen from {50, 100, 150}. The lower bound of the niche cell type proportion, β_1, is selected from {0.3, 0.5, 0.7}, while the upper bound β_2 is fixed at 0.8. The Dirichlet parameter α used to generate the cell type proportions is selected from {1, 10, 25}. The noise level σ is selected from {0, 0.05, 0.1}. We obtain a total of 162 different settings, and we generate three replicates of simulated data under each setting.

Comparison with existing methods

NMF [56] is a matrix decomposition technique that factorizes the celltype abundance matrix into non-negative low-dimensional modules, which can be interpreted as distinct niches representing different microenvironments. Pearson correlation-related method [57], in contrast, groups cell types with highly similar distribution patterns and considers them as belonging to the same microenvironment. Misty [16] employs a multiview machine learning framework to decompose gene expression variation into contributions from different spatial scales, including cell-intrinsic, local neighborhood, and broader tissue contexts, thereby quantifying the influence of the microenvironment on cellular states. In comparison, STopover [15] adopts a topology-based strategy to identify and quantify spatial co-localization patterns between cell types or gene features. By detecting

connected spatial components and applying permutation testing to evaluate the significance of overlaps, STopover provides a robust measure of spatial interactions within the tumor microenvironment. All methods are applied with their default parameter settings in this study.

Selection of topic numbers for different methods

In the benchmark using simulated data, the number of topics for SpaNiche and NMF is uniformly set to 15. For methods based on Correlation, STopover, and MISTy, the similarity matrices of cell-type spatial distributions do not directly provide colocalization patterns between cells, so we further perform community clustering to obtain colocalization results. Default clustering parameters are used in these analyses.

Evaluation methods

We evaluate SpaNiche from three aspects: the ability to identify niches, the accuracy of localizing niche-related spots, and the capability to discover niche-specific ligand-receptor pairs.

We first evaluate the ability to identify niches. Both SpaNiche and NMF can generate a topic-by-cell type score matrix. In practice, we focus on the cell types with higher scores in each topic and consider these cell types as forming a potential microenvironment. Denote the topic matrix as $H^{\text{topic}} \in \mathbb{R}^{N_{\text{topic}} \times t}$. Therefore, we say that niche l can be identified by the topic i (i.e., the i-th row of H^{topic}), if and only if the following condition is met:

$$H^{\text{topic}}_{ij_1} > H^{\text{topic}}_{ij_2}, \forall c_{j_1} \in C_l, c_{j_2} \notin C_l.$$

As for correlation-based methods such as Pearson correlation and the Jaccard index (e.g. STopover [15]), we first apply community clustering (louvain) to cluster cell types based on the correlation matrix (Additional file 1: Fig. S4a, b). If one of the clusters matches the cell types in C_l, we consider that this method has identified niche l. In this way, we count the number of niches identified by each method, $N_{\text{niche}'}$, and define sensitivity as:

$$\text{SEN}_{\text{niche}} = \frac{N_{\text{niche}'}}{N_{\text{niche}}}.$$

In addition to niche identification, both SpaNiche and NMF can obtain a spot-by-topic weight matrix $W \in \mathbb{R}^{N \times N_{\text{topic}}}$, allowing us to deduce which spots correspond to each niche. If topic i corresponds to niche l, we take its weight vector $\boldsymbol{w}_i = (W_{1i}, \ldots, W_{Ni})$, then evaluate whether it can distinguish between spots within S_l and those outside S_l. This evaluation is framed as a binary classification problem, and we use the area under the receiver operating characteristic curve (ROC AUC) as the evaluation metric. If a niche is not identified by any topic, we calculate the AUC for each column of W and take the maximum value. If multiple topics correspond to one niche, we sum all the topic weight vectors spot-wise to obtain a new weight vector, then use it to calculate the AUC.

Finally, we compute the AUC for each niche and take the average to measure the overall performance.

SpaNiche is also capable of discovering marker ligand-receptor pairs specific to each niche. Based on the topic-by-ligand-receptor score matrix $H^{\text{lr}} \in \mathbb{R}^{N_{\text{topic}} \times p}$, we evaluate whether the topic corresponding to niche l can reflect its ligand-receptor markers M_l^{lr}. If topic i corresponds to niche l, and the top 10 ligand-receptor pairs in the score vector $\left(H_{i1}^{\text{lr}}, \ldots, H_{ip}^{\text{lr}}\right)$ include an element from M_l^{lr}, we consider that the method has identified this marker ligand-receptor pair for niche l. Denote the number of the identified ligand-receptor pairs by N_l^{lr}, we define sensitivity as:

$$\text{SEN}_l^{\text{lr}} = \frac{N_l^{\text{lr}}}{10}, l = 1, \ldots, N_{\text{niche}}.$$

Finally, we compute the average sensitivity across all niches.

Benchmarking with real dataset

In performance validation using real ST data from a human lymph node, we include comparisons with NicheCompass [26] and EcoTyper [27]. NicheCompass constructs gene programs and clusters spatial observations based on their scores across these gene programs. Following the recommended workflow of NicheCompass, we obtain its latent embeddings and perform Leiden clustering to derive niche assignments. During ecotype identification using EcoTyper, InstaPrism is first applied to infer the cell type composition of each spot, as well as celltype specific gene expression profiles [58]. These outputs are used as the input to EcoTyper; therefore, step 1 and 2 of the standard EcoTyper pipeline are skipped. NMF is performed with 20 restarts, allowing a maximum of 10 states per cell type and requiring a minimum of two states to define an ecotype. EcoTyper infers co-occurrence patterns exclusively at the cell-state level and does not directly resolve cell subtypes, which limits the biological interpretability of the resulting ecotypes. To address this limitation, we approximate the subtype composition within each ecotype by assessing expression similarity between inferred cell states and predefined cell subtypes.

Selection of topic numbers for different methods

In the benchmark using human lymph node data, for SpaNiche, 10–14 candidate topics are initially considered. After manually inspecting the characteristics of each decomposed topic, 13 topics are found to best capture the intrinsic biological properties of the data. For NMF, we adopt the parameters from the cell2location study [12], setting the number of topics to 12.

For methods based on Pearson correlation or STopover, after obtaining the similarity matrices of cell-type spatial distributions, colocalization patterns are derived using community detection on undirected similarity networks. Specifically, weighted undirected graphs are constructed where edges represent pairwise similarity between cell types. To retain only biologically meaningful associations, edges with weights below a predefined threshold (0.65 for correlations and 0.37 for STopover scores) are removed. Community detection is then performed using the Louvain algorithm to identify groups of cell types exhibiting similar spatial patterns (cluster_louvain function in R).

For MISTy-derived interaction importance scores, a weighted directed graph is constructed in which each edge points from a predictor cell type to its target. Edges with importance < 2 are excluded to focus on high-confidence relationships. Community detection on this directed network is performed using the Infomap algorithm, which is well suited for capturing directional information flow (cluster_infomap function in R).

For EcoTyper, the number of ecotypes is automatically determined by the method. For NicheCompass, spatial clustering numbers between 5 and 11 produce similar results, and 11 is ultimately selected as the number of niches detected by this method.

Calculation of cell composition difference scores

Based on the spatial distribution characteristics of different cell types and previous analyses of this dataset, we define five spatial niches and the key cell types that should be present in each: (1) "B_activated and B_naive cells" should include at least B_activated and B_naive cells; (2) "vascular niche" should include at least Endo and VSMC; (3) "germinal center-associated niche" should include at least B_Cycling, B_GC_DZ, B_GC_LZ, B_GC_prePB, FDC, and T_CD4_TfH_GC cells; (4) "B_activated and B_plasma cells" should include at least B_activated and B_plasma cells; and (5) "germinal centers and B cell follicles" should include the six germinal center cell types plus B_activated and B_naive cells detectable at larger spatial scales. These predefined cell type-niche relationships constitute our reference standard (a binary matrix of reference niches × cell types).

Next, we process the results from each method.

(1) Correlation, STopover, and MISTy: These methods detect colocalization patterns based on community clustering, where each cell type belongs to only one cluster (topic). We reconstruct their results into a binary cell type × topic matrix, where a value of 1 indicates the inclusion of a cell type in a topic and 0 otherwise.
(2) NMF: NMF outputs an abundance matrix of cell type × factor (topic). We compute the global median of the matrix as a threshold and binarize the matrix, assigning 1 to values above the threshold and 0 to values below, to ensure consistency with other methods.
(3) SpaNiche, EcoTyper, and NicheCompass: Similarly, the final matrices are binarized to match the structure of the other methods.

After obtaining binary matrices for all methods, each topic from every method is compared pairwise with all reference niches. For each predefined niche, we calculate the difference only for the set of cell types labeled as 1 in the reference, using the sum of squared differences between the topic vector and the reference vector as the initial difference score. For a given method and niche, the minimum difference score among all topics is assigned as the method's difference score for that niche.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04069-z.

Additional file 1: Supplementary figures. Figures S1-11.

Additional file 2: Supplementary tables. Tables S1-2.

Acknowledgements
Not applicable.

Peer review information
Mingyao Li and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
R.X. and S.M. conceived and designed the project. S.H. performed the analyses and created the SpaNiche code, with help from J.T. on the algorithmic component. X.W. and Q.R. benchmarked it. S.H., Q.R., and X.W. wrote the manuscript with comments and input from all the co-authors. J.X. provided critical feedback for this project. R.X. and S.M. supervised the study. All authors reviewed the paper and approved the final version.

Funding
This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM118 to R.X.), the National Natural Science Foundation of China (12425110 and 12371286 to R.X.) the National Key Research and Development Program of China (2026ZD0553802 and 2024YFF0507404 to R.X., 2023YFC3603200 to S.M.), the Shanghai Rising-Star Program (23YF1421000 to S.M.), the Shanghai Municipal Commission of Education (JWAIYB-3 to S.M.), the Clinical Research Project of Shanghai Municipal Health Commission in Health Industry (20234Y0285 to S.M.), the Fundamental Research Funds for the Central Universities (YG2023QNA01 to S.M.), the Sichuan Science and Technology Program (2025YFHZ0069 to S.M.), the Chengdu Science and Technology Program (2024-YF05-01784-SN to S.M.) and the Sino-Russian Mathematics Center.

Data availability
All datasets analyzed in this study are publicly available as described below. Detailed information for all datasets is summarized in Additional file 2: Table S1–2.
Human lymph node datasets
The human lymph node ST dataset is available from the 10x Genomics website (https://www.10xgenomics.com/cn/datasets/human-lymph-node-1-standard-1-1-0). The scRNA-seq reference data from human secondary lymphoid organs (lymph nodes, spleen, and tonsils) used in this study is integrated and distributed as part of the cell2location [12] tutorial resources (https://cell2location.readthedocs.io/en/latest/).
Colorectal cancer datasets
The colorectal cancer ST datasets are obtained from six sources: 14 slides from Valdeolivas et al. [31] (https://zenodo.org/records/7760264) [62], 4 slides from Liu et al. [34], 2 slides from the 10x Genomics data repository (https://www.10xgenomics.com/datasets), 4 slides from Qi et al. [32], 4 slides from Wu et al. [30] (http://www.cancerdiversity.asia/scCRLM/), 4 slides from Wang et al. [33] (GEO accession: GSE225857) [63]. The processed scRNA-seq data and metadata for the human colorectal cohorts (SMC and KUL3) [59] are available in the GEO under accession numbers GSE132465 [64], GSE132257 [65], and GSE144735 [66].
Colorectal cancer liver metastasis datasets
The colorectal cancer liver metastasis ST datasets used in this study are obtained from three sources: 4 slides from Garbarino et al. [38] (GEO accession: GSE206552) [67], 4 slides from Wu et al. [30] (http://www.cancerdiversity.asia/scCRLM/), and 2 slides from Wang et al. [33] (GEO accession: GSE225857) [63]. The scRNA-seq reference data [60] of the colorectal cancer liver metastasis is available in the GEO under accession numbers GSE178318 [68].
Prostate cancer datasets
The prostate cancer dataset [39] used in this study, including both ST data and the reference scRNA-seq data, is available from Zenodo (https://zenodo.org/records/7526696) [69].
Brain datasets
The processed human dorsolateral prefrontal cortex (DLPFC) ST dataset is available from the Bioconductor package spatialLIBD [40, 61]. The human DLPFC single-nucleus transcriptome reference data [42] are available via the LIBD Globus share endpoints (https://research.libd.org/globus/). The human middle temporal gyrus (MTG) ST dataset [41] is available from the GEO under accession number GSE220442 [70].
Besides, SpaNiche is implemented as a computationally efficient R package for the general analysis of ST data. The package is available under an MIT license on GitHub (https://github.com/SiyuanHuang1/SpaNiche) [71] and Zenodo (https://doi.org/10.5281/zenodo.19179447) [72].

Declarations

Ethics approval and consent to participate
Ethics approval was not needed for this study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 8 September 2025 Accepted: 2 April 2026
Published online: 21 April 2026

References

1. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet. 2021;22:71–88.
2. Lovinger DM. Communication networks in the brain: neurons, receptors, neurotransmitters, and alcohol. Alcohol Res Health. 2008;31:196–214.
3. Mbeunkui F, Johann DJ Jr. Cancer and the tumor microenvironment: a review of an essential relationship. Cancer Chemother Pharmacol. 2009;63:571–82.
4. Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.
5. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell. 2022;185:1777-1792.e1721.
6. Russell AJC, Weir JA, Nadaf NM, Shabet M, Kumar V, Kambhampati S, et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. Nature. 2024;625:101–9.
7. Andersson A, Larsson L, Stenbeck L, Salmen F, Ehinger A, Wu SZ, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. Nat Commun. 2021;12:6012.
8. Liu Y, Xun Z, Ma K, Liang S, Li X, Zhou S, et al. Identification of a tumour immune barrier in the HCC microenvironment that determines the efficacy of immunotherapy. J Hepatol. 2023;78:770–82.
9. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol. 2021;39:1375–84.
10. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021;18:1342–51.
11. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun. 2022;13:1739.
12. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40:661–71.
13. Tang J, Chen Z, Qian K, Huang S, He Y, Yin S, et al. The interpretable multimodal dimension reduction framework SpaH-Dmap enhances resolution in spatial transcriptomics. Nat Cell Biol. 2026. https://doi.org/10.1038/s41556-025-01838-z.
14. Ru B, Huang J, Zhang Y, Aldape K, Jiang P. Estimation of cell lineages in tumors from spatial transcriptomics data. Nat Commun. 2023;14:568.
15. Bae S, Lee H, Na KJ, Lee DS, Choi H, Kim YT. Stopover captures spatial colocalization and interaction in the tumor microenvironment using topological analysis in spatial transcriptomics data. bioRxiv. 2022:2022.2011. 2016.516708.
16. Tanevski J, Flores ROR, Gabor A, Schapiro D, Saez-Rodriguez J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. Genome Biol. 2022;23:97.
17. Ren X, Zhong G, Zhang Q, Zhang L, Sun Y, Zhang Z. Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. Cell Res. 2020;30:763–78.
18. Armingol E, Baghdassarian HM, Lewis NE. The diversification of methods for studying cell-cell interactions and communication. Nat Rev Genet. 2024;25:381–400.
19. Wu L, Yan J, Bai Y, Chen F, Zou X, Xu J, et al. An invasive zone in human liver cancer identified by Stereo-seq promotes hepatocyte-tumor cell crosstalk, local immunosuppression and tumor progression. Cell Res. 2023;33:585–603.
20. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171:1611-1624. e1624.
21. Baruzzo G, Cesaro G, Di Camillo B. Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scseqcomm. Bioinformatics. 2022;38:1920–9.
22. Wilkerson MD, Hayes DN. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.
23. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53:1334–47.
24. 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;14:1155.
25. LeBien TW, Tedder TF. B lymphocytes: how they develop and function. Blood. 2008;112:1570–80.
26. Birk S, Bonafonte-Pardas I, Feriz AM, Boxall A, Agirre E, Memi F, et al. Quantitative characterization of cell niches in spatially resolved omics data. Nat Genet. 2025;57:897–909.
27. Luca BA, Steen CB, Matusiak M, Azizi A, Varma S, Zhu C, et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell. 2021;184:5482-5496 e5428.
28. Yin Y, Wu S. B cells recruitment promotes M2 macrophage polarization to inhibit inflammation during wound healing. Clin Exp Immunol. 2025. https://doi.org/10.1093/cei/uxaf002.
29. Kim YU, Liu X, Tanaka S, Tran DQ, Chung Y. Regulation of germinal center reactions by B and T cells. Antibodies. 2013;2:554–86.
30. Wu Y, Yang S, Ma J, Chen Z, Song G, Rao D, et al. Spatiotemporal immune landscape of colorectal cancer liver metastasis at single-cell level. Cancer Discov. 2022;12:134–53.
31. Valdeolivas A, Amberg B, Giroud N, Richardson M, Galvez EJC, Badillo S, et al. Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. NPJ Precis Oncol. 2024;8:10.
32. Qi J, Sun H, Zhang Y, Wang Z, Xun Z, Li Z, et al. Single-cell and spatial analysis reveal interaction of FAP(+) fibroblasts and SPP1(+) macrophages in colorectal cancer. Nat Commun. 2022;13:1742.
33. Wang F, Long J, Li L, Wu ZX, Da TT, Wang XQ, et al. Single-cell and spatial transcriptome analysis reveals the cellular heterogeneity of liver metastatic colorectal cancer. Sci Adv. 2023;9:eadf5464.
34. Liu HT, Chen SY, Peng LL, Zhong L, Zhou L, Liao SQ, et al. Spatially resolved transcriptomics revealed local invasion-related genes in colorectal cancer. Front Oncol. 2023;13:1089090.

35. Taams LS, van Amelsfort JM, Tiemessen MM, Jacobs KM, de Jong EC, Akbar AN, et al. Modulation of monocyte/macrophage function by human CD4+CD25+ regulatory T cells. Hum Immunol. 2005;66:222–30.
36. Kraaij MD, Savage ND, van der Kooij SW, Koekkoek K, Wang J, van den Berg JM, et al. Induction of regulatory T cells by macrophages is dependent on production of reactive oxygen species. Proc Natl Acad Sci U S A. 2010;107:17686–91.
37. Sautès-Fridman C, Petitprez F, Calderaro J, Fridman WH. Tertiary lymphoid structures in the era of cancer immunotherapy. Nat Rev Cancer. 2019;19:307–25.
38. Garbarino O, Lambroia L, Basso G, Marrella V, Franceschini B, Soldani C, et al. Spatial resolution of cellular senescence dynamics in human colorectal liver metastasis. Aging Cell. 2023;22:e13853.
39. Hirz T, Mei S, Sarkar H, Kfoury Y, Wu S, Verhoeven BM, et al. Dissecting the immune suppressive human prostate tumor microenvironment via integrated single-cell and spatial transcriptomic analyses. Nat Commun. 2023;14:663.
40. 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.
41. Chen S, Chang Y, Li L, Acosta D, Li Y, Guo Q, et al. Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in Alzheimer's disease. Acta Neuropathol Commun. 2022;10:188.
42. Tran MN, Maynard KR, Spangler A, Huuki LA, Montgomery KD, Sadashivaiah V, et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. Neuron. 2021;109:3088-3103. e3085.
43. Bloom GS. Amyloid-beta and tau: the trigger and bullet in Alzheimer disease pathogenesis. JAMA Neurol. 2014;71:505–8.
44. Ranasinghe KG, Verma P, Cai C, Xie X, Kudo K, Gao X, et al. Altered excitatory and inhibitory neuronal subpopulation parameters are distinctly associated with tau and amyloid in Alzheimer's disease. Elife. 2022. https://doi.org/10.7554/eLife.77850.
45. Palop JJ, Chin J, Roberson ED, Wang J, Thwin MT, Bien-Ly N, et al. Aberrant excitatory neuronal activity and compensatory remodeling of inhibitory hippocampal circuits in mouse models of Alzheimer's disease. Neuron. 2007;55:697–711.
46. Shiga M, Seno S, Onizuka M, Matsuda H. SC-JNMF: single-cell clustering integrating multiple quantification methods based on joint non-negative matrix factorization. PeerJ. 2021;9:e12087.
47. Jiang H, Wang M-N, Huang Y-A, Huang Y. Graph-regularized non-negative matrix factorization for single-cell clustering in scrna-seq data. IEEE J Biomed Health Inform. 2024. https://doi.org/10.1109/JBHI.2024.3400050.
48. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177:1873-1887 e1817.
49. Elosua-Bayes M, Nieto P, Mereu E, Gut I, Heyn H. Spotlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res. 2021;49:e50.
50. Choi H, Na KJ, Jung Y, Lim M, Lee D, Lee JE, et al. Spatial transcriptomics reveals spatially diverse cancer-associated fibroblast in lung squamous cell carcinoma linked to tumor progression. bioRxiv. 2024:2024.2005. 2016.594592.
51. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40:517–26.
52. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12:1088.
53. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS ONE. 2017;12:e0176278.
54. Page L, Brin S, Motwani R, Winograd T: The PageRank citation ranking: Bringing order to the web. Stanford, CA, USA: Stanford infolab; 1999.
55. Senbabaoglu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. Sci Rep. 2014;4:6207.
56. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401:788–91.
57. Pearson KVII. Note on regression and inheritance in the case of two parents. Proc R Soc Lond B Biol Sci. 1895;58:240–2.
58. Hu M, Chikina M. Instaprism: an R package for fast implementation of BayesPrism. Bioinformatics. 2024. https://doi.org/10.1093/bioinformatics/btae440.
59. Valdeolivas A, et al. Profiling the Heterogeneity of Colorectal Cancer Consensus Molecular Subtypes using Spatial Transcriptomics: datasets. 2023. Zenodo. https://doi.org/10.5281/zenodo.7760264.
60. Wang F, et al. Single-cell and spatial transcriptome analysis reveals the cellular heterogeneity of liver metastatic colorectal cancer. Datasets. Gene Expression Omnibus. 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE225857.
61. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52:594–603.
62. Lee HO, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Datasets. Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132465.
63. Lee HO, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Datasets. Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132257.
64. Lee HO, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Datasets. Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144735.
65. Garbarino O, et al. Spatial resolution of cellular senescence dynamics in human colorectal liver metastasis. Datasets. Gene Expression Omnibus. 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE206552.
66. Che LH, Liu JW, Huo JP, Luo R, Xu RM, He C, et al. A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative chemotherapy. Cell Discov. 2021;7:80.
67. Che LH, et al. A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative chemotherapy. Datasets. Gene Expression Omnibus. 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178318.
68. Hirz T, et al. Prostate cancer single-cell and spatial transcriptomic data analyses. 2023. Zenodo. https://doi.org/10.5281/zenodo.7526696.
69. 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.

70. Chen S, et al. Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in Alzheimer's disease. Datasets. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE220442.
71. Huang S. SpaNiche. Github. 2026. https://github.com/SiyuanHuang1/SpaNiche.
72. Huang S. SpaNiche. Zenodo. 2026. https://doi.org/10.5281/zenodo.19179447.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.