hoodscanR: profiling single-cell neighborhoods in spatial transcriptomics data

Ning Liu[1,2,3,*], Jarryd Martin[2,3], Dharmesh D Bhuva[1,2,3], Jinjin Chen[2,3], Mengbo Li[2,3], Samuel C. Lee[2,3], Malvika Kharbanda[1], Jinming Cheng[2,3], Ahmed Mohamed[2,3], Arutha Kulasinghe[4], Yunshun Chen[2,3], Chin Wee Tan[2,3,4], Melissa J Davis[1,2,3]

[1] South Australian immunoGENomics Cancer Institute (SAiGENCI), Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, SA 5005, Australia

[2] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, Victoria 3052, Australia

[3] Department of Medical Biology, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, VIC 3010, Australia.

[4] Frazer Institute, Faculty of Medicine, The University of Queensland, Brisbane, Queensland 4102, Australia.

* whom correspondence should be addressed. Email: ning.liu@adelaide.edu.au

Abstract

Understanding complex cellular niches and neighborhoods are giving us new insights into tissue biology. Accurate neighborhood identification is crucial, yet existing methodologies often struggle to detect mixed neighborhoods and generate cell-specific neighborhood profiles. To address these limitations, we introduce hoodscanR, a Bioconductor package designed for neighborhood identification and downstream analyses using spatial data. Applying hoodscanR to breast and lung cancer datasets, we showcase its efficacy in conducting detailed neighborhood analyses and identify subtle transcriptional changes in tumor cells from different neighborhoods. Such analyses can help researchers gain valuable insights into disease mechanisms and potential therapeutic targets.

Keywords

Spatial transcriptomics, Cellular neighborhood, Cancer micro-environment, Bioconductor package.

Introduction

Spatial transcriptomics stands out as a powerful technology, offering a distinctive perspective that goes beyond traditional bulk RNA-seq and single-cell RNA-seq (scRNA-seq) methods. It yields valuable insights into the complex molecular landscapes of tissues, uncovering spatial variations often overlooked by conventional approaches. Recent advancements in spatial platforms, including Nanostring CosMx Spatial Molecular Imager[1], 10X Genomics Xenium[2], Vizgen Merscope[3], Akoya CODEX[4, 5], and others, have facilitated the generation of single-cell level spatial data. This shift towards high-resolution spatial profiling has created a pressing demand for innovative analytical tools capable of fully exploiting these datasets. However, despite the potential of spatial transcriptomics, the field is still in its early stages, with many analyses resembling conventional scRNA-seq-like approaches. These analyses often disregard the rich spatial context of the data, failing to harness the cellular coordinates. Cellular neighborhood analysis, a powerful approach to fully utilize cell spatial information, becomes particularly important when applied to single-cell level spatial transcriptomics data. Tools are needed to identify and characterize the niches or neighborhoods in which cells reside, as these regions may harbor crucial tissue micro-environment (TME) biology that influences responses to therapy and disease progression. Understanding these neighborhoods provides researchers with novel insights into cellular interactions and communications within the TME, offering a nuanced understanding of the complex biological processes at play. Such insights hold the potential to enhance our understanding of complex diseases like cancer and contribute to the development of more effective therapies.

In recent years, there has been a growing trend in the development of methods dedicated to conducting neighborhood analyses to interpret complex cellular neighborhoods within spatial transcriptomics data. These methods range from clustering-based approaches that leverage frequency matrices of k-nearest cells[1] to graph network-based strategy that is built into interactive viewer[6]. Widely used toolkits Squidpy[7] and Giotto[8] have made substantial contributions to the field by facilitating neighborhood analysis using enrichment tests using a graph-based approaches compatible across multiple spatial technology platforms. Nevertheless, despite these advancements, there are critical gaps in existing methodologies (Table 1). Most notably, the existing methods exhibit a limitation in detecting mixed neighborhoods, wherein cellular surroundings comprise two or more distinct cell types or groups. For example, when cells reside in neighborhoods characterized by a mixture of B cells and stromal cells, current methods tend to categorize such neighborhoods as either exclusively B cell or stromal cell neighborhoods, failing to capture the nuanced composition of mixed cellular environments. Furthermore, current tools lack the capability to provide cell-level neighborhood annotations, meaning detailed neighborhood profiles for individual cells are unavailable. This critical feature is essential for a comprehensive characterization of the spatial context surrounding each cell. In response to these unaddressed challenges, we developed hoodscanR, a Bioconductor R package designed to perform comprehensive neighborhood analyses on spatial transcriptomics data. hoodscanR aims to bridge critical gaps by enabling the exploration of mixed neighborhoods and providing cell-level neighborhood profiles. In this study, we introduce the functionalities and capabilities of hoodscanR and demonstrate its utility in investigating the cellular neighborhoods within publicly available spatial transcriptomics datasets.

Table1. Features of existing neighborhood analysis methods for spatial transcriptomics data.

Methods / Features	Giotto[8]	Squidpy[7]	He et al.[1]	Visinity[6]
Language	R	Python	R	Javascript, Python
Infrastructure	Giotto object	AnnData	NA	Interactive viewer
Input	Network graph	Network graph	Frequency matrix	Image data
Multi-platforms compatible	Yes	Yes	Yes	No
Co-localization	Available	Available	NA	NA
Mixed neighborhood	No	No	No	No
Cell-level neighborhood profiles	No	No	No	No

Results

Development of hoodscanR

hoodscanR uses an efficient computational pipeline to investigate spatial neighborhood relationships among cells within spatial transcriptomics datasets (Figure 1). One of the hallmark features of hoodscanR is using the Bioconductor spatialExperiment infrastructure as the backbone of the analysis. This significantly increases the compatibility of intermediate results from hoodscanR with diverse Bioconductor packages tailored for preprocessing, quality control, normalization, cell type annotation, and various downstream analyses specifically crafted for spatial transcriptomics data.

At the core of hoodscanR, the searching process for nearest cells is initiated using an Approximate Nearest Neighbor (ANN) search algorithm[9], facilitating the identification of the k-nearest neighboring cells for each cell in the dataset. Subsequently, a Euclidean distance matrix between cells and their k-nearest neighboring cells is then calculated (Figure 1). Simultaneously, cell-level annotations provided by users, such as cell types, are used to construct a cell annotation matrix, describing the organization of cells based on their distances to neighboring cells.

The fundamental function of hoodscanR is to identify cellular neighborhoods within spatial transcriptomics data. It achieves this by using the SoftMax function, enhanced by a hyperparameter τ (tau), which governs the shape of the resulting probability distribution and provides control over the influence of neighboring cells. The algorithm is expressed as follows:

$$P(x, h_a \mid \tau) = \frac{\sum_i \mathbb{1}_{h_a}(x_i) \cdot e^{(-d(x,x_i)^2/\tau)}}{\sum_i e^{(-d(x,x_i)^2/\tau)}}$$

Where:

$P(x, h_a \mid \tau)$ denotes the probability of cell x residing within the local neighborhood h_a.

$d(x, x_i)$ signifies the spatial Euclidean distance between cell x and its neighboring cell x_i.

τ stands as the hyperparameter, facilitating fine-tuned modulation of the impact of neighboring cells.

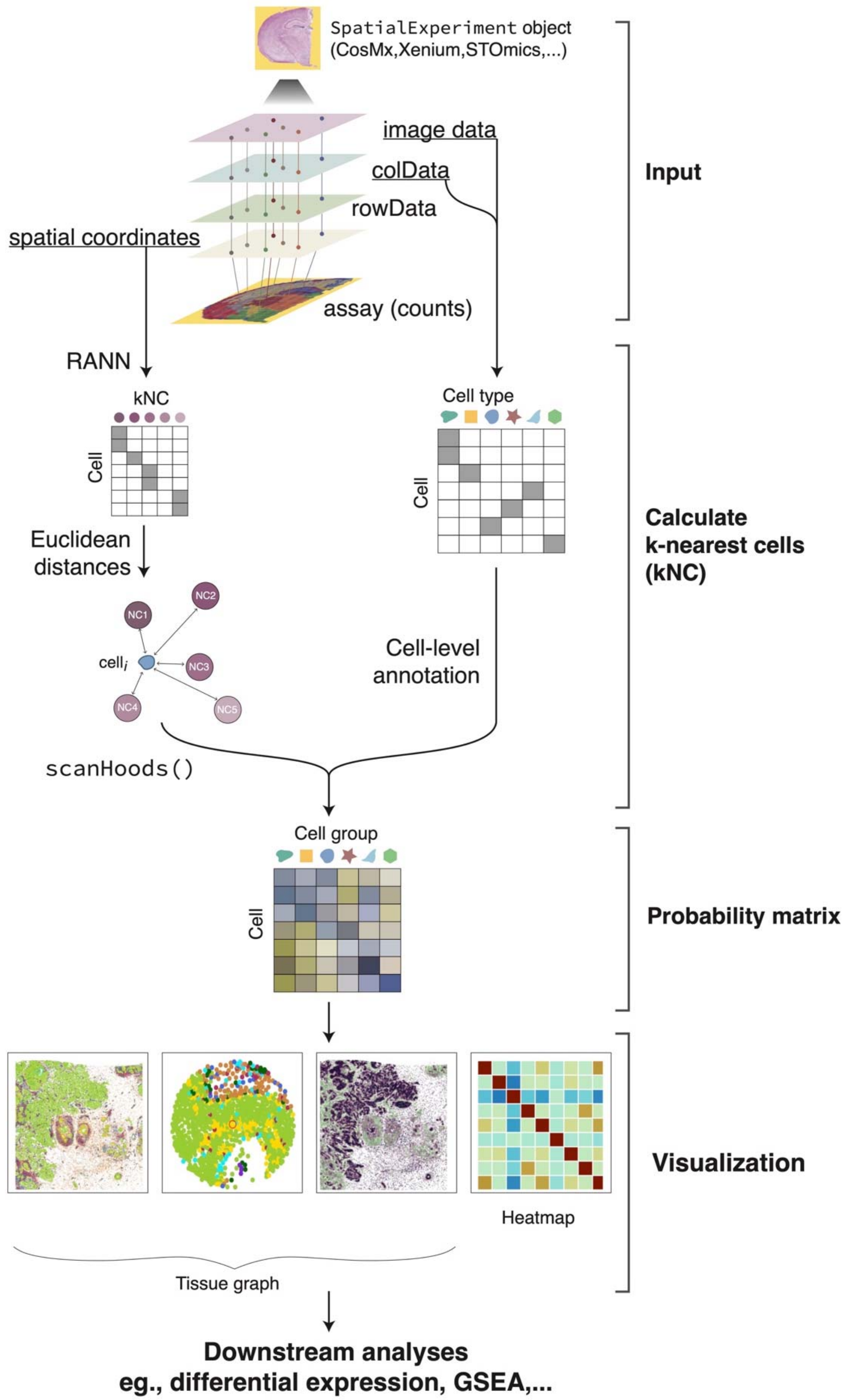
h_a denotes the cell neighborhood a, defined by the cell-level annotations provided by users. For example, if cell types were provided, h_a means cell type a neighborhood.

$\mathbb{1}(.)$ is the indicator function.

Upon the aggregation of probabilities by user-defined cell-level annotation groups, such as cell type annotations, hoodscanR generates a comprehensive probability matrix in which each value represents the probability of each cell belonging to a specific cell neighborhood (Figure 1). This matrix describes the cellular neighborhood profiles for all cells, serves as the backbone for downstream analyses, enabling researchers to delve into spatial patterns and relationships.

To investigate how the hyperparameter τ affects the results generated by hoodscanR, we conducted an extensive examination of the probability matrix across a range of τ values. This analysis revealed that smaller τ values assign greater weights to nearby cells, while higher τ values also consider more distant cells as contributors to the neighborhood (Suppl. Fig. 1). Thus, the choice of τ becomes essentially linked to the specific biological questions being addressed. For instance, when the research focus is on capturing a global overview of neighborhood estimation, higher τ values may be more suitable, as they incorporate further-away cells as essential components of the neighborhood. Notably, the default τ value in the hoodscanR package is set to one-fifth of the median of the distance matrix, a choice that results in a gradual decay of probabilities from the nearest to the furthest cells (Suppl. Fig. 1).

After neighborhood identification, hoodscanR extends its capabilities to offer a diverse suite of downstream neighborhood analysis tools (Figure 1). Users can apply these tools to visualize spatial relationships, evaluate co-localization patterns, perform spatial neighborhood clustering of cells, and obtain cell-level neighborhood annotations. These functionalities allow researchers to gain insights within the spatial transcriptomic landscape, facilitating the discovery of novel biological knowledge.



SpatialExperiment object
(CosMx,Xenium,STOmics,...)
image data
colData
rowData
spatial coordinates
assay (counts)
Input
RANN
kNC
Cell
Euclidean distances
NC1
NC2
cell_i
NC3
NC4
NC5
Cell type
Cell
Cell-level annotation
scanHoods()
Calculate k-nearest cells (kNC)
Cell group
Cell
Probability matrix
Heatmap
Tissue graph
Visualization
Downstream analyses
eg., differential expression, GSEA,...

Figure 1. Schematic visualization showing the main components of the hoodscanR package. From the top to bottom, it describes the computational steps of inputting SpatialExperiment

object, calculating k-nearest cells, generating probability matrix, and performing visualizations and downstream neighborhood-based analyses.

hoodscanR identifies celluar neighborhoods in cancer

To demonstrate the power of hoodscanR in detecting spatial cellular neighborhoods, we performed neighborhood identification on two publicly available spatial transcriptomics datasets obtained from two *in situ* transcriptomic platforms: breast cancer data obtained from the 10X Genomics Xenium (Figure 2A) and non-small cell lung cancer (NSCLC) data acquired from the Nanostring CosMx Spatial Molecular Imager (Figure 2B). hoodscanR allow us to perform neighborhood identification by profiling neighborhood distributions for each cell, representing the probability of a cell being situated within each distinct cell-type neighborhood (Figure 2C and 2F).

To validate the accuracy of hoodscanR in characterizing these cellular neighborhoods, we focus on one randomly selected cell for each dataset by exploring the distribution of cell types within their spatial area (Figure 2D, E, G, and H). For instance, during an examination of a ductal carcinoma in situ (DCIS) grade 2 cell, an early form of breast cancer cells, from the Xenium data (Figure 2C: cell ID 27620), we observed its proximity to multiple DCIS grade 2 cells and a number of ACTA2+ myoepithelial cells (Figure 2D). As a result, hoodscanR assigned probabilities of 61.02% for residing in the DCIS grade 2 neighborhood and 30.96% for the ACTA2+ myoepithelial neighborhood for this specific cell (Figure 2E). Similarly, when assessing a stromal cell within the CosMx NSCLC data (Figure 2F: cell ID 16_4233), we observed that hoodscanR assigned probabilities of 55.95% for the B cell neighborhood and 27.17% for the plasma cell neighborhood (Figure 2G and H). These examples demonstrate the power of hoodscanR in characterizing cellular neighborhoods within spatial transcriptomics data, regardless of the platform, and its capacity to accommodate scenarios where cells may belong to mixed neighborhoods. The identification of B cell neighborhoods is particularly noteworthy in the context of cancer therapy responses. B cell neighborhoods serve as crucial sites for antibody production, contributing to the immune response against tumor cells and influencing therapeutic efficacy[10]. Furthermore, in lung cancer, the presence of tertiary lymphoid structures (TLS), characterized by highly organized T and B lymphocyte colonies within nonlymphoid tissues, has been associated with favorable clinical outcomes in non-small cell lung cancer (NSCLC)[11]. These structures, resembling secondary lymphoid organs, play an important role in regulating antitumor immune responses and are emerging as potential targets for novel therapeutic interventions. By delineating cellular neighborhoods, including B-cell-rich TME, hoodscanR offers the potential for investigating the relationship between immune cells and tumor cells within the TME, providing insights that could inform the development of more effective cancer therapies.

Furthermore, hoodscanR introduces an additional analytical dimension by enabling the computation of entropy and perplexity metrics for each cell (See methods). These metrics are derived from the probability distributions, capturing the spatial relationships among cells and their respective neighborhoods. Entropy offers a measure of the diversity and variability within these neighborhoods (Figure 2I and K), while perplexity provides insights into the uncertainty and complexity of cell neighborhoods (Figure 2J and L). This reveals regions of the TME with distinct cellular compositions and areas with complex interactions between cell types. These metrics provide an understanding of the heterogeneity and complexity present within tissues, allowing researchers to gain novel insights and make discoveries in the spatial transcriptomics landscape.

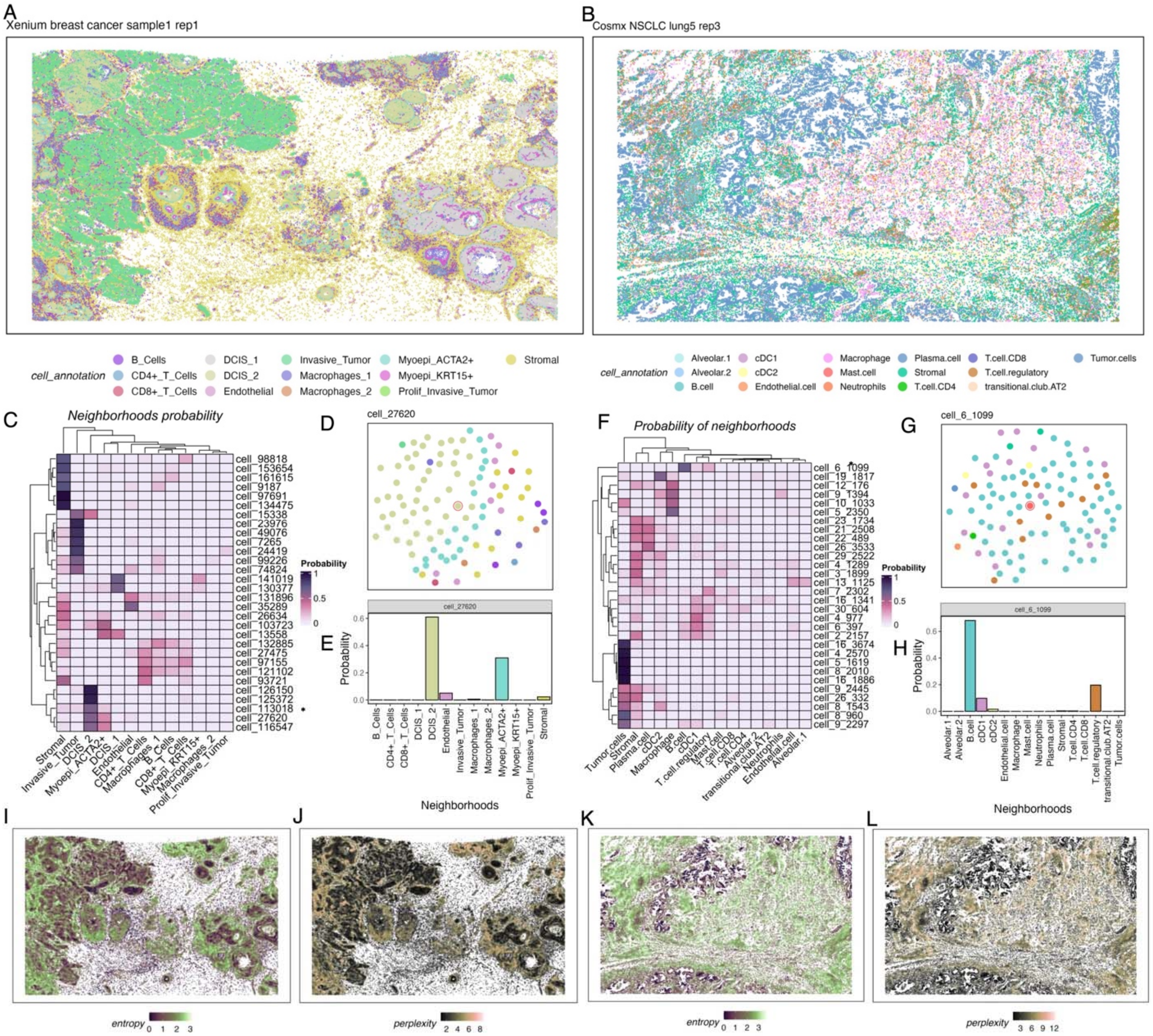


Figure 2. Neighborhood identification in 10X Xenium breast cancer data and Nanostring CosMx NSCLC data. Cell type spatial distribution in the breast cancer data (A) and NSCLC data (B). Neighborhood distribution visualization via heatmap of randomly selected 30 cells from breast cancer data (C) and NSCLC data (F), darker color means higher probability of the cell located in specific cell type neighborhood. The cell type spatial distribution in the spatial area around the selected cells (marked by * in the heatmap) in breast cancer data (D & E) and NSCLC data (G & H). Entropy spatial distribution of cells in the breast cancer data (I) and NSCLC data (K). Perplexity spatial distribution of cells in the breast cancer data (J) and NSCLC data (L).

hoodscanR performs neighborhood-based downstream analyses

Existing neighborhood identification methods, such as Squidpy and Giotto, predominantly focus on neighborhood co-localization analyses. What sets hoodscanR apart is its unique ability to generate annotations at the cell level and to carry out neighborhood-based downstream analyses, features notably absent in other existing tools. To demonstrate the versatility of hoodscanR, we use the CosMx NSCLC dataset as an example.

Firstly, as with other spatial analysis tools, hoodscanR can perform neighborhood co-localization analysis by computing Pearson correlations on the neighborhood distribution of cells. Hence, the co-localization status of each cell type neighborhood within this specific tissue slide can then be visualized (Figure 3A). To benchmark the ability of hoodscanR, Squidpy and Giotto when carrying out co-localization analysis, we subset the Xenium breast cancer data and CosMx NSCLC into different resolution, followed by applying these tools to the subsets. As a result, hoodscanR demonstrated superior computational performance (Suppl. Fig. 2), while delivering similar outcome (mean Pearson correlation coefficient of 0.781) of neighborhood co-localization compared to both Xenium and CosMx data (Suppl. Fig. 3 and 4). The computational efficiency gains significance, particularly in the context of the growing spatial data resolutions and larger tissue areas.

To validate the co-localization results, we present a subset of cell types alongside their spatial distribution within the breast cancer tissue slide, showing that immune cell types such as B cells, T cells, and macrophages exhibit co-localization, while they are distinctly separated from tumor cells (Figure 3B). This observation serves as robust validation of the co-localization analysis results generated by hoodscanR, thus reinforcing the effectiveness and reliability of hoodscanR in revealing spatial relationships within various tissue environments, particularly when dealing with complex spatial transcriptomics data.

Building upon the foundation of cell-level neighborhood annotations, hoodscanR allows users to perform unsupervised clustering, grouping cells with similar neighborhood distribution patterns into cohesive clusters. This data-driven approach enables the classification of cells based on their spatial relationships within the tissue slide. In the

CosMx NSCLC data, we applied unsupervised clustering to delineate 10 distinct clusters (see Methods), each representing a unique spatial pattern within the tissue (Figure 3C), demonstrating complex spatial associations. For instance, cluster 1, a candidate cluster for TLS, corresponds to a mixed neighborhood including B cells, cDC1 cells, and stromal cells (Figure 3D - 1), cluster 3 aligns with macrophages and cDC2 cells (Figure 3D - 3), and cluster 6 corresponds to tumor cells (Figure 3D - 6). This unsupervised clustering approach facilitates the identification of diverse cellular neighborhoods and their unique spatial signatures, providing a comprehensive view of the complexity of TME.

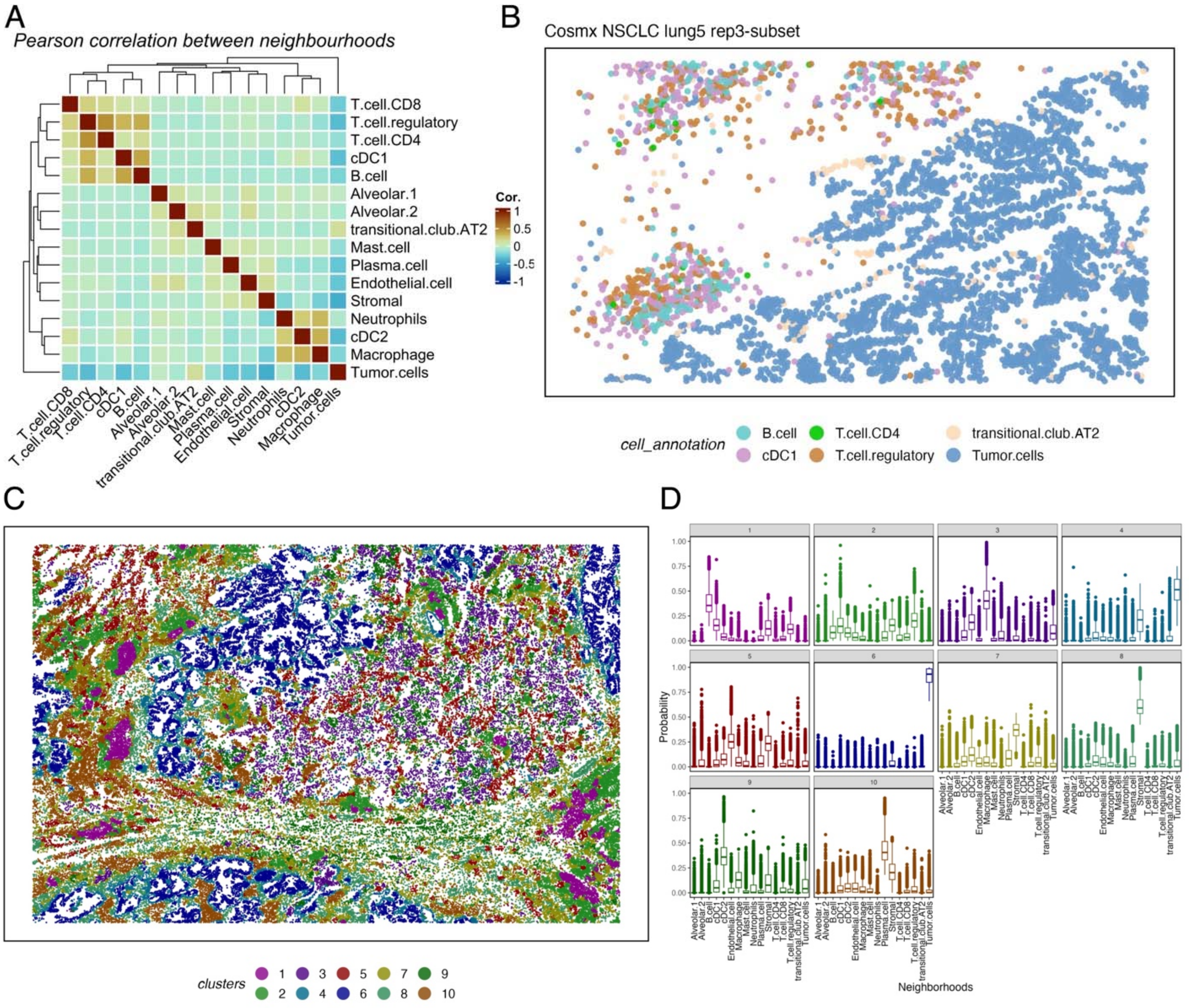


Figure 3. Neighborhood-based downstream analyses in CosMx NSCLC data performed by hoodscanR. A: heatmap representing the co-localization status of cell type neighborhood in this tissue slide. Colors denote positive (dark red) and negative (blue) correlations. B: Spatial location plot of a subset of cell types in this slide. Colors denote cell types. C: Spatial location plot of the slide with colors stratified by neighborhood-based clusters. D: The neighborhood distribution profiles of each neighborhood-based cluster identified in the tissue slide.

We then used uniform manifold approximation and projection (UMAP) to perform dimension reduction on the expression data within the CosMx NSCLC data, enabling the projection of gene expression profiles into a lower-dimensional space. This facilitates the visualization of cell lineage (Figure 4A) alongside identified neighborhood-based clusters (Figure 4B). A distinctive feature observed is the dispersion of cells of the same cell type across different neighborhood clusters, signifying diverse spatial neighborhood profiles. For examples, a substantial proportion (76.14%) of macrophages (pink points in Figure 4A) are distributed across various neighborhood clusters, including cluster 3 (43.63%), indicative of the macrophage + cDC2 + tumor neighborhood, clusters 7 (12.31%), and cluster 9 (20.2%), representing the stromal + cDC2 neighborhood (Figure 3D and 4B).

An important aspect of hoodscanR lies in its ability to investigate the relationship between spatial neighborhoods and transcriptional changes. To demonstrate this, we conducted a nuanced analysis by extracting and pseudo-bulking tumor cells from two distinct neighborhood clusters: the stromal cluster and the macrophage cluster across three consecutive slides (Suppl. Fig. 5). Interestingly, different spatial neighborhoods contribute significantly to the variation observed in the first dimension from a principal component analysis (PCA) of expression of pseudo-bulk samples (Figure 4C). This signifies that hoodscanR can capture transcriptional changes attributed to diverse spatial neighborhoods.

Subsequently, we performed a differential expression (DE) analysis using the limma-voom pipeline[12], identifying 220 DE genes from 832 genes, including 73 up-regulated and 147 down-regulated genes when comparing tumor cells from the macrophage neighborhood to those from the stromal neighborhood (Figure 4D and Supplementary file 2). Additionally, we performed gene-set enrichment analysis on the identified DE genes with MSigDB gene-sets[13, 14], identifying significantly down-regulated pathways that are relevant to collagen, such as *collagen-activated tyrosine kinase receptor signaling pathway* and *collagen metabolic and catabolic process* (Figure 4E). Previous studies have found that the overexpression of collagen genes such as COL11A1[15] and COL3A1[16] in NSCLC may indicate poor prognosis and drug resistance, and COL1A1 is correlated with immune infiltration in NSCLC[17]. Our finding of these genes that are expressed significantly more in the tumor cells from the macrophage neighborhood than tumor cells from the stromal neighborhood can potentially lead a more detailed investigation of the biological mechanism about transcriptional changes within the context of cancer spatial TME studies. In essence, hoodscanR introduces a novel perspective in spatial transcriptomics, allowing the identification of transcriptomic changes across subtly different spatial neighborhoods and providing insights into the complex cell-cell interactions within the TME.

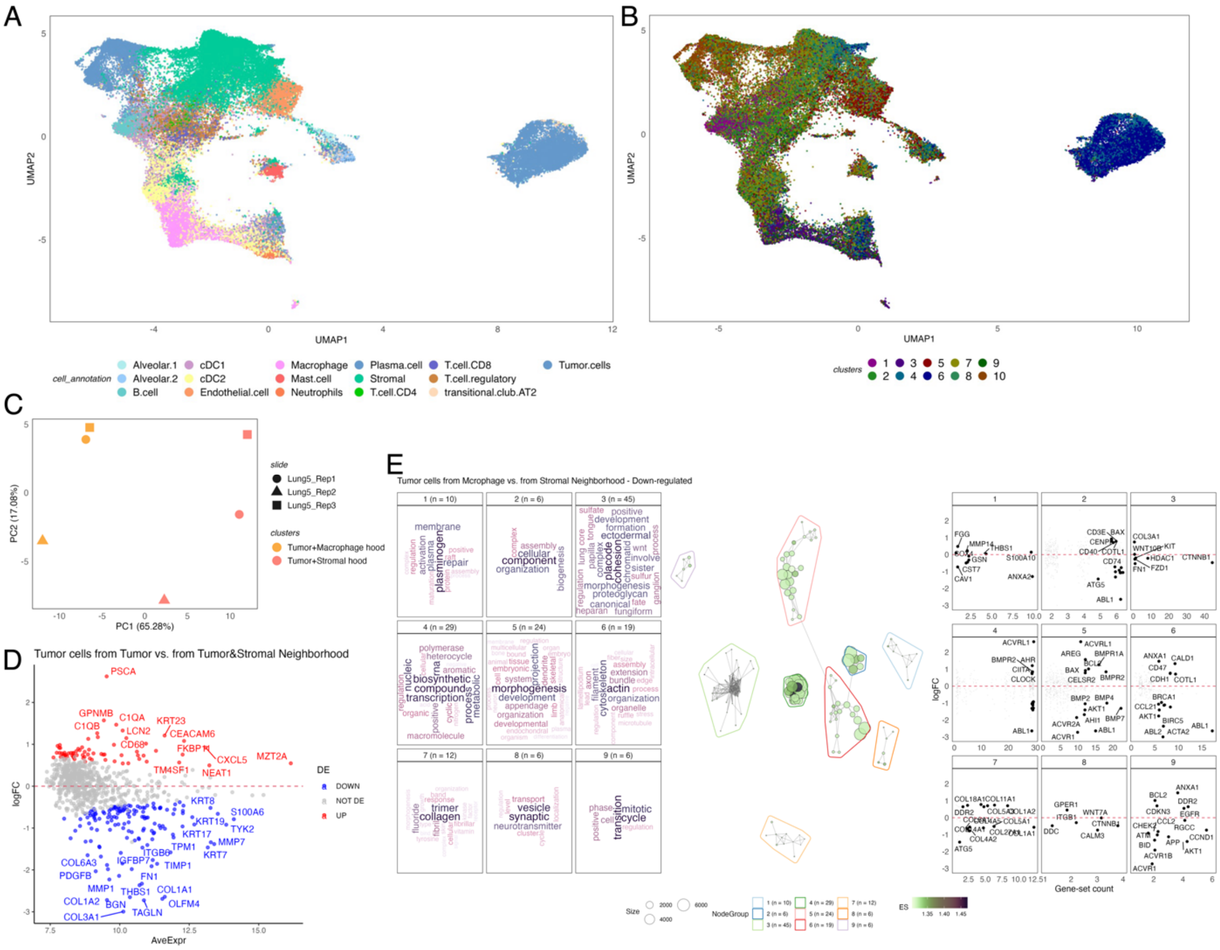


Figure 4. Neighborhood-based transcriptomics analysis in CosMx NSCLC data. A: UMAP of expression of cells in the data, colors denote cell types. B: UMAP of expression of cells in the data, colors denote the identified neighborhood-based clusters. C: PCA of pseudo-bulk samples of tumor cells from two different neighborhood clusters across three consecutive slides. Colors denote clusters and shapes denote replicates. D: MA plot describing the outcome of the DE analysis. Colors indicate up- (red) or down-regulated (blue) genes. E: vissE visualization of significantly enriched gene-sets from the down-regulated DE genes in the comparison between tumor cells from macrophage neighborhood and tumor cells from stromal neighborhood. Left panel are word cloud plots describing gene-set clusters of different biological themes, middle panel is the gene-set overlap network graph of gene-sets, and right panel is the fold-change (log2-scaled) for genes belonging to each gene-set cluster.

hoodscanR supports different gene annotations

The accuracy of cell type-based cellular neighborhoods, as identified in the previous case, is inherently tied to the cell type annotation methods. hoodscanR therefore offers the flexibility to detect various cellular neighborhoods based on different gene annotation inputs. A particularly valuable application is gene expression-based neighborhoods detection. For example, in a breast cancer tissue slide, we can focus on important breast cancer-related hormone receptor genes, including androgen receptor gene (AR), estrogen receptor gene (ESR1), and progesterone receptor gene (PGR). By assessing if these genes are expressed or not, we classified 574,527 cells from a Xenium Invasive ductal carcinoma (IDC) dataset into eight distinct groups (Figure 5A).

Similar to the previous neighborhood identification based on cell types, hoodscanR can identify gene expression-specific neighborhoods with measurement of neighborhood perplexity for each cell (Figure 5B). We can then proceed to cluster cells into gene-specific neighborhood clusters based on the neighborhood distribution (Figure 5C and D). These analyses lead to a nuanced understanding which adeptly discover tumor cells located within neighborhoods characterized by varying combinations of hormone receptors. This not only suggests a spatial perspective on the progression of DCIS, influenced by distinct combinations of hormone receptors but also sheds light on the higher-order spatial structure of cells with different hormone receptor expression profiles. More importantly, interesting insights can be observed by comparing the pathological annotations (Figure 5E) of this slide[18] with the identified neighborhood clusters. DCIS regions are notably associated with cells from cluster 5, forming the inner layer, surrounded spatially by cells from cluster 3, comprising the outer layer (Figure 5F). On the other hand, invasive tumors exhibit a distinctive pattern with cells from cluster 3 scattered throughout, accompanied by cells from cluster 6 (Figure 5F). These observations hint at a potential relationship between the tumor type (non-invasive or invasive) and the higher-order spatial organization of cells with diverse hormone receptor expression profiles. In summary, these findings reinforce hoodscanR's capability to identify subtle spatial changes in expression-based cellular neighborhoods, providing novel insights into the complex spatial dynamics of gene expression within the context of breast cancer.

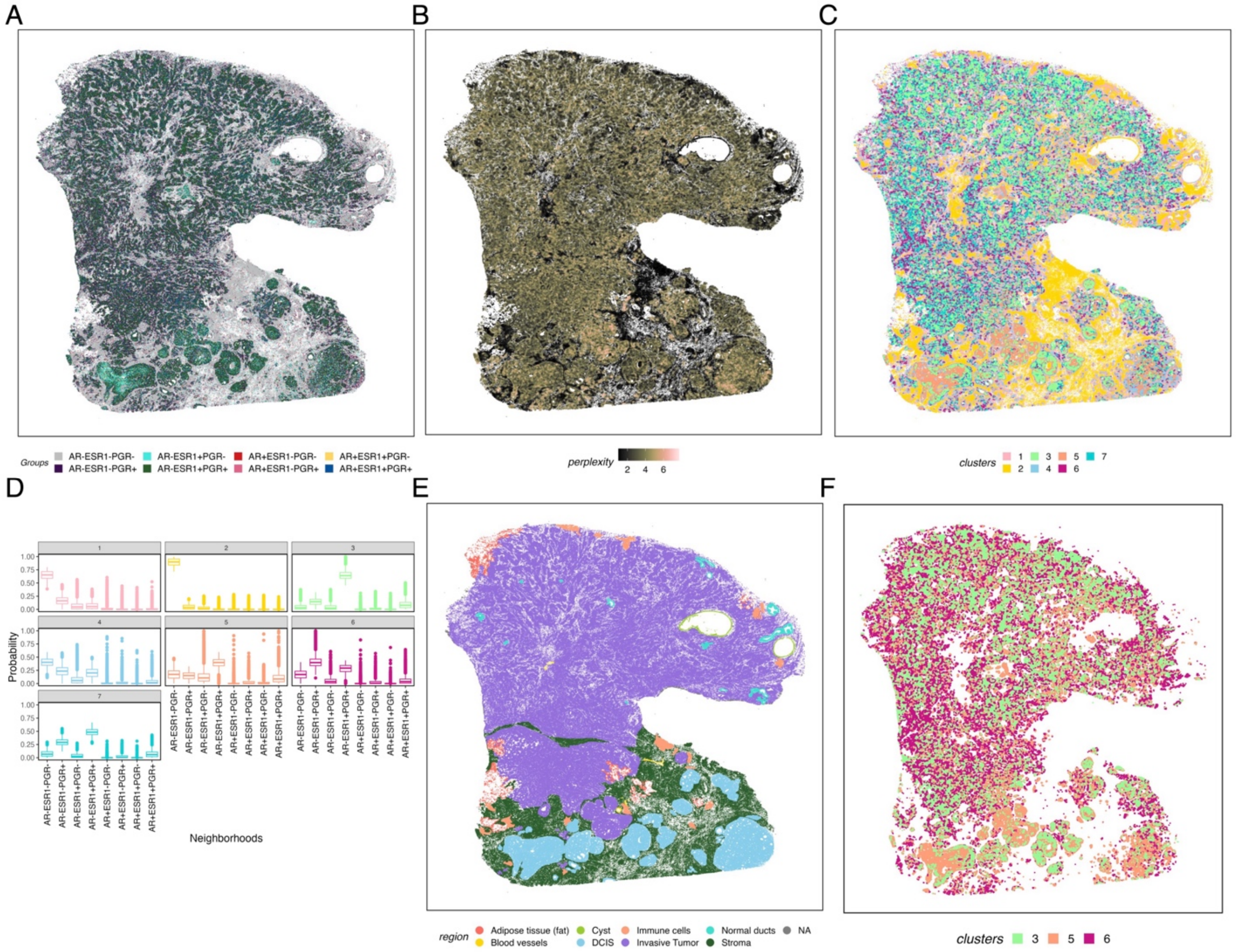


Figure 5. Gene expression-based neighborhood analysis in 10X Xenium IDC data performed by hoodscanR. A: spatial plot showing the hormone gene expression-based cell grouping on the tissue slide. B: perplexity spatial distribution of cells. C: spatial plot showing the identified gene expression-based neighborhood clusters. D: neighborhood distribution of each cluster. E: pathological annotation of regions on the tissue. F: spatial plot showing specific neighborhood clusters, including cluster 3, 5 and 6.

Discussion

Spatial technologies have pushed their limit toward profiling spatial transcriptomics at single-cell level. To make the best use of these cutting-edge technologies, we developed hoodscanR, a powerful yet user-friendly Bioconductor package for revealing spatial cellular relationships within high-dimensional spatial transcriptomics datasets via spatial cellular neighborhoods identification and neighborhood-based downstream analyses.

While we have shown hoodscanR can identify biologically meaningful cellular neighborhoods, as with all methods it is not without limitations. Firstly, the preprocessing of spatial transcriptomics data is essential. Current quality control procedures of spatial transcriptomics data predominantly operate at cell level, potentially missing crucial details detectable only at a transcript (subcellular) level. A more comprehensive strategy for data preprocessing, accounting for information at transcript level, filtering out uninformative cells accurately and enhancing the precision of neighborhood identification, becomes imperative. Secondly, hoodscanR is a cell-based method, emphasizing the critical role of accurate cell segmentation. A systematic review of existing segmentation methods is lacking, necessitating future research to evaluate and compare methods under diverse spatial transcriptomic platforms. Lastly, while hoodscanR enables exploration of spatial gene expression patterns, users must interpret results cautiously. Recognizing that spatial context alone may not capture the full complexity of molecular interactions within a tissue. Integrating multi-omics data can provide a more comprehensive understanding, ensuring spatial analyses are embedded within a broad molecular context. Strategically addressing these considerations allows researchers to fine-tune the utilization of hoodscanR, strengthening the integrity of their analyses and facilitating the discovery of novel insights.

Three of the major strengths of hoodscanR are its compatibility, adaptability and flexibility. Developed based on the Bioconductor SpatialExperiment infrastructure, hoodscanR exhibits compatibility with many other spatial and single-cell RNA-seq-based tools from Bioconductor, amplifying its utility for downstream analyses. Importantly, its adaptability makes it platform agnostic, demonstrated by successful applications on Nanostring CosMx and 10X Genomics Xenium. Spatial datasets from various platforms, including Vizgen MERSCOPE, 10X Visium HD, BGI STOmics, and Akoya Biosciences CODEX, can also undergo comprehensive analysis using hoodscanR, given the availability of cell-based coordinates. The flexibility of hoodscanR is demonstrated by the types of annotations it can use. With the Xenium breast cancer data, hoodscanR showcased this by not only using cell type annotations but also accommodating gene expression level grouping of cells, suggesting the potential for exploring additional annotation options such as ligand-receptor or growth factor-receptor annotations. This flexibility provides providing researchers the flexibility to generate hypotheses on the basis of integrating spatial localization and gene expression before testing statistical associations between these.

The significance of hoodscanR in identifying and analyzing neighborhoods becomes particularly crucial in the context of complex diseases, especially in cancer research. Neighborhood information is indispensable for unraveling complex disease etiology, especially so for understanding disease progression. In cancer research, where the TME plays an important role in dictating therapy responses, the ability offered by hoodscanR to identify neighborhoods offers a unique perspective to investigate novel mechanisms underlying the transition of cancer cells at both transcriptomic and proteomic level. Notably, our findings from the NanoString CosMx NSCLC dataset identified cellular neighborhoods that potentially associated with the presence of TLS, which correlated with positive clinical results[11]. This capability offers the potential to contribute valuable insights into the progression of cancer, paving the way for the development of novel therapeutic strategies.

In conclusion, our study introduces hoodscanR, a Bioconductor package designed for comprehensive neighborhood analysis in spatial transcriptomics datasets. Through its integration with the SpatialExperiment infrastructure and efficient algorithms, hoodscanR offers fine-tuned control over neighborhood identification, allowing researchers to investigate complex cellular relationships within spatially resolved datasets. By demonstrating hoodscanR's efficacy on the 10X Genomics Xenium breast cancer and Nanostring Technologies CosMx non-small cell lung cancer datasets, we showcase its ability in identifying cellular neighborhoods and elucidating spatial gene expression patterns. Furthermore, our findings emphasize the significance of neighborhood analysis in understanding the complex TME of cancer tissues, which can potentially lead to the identification of novel biological mechanisms underlying disease progression and therapeutic responses. Importantly, hoodscanR's flexibility in handling diverse spatial datasets and its ability to accommodate different types of cell annotations enhance its utility for a wide range of spatial transcriptomic studies. Overall, hoodscanR contributes to advancing the field of spatial transcriptomics by providing researchers with a powerful tool, thereby paving the way for deeper insights into tissue biology and disease mechanisms.

Methods

Data pre-processing

Both CosMx and Xenium datasets underwent a rigorous quality control process to ensure the inclusion of high-quality cells in the neighborhood analysis. For the CosMx data, thresholds were set at the 0.1 quantile to filter out cells with low transcript count or low gene detection count across all cells per tissue slide. Additionally, genes with mean expression (log-scaled count per million) and variance lower than the negative probes were excluded from further analyses. In the case of the Xenium data, filtering followed the guidelines outlined in the Squidpy[7] toolkit tutorial. Cells with a transcript count less than 10 and genes detected in fewer than 5 cells were removed from the neighborhood and downstream analyses. As a result, almost 90,000 cells per slide with 870 genes and 156,224 cells with 313 genes were kept for the NSCLC and breast cancer datasets, respectively.

Cell type annotation

Cell type annotations for the Nanostring CosMx NSCLC data was carried out with modifications as previosuly described in Tan et al 2024[19]. Briefly, specific modifications include using *SCTransform* from the Seurat v4 package[20] to normalise filtered counts from the quality control step. By modelling negative probe detection as a fixed factor, we regressed out the confounding effects caused by background. The annotation process of the data were performed using InSituType[21], with the Single-cell Lung Cancer Atlas (LuCA)[22] as the reference. In terms of 10X Xenium data, cell type annotation was obtained from the Janesick A*, et al* paper[2].

Metrics calculation for neighborhood probability distribution

Entropy and perplexity serve as fundamental metrics for summarizing the neighborhood probability distribution generated by hoodscanR. Entropy quantifies the uncertainty or randomness in the distribution of cell types within a cell's neighborhood, calculated mathematically using the Shannon entropy formula[23]. Conversely, perplexity reflects the effective number of neighbors contributing to the neighborhood distribution, calculated as 2 raised to the power of entropy. Higher entropy or perplexity values indicate greater diversity or complexity within the cellular neighborhoods, while lower values signify more uniform distributions. In hoodscanR, these metrics can be calculated using the *calcMetrics* function.

Benchmarking of co-localization analysis

Both 10X Xenium breast cancer and Nanostring CosMx datasets were utilized for benchmarking the co-localization analysis among hoodscanR, Squidpy, and Giotto. To ensure robustness, the data were randomly subset into ten different sizes ranging from 0.1 to 1, after which each package's respective methodologies were applied. Specifically, for hoodscanR, neighborhood identification and co-localization analysis were conducted using the *plotColocal* function. In contrast, Squidpy and Giotto performed network graph construction using the *gr.spatial_neighbors* and *createSpatialDelaunayNetwork* functions, followed by co-localization analysis using the *gr.nhood_enrichment* and *cellProximityEnrichment* functions, respectively.

Unsupervised clustering of neighborhood distribution

To perform unsupervised clustering of the neighborhood distribution, the identified neighborhood distribution of each cell was utilized as input data. The K-means clustering algorithm, a widely used method for partitioning data into distinct clusters based on dissimilarity, were used. Specifically, we set the parameters for K-means clustering as iter_max = 1000, nstart = 5, and algo = "Hartigan-Wong". The iter_max parameter determines the maximum number of iterations allowed to converge to a solution, while nstart specifies the number of initial cluster centers to use in the algorithm. Additionally, the "Hartigan-Wong" algorithm was chosen as the method for center initialization. To determine the optimal number of clusters (k) for the K-means clustering algorithm, we used the elbow method. In a nutshell, we first calculated the within-cluster sum of squares for a range of k values and then identified the point at which the distortion or inertia starts decreasing in a linear fashion.

Differential expression analysis

To conduct the differential expression (DE) analysis on the CosMx NSCLC dataset and identify DE genes among tumor cells from distinct cellular neighborhoods, pseudo-bulk via summation samples were initially generated from cells within the identified neighborhood-based clusters. Subsequently, the standR package was used to assess relative log expression (RLE) and perform principal component analysis (PCA) to explore the technical and biological variation in the pseudo-bulk data. Following this, the limma-voom pipeline was utilized for DE analysis with TMM normalisation[24], incorporating slide information as a covariate in the linear model to account for slide-related variations. The resulting statistic was an empirical Bayes moderated t-statistic. Multiple testing adjustment using the Benjamini–Hochberg procedure was then applied to identify DE genes that reached statistical significance (FDR < 0.05).

Gene-set enrichment analysis and visualization

Gene-sets from the Molecular Signatures Database (MsigDB, v7.2), including Hallmarks, C2 (curated gene-sets), and C5 (gene ontology terms) categories, along with KEGG pathway gene-sets, were obtained using the *getMisgdb* and *appendKEGG* functions from the msigdb R package (v1.1.5). Gene-set enrichment analysis (GSEA) was performed using *fry* from the limma package (v3.58.1). A false discovery rate of 0.05 was used as the threshold for determining significantly enriched gene-sets. The results of GSEA were systematically examined and visualized using an unbiased approach through the novel network enrichment and visualization R package vissE[25].

Data availability

The Nanostring CosMx non-small cell lung cancer (NSCLC) data utilized in this study was sourced from the official Nanostring website: https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/. The 10X Xenium breast cancer datasets used in this study were retrieved from the 10X publicly available database at https://www.10xgenomics.com/datasets. The

Xenium *in situ* dataset comprises human breast cancer FFPE sections and utilizes a 280-gene Xenium Human Breast Gene Expression Panel supplemented with 33 additional custom genes. The CosMx dataset was generated from FFPE NSCLC tissue samples using a 960-plex CosMx RNA panel.

Code availability

The code used for the performing the described analyses is available from the corresponding author upon reasonable request. The hoodscanR package is freely available in Bioconductor (release > 3.18) at https://bioconductor.org/packages/release/bioc/html/hoodscanR.html.

Reference

1. He S*, et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol* **40**, 1794-1806 (2022).

2. Janesick A*, et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* **14**, 8353 (2023).

3. Yao Z*, et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317-332 (2023).

4. Parra ER, Francisco-Cruz A, Wistuba, II. State-of-the-Art of Profiling Immune Contexture in the Era of Multiplexed Staining and Digital Analysis to Study Paraffin Tumor Tissues. *Cancers (Basel)* **11**, (2019).

5. Jhaveri N*, et al.* Mapping the spatial proteome of head and neck tumors: key immune mediators and metabolic determinants in the tumor microenvironment. *GEN Biotechnology* **2**, 418-434 (2023).

6. Warchol S*, et al.* Visinity: Visual Spatial Neighborhood Analysis for Multiplexed Tissue Imaging Data. *IEEE Trans Vis Comput Graph* **29**, 106-116 (2023).

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

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

9. Arya S, Mount DM, Netanyahu NS, Silverman R, Wu AY. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)* **45**, 891-923 (1998).

10. Tokunaga R*, et al.* B cell and B cell-related pathways for novel cancer treatments. *Cancer Treat Rev* **73**, 10-19 (2019).

11. Ren F*, et al.* Tertiary lymphoid structures in lung adenocarcinoma: characteristics and related factors. *Cancer Med* **11**, 2969-2977 (2022).

12. Law CW, Chen YS, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, (2014).

13. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740 (2011).

14. Subramanian A*, et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P Natl Acad Sci USA* **102**, 15545-15550 (2005).

15. Shen L, Yang M, Lin Q, Zhang Z, Zhu B, Miao C. COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. *Oncology reports* **36**, 877-885 (2016).

16. Wang L, Sun Y, Guo Z, Liu H. COL3A1 Overexpression Associates with Poor Prognosis and Cisplatin Resistance in Lung Cancer. *Balkan Med J* **39**, 393-400 (2022).

17. Geng Q, Shen Z, Li L, Zhao J. COL1A1 is a prognostic biomarker and correlated with immune infiltrates in lung cancer. *PeerJ* **9**, e11145 (2021).

18. Bhuva DD*, et al.* Library size confounds biology in spatial transcriptomics data. *bioRxiv*, 2023.2003. 2015.532733 (2023).

19. Tan CW*, et al.* In situ single-cell profiling sheds light on IFI27 localisation during SARS-CoV-2 infection. *EBioMedicine* **101**, 105016 (2024).

20. Hao Y*, et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).

21. Danaher P*, et al.* Insitutype: likelihood-based cell typing for single cell spatial transcriptomics. *bioRxiv*, 2022.2010. 2019.512902 (2022).

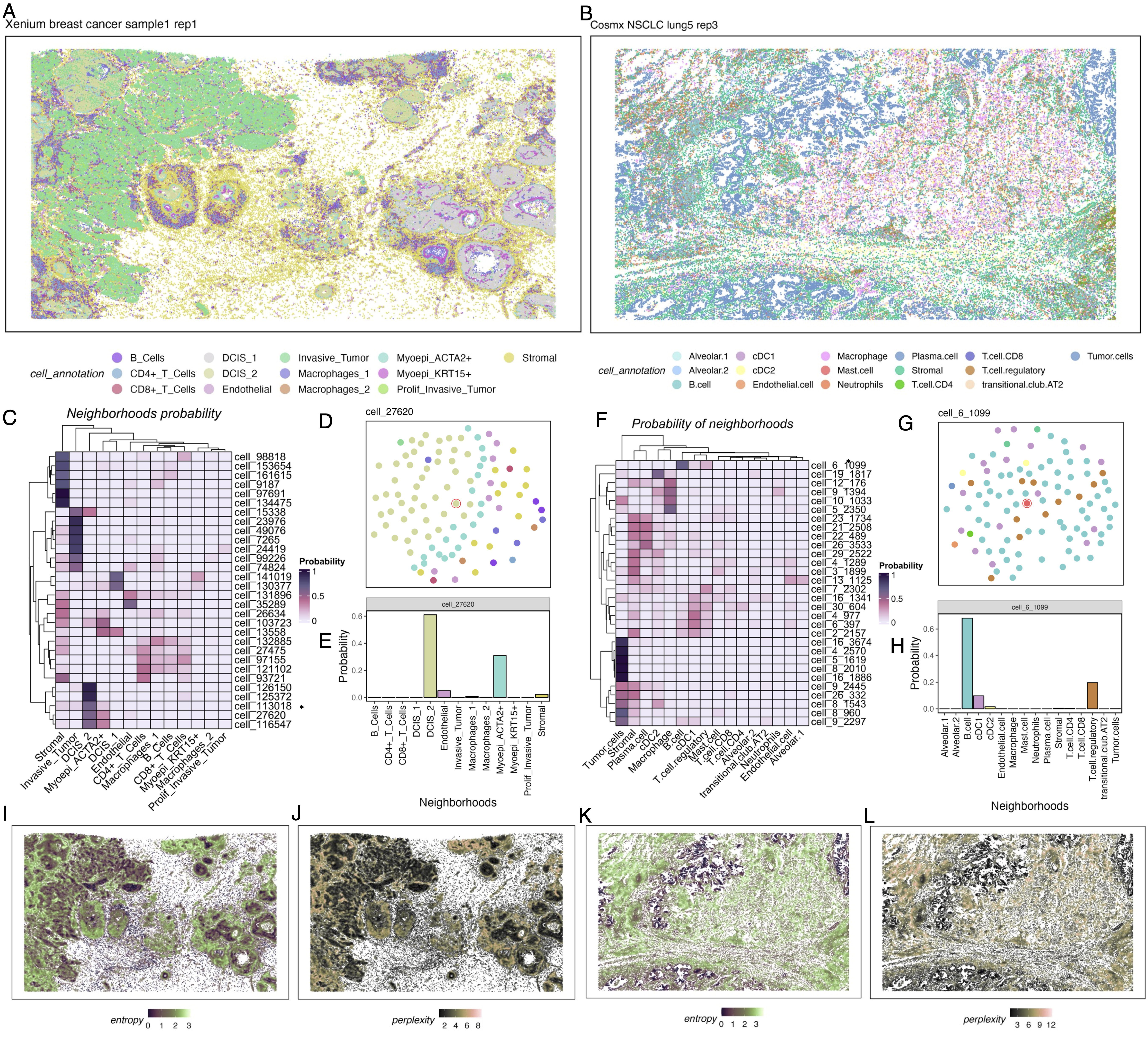
22. Salcher S*, et al.* High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503-1520 e1508 (2022).

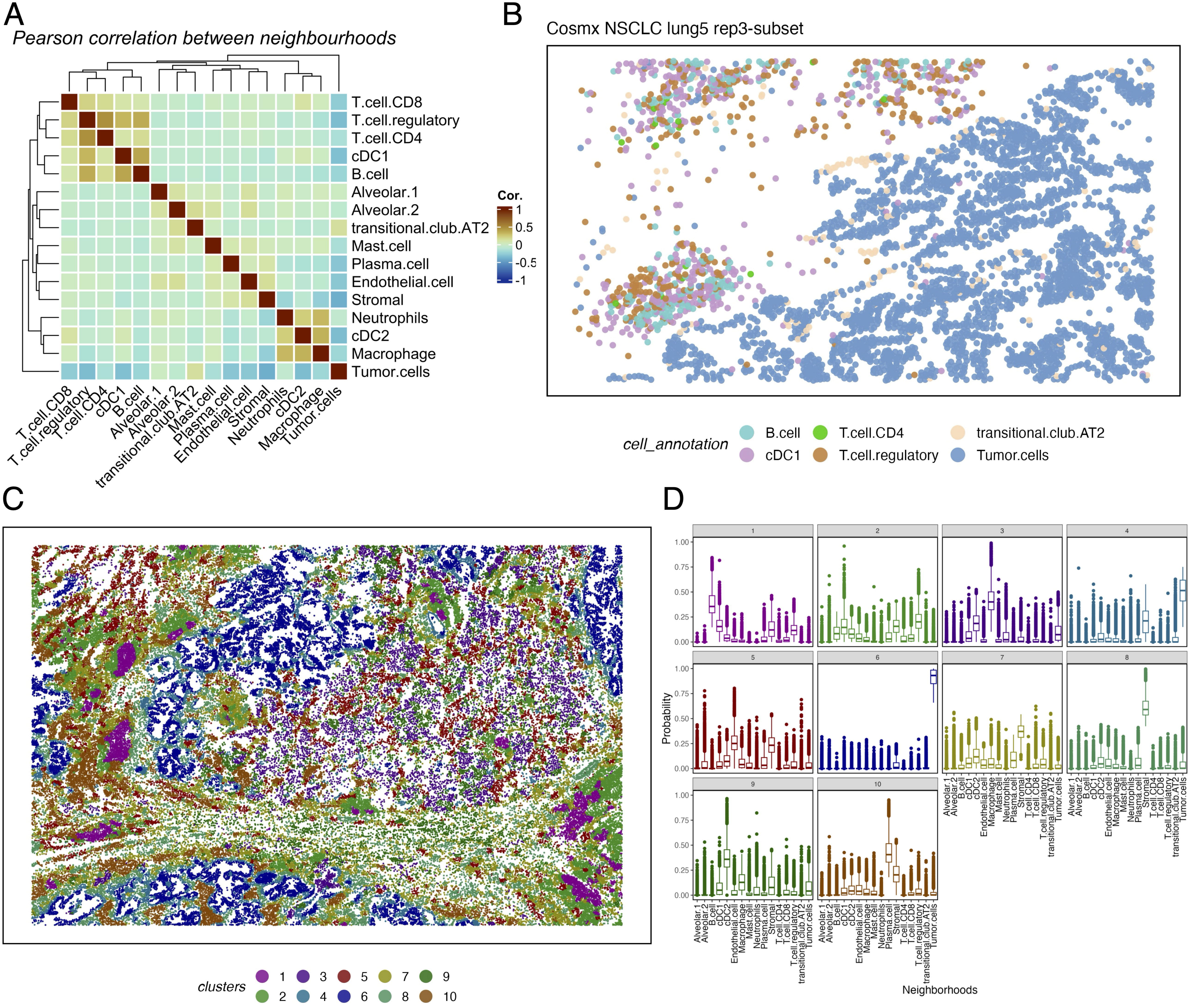
23. Shannon CE. The mathematical theory of communication. 1963. *MD Comput* **14**, 306-317 (1997).

24. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, (2010).

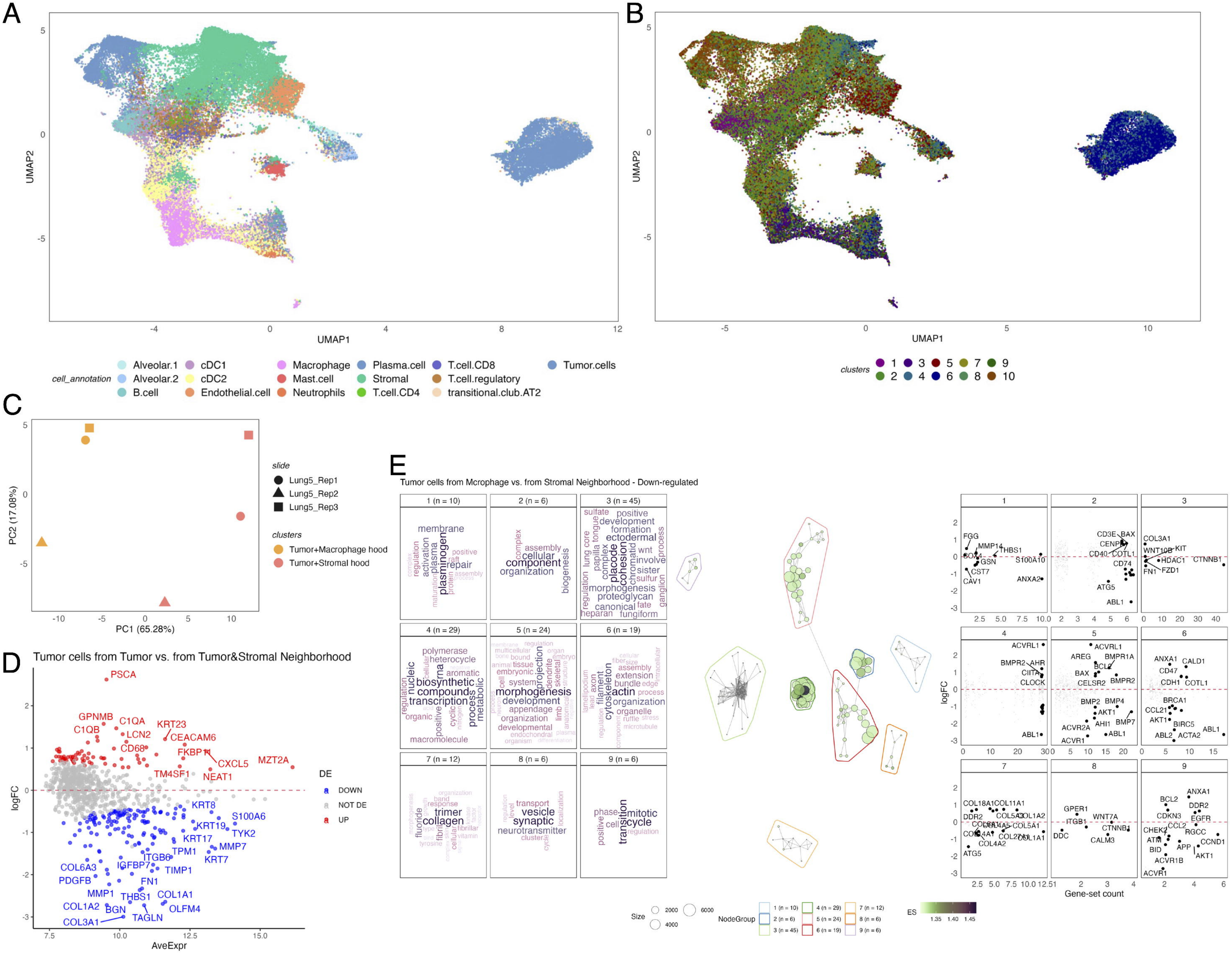
25. Bhuva DD*, et al.* vissE: a versatile tool to identify and visualise higher-order molecular phenotypes from functional enrichment analysis. *Bmc Bioinformatics* **25**, (2024).

A
B
C
Groups
AR-ESR1-PGR-
AR-ESR1-PGR+
AR-ESR1+PGR-
AR-ESR1+PGR+
AR+ESR1-PGR-
AR+ESR1-PGR+
AR+ESR1+PGR-
AR+ESR1+PGR+
perplexity
2
4
6
clusters
1
2
3
4
5
6
7
D
1
2
3
4
5
6
7
Probability
1.00
0.75
0.50
0.25
0.00
Neighborhoods
AR-ESR1-PGR-
AR-ESR1-PGR+
AR-ESR1+PGR-
AR-ESR1+PGR+
AR+ESR1-PGR-
AR+ESR1-PGR+
AR+ESR1+PGR-
AR+ESR1+PGR+
E
region
Adipose tissue (fat)
Blood vessels
Cyst
DCIS
Immune cells
Invasive Tumor
Normal ducts
Stroma
NA
F
clusters
3
5
6


A
Xenium breast cancer sample1 rep1
B
Cosmx NSCLC lung5 rep3
cell_annotation
B_Cells
CD4+_T_Cells
CD8+_T_Cells
DCIS_1
DCIS_2
Endothelial
Invasive_Tumor
Macrophages_1
Macrophages_2
Myoepi_ACTA2+
Myoepi_KRT15+
Prolif_Invasive_Tumor
Stromal
cell_annotation
Alveolar.1
Alveolar.2
B.cell
cDC1
cDC2
Endothelial.cell
Macrophage
Mast.cell
Neutrophils
Plasma.cell
Stromal
T.cell.CD4
T.cell.CD8
T.cell.regulatory
transitional.club.AT2
Tumor.cells
C
Neighborhoods probability
cell_98818
cell_153654
cell_161615
cell_9187
cell_97691
cell_134475
cell_15338
cell_23976
cell_49076
cell_7265
cell_24419
cell_99226
cell_74824
cell_141019
cell_130377
cell_131896
cell_35289
cell_26634
cell_103723
cell_13558
cell_132885
cell_27475
cell_97155
cell_121102
cell_93721
cell_126150
cell_125372
cell_113018
cell_27620
cell_116547
*
Probability
1
0.5
0
Stromal
Invasive_Tumor
DCIS_2
Myoepi_ACTA2+
DCIS_1
Endothelial
CD4+_T_Cells
Macrophages_1
B_Cells
CD8+_T_Cells
Myoepi_KRT15+
Macrophages_2
Prolif_Invasive_Tumor
D
cell_27620
E
cell_27620
Probability
0.6
0.4
0.2
0.0
B_Cells
CD4+_T_Cells
CD8+_T_Cells
DCIS_1
DCIS_2
Endothelial
Invasive_Tumor
Macrophages_1
Macrophages_2
Myoepi_ACTA2+
Myoepi_KRT15+
Prolif_Invasive_Tumor
Stromal
Neighborhoods
F
Probability of neighborhoods
cell_6_1099
*
cell_19_1817
cell_12_176
cell_9_1394
cell_10_1033
cell_5_2350
cell_23_1734
cell_21_2508
cell_22_489
cell_26_3533
cell_29_2522
cell_4_1289
cell_3_1899
cell_13_1125
cell_7_2302
cell_16_1341
cell_30_604
cell_4_977
cell_6_397
cell_2_2157
cell_16_3674
cell_4_2570
cell_5_1619
cell_8_2010
cell_16_1886
cell_9_2445
cell_26_332
cell_8_1543
cell_8_960
cell_9_2297
Probability
1
0.5
0
Tumor.cells
Stromal
Plasma.cell
cDC2
Macrophage
B.cell
cDC1
T.cell.regulatory
Mast.cell
T.cell.CD8
T.cell.CD4
Alveolar.2
transitional.club.AT2
Neutrophils
Endothelial.cell
Alveolar.1
G
cell_6_1099
H
cell_6_1099
Probability
0.6
0.4
0.2
0.0
Alveolar.1
Alveolar.2
B.cell
cDC1
cDC2
Endothelial.cell
Macrophage
Mast.cell
Neutrophils
Plasma.cell
Stromal
T.cell.CD4
T.cell.CD8
T.cell.regulatory
transitional.club.AT2
Tumor.cells
Neighborhoods
I
J
K
L
entropy
0
1
2
3
perplexity
2
4
6
8
entropy
0
1
2
3
perplexity
3
6
9
12


A
Pearson correlation between neighbourhoods
T.cell.CD8
T.cell.regulatory
T.cell.CD4
cDC1
B.cell
Alveolar.1
Alveolar.2
transitional.club.AT2
Mast.cell
Plasma.cell
Endothelial.cell
Stromal
Neutrophils
cDC2
Macrophage
Tumor.cells
Cor.
1
0.5
0
-0.5
-1
B
Cosmx NSCLC lung5 rep3-subset
cell_annotation
B.cell
cDC1
T.cell.CD4
T.cell.regulatory
transitional.club.AT2
Tumor.cells
C
clusters
1
2
3
4
5
6
7
8
9
10
D
Probability
0.00
0.25
0.50
0.75
1.00
Neighborhoods
Alveolar.1
Alveolar.2
B.cell
cDC1
cDC2
Endothelial.cell
Macrophage
Mast.cell
Neutrophils
Plasma.cell
Stromal
T.cell.CD4
T.cell.CD8
T.cell.regulatory
transitional.club.AT2
Tumor.cells

SpatialExperiment object
(CosMx,Xenium,STOmics,...)
image data
colData
rowData
spatial coordinates
assay (counts)
Input
RANN
kNC
Cell
Cell type
Cell
Euclidean distances
NC1
NC2
cell$_i$
NC3
NC4
NC5
Cell-level annotation
Calculate k-nearest cells (kNC)
scanHoods()
Cell group
Cell
Probability matrix
Heatmap
Tissue graph
Visualization
Downstream analyses
eg., differential expression, GSEA,...


A
UMAP1
UMAP2
cell_annotation
Alveolar.1
Alveolar.2
B.cell
cDC1
cDC2
Endothelial.cell
Macrophage
Mast.cell
Neutrophils
Plasma.cell
Stromal
T.cell.CD4
T.cell.CD8
T.cell.regulatory
transitional.club.AT2
Tumor.cells
B
clusters
1 2 3 4 5 6 7 8 9 10
C
PC1 (65.28%)
PC2 (17.08%)
slide
Lung5_Rep1
Lung5_Rep2
Lung5_Rep3
clusters
Tumor+Macrophage hood
Tumor+Stromal hood
D
Tumor cells from Tumor vs. from Tumor&Stromal Neighborhood
AveExpr
logFC
DE
DOWN
NOT DE
UP
PSCA
GPNMB
C1QB
C1QA
LCN2
KRT23
CEACAM6
CD68
FKBP11
CXCL5
MZT2A
TM4SF1
NEAT1
KRT8
S100A6
KRT19
TYK2
KRT17
MMP7
TPM1
KRT7
ITGB6
IGFBP7
TIMP1
COL6A3
PDGFB
FN1
MMP1
THBS1
COL1A1
COL1A2
BGN
OLFM4
COL3A1
TAGLN
E
Tumor cells from Mcrophage vs. from Stromal Neighborhood - Down-regulated
1 (n = 10)
2 (n = 6)
3 (n = 45)
4 (n = 29)
5 (n = 24)
6 (n = 19)
7 (n = 12)
8 (n = 6)
9 (n = 6)
Size
2000
4000
6000
NodeGroup
ES
1.35 1.40 1.45
logFC
Gene-set count
FGG
MMP14
THBS1
SOX4
GSN
S100A10
CST7
CAV1
ANXA2
CD3E
BAX
CENPF
CD40
COTL1
CD74
ATG5
ABL1
COL3A1
WNT10B
KIT
HDAC1
CTNNB1
FN1
FZD1
ACVRL1
BMPR2
AHR
CIITA
CLOCK
AREG
BMPR1A
BCL2
CELSR2
BMP4
AKT1
AHI1
BMP7
ACVR2A
ACVR1
ANXA1
CALD1
CD47
CDH1
BRCA1
CCL21
BIRC5
ABL2
ACTA2
COL18A1
COL11A1
DDR2
COL5A3
COL1A2
COL6A1
COL4A5
COL5A1
COL4A1
COL27A1
COL4A2
GPER1
ITGB1
WNT7A
CTNNB1
DDC
CALM3
CDKN3
EGFR
CCL2
CHEK2
RGCC
ATM
APP
CCND1
BID
ACVR1B
membrane
plasminogen
activation
repair
cellular
component
organization
assembly
biogenesis
development
morphogenesis
proteoglycan
heparan
cohesion
polymerase
heterocycle
biosynthetic
compound
transcription
nucleic
macromolecule
metabolic
process
cytoskeleton
actin
organization
organelle
collagen
trimer
fibril
vesicle
synaptic
neurotransmitter
transport
mitotic
cycle
transition
phase