hoodscanR: profiling single-cell neighborhoods in spatial transcriptomics data

Ning Liu[1,2,3,4,*], 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[5], Yunshun Chen[2,3], Chin Wee Tan[2,3,5], Fuyi Li[1], Jose M Polo[1,4,6,*], Melissa J Davis[1,2,3,7,*]

[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] Adelaide Centre for Epigenetics, School of Biomedicine, Faculty of Health and Medical Sciences, The University of Adelaide, SA 5005, Australia

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

[6] Department of Anatomy and Developmental Biology, Monash University, Wellington Road, Clayton, VIC 3800, Australia

[7] Computational Biology, Isomorphic Labs, Bishopsgate, London, England, EC2M 4RB, United Kingdom

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

Abstract

Understanding complex cellular niches and neighborhoods have provided new insights into tissue biology. Thus, accurate neighborhood identification is crucial, yet existing methodologies often struggle to detect informative neighborhoods and generate cell-specific neighborhood profiles. To address these limitations, we developed 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.

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. Since it conserves the spatial information of a tissue, it yields valuable insights into the complex molecular and cellular landscapes, uncovering spatial variations and relationships often overlooked by conventional approaches. Recent advancements in spatial-omics 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. 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. Thus, this shift towards high-resolution spatial profiling and the lack of appropriate methods has created a pressing demand for innovative analytical tools capable of fully exploiting these datasets. Cellular neighborhood analysis, a powerful approach to fully utilize cell spatial information, becomes particularly important when applied to single-cell level spatial transcriptomics data. Bioinformatics 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 the fundamental tissue biology, physiology as well as responses to therapy and disease progression. Therefore, understanding these neighborhoods is key for the full utilization of the spatial data and to provide 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 (Table 1). 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 via enrichment tests using a graph-based approaches compatible across multiple spatial technology platforms. Additionally, many tools have been developed to detect spatial domains from spatial transcriptomics datasets by accounting for the spatial information, i.e. cellular neighborhood when clustering data via various approaches, including BuildNicheAssay from Seurat (9), MERINGUE (10), BANKSY (11), BayesSpace (12), STAGATE (13), SpaGCN (14) and UTAG (15). Nevertheless, despite these advancements, there are critical gaps in existing methodologies. Most notably, while some existing tools can detect spatial domains that comprise multiple cell types, such as UTAG, SpaGCN and Giotto's HMRF-based approach, they do not provide partial membership at a single-cell level. 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 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. Unlike existing methods, hoodscanR aims to bridge critical gaps by enabling per-cell partial membership across multiple neighborhoods, providing a more precise and detailed understanding of the tissue microenvironments. Additionally, hoodscanR generates cell-level neighborhood profiles, a unique feature that allows for an in-depth summarization of the spatial context at a single-cell resolution. Moreover, hoodscanR can identify neighborhood-based spatial domains, offering insights into the higher-order organisation of tissues. 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/domain identification methods for spatial transcriptomics data.

Features	Giotto(8)	Squidpy(7)	Seurat	Banksy	BayesSpace	MERINGUE	SpaGCN	Stagate	Utag	hoodscanR
Language	R	Python	R	R/Python	R	R	Python	Python	Python	R
Infrastructure	Giotto object	AnnData	Seurat object	Spatial Experiment	Spatial Experiment	NA	AnnData	AnnData	AnnData	Spatial Experiment
Multi-platforms compatible	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Co-localization	Yes	Yes	No	No	No	No	No	No	No	Yes
Multi-neighborhood membership	No	No	No	No	No	No	No	No	No	Yes
Cell-level neighborhood profiles	No	No	No	No	No	No	No	No	No	Yes
Spatial domain detection	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

108

Materials and 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 (16). Briefly, specific modifications include using *SCTransform* from the Seurat package (17) 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 (18), with the Single-

cell Lung Cancer Atlas (LuCA)(19) 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

Perplexity serves as a fundamental metrics for summarizing the neighborhood probability distribution generated by hoodscanR. It provides a measure of effective diversity of complexity within a cell's neighborhood. The perplexity $P(x)$ for a given cell x is calculated as:

$$P(x) = 2^{H(x)}$$

Where $H(x)$ represents the Shannon entropy (20) of the neighborhood probability distribution of cell x, defined as:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Where $p(x_i)$ is the probability of cell x located in the i-th neighborhood and n is the total number of distinct neighborhoods. Higher perplexity values indicate greater diversity or complexity within the cellular neighborhoods, suggesting that a larger number of distinct cell types are contributing to the neighborhood. In hoodscanR, perplexity can be calculated using the *calcMetrics* function.

To assess the statistical significance of the observed perplexity values within cell neighborhoods, we employed an empirical permutation test. For each neighborhood, we generated a distribution of perplexity values by randomly shuffling the spatial

coordinates of cells and recalculating the perplexity across 1,000 permutations. The empirical p-value for each neighborhood was then calculated as the proportion of permuted perplexity values that were greater than or equal to the observed perplexity value, adjusted for the finite number of permutations:

$$P_{empirical} = \frac{\sum_{i=1}^{N} \mathbb{1}(P_{obs} \geq P_i) + 1}{N + 1}$$

where P_{obs} is the observed perplexity for a given neighborhood, P_i represents the perplexity from the i-th permutation, $\mathbb{1}(.)$ is an indicator function that equals 1 when the condition inside is true and 0 otherwise, and N is the total number of permutations. This correction ensures that the empirical p-values are properly calibrated, even with a limited number of permutations, thus providing a robust measure of statistical significance.

Hyperparameter k and τ testing

In order to test the effect different parameter k values have on the results of hoodscanR, we first test a range of k values (10, 50, 100, 200, 500, 1000), using the default τ setting from the *scanHoods* function. We then computed the Pearson correlation between the resulting probability matrices to assess consistency of the outcomes across different k. For testing the τ parameter, we fixed k = 100 and examined a range of τ values, which were derived from different scaling of the distance matrix (see Availability of data and materials).

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.

182

183 ## Benchmarking spatial domain identification

184 To benchmark hoodscanR against other state-of-the-art methods in identifying spatial
185 domains from spatial transcriptomics datasets, we selected 12 publicly available
186 datasets. These included 3 CosMx NSCLC, 6 MERFISH mouse colon, 1 STARmap
187 mouse cortex, and 2 Xenium breast cancer slides. Due to the large size of the Xenium
188 breast cancer datasets (555,579 cells and 353,783 cells, respectively), some tools were
189 unable to process these datasets efficiently. To address this, we downsampled these
190 slides to 10,000 and 50,000 cells, while maintaining the original cell type proportion
191 distributions, resulting in a total of 16 datasets. This was achieved by randomly
192 sampling cells from each cell type cluster based on their proportional weighting (see
193 Availability of data and materials). Annotated region labels or pathological annotations
194 in these datasets were used as the ground truth for spatial domains.
195 Normalization of the gene expression data was handled differently depending on the
196 dataset. For the MERFISH dataset, we directly used the provided log-normalized
197 counts. For all other datasets, we used the *quickCluster* and *calculateSumFactors*
198 functions from the *scran* R package (21) to estimate size factors, followed by the
199 *logNormCounts* function from the scuttle R package (22) to normalize the counts.

200 We compared hoodscanR against seven other methods capable of performing spatial
201 domain detection: BuildNicheAssay from Seurat, Banksy, BayesSpace, MERINGUE,
202 SpaGCN, Stagate, and Utag. For each method, we calculated a composite performance
203 score to assess the accuracy of spatial domain identification. This score was computed

as the mean of four key metrics: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), purity, and homogeneity. It is defined as:

$$Score = mean(ARI + NMI + Purity + Homogeneity)$$

These metrics were chosen for their ability to quantify different aspects of clustering accuracy. ARI measures the similarity between the predicted and true clusters, adjusting for chance. It is defined as:

$$ARI = \frac{RI - Expected\ RI}{Max\ RI - Expected\ RI}$$

Where RI is the Rand Index, which counts the number of correct pairwise classifications between the predicted and true labels. ARI in this paper is calculated using the *aricode* R package.

NMI quantifies the amount of information shared between the predicted and true clusters. It is defined as:

$$NMI = \frac{2 \times I(U,V)}{H(U) + H(V)}$$

where $I(U,V)$ is the mutual information between clusters U and V, and $H(U)$ and $H(V)$ are the entropies of the true and predicted clusters, respectively. NMI in this paper is calculated using the *aricode* R package.

Purity measures the extent to which each cluster contains only members of a single class. It is calculated as:

$$Purity = \frac{1}{N}\sum_{k=1}^{K} \max_{j} |C_k \cap T_j|$$

where C_k are the predicted clusters, T_j are the true clusters, and N is the total number of samples.

Homogeneity ensures that all the clusters contain only data points which are members of a single class. It is defined as:

$$Homogeneity = 1 - \frac{H(C|T)}{H(C)}$$

where $H(C|T)$ is the conditional entropy of the true clusters given the predicted clusters, and $H(C)$ is the entropy of the predicted clusters.

To ensure fairness in the benchmarking process, all methods were executed using their default settings (see Availability of data and materials). Additionally, a time penalty was applied: if a method failed to complete the processing of a dataset within 24 hours, the process was terminated. This ensures that the comparison accounts not only for accuracy but also for computational efficiency.

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

247 within-cluster sum of squares for a range of k values and then identified the point at
248 which the distortion or inertia starts decreasing in a linear fashion.

249

250 Differential expression analysis

251 To conduct the differential expression (DE) analysis on the CosMx NSCLC dataset and
252 identify DE genes among tumor cells from distinct cellular neighborhoods, pseudo-bulk
253 via summation samples were initially generated from cells within the identified
254 neighborhood-based clusters using *summarizeAssayByGroup* function from the *scuttle*
255 R package (22). Subsequently, the *standR* package (23) was used to assess relative
256 log expression (RLE) and perform principal component analysis (PCA) to explore the
257 technical and biological variation in the pseudo-bulk data. Following this, the limma-
258 voom pipeline (24) was utilized for DE analysis with TMM normalisation (25),
259 incorporating slide information as a covariate in the linear model to account for slide-
260 related variations. The resulting statistic was an empirical Bayes moderated t-statistic.
261 Multiple testing adjustment using the Benjamini–Hochberg procedure was then applied
262 to identify DE genes that reached statistical significance (FDR < 0.05). To identify DE
263 genes in the 10X VisiumHD mouse brain dataset, we first extracted neurons from the
264 *Neurod6*+ and *Calb2*+ neighborhoods and then applied Seurat's *FindAllMarkers*
265 function (9). We retained only genes meeting two criteria: a log2 fold change greater
266 than 1 and an adjusted p-value less than 0.05. The top ten genes from each group,
267 based on these filters, were subsequently selected for heatmap visualization using the
268 *DoHeatmap* function from Seurat.

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 (26).

Results

Development of hoodscanR

hoodscanR uses an efficient computational pipeline to investigate spatial neighborhood relationships among cells within spatial transcriptomics datasets (Figure 1). At the core of hoodscanR, the searching process for nearest cells is initiated using an Approximate Nearest Neighbor (ANN) search algorithm (27), which uses k-dimensional tree to efficiently manage the two-dimensional spatial coordinates of spatial transcriptomics data, providing rapid identification of nearest neighbors while maintaining high accuracy. This facilitates the identification of the k-nearest neighboring cells for each cell in the dataset. This process outputs a list of indices representing the nearest neighbors of each cell, denoted as:

$$ANN(x) = \{x_1, x_2, \dots, x_k\}$$

Following the identification of nearest neighbors, hoodscanR calculates the distance between each cell and its k-nearest neighbors. Here Euclidean distance is used due to its simplicity and effectiveness in measuring distances between points in a two-dimensional space. This results in a distance matrix D, where each element D_{ij} represents the distance between cell x_i and its neighbor x_j from $ANN(x)$ (Figure 1). Simultaneously, cell-level annotations provided by users, such as cell types, are used to construct a cell annotation matrix A, which describes the organisation of cells based on

their distances to neighboring cells. Each entry A indicates whether cell x_i belongs to annotation group j:

$$A = \begin{cases} 1 & if\ cell\ x_i \in \{x_1, x_2, \dots, x_k\}\ belongs\ to\ annotation\ group\ j \\ 0 & otherwise \end{cases}$$

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:

$$p_{h_j}(x;\tau) = \frac{\sum_{i=1}^{k} \mathbb{1}_{h_j}(x_i) \cdot exp\left\{-\frac{d^2(x, x_i)}{\tau}\right\}}{\sum_{i=1}^{k} exp\left\{-\frac{d^2(x, x_i)}{\tau}\right\}}$$

where $x, x_i \in \{x_1, x_2, \dots, x_k\}$, and

$$\mathbb{1}_{h_j}(x_i) = \begin{cases} 1 & if\ A[i,j]\ is\ 1 \\ 0 & otherwise \end{cases}$$

Where:

$p_{h_j}(x;\tau)$ denotes the probability of cell x residing within the local neighborhood h_j.

$d(x, x_i)$ 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_j denotes the cell neighborhood a, defined by the cell-level annotations provided by users. For example, if cell types were provided, h_j means cell type j neighborhood.

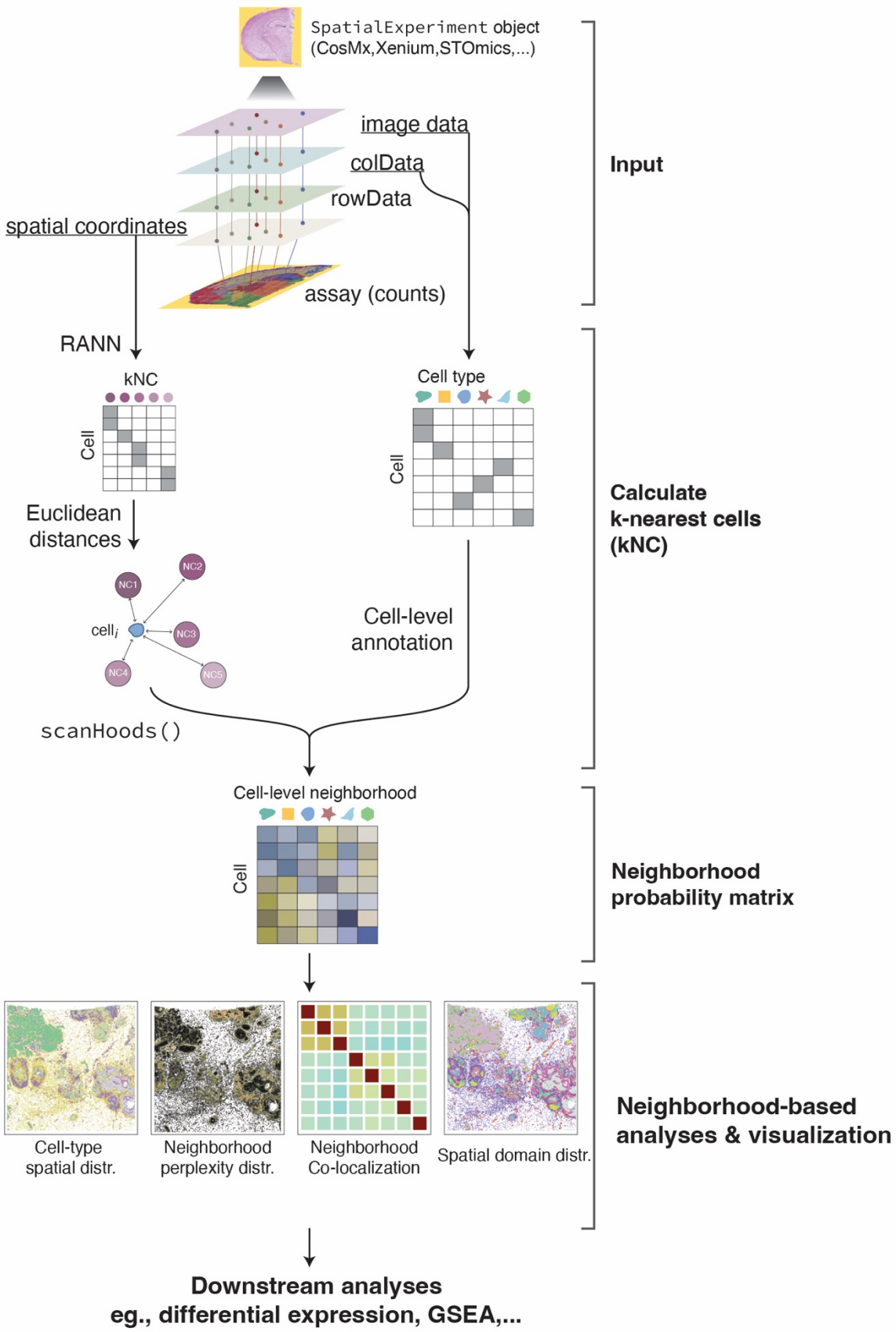
$\mathbb{1}(.)$ is the indicator function, which checks whether cell x_i belongs to the neighborhood h_j as per the annotation matrix A.

Upon the aggregation of probabilities by user-defined cell-level annotation groups, such as cell type annotations, hoodscanR generates a comprehensive probability matrix P, where 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 k and τ affects the results generated by hoodscanR, we conducted an extensive examination of the probability matrix across a range of k and τ values (see Methods). This analysis revealed that different k values generate highly similar results, with a mean Pearson correlation coefficient of 0.93 (Supplementary figure S1). Regarding τ, smaller values assign greater weights to nearby cells, while larger τ values consider more distant cells as contributors to the neighborhood (Supplementary figure S2). Thus, the choice of τ becomes essentially linked to the specific biological questions being addressed. For example, smaller τ values, such as one-fifth of the median of the distance matrix, which is set to the default τ value in the hoodscanR package, are more suitable for analyses focused on local interactions, where nearby cells have stronger influence on the neighborhood

calculation. In contrast, larger τ values, such as the median of the distance matrix, are ideal for capturing more global spatial relationships, incorporating cells that are further away as significant components of the neighborhood.

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. Last but not least, 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. In conclusion, hoodscanR provides a powerful and flexible method for spatial neighborhood identification and analysis.


SpatialExperiment object
(CosMx,Xenium,STOmics,...)
image data
colData
rowData
spatial coordinates
assay (counts)
Input
RANN
kNC
Cell
Cell type
Cell
Euclidean
distances
NC1
NC2
NC3
NC4
NC5
cell$_i$
Cell-level
annotation
scanHoods()
Calculate
k-nearest cells
(kNC)
Cell-level neighborhood
Cell
Neighborhood
probability matrix
Cell-type
spatial distr.
Neighborhood
perplexity distr.
Neighborhood
Co-localization
Spatial domain distr.
Neighborhood-based
analyses & visualization
Downstream analyses
eg., differential expression, GSEA,...

357

Figure 1. Schematic visualization showing the main components and computational workflow of the hoodscanR package. The process begins with inputting a SpatialExperiment object, which contains spatial data and associated metadata. Next, the package calculates the k-nearest cells based on spatial proximity. This step follows by generating a neighborhood probability matrix, which quantifies the likelihood of cell interactions within their local neighborhood. Finally, the package performs visualizations and downstream neighborhood-based analyses to provide insights into spatial patterns and relationships.

Benchmarking hoodscanR in spatial domains identification

Building upon the foundation of cell-level neighborhood probability matrix (Figure 1), 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, identifying neighborhood-driven spatial domains.

To evaluate the effectiveness and robustness of the neighborhood-based spatial domain identification function in hoodscanR, and compared with other tools, we conducted a benchmarking experiment against several state-of-the-art methods (see Methods). This benchmark experiment involved detecting spatial domains of 16 publicly available datasets, covering a range of spatial platforms and tissue types, including CosMx NSCLC (1), MERFISH mouse colon (28), STARmap mouse cortex (29) and Xenium breast cancer (2,30). The datasets were chosen because the tissue is well-annotated with region labels or there are pathological annotataion that can be served as

381 ground truth of spatial domains. hoodscanR was benchmarked against seven other
382 methods that can perform spatial domain detection: BuildNicheAssay from Seurat,
383 Banksy, BayesSpace, MERINGUE, SpaGCN, Stagate and Utag (Figure 2A and
384 supplementary figure S3-7). As a result from these 128 experiments, hoodscanR
385 outperform other methods, followed by Utag and Banksy by achieving the highest
386 performance score on average across all tested datasets (Figure 2B). More importantly,
387 hoodscanR outperforms all others in computing efficiency (Figure 2C), being
388 approximately 21-fold faster on average than Banksy, which ranks second in speed.
389 This advantage in computational speed is particularly important as increasingly large
390 and high-resolution datasets will be generated with the advancements in spatial
391 transcriptomics technologies. Additionally, hoodscanR can recapitulate tissue spatial
392 architecture in a biologically coherent manner. For example, in the MERFISH mouse
393 colon dataset (Figure 2A), hoodscanR accurately delineates four concentric layers,
394 mucosa (MU), submucosa (SM), muscularis externa (ME), and other, all of which
395 contain multiple cell types. This result closely mirrors the ground truth. By clustering
396 cells with similar neighborhood distributions, hoodscanR captures subtle boundaries
397 more effectively than many alternative methods, preserving the colon's characteristic
398 concentric organisation, such as the ME structure. Taken together, these results
399 highlight strengths of hoodscanR in domain identification across large-scale spatial
400 transcriptomics datasets.

401 Additionally, to evaluate the robustness of hoodscanR across different resolutions of cell
402 type annotations, we conducted an experiment using high-resolution, medium-
403 resolution, and low-resolution annotations as inputs. The high-resolution annotations

included detailed cell types, such as CD4+ T cells, CD8+ T cells, and macrophages. The medium-resolution annotations combined all T cells into a single category, and the low-resolution annotations further grouped all immune cells into a single “Immune” category. Despite the reduction in annotation resolution, the identified neighborhood-based spatial domains have a Normalized Mutual Information (NMI) score of greater than 0.8 when comparing using the high-resolution results as the reference (Supplementary figure S8). Taken together, these results showcase the power of hoodcanR in accurately identifying neighborhood-based spatial domains in a scalable and efficient manner. They also indicate that hoodscanR is robust to variations in annotation granularity, maintaining the integrity of the spatial relationships even when the resolution of cell type annotations is reduced.

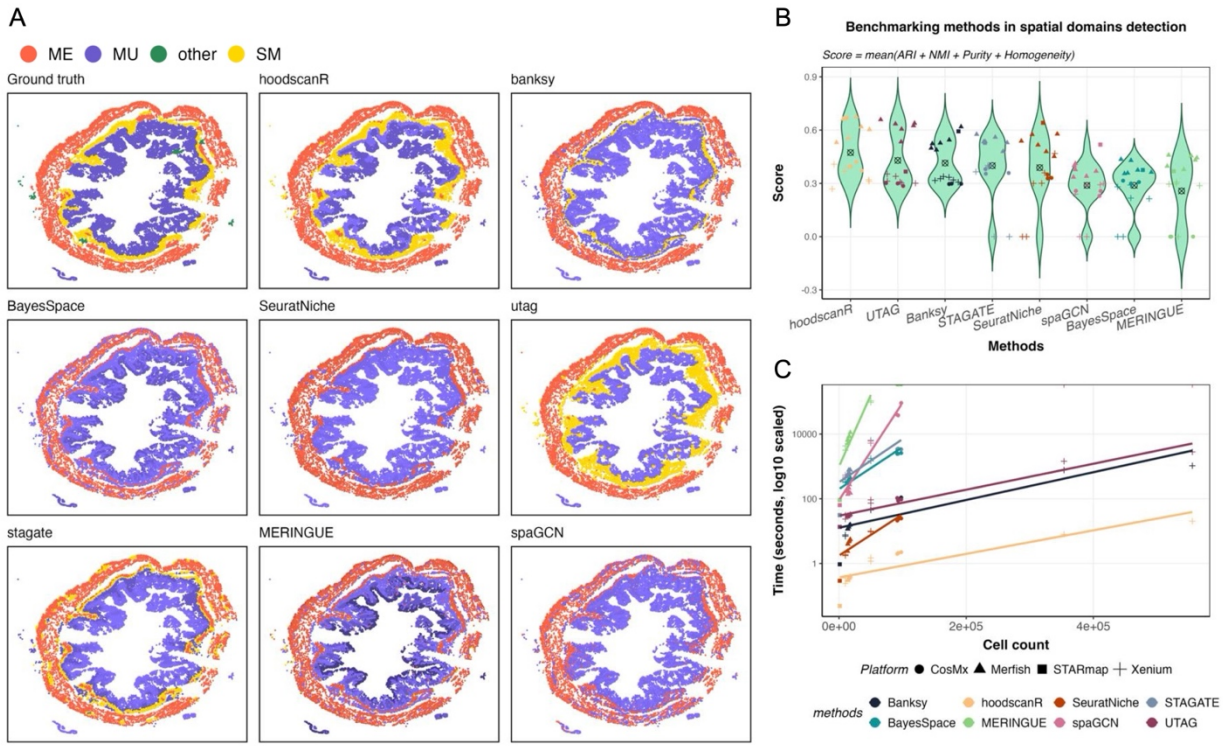


Figure 2. Benchmarking hoodscanR against other methods in detecting spatial domains. A: Spatial maps of the MERFISH mouse colon data, colored by the spatial domains detected from different methods compared to the ground truth domain annotations, including muscularis externa (ME), mucosa (MU), submucosa (SM) and other (top left corner). B: Violin plots showing the performance score of each method across all tested datasets, ordered from the highest to the lowest mean score. C: Computational efficiency of each method, plotted as the log10-scaled time (in seconds) required to process datasets. Shapes represent the platform of the dataset, colors denote the methods, and the lines are generalized linear smooths indicating overall trends for each method.

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 different *in situ* transcriptomic platforms: breast cancer data obtained from the 10X Genomics Xenium (Figure 3A) and non-small cell lung cancer (NSCLC) data acquired from the Nanostring CosMx Spatial Molecular Imager (Figure 3B). We first applied hoodscanR onto the breast cancer dataset using the default parameters (k=100 and τ=median(dist^2)/5). hoodscanR allows us to perform neighborhood identification by profiling neighborhood distributions for each cell within 6 seconds, representing the probability of a cell being situated within each distinct cell-type neighborhood (Figure 3C and 3F).

439 To validate the accuracy of hoodscanR in characterizing these cellular neighborhoods,
440 we focus on one randomly selected cell for each dataset by exploring the distribution of
441 cell types within their spatial area (Figure 3D, E, G, and H). For instance, we examined
442 a ductal carcinoma in situ (DCIS) grade 2 cell, an early form of breast cancer cells, from
443 the Xenium data (Figure 3C and D: cell ID 27620), where we observed 47 DCIS grade 2
444 cells and 21 ACTA2+ myoepithelial cells from the nearest 100 neighboring cells (Figure
445 3D). Consistently, hoodscanR assigned probabilities of 61.02% for residing in the DCIS
446 grade 2 neighborhood and 30.96% for the ACTA2+ myoepithelial neighborhood for this
447 specific cell (Figure 3E). Similarly, when assessing a stromal cell within the CosMx
448 NSCLC data (Figure 3F: cell ID 6_1099), we observed that hoodscanR assigned
449 probabilities of 68.2% for the B cell neighborhood and 19.67% for the regulatory T cell
450 neighborhood while there are 66 and 11 B cells and regulatory T cells in the nearest
451 100 neighboring cells (Figure 3G and H). These examples demonstrate the power of
452 hoodscanR in accurately characterizing cellular neighborhoods within spatial
453 transcriptomics data, regardless of the platform, and its capacity to accommodate
454 scenarios where cells may belong to neighborhoods of multiple cell types. The
455 identification of B cell neighborhoods is particularly noteworthy in the context of cancer
456 therapy responses. B cell neighborhoods serve as crucial sites for antibody production,
457 contributing to the immune response against tumor cells and influencing therapeutic
458 efficacy (31). Furthermore, in lung cancer, the presence of tertiary lymphoid structures
459 (TLS), characterized by highly organized T and B lymphocyte colonies within
460 nonlymphoid tissues, has been associated with favorable clinical outcomes in non-small
461 cell lung cancer (NSCLC) (32). 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 uncertainty, which is measured by perplexity, and performing permutation test for each cell (see Methods). Perplexity is calculated from the probability matrix, capturing the spatial relationships among cells and their respective neighborhoods. Perplexity provides overall measurement of the uncertainty and complexity of cell neighborhoods (Figure 3I and K). This in turn reveals regions of the TME with distinct cellular compositions and areas with complex interactions between cell types. P-values of perplexity (Figure 3J and L) can be obtained via an empirical permutation test (see Methods). This allows users to identify regions with significant higher perplexity from tissue. Altogether, 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.

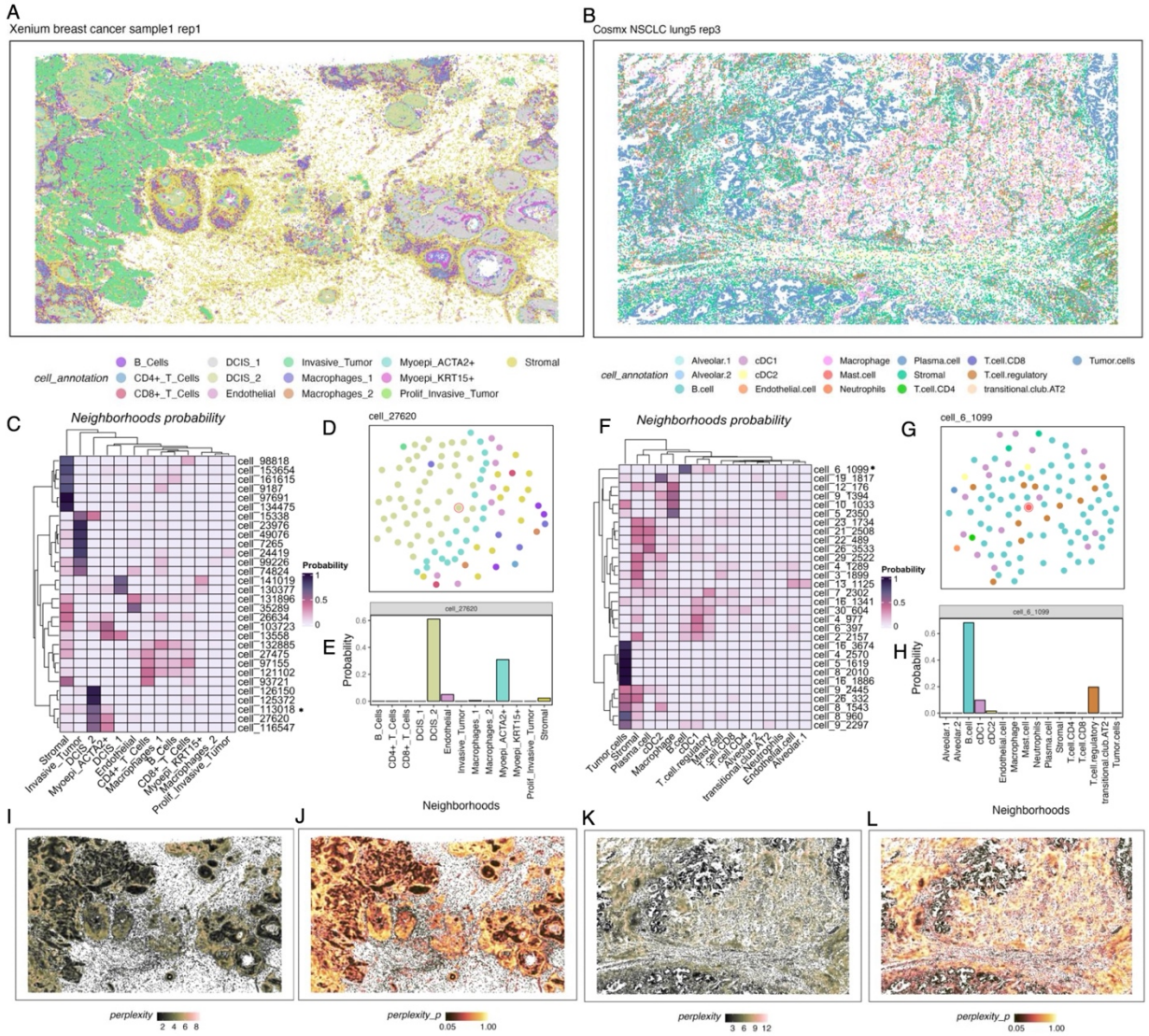


Figure 3. Neighborhood identification in 10X Xenium breast cancer data and Nanostring CosMx NSCLC data. Cell type spatial distribution in the breast cancer data (A) and NSCLC data (B). Neighborhood distribution visualization via heatmap of randomly selected 30 cells from breast cancer data (C) and NSCLC 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). Perplexity spatial distribution of cells in the breast cancer data (I) and NSCLC data (K). P-value distribution of perplexity in the breast cancer data (J) and NSCLC data (L).

492

493 ## hoodscanR performs neighborhood-based downstream analyses

494

495 Existing neighborhood identification methods, such as Squidpy and Giotto,
496 predominantly focus on neighborhood co-localization analyses. Another key function of
497 hoodscanR is to generate neighborhood profile at single-cell level and to carry out
498 neighborhood-based downstream analyses, features notably absent in other existing
499 tools. To demonstrate the versatility of hoodscanR, we use the CosMx NSCLC dataset
500 as an example.

501

502 Firstly, as with other spatial analysis tools, hoodscanR can perform neighborhood co-
503 localization analysis by computing Pearson correlations on the neighborhood
504 distribution of cells. Hence, the co-localization status of each cell type neighborhood
505 within this specific tissue slide can then be visualized (Figure 4A). To benchmark the
506 ability of hoodscanR, Squidpy and Giotto when carrying out co-localization analysis, we
507 subset the Xenium breast cancer data and CosMx NSCLC into different resolution,
508 followed by applying these tools to the subsets. As a result, hoodscanR demonstrated
509 superior computational performance (Supplementary figure S9), while delivering similar
510 outcome (mean Pearson correlation coefficient of 0.781) of neighborhood co-
511 localization compared to both Xenium and CosMx data (Supplementary figure S10 and
512 S11). The computational efficiency gains significance, particularly in the context of the
513 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 4B). 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.

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 4C), demonstrating complex spatial associations. For instance, cluster 1, a candidate cluster for TLS, corresponds to a neighborhood including B cells, cDC1 cells, and stromal cells (Figure 4D - 1), cluster 3 aligns with macrophages and cDC2 cells (Figure 4D - 3), and cluster 6 corresponds to tumor cells (Figure 4D - 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.

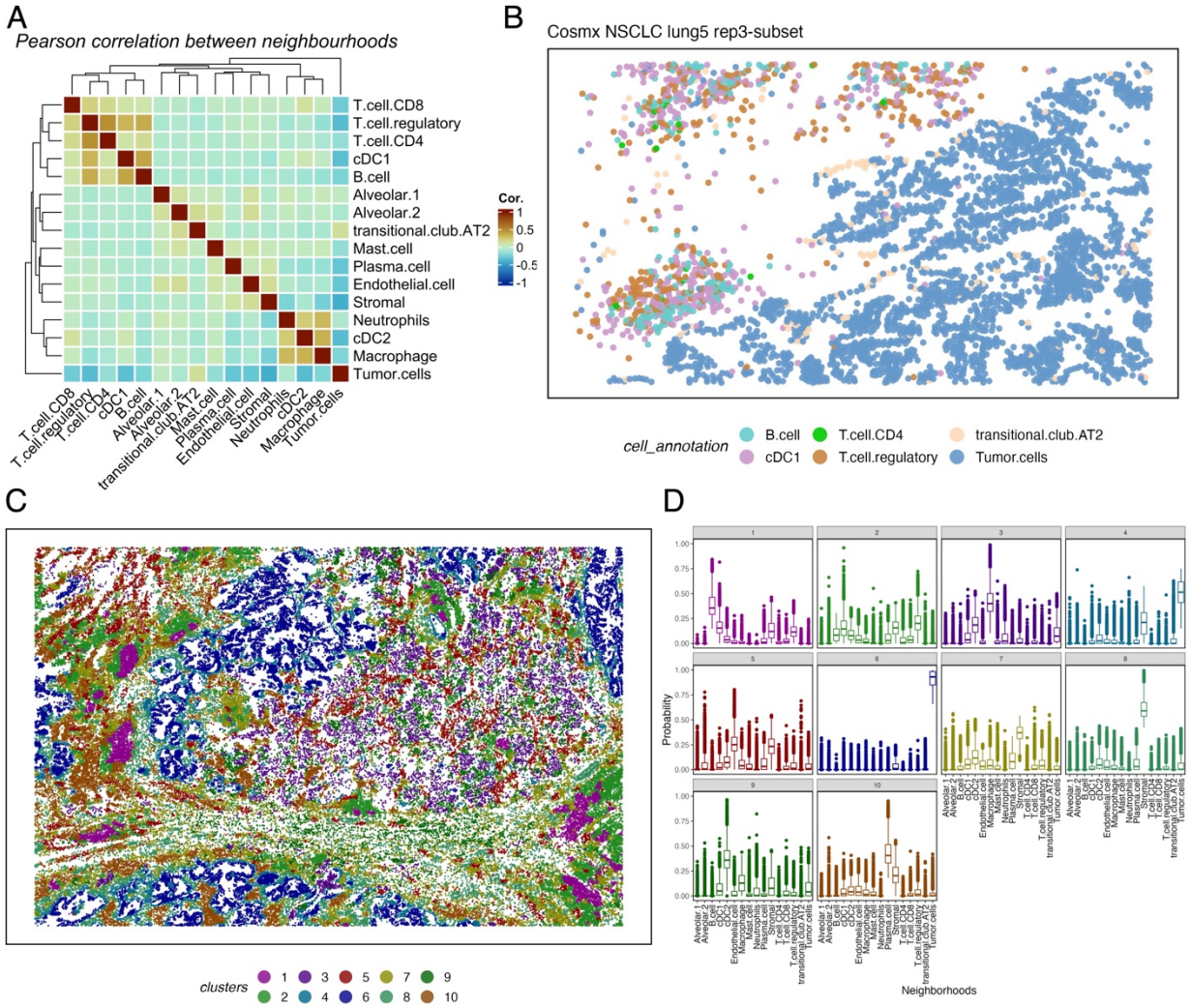


Figure 4. Neighborhood-based downstream analyses in CosMx NSCLC data performed by hoodscanR. A: heatmap representing the co-localization status of cell type neighborhood in this tissue slide. Colors denote positive (dark red) and negative (blue) correlations. B: Spatial location plot of a subset of cell 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.

542 ## hoodscanR detects changes between tumor cells from different neighborhoods

543

544 Building on these findings, we then used uniform manifold approximation and projection
545 (UMAP) to perform dimension reduction visualization on the expression data of the
546 CosMx NSCLC data, enabling the projection of gene expression profiles into a lower-
547 dimensional space. This facilitates the visualization of cell lineage (Figure 5A) alongside
548 identified neighborhood-based clusters (Figure 5B). A distinctive feature observed is the
549 dispersion of cells of the same cell type across different neighborhood clusters,
550 signifying diverse spatial neighborhood profiles. For example, a substantial proportion
551 (76.14%) of macrophages (pink points in Figure 5A) are distributed across various
552 neighborhood clusters, including cluster 3 (43.63%), indicative of the macrophage +
553 cDC2 + tumor neighborhood, clusters 7 (12.31%), and cluster 9 (20.2%), representing
554 the stromal + cDC2 neighborhood (Figure 3D and 5B).

555

556 An important aspect of hoodscanR lies in its ability to investigate the relationship
557 between spatial neighborhoods and transcriptional changes. To demonstrate this, we
558 conducted a nuanced analysis by extracting and pseudo-bulking tumor cells from two
559 distinct neighborhood clusters: the stromal cluster and the macrophage cluster across
560 three consecutive slides (Supplementary figure S12). Interestingly, different spatial
561 neighborhoods contribute significantly to the variation observed in the first dimension
562 from a principal component analysis (PCA) of expression of pseudo-bulk samples

563 (Figure 5C). This signifies that hoodscanR can capture transcriptional changes
564 attributed to diverse spatial neighborhoods.

565

566 Subsequently, we performed a differential expression (DE) analysis using the limma-
567 voom pipeline (24), identifying 220 DE genes from 832 genes, including 73 up-regulated
568 and 147 down-regulated genes when comparing tumor cells from the macrophage
569 neighborhood to those from the stromal neighborhood (Figure 5D and supplementary
570 file 1). Additionally, we performed gene-set enrichment analysis on the identified DE
571 genes with MSigDB gene-sets (33,34), detecting 384 significantly enriched gene-sets
572 (Supplementary file 2). We further perform unsupervised clustering on gene-sets using
573 vissE (26), identifying clusters of gene-sets networks (Figure 5E and supplementary
574 figure S13, middle panel). Notably, we observe pathways enriched in down-regulated
575 DE genes related to collagen, such as *collagen-activated tyrosine kinase receptor*
576 *signalling pathway* and *collagen metabolic and catabolic process* (Figure 5E left panel).
577 These pathways are accompanied by the differential expression of key collagen-related
578 genes, such as *COL1A1*, *COL11A1* and *COL5A1* (Figure 5E right panel). Previous
579 studies have found that the overexpression of collagen genes such as *COL11A1* (35)
580 and *COL3A1* (36) in NSCLC may indicate poor prognosis and drug resistance, and
581 *COL1A1* is correlated with immune infiltration in NSCLC (37). Our finding of these
582 genes that are expressed significantly more in the tumor cells from the macrophage
583 neighborhood than tumor cells from the stromal neighborhood can potentially lead a
584 more detailed investigation of the biological mechanism about transcriptional changes
585 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 spatial organisation within the TME.

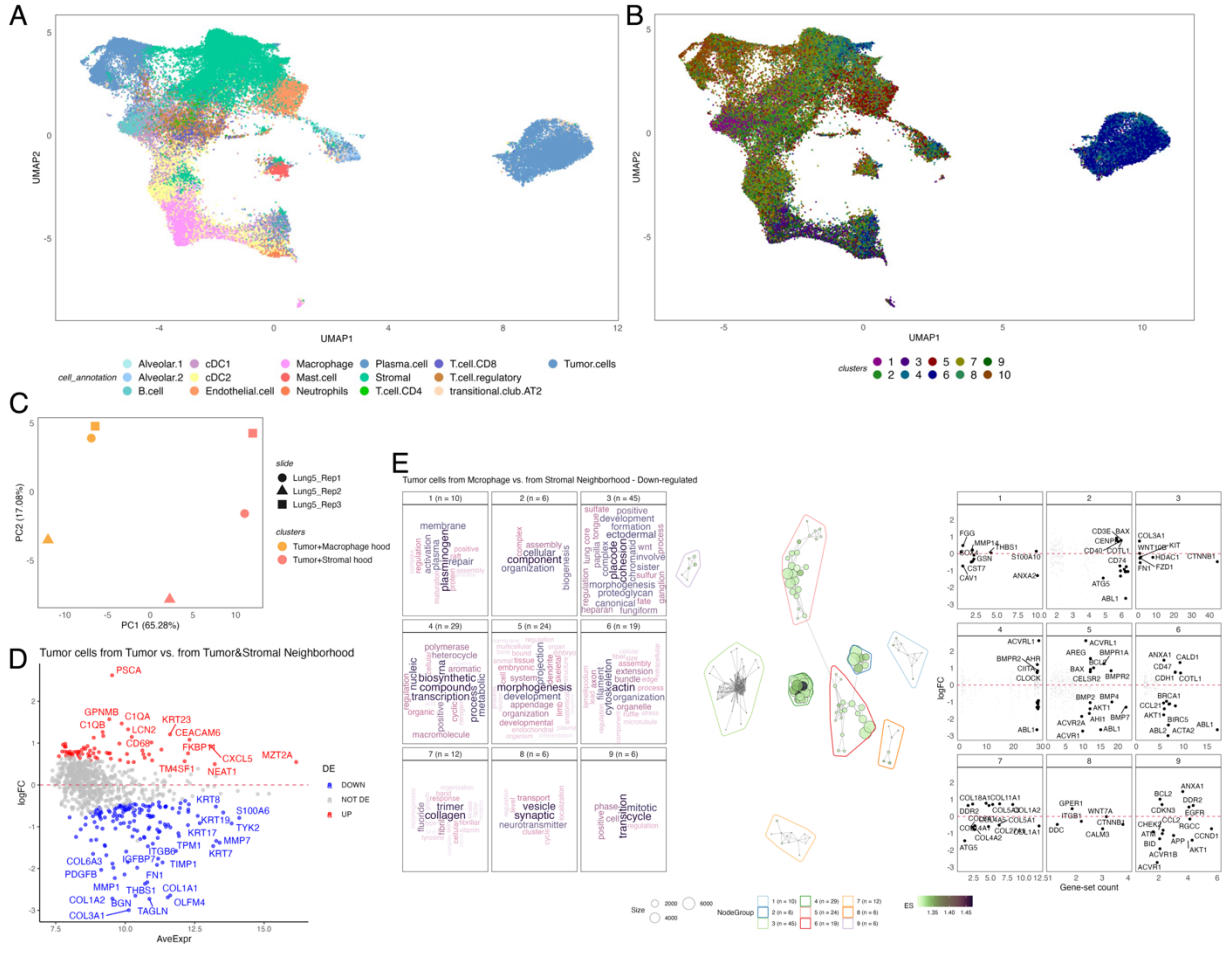


Figure 5. Neighborhood-based transcriptomics analysis in CosMx NSCLC data. A: UMAP of expression of cells in the data, colors denote cell types. B: UMAP of expression of cells in the data, colors denote the identified neighborhood-based clusters. C: PCA of pseudo-bulk samples of tumor cells from two 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

597 tumor cells from macrophage neighborhood and tumor cells from stromal neighborhood. Left
598 panel are word cloud plots describing gene-set clusters of different biological themes, middle
599 panel is the gene-set overlap network graph of gene-sets, and right panel is the fold-change
600 (log2-scaled) for genes belonging to each gene-set cluster.

601

602 hoodscanR supports different gene annotations

603

604 The accuracy of cell type-based cellular neighborhoods, as identified in the previous
605 case, is inherently tied to the accuracy of cell type annotation. Thus, we built
606 hoodscanR to be flexible and capable to detect various cellular neighborhoods based
607 on different gene annotation inputs. A particularly valuable application is gene
608 expression-based neighborhoods detection. To showcase this, we utilised a breast
609 cancer tissue, where we can focus on breast cancer-related hormone receptor genes,
610 including androgen receptor gene (*AR*), estrogen receptor gene (*ESR1*), and
611 progesterone receptor gene (*PGR*). By assessing if these genes are expressed or not,
612 we classified 574,527 cells from a Xenium Invasive ductal carcinoma (IDC) dataset into
613 eight distinct groups (Figure 6A). Similar to the previous neighborhood identification
614 based on cell types, hoodscanR can identify spatial domains based on gene
615 expression-specific neighborhoods (Figure 6B and supplementary figure S14). These
616 analyses lead to a nuanced understanding which adeptly discover tumor cells located
617 within neighborhoods characterized by varying combinations of hormone receptors. This
618 not only suggests a spatial perspective on the progression of DCIS, influenced by

619 distinct combinations of hormone receptors but also sheds light on the higher-order
620 spatial structure of cells with different hormone receptor expression profiles.

621

622 Moreover, interesting insights can be observed by comparing the pathological
623 annotations (Figure 6C) of this slide (30), where invasive and non-invasive (DCIS)
624 cancer phenotype regions were annotated, with the identified neighborhood clusters.
625 Invasive tumors exhibit a distinctive pattern with cells from cluster 3 scattered
626 throughout, accompanied by cells from cluster 6 (Figure 6D). Both clusters are
627 associated with *ESR1+PGR+* and *ESR1-PGR+* cell neighborhoods, respectively. This
628 finding aligns with previous findings indicating that nearly 80% of invasive breast
629 cancers are ER-positive, and PR is overexpressed in ER+ tumors (38). Conversely,
630 DCIS regions are notably associated with cells from cluster 5, which form the inner layer
631 surrounded by cells from cluster 3, comprising the outer layer (Figure 6D). Cluster 5
632 includes cells expressing *AR*, which is consistent with previous findings showing *AR*
633 expression in DCIS components adjacent to invasive cancer (39). Additionally, the
634 expression of *AR* has been reported to decrease as the disease progresses from DCIS
635 to invasive cancer (40). These observations suggest a potential relationship between
636 the tumor type (non-invasive or invasive) and the higher-order spatial organisation of
637 cells with diverse hormone receptor expression profiles.

638

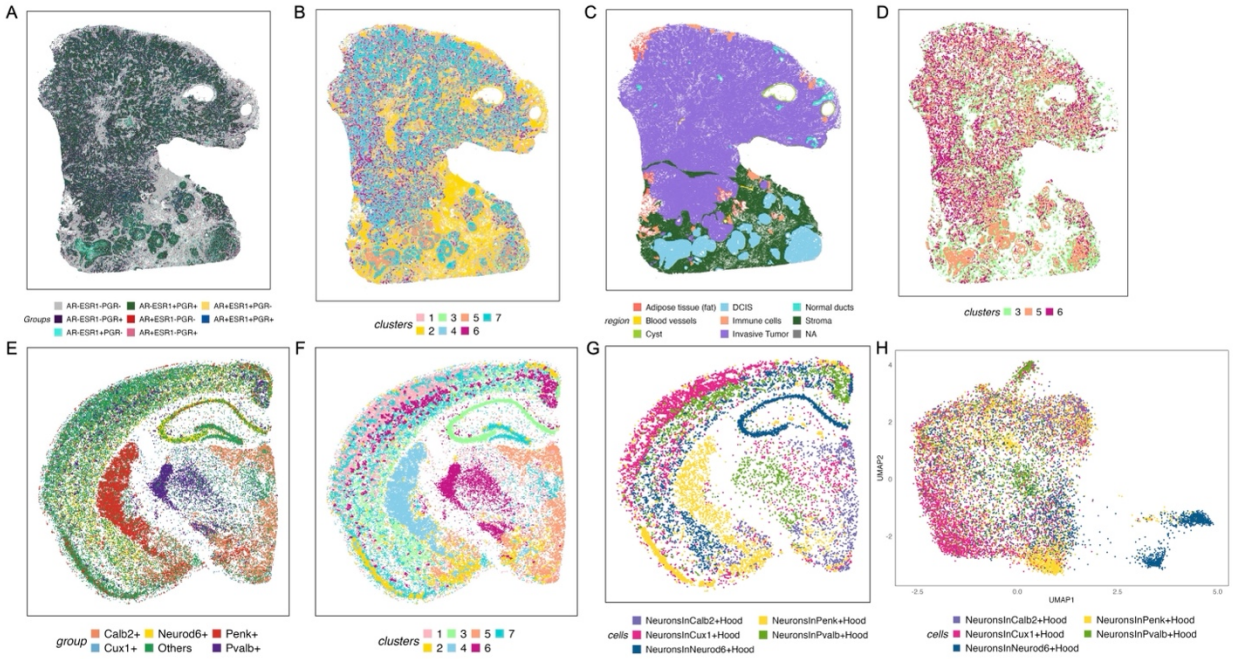


Figure 6. Gene expression-based neighborhood analysis in 10X Xenium IDC (A-D) and VisiumHD mouse brain (E-H) data using hoodscanR. A: spatial plot showing the hormone gene expression-based cell grouping on the tissue slide. B: gene expression-based domains in IDC tissue. C: pathological annotation of regions on the tissue. D: spatial plot showing specific neighborhood clusters, including cluster 3, 5 and 6. E: spatial plot showing the selected gene expression-based cell grouping. F: gene expression-based domains in mouse brain tissue. G: neurons spatial distribution in different gene-expression neighborhoods. H: a dimension reduction UMAP visualization of the neurons from different neighborhoods.

To further demonstrate the broad application of hoodscanR beyond cancer and human data, we analysed the publicly available Visium HD mouse brain data. In this case, we selected five marker genes that can divide the brain into different regions: *Calb2* for paraventricular nucleus of the thalamus (PVT) (41), *Neurod6* for deeper layers of the cortex (42), *Penk* for striatal medium spiny neuron (43), *Cux1* for upper layer of the

654 cortex (44) and *Pvalb* for hippocampus (45). Additionally, we included the gene *Rbfox3*,
655 which is exclusively expressed in neuron cells (46). The spatial map of these cells
656 reveals their distinct regional expression patterns within the brain tissue (Figure 6E and
657 supplementary figure S15). Applying hoodscanR, we can identify neighborhood-based
658 clusters (Figure 6F and supplementary figure S16). These clusters largely correspond to
659 specific brain regions (Supplementary figure S17), reflecting the well-structured nature
660 of the brain tissue, with each cluster predominately associated with one of the maker
661 genes.

662

663 Focusing on neuron cells by filtering the dataset for cells expression *Rbfox3*, we
664 explored how these neurons are distributed across different neighborhoods (Figure 6G).
665 By conducting dimension reduction via UMAP on the expression data of these neurons
666 (Figure 6H), we can visualise the neurons from different neighborhoods tend to cluster
667 differently, indicating variance in expression between these neurons, especially
668 between those within Neurod6+ neighborhoods and Calb2+ neighborhoods
669 (Supplementary figure S18). Moreover, performing a differential expression analysis
670 between neurons from these two neighborhoods revealed a set of genes that are
671 significantly upregulated in the Neurod6+ region compared to the Calb2+ region
672 (Supplementary figure S19), indicating distinct microenvironmental influences and
673 potential functional specialization. For example, *Nrgn*, *Hpca*, and *Rasgrp1*, which are all
674 involved in regulating intracellular calcium signalling and synaptic plasticity, are up-
675 regulated in the neurons from the Neurod6+ neighborhood, predominantly localized to
676 the hippocampal area. This observation aligns with previous studies showing that these

genes are critical for synaptic plasticity (47-50), which underlies learning and memory. Taken together, these patterns illustrate how local spatial neighborhood composition can shape neuronal identity and function, highlighting hoodscanR's ability to detect spatially restricted transcriptional differences even in highly structured tissues like the brain.

In summary, these findings highlight hoodscanR's capability to identify subtle spatial changes in expression-based cellular neighborhoods, providing novel insights into the complex spatial dynamics of gene expression in both cancerous and non-cancerous tissues.

Discussion

Spatial technologies are pushing the limits 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. In our benchmarking experiments, hoodscanR demonstrated robust accuracy and computation efficiency compared to several state-of-the-art methods, further validating its utility in diverse spatial transcriptomics applications.

While we have shown hoodscanR can identify biologically meaningful cellular neighborhoods, as with all methods it is not without limitations. Firstly, although hoodscanR is flexible in relation to gene annotations, 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, which hoodscanR depends upon other methods. The segmentation process, which determines how individual cells

710 are identified and their spatial coordinates are established, is fundamental to the
711 accurate detection of cellular neighborhoods. Variations in segmentation methods, such
712 as differences in how cell boundaries are defined or how centroids are calculated, can
713 lead to significant differences in cell type distributions and spatial relationships within
714 the tissue. Such variations can impact the neighborhood detection results produced by
715 hoodscanR, potentially leading to different biological interpretations. A systematic
716 review of existing segmentation methods is lacking, necessitating future research to
717 evaluate and compare methods under diverse spatial transcriptomic platforms. Lastly,
718 while hoodscanR enables exploration of spatial gene expression patterns, users must
719 interpret results cautiously. Recognizing that spatial context alone may not capture the
720 full complexity of molecular interactions within a tissue. Integrating multi-omics data can
721 provide a more comprehensive understanding, ensuring spatial analyses are embedded
722 within a broad molecular context. Strategically addressing these considerations allows
723 researchers to fine-tune the utilization of hoodscanR, strengthening the integrity of their
724 analyses and facilitating the discovery of novel insights.

725

726 Three of the major strengths of hoodscanR are its compatibility, adaptability and
727 flexibility. Developed based on the Bioconductor SpatialExperiment infrastructure,
728 hoodscanR exhibits compatibility with many other spatial and single-cell RNA-seq-
729 based tools from Bioconductor, amplifying its utility for downstream analyses.
730 Importantly, its adaptability makes it platform agnostic, demonstrated by successful
731 applications on Nanostring CosMx, 10X Genomics Xenium, MERFISH and STARmap.
732 Spatial datasets from various platforms, including Vizgen MERSCOPE, 10X Visium HD,

733 BGI STOmics, and Akoya Biosciences CODEX, can also undergo comprehensive
734 analysis using hoodscanR, given the availability of cell-based coordinates. The flexibility
735 of hoodscanR is demonstrated by the types of annotations it can use. With the Xenium
736 breast cancer data, hoodscanR showcased this by not only using cell type annotations
737 but also accommodating gene expression level grouping of cells, suggesting the
738 potential for exploring additional annotation options such as ligand-receptor or growth
739 factor-receptor annotations. This design provides researchers the flexibility to generate
740 hypotheses on the basis of integrating spatial localization and gene expression before
741 testing statistical associations between these. Lastly, while hoodscanR is not designed
742 for direct cell-cell communication analysis, it plays a crucial role in accurately identifying
743 and characterizing spatial neighborhoods. This capability can complement existing cell-
744 cell communication tools, such as COMMOT (51), CellChat (52) or CellPhoneDB (53),
745 by providing a more refined spatial context that may enhance the accuracy and
746 robustness of cell-cell communication network identification in complex tissue
747 environments.

748

749 The significance of hoodscanR in identifying and analyzing neighborhoods becomes
750 particularly crucial in the context of complex diseases, notably in cancer research.
751 Neighborhood information is indispensable for unraveling complex disease etiology,
752 especially so for understanding disease progression. In cancer research, where the
753 TME plays an important role in dictating therapy responses, the ability offered by
754 hoodscanR to identify neighborhoods offers a unique perspective to investigate novel
755 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 are potentially associated with the presence of TLS, which correlated with positive clinical results (32). 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.

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 CosMx dataset was generated from FFPE NSCLC tissue samples using a 960-plex CosMx RNA panel. 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 MERFISH mouse colon data was downloaded from https://doi.org/10.5061/dryad.rjdfn2zh3, originated from Cadinu et al., 2024 (28). The STARmap mouse cortex data was sourced from the Wang et al., 2018 study (29). The 10X VisiumHD mouse brain dataset was sourced from the official 10X Genomics website: https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he.

Code availability

The code used for performing the described analyses is available in GitHub at https://github.com/ningbioinfo/hoodscanR_manuscript_code. The hoodscanR package is freely available in Bioconductor (release > 3.18) at https://bioconductor.org/packages/release/bioc/html/hoodscanR.html.

Acknowledgements

The authors thank both Nanostring Technologies and 10X Genomics for releasing publicly available CosMx and Xenium datasets, respectively.

Funding

Ning Liu and Dharmesh Bhuva are supported by the South Australian immunoGENomics Cancer Institute (SAiGENCI) received grant funding from the Australian Government. WEHI acknowledges the support of the Operational Infrastructure Program of the Victorian Government.

Author information

Conceptualization: N.L., J.M., C.W.T. and M.J.D. Method development: N.L., D.D.B., A.M., M.L. Writing – original draft: N.L. Writing – draft revision: N.L., J.C., S.C.L., M.K., J.C., A.K., Y.C., C.W.T., F.L, J.M.P.

Ethic declarations

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

819 # Reference

820

821 1. He, S., Bhatt, R., Brown, C., Brown, E.A., Buhr, D.L.,
822 Chantranuvatana, K., Danaher, P., Dunaway, D., Garrison, R.G.,
823 Geiss, G. *et al.* (2022) High-plex imaging of RNA and proteins at
824 subcellular resolution in fixed tissue by spatial molecular imaging.
825 *Nat Biotechnol*, **40**, 1794-1806.
826 2. Janesick, A., Shelansky, R., Gottscho, A.D., Wagner, F., Williams,
827 S.R., Rouault, M., Beliakoff, G., Morrison, C.A., Oliveira, M.F.,
828 Sicherman, J.T. *et al.* (2023) High resolution mapping of the tumor
829 microenvironment using integrated single-cell, spatial and in situ
830 analysis. *Nat Commun*, **14**, 8353.
831 3. Yao, Z., van Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D.,
832 Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M. *et al.* (2023) A
833 high-resolution transcriptomic and spatial atlas of cell types in the
834 whole mouse brain. *Nature*, **624**, 317-332.
835 4. Parra, E.R., Francisco-Cruz, A. and Wistuba, II. (2019) State-of-the-
836 Art of Profiling Immune Contexture in the Era of Multiplexed
837 Staining and Digital Analysis to Study Paraffin Tumor Tissues.
838 *Cancers (Basel)*, **11**.
839 5. Jhaveri, N., Ben Cheikh, B., Nikulina, N., Ma, N., Klymyshyn, D.,
840 DeRosa, J., Mihani, R., Pratapa, A., Kassim, Y. and Bommakanti, S.
841 (2023) Mapping the spatial proteome of head and neck tumors:
842 key immune mediators and metabolic determinants in the tumor
843 microenvironment. *GEN Biotechnology*, **2**, 418-434.
844 6. Warchol, S., Krueger, R., Nirmal, A.J., Gaglia, G., Jessup, J., Ritch,
845 C.C., Hoffer, J., Muhlich, J., Burger, M.L., Jacks, T. *et al.* (2023)
846 Visinity: Visual Spatial Neighborhood Analysis for Multiplexed
847 Tissue Imaging Data. *IEEE Trans Vis Comput Graph*, **29**, 106-116.
848 7. Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A.C.,
849 Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I.

850 *et al.* (2022) Squidpy: a scalable framework for spatial omics
851 analysis. *Nat Methods*, **19**, 171-178.
852 8. Dries, R., Zhu, Q., Dong, R., Eng, C.L., Li, H., Liu, K., Fu, Y., Zhao, T.,
853 Sarkar, A., Bao, F. *et al.* (2021) Giotto: a toolbox for integrative
854 analysis and visualization of spatial expression data. *Genome Biol*,
855 **22**, 78.
856 9. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P.,
857 Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-
858 Granda, C. and Satija, R. (2024) Dictionary learning for integrative,
859 multimodal and scalable single-cell analysis. *Nat Biotechnol*, **42**,
860 293-304.
861 10. Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X. and Fan, J.
862 (2021) Characterizing spatial gene expression heterogeneity in
863 spatially resolved single-cell transcriptomic data with nonuniform
864 cellular densities. *Genome research*, **31**, 1843-1855.
865 11. Singhal, V., Chou, N., Lee, J., Yue, Y., Liu, J., Chock, W.K., Lin, L.,
866 Chang, Y.C., Teo, E.M.L., Aow, J. *et al.* (2024) BANKSY unifies cell
867 typing and tissue domain segmentation for scalable spatial omics
868 data analysis. *Nat Genet*, **56**, 431-441.
869 12. Zhao, E., Stone, M.R., Ren, X., Guenthoer, J., Smythe, K.S., Pulliam,
870 T., Williams, S.R., Uytingco, C.R., Taylor, S.E. and Nghiem, P. (2021)
871 Spatial transcriptomics at subspot resolution with BayesSpace.
872 *Nature biotechnology*, **39**, 1375-1384.
873 13. Dong, K. and Zhang, S. (2022) Deciphering spatial domains from
874 spatially resolved transcriptomics with an adaptive graph attention
875 auto-encoder. *Nat Commun*, **13**, 1739.
876 14. Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., Lee,
877 E.B., Shinohara, R.T. and Li, M. (2021) SpaGCN: Integrating gene
878 expression, spatial location and histology to identify spatial
879 domains and spatially variable genes by graph convolutional
880 network. *Nat Methods*, **18**, 1342-1351.
881 15. Kim, J., Rustam, S., Mosquera, J.M., Randell, S.H., Shaykhiev, R.,
882 Rendeiro, A.F. and Elemento, O. (2022) Unsupervised discovery of

883 tissue architecture in multiplexed imaging. *Nat Methods*, **19**, 1653-
884 1661.
885 16. Tan, C.W., Chen, J., Liu, N., Bhuva, D.D., Blick, T., Monkman, J.,
886 Cooper, C., Kharbanda, M., Feher, K., Phipson, B. *et al.* (2024) In
887 situ single-cell profiling sheds light on IFI27 localisation during
888 SARS-CoV-2 infection. *EBioMedicine*, **101**, 105016.
889 17. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S.,
890 Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M. *et al.* (2021)
891 Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573-
892 3587 e3529.
893 18. Danaher, P., Zhao, E., Yang, Z., Ross, D., Gregory, M., Reitz, Z., Kim,
894 T.K., Baxter, S., Jackson, S. and He, S. (2022) Insitutype: likelihood-
895 based cell typing for single cell spatial transcriptomics. *bioRxiv*,
896 2022.2010. 2019.512902.
897 19. Salcher, S., Sturm, G., Horvath, L., Untergasser, G., Kuempers, C.,
898 Fotakis, G., Panizzolo, E., Martowicz, A., Trebo, M., Pall, G. *et al.*
899 (2022) High-resolution single-cell atlas reveals diversity and
900 plasticity of tissue-resident neutrophils in non-small cell lung
901 cancer. *Cancer Cell*, **40**, 1503-1520 e1508.
902 20. Shannon, C.E. (1997) The mathematical theory of communication.
903 1963. *MD Comput*, **14**, 306-317.
904 21. Lun, A.T., McCarthy, D.J. and Marioni, J.C. (2016) A step-by-step
905 workflow for low-level analysis of single-cell RNA-seq data with
906 Bioconductor. *F1000Res*, **5**, 2122.
907 22. McCarthy, D.J., Campbell, K.R., Lun, A.T. and Wills, Q.F. (2017)
908 Scater: pre-processing, quality control, normalization and
909 visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**,
910 1179-1186.
911 23. Liu, N., Bhuva, D.D., Mohamed, A., Bokelund, M., Kulasinghe, A.,
912 Tan, C.W. and Davis, M.J. (2024) standR: spatial transcriptomic
913 analysis for GeoMx DSP data. *Nucleic Acids Res*, **52**, e2.

914 24. Law, C.W., Chen, Y.S., Shi, W. and Smyth, G.K. (2014) voom:
915 precision weights unlock linear model analysis tools for RNA-seq
916 read counts. *Genome Biology*, **15**.
917 25. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization
918 method for differential expression analysis of RNA-seq data.
919 *Genome Biology*, **11**.
920 26. Bhuva, D.D., Tan, C.W., Liu, N., Whitfield, H.J., Papachristos, N.,
921 Lee, S.C., Kharbanda, M., Mohamed, A. and Davis, M.J. (2024)
922 vissE: a versatile tool to identify and visualise higher-order
923 molecular phenotypes from functional enrichment analysis. *Bmc*
924 *Bioinformatics*, **25**.
925 27. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R. and Wu, A.Y.
926 (1998) An optimal algorithm for approximate nearest neighbor
927 searching fixed dimensions. *Journal of the ACM (JACM)*, **45**, 891-
928 923.
929 28. Cadinu, P., Sivanathan, K.N., Misra, A., Xu, R.J., Mangani, D., Yang,
930 E., Rone, J.M., Tooley, K., Kye, Y.C., Bod, L. *et al.* (2024) Charting
931 the cellular biogeography in colitis reveals fibroblast trajectories
932 and coordinated spatial remodeling. *Cell*, **187**, 2010-2028 e2030.
933 29. Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N.,
934 Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J. *et al.* (2018)
935 Three-dimensional intact-tissue sequencing of single-cell
936 transcriptional states. *Science*, **361**.
937 30. Bhuva, D.D., Tan, C.W., Salim, A., Marceaux, C., Pickering, M.A.,
938 Chen, J., Kharbanda, M., Jin, X., Liu, N., Feher, K. *et al.* (2024)
939 Library size confounds biology in spatial transcriptomics data.
940 *Genome Biol*, **25**, 99.
941 31. Tokunaga, R., Naseem, M., Lo, J.H., Battaglin, F., Soni, S., Puccini,
942 A., Berger, M.D., Zhang, W., Baba, H. and Lenz, H.J. (2019) B cell
943 and B cell-related pathways for novel cancer treatments. *Cancer*
944 *Treat Rev*, **73**, 10-19.
945 32. Ren, F., Xie, M., Gao, J., Wu, C., Xu, Y., Zang, X., Ma, X., Deng, H.,
946 Song, J., Huang, A. *et al.* (2022) Tertiary lymphoid structures in lung

947 adenocarcinoma: characteristics and related factors. *Cancer Med*,
948 **11**, 2969-2977.
949 33. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H.,
950 Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database
951 (MSigDB) 3.0. *Bioinformatics*, **27**, 1739-1740.
952 34. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert,
953 B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R.,
954 Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis:
955 A knowledge-based approach for interpreting genome-wide
956 expression profiles. *P Natl Acad Sci USA*, **102**, 15545-15550.
957 35. Shen, L., Yang, M., Lin, Q., Zhang, Z., Zhu, B. and Miao, C. (2016)
958 COL11A1 is overexpressed in recurrent non-small cell lung cancer
959 and promotes cell proliferation, migration, invasion and drug
960 resistance. *Oncology reports*, **36**, 877-885.
961 36. Wang, L., Sun, Y., Guo, Z. and Liu, H. (2022) COL3A1
962 Overexpression Associates with Poor Prognosis and Cisplatin
963 Resistance in Lung Cancer. *Balkan Med J*, **39**, 393-400.
964 37. Geng, Q., Shen, Z., Li, L. and Zhao, J. (2021) COL1A1 is a prognostic
965 biomarker and correlated with immune infiltrates in lung cancer.
966 *PeerJ*, **9**, e11145.
967 38. Wei, S. (2023) Hormone receptors in breast cancer: An update on
968 the uncommon subtypes. *Pathol Res Pract*, **250**, 154791.
969 39. Yu, Q., Niu, Y., Liu, N., Zhang, J., Liu, T., Zhang, R., Wang, S., Ding, X.
970 and Xiao, X. (2011) Expression of androgen receptor in breast
971 cancer and its significance as a prognostic factor. *Annals of*
972 *oncology*, **22**, 1288-1294.
973 40. Salvi, S., Bonafe, M. and Bravaccini, S. (2020) Androgen receptor in
974 breast cancer: A wolf in sheep's clothing? A lesson from prostate
975 cancer. *Semin Cancer Biol*, **60**, 132-137.
976 41. Gao, C., Gohel, C.A., Leng, Y., Ma, J., Goldman, D., Levine, A.J. and
977 Penzo, M.A. (2023) Molecular and spatial profiling of the
978 paraventricular nucleus of the thalamus. *Elife*, **12**, e81818.

979 42. Tutukova, S., Tarabykin, V. and Hernandez-Miranda, L.R. (2021)
980 The Role of Neurod Genes in Brain Development, Function, and
981 Disease. *Front Mol Neurosci*, **14**, 662774.
982 43. Yang, L., Su, Z., Wang, Z., Li, Z., Shang, Z., Du, H., Liu, G., Qi, D.,
983 Yang, Z., Xu, Z. and Zhang, Z. (2021) Transcriptional profiling
984 reveals the transcription factor networks regulating the survival of
985 striatal neurons. *Cell Death Dis*, **12**, 262.
986 44. Cubelos, B., Sebastian-Serrano, A., Beccari, L., Calcagnotto, M.E.,
987 Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo,
988 J.M., Bovolenta, P. *et al.* (2010) Cux1 and Cux2 regulate dendritic
989 branching, spine morphology, and synapses of the upper layer
990 neurons of the cortex. *Neuron*, **66**, 523-535.
991 45. Xia, F., Richards, B.A., Tran, M.M., Josselyn, S.A., Takehara-
992 Nishiuchi, K. and Frankland, P.W. (2017) Parvalbumin-positive
993 interneurons mediate neocortical-hippocampal interactions that
994 are necessary for memory consolidation. *Elife*, **6**, e27868.
995 46. Lin, Y.S., Wang, H.Y., Huang, D.F., Hsieh, P.F., Lin, M.Y., Chou, C.H.,
996 Wu, I.J., Huang, G.J., Gau, S.S. and Huang, H.S. (2016) Neuronal
997 Splicing Regulator RBFOX3 (NeuN) Regulates Adult Hippocampal
998 Neurogenesis and Synaptogenesis. *PLoS One*, **11**, e0164164.
999 47. Chiu, S.-L., Diering, G.H., Ye, B., Takamiya, K., Chen, C.-M., Jiang, Y.,
1000 Niranjan, T., Schwartz, C.E., Wang, T. and Huganir, R.L. (2017)
1001 GRASP1 regulates synaptic plasticity and learning through
1002 endosomal recycling of AMPA receptors. *Neuron*, **93**, 1405-1419.
1003 e1408.
1004 48. Zhong, L. and Gerges, N.Z. (2010) Neurogranin and synaptic
1005 plasticity balance. *Commun Integr Biol*, **3**, 340-342.
1006 49. Palmer, C.L., Lim, W., Hastie, P.G., Toward, M., Korolchuk, V.I.,
1007 Burbidge, S.A., Banting, G., Collingridge, G.L., Isaac, J.T. and Henley,
1008 J.M. (2005) Hippocalcin functions as a calcium sensor in
1009 hippocampal LTD. *Neuron*, **47**, 487-494.

1010 50. Neves, G., Cooke, S.F. and Bliss, T.V. (2008) Synaptic plasticity,
1011 memory and the hippocampus: a neural network approach to
1012 causality. *Nat Rev Neurosci*, **9**, 65-75.
1013 51. Cang, Z., Zhao, Y., Almet, A.A., Stabell, A., Ramos, R., Plikus, M.V.,
1014 Atwood, S.X. and Nie, Q. (2023) Screening cell-cell communication
1015 in spatial transcriptomics via collective optimal transport. *Nat*
1016 *Methods*, **20**, 218-228.
1017 52. Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan,
1018 C.H., Myung, P., Plikus, M.V. and Nie, Q. (2021) Inference and
1019 analysis of cell-cell communication using CellChat. *Nat Commun*,
1020 **12**, 1088.
1021 53. Efremova, M., Vento-Tormo, M., Teichmann, S.A. and Vento-
1022 Tormo, R. (2020) CellPhoneDB: inferring cell-cell communication
1023 from combined expression of multi-subunit ligand-receptor
1024 complexes. *Nat Protoc*, **15**, 1484-1506.

1025
1026