STHD: probabilistic cell typing of single Spots in whole Transcriptome spatial data with High Definition

Chuhanwen Sun[1]*, Yi Zhang[1-4*#]

1. Department of Neurosurgery, Duke University

2. Department of Biostatistics and Bioinformatics, Duke University

3. Brain Tumor Omics Program, The Preston Robert Tisch Brain Tumor Center, Duke University

4. Duke Cancer Institute

*These authors contributed equally.

#Corresponding author: Yi Zhang, Ph.D. yi.zhang@duke.edu

Abstract

Recent spatial transcriptomics (ST) technologies have enabled sub-single-cell resolution profiling of gene expression across the whole transcriptome. However, the transition to high-definition ST significantly increased sparsity and dimensionality, posing computational challenges in discerning cell identities, understanding neighborhood structure, and identifying differential expression - all are crucial steps to study normal and disease ST samples. Here we present STHD, a novel machine learning method for probabilistic cell typing of single spots in whole-transcriptome, high-resolution ST data. Unlike current binning-aggregation-deconvolution strategy, STHD directly models gene expression at single-spot level to infer cell type identities. It addresses sparsity by modeling count statistics, incorporating neighbor similarities, and leveraging reference single-cell RNA-seq data. We demonstrated that STHD accurately predicts cell type identities at single-spot level, which automatically achieved precise segmentation of global tissue architecture and local multicellular neighborhoods. The STHD labels facilitated various downstream analyses, including cell type-stratified bin aggregation, spatial compositional comparison, and cell type-specific differential expression analyses. These high-resolution labels further defined frontlines of inter-cell type interactions, revealing direct cell-cell communication activities at immune hubs of a colon cancer sample. Overall, computational

modeling of high-resolution spots with STHD uncovers precise spatial organization and deeper biological insights for disease mechanisms.

Main

Spatial transcriptomics (ST) technologies have enabled gene expression profiling in the original spatial context of tissues. It provides a systematic approach to profile cell type distribution, regional variations, and cell-cell communications in normal and disease samples. Current ST technologies are based on next-generation sequencing or in situ hybridization imaging, in short, seq-ST and image-ST[1]. Often, seq-ST approaches can cover whole-transcriptome with compromised resolution or depth (e.g. 10X Visium, Slide-seq[2], Stereo-seq[3]), whereas image-based approaches can provide sub-cellular resolution for genes in a designed panel (e.g. 10X Xenium, STARmap[4], MERFISH[5]).

Recent ST technologies have rapidly advanced with image-ST increasing gene coverage and seq-ST enhancing resolution to subcellular scale. For example, VisiumHD from 10X Genomics recently become available, capable of whole transcriptome profiling in spots of size 2x2um, which is sub-cellular for many cell types and tissue contexts. However, analyses of this data type present significant computational challenges. The read coverage in each HD spot is often highly sparse and usually requires binning of adjacent spots at a chosen size (e.g. aggregating 4x4 spots into 8x8um bins, single-cell size). The binning process then computationally merges spots for sufficient read depth in downstream analytical tasks[6]. However, binning compromises resolution and mixes cell types, which further requires cell type deconvolution as in low-resolution seq-ST (using e.g. RCTD[7], CARD[8], Cell2Location[9]). Additionally, the vast number of spots or bins in high-resolution ST data poses challenges in computational efficiency.

We here develop STHD (probabilistic cell typing of single Spots in whole Transcriptome spatial data with High Definition), a machine learning method for cell typing of individual high-resolution spots in whole-transcriptome spatial data. The STHD model leverages cell type-specific gene expression from reference single-cell RNA-seq data, constructs a statistical model on spot gene counts, and employs regularization from neighbor similarity. The STHD model outputs cell type labels and probabilities for each single spot, which addresses the compromised resolution resulted from the current binning approach. Our results demonstrate that STHD labels can further stratify cell types in bin aggregation, which facilitates cell type-specific analyses like compositional comparison and differential expression analyses. Additionally, STHD identifies frontlines of inter-cell type interactions to enable direct cell-cell communication analysis.

An overview of STHD model is illustrated in **Fig.1a**. STHD infers latent cell type identities of each spot using a loss function that simultaneously optimizes two components: likelihood of spot gene counts and similarity with neighbor spots (**Fig.1a**, **Methods**). The first part optimizes log likelihood of spot gene counts following Poisson distribution, where parameters are based on normalized gene expression derived from a reference scRNA-seq dataset with cell type annotation from a matched disease or system. The second part denoises sparsity using cross entropy loss based on neighborhood similarity in cell type probabilities. The contribution from neighbors is controlled by tuning the neighbor parameter β. For each high-resolution spot, STHD outputs cell type probabilities and labels based on Maximum a Posterior (MAP). The spot cell types enable multiple downstream analytical tasks, including cell type-stratified bin aggregation, cell type-specific compositional and differential gene expression analyses, and cell-cell interaction analyses. STHD implements fast optimization enabled by efficient gradient descent[10,11]. STHD also includes local low-count detection to filter cell and tissue gaps (**Methods, Supplementary Fig.1**). To be compatible with the vast number of spots in one sample, we parallelized patch-level inference followed by whole-sample integration

(**Supplementary Fig.2**). Finally, to enable close examination of cellular neighborhoods, we developed STHDviewer for interactive and scalable visualization of spots in the entire sample (**Fig.2**).

Different from current standards that start from binning, we propose that analyses of whole-transcriptome, high-resolution ST data can begin with modeling the subcellular HD spots as STHD. This strategy offers advantages compared to alternative approaches. First, direct unsupervised clustering of spot-level gene expression results in false grouping of spots due to extreme sparsity (**Fig.1b**, left). Second, the current bin-aggregation approach can increase depth, but it mixes cell types and compromise resolution (**Fig.1b**, middle). Lastly, segmenting cell areas based on the hematoxylin and eosin (H&E) histological images may provide precise cell boundaries, but whole-cell segmentation has remained a challenging task[12](**Fig.1b**, right).

We demonstrated STHD on the recent available VisiumHD sample from 10X Genomics, where a 6.5x6.5mm of human colorectal tumor tissue was profiled by 8,726,600 HD spots covering 18,085 genes. We constructed the normalized gene expression reference from an independent human colon cancer atlas study that profiled 43,113 genes from 370,115 single cells annotated into 98 fine-grained cell types from GSE178341[13] (**Supplementary Fig.3**). The reference gene profile construction step of STHD selected 4618 cell type-informative genes (**Methods, Supplementary Fig.3**). **Fig.1c** illustrated STHD label for a patch in the colorectal cancer sample containing 22,336 spots in a 300x300um area or 1100x1100 pixels in the full-resolution H&E image. First, we observed that STHD-predicted cell type labels collectively captured the compartmentalization of colon crypt-like structure with tumor-like epithelial cell type signatures (**Fig.1c**). In detail, stem-like proliferating cells were positioned inside, circled by the abundant enterocyte cells with scattered tumor stem and transit amplifying (TA)-like cells[13,14]. The intercrypt regions were composed of abundant endothelial and fibroblast cells, with diverse

immune cell populations such as plasma cells, monocytes, macrophages, dendritic cells, and T cells (**Fig.1c**). Interestingly, the cell type probabilities displayed high confidence for abundant epithelial spots as well as for rarer cells like macrophages and plasma IgG B cells (**Fig.1d**, **Extended Fig.1**). By default, spots with maximum posterior probability below 0.8 were assigned as ambiguous, which mainly located at boundaries of different cell types and tissue structures, which could overlap cell mixtures or cell gaps (**Fig.1c, d**). In comparison, firstly, clustering of spot-level gene expression vaguely reflected colon crypt structure with mingled cell identities (**Fig.1e, Supplementary Fig.4a**). Secondly, clustering of gene expression of bins aggregated from 4x4 spots (8x8um bins) within patch resulted in less noise in crypt center and intercrypt regions (**Fig.1f** left**, Supplementary Fig.4b**). Similar bin clustering for the entire sample could clearly group the stemness center, enterocytes, and intercrypt regions, but the boundaries are assigned as one separated cluster that mixed intercrypt immune cells (**Fig.1f** right). Further deconvolution on the aggregated bins using RCTD separated epithelial structure and immune cells but lost spot-level resolution (**Fig.1g**). Finally, the H&E image contains densely organized cells where sensitive whole-cell segmentation has been challenging (**Fig.1h**). In contrast, STHD-guided binning aggregated spots of the same cell type and separated cell types from the same bin areas; thus, tumor epithelial cells and intercrypt endothelial, fibroblasts and immune cells were clearly separated on the STHD-guided binning spatial maps (**Fig.1i, Supplementary Fig.4c, d**).

Since high-resolution ST data lacks ground truth of spot labels, we evaluated performance from aspects of histopathology, comparison to other methods, simulation, and marker gene expression. First, the spatial distribution of STHD cell type labels aligns with cell organization morphology based on H&E staining (**Fig.1h**). Second, we compared to cell type deconvolution on 8x8um bins by RCTD doublet mode or full mode using the same reference human colon cancer scRNA-seq reference (**Methods**). RCTD also resolved cell type proportions that match

the colon crypt organization and the distribution of STHD labels for enterocytes and Stem/TA-like proliferating cells (**Fig.1g**, **Extended Fig.2a,b**). At the raw 2x2um spot level, we next tested computational methods that were able to label cell types by reference-based deconvolution like RCTD[7] and CARD[8], Cell2Location[9], Celloscope[15], Stereoscope[16], or by neighborhood-augmented clustering like Banksy[17]. We noted that the common gene count filters in these tools need to be removed to avoid filtering of most spots; for example, 20436 of the 22336 spots are removed by RCTD default UMI cutoff 100 (**Extended Fig.2c**). At single-spot level, RCTD, CARD, Cell2Location, and Stereoscope can vaguely reveal the crypt structure while with highly mixed cell types from different lineages (**Extended Fig.3a-e**). The unsupervised Banksy efficiently separated epithelial crypt and intercrypt endothelial structures, but mixed the rarer immune cells inside the endothelial cluster and called the boundary as a separate cluster (**Extended Fig.3f**). Moreover, STHD has highest computational efficiency among all methods that enables whole-sample inference (**Extended Fig.4**).

To quantitatively assess performance of STHD, we simulated a comprehensive *in silico* whole-transcriptome, high-resolution ST data (**Methods, Fig.1j, Supplementary Fig.5**). As shown in **Fig.1j**, we simulated 500 cells with heterogeneous cell sizes and cell types on 22,500 spots, each with a true cell type label and gene counts following expression profile from the same human colon cancer reference. STHD achieved average area under the curve (AUC) 0.9977 with overall accuracy 93.97% for this multi-class classification task with 98 categories (**Fig.1j, Supplementary Fig.5**). Next, we systematically investigated cell type-specific expression of marker genes based on STHD prediction across the entire colon cancer sample. Despite spot sparsity (**Supplementary Fig.4e, Supplementary Fig.3d**), STHD labels displayed similar cell type-specific expression patterns of marker genes as in scRNA-seq reference data in normal epithelial cells (**Fig.1k**, left two panels, **Supplementary Fig.7a**) and all other cell lineages: tumor epithelial cells (**Supplementary Fig.7b**), myeloid cells (**Supplementary Fig.8a**), B and

plasma cells (**Supplementary Fig.8b**), CD8T, CD4T, and other T cells (**Supplementary Fig.9, Supplementary Fig.10a**), fibroblasts (**Supplementary Fig.10b**), and endothelial cells and pericytes (**Supplementary Fig.11**). STHD-guided binning effectively increased marker gene coverage and meanwhile retained cell-type specificity, shown in two example bin sizes: 8x8um bins (**Fig. 1k**, third panel, **Supplementary Figs.7-11**) and 16x16um bins (**Fig.1k**, fourth panel, **Supplementary Figs.7-11**).

STHD enabled scalable inference across the entire ST sample through parallelized inference of spatial patches followed by whole-sample integration (**Supplementary Fig.2**). To facilitate close examination of spot identities and scalable investigation of cellular neighborhoods, we developed STHDviewer, an interactive and scalable visualization toolkit capable of rendering millions of spots in a webpage (**Fig.2a, Extended Fig.5**). As demonstrated in **Fig.2a**, STHD prediction automatically achieved precise segmentation of global tissue architecture and local multicellular neighborhoods. Specifically, fibroblasts separated regions of normal-like area with colon gland morphology (named Epi-Normal-like, **Extended Fig.5, Extended Fig.6**) and tumor epithelial areas at the margins (Epi-Tumor-left, Epi-Tumor-top, Epi-Tumor-right) compared to tumor in deposit (Epi-Tumor-inside)[18,19] (**Fig.2b**). Across tumor regions of interest (ROI), STHD labels delineated compositional heterogeneity of cell lineages and types (**Fig.2c**). Notably, spatial heterogeneity of immune cell populations was also observed; for instance, the Epi-Tumor-inside region contained macrophages[20,21] but few T cells compared to tumor regions at the margin, indicating distinct “hot” verses “cold” T cell infiltration within the same sample (**Fig.2c**, top, **Extended Fig.7**).

Besides cell type compositional comparisons, STHD facilitates genome-wide and cell type-specific differential expression (DE) analyses across spatial regions. Comparing different tumor ROI, epithelial cells in Epi-Tumor-inside expressed highest stemness gene LGR5[22], whereas

182 epithelial cells in Epi-Tumor-top highly express CCL20[23] and enrich pathways like TNF-alpha
183 signaling (**Fig.2e, Supplementary Fig.12, Extended Fig.8**). Since the Epi-Tumor-inside region
184 enrich macrophages, we compared differentially expressed genes specific to macrophage spots
185 and observed that macrophage infiltrating inside tumor expressed higher extracellular matrix
186 gene SPP1 and lower metalloelastase MMP12, cytokine CXCL8, and antigen presentation gene
187 B2M, indicating that macrophage states segregated spatially in the same tumor[21,24,25](**Fig.2f**,
188 **Supplementary Fig.13**). The cell type-specific DE analyses can also be performed on the level
189 of STHD-guided bins with higher depth, which resulted in a higher number of significant DE
190 genes with well correlated effects (**Fig.2f**, **Extended Fig.8, Supplementary Fig.13**).
191
192 We next demonstrated STHD in analyzing the less abundant immune cells after using
193 STHDviewer to examine immune cell hubs at different locations in the colon cancer sample. An
194 advantage of STHD is that rarer cell types and fine-grained cellular neighborhoods are
195 maintained on top of the denoising effect for those abundant cell types forming continuous
196 tissue architecture; an example of immune-endothelial cell interactions was demonstrated in
197 **Extended Fig.9**. **Fig.2g** highlighted two T cell-rich areas marked in **Fig.2b**. The Immune-Trich-
198 left region contained CD4T and CD8T cells with abundant macrophages and C1Q-high dendritic
199 cells (DC)[26], while the Immune-Trich-right area contained CD4T, CD8T cells, mature DCs
200 enriched in immunoregulatory molecules (mregDC)[27], and Plasma IgG B cells (**Fig.2h**). In
201 neighborhood enrichment analyses, each area displayed unique co-occurrence patterns among
202 immune cell types; for example, T cells in Immune-Trich-left area frequently interacted with
203 macrophage and DCs, while T cells in Immune-Trich-right area co-occurred with DCs (**Fig.2g,i**).
204 Utilizing cell type-specific DE analysis, we observed that T cells in the Immune-Trich-left region
205 expresses higher cytotoxicity genes like GZMB comparing to T cells in the Immune-Trich-right
206 region, which can be associated with the immunosuppressive function of mregDCs in tumor[28,29]
207 (**Extended Fig.10**).

Cell-cell communications between cell types in disease microenvironment could drive disease progression like tumorigenesis[30,31]. Molecular interactions between cells types in single-cell transcriptomics analyses are traditionally inferred by evaluating averaged expression of communicating gene pairs, which lacks support on direct and physical interactions[32]. With the sub-cellular resolution spots labeled, we observed that STHD could pinpoint the precise frontlines of inter-cell type interaction. As shown in **Fig.2j**, we highlighted the frontlines of T cell-macrophage interaction and T cell-DC cell interaction in each of the T cell-rich regions. We hypothesized that cell-cell interactions near-frontline could be stronger than that in far-frontline spots. For a given cell type pair, we identified spots at boundary of two cell types based on posterior probabilities and considered STHD-labeled spots within 4 spots (comparable to single cell size) in distance as near-frontline spots (**Methods**). We calculated ligand-receptor interaction scores (LRIS)[21,22] between the near-frontline spots and compared with background interaction scores from far-frontline spots, conditioning on the same cell type pair. As shown in **Fig.2k**, near-frontline spots (grouped as 3,4) harbored higher LRIS activities compared to the far-frontline spots (grouped as 1,2) in both cell type pairs of T-Macrophage (**Fig.2j,k** top) and T-DC interactions (**Fig.2j,k** bottom). Specifically, the T-Macrophage communications in Immune-Trich-left were largely driven by CD74-IFNG interaction, while the T-DC communications in Immune-Trich-right were driven by CCL17/CCL19-CCR7, which indicated the immunoregulatory effect of mature DCs on T cells[33].

In summary, we present STHD, a method for cell typing of individual spots for genome-wide spatial transcriptomics with high resolution. STHD addresses the computational challenge caused by data sparsity and proposes a fundamental strategy to enhance resolution compared to current standards. We demonstrated the accuracy, precision, and efficiency of STHD and its usability in multiple downstream analyses for global disease architecture, local cellular

neighborhoods, cell type-specific differential expression, cell type-augmented binning, and cell-cell interaction studies. STHD and STHDviewer provided efficient modeling, analyses, and examination for the entire sample. STHD allows incorporation of external tissue gap masks, better reference cell type signatures, and additional cell mixture deconvolution of ambiguous spots. Overall, we propose that modeling of individual high-resolution spots by borrowing neighborhood information can computationally eliminate the dilemma in resolution-throughput tradeoff for emerging ST technologies, which enables genome-wide and high-resolution discoveries in any tissues and diseases of intricate cellular heterogeneity.

Data Availability

The reference scRNA-seq atlas data and cell type annotation from Human Colon Cancer Atlas was downloaded from Pelka et. al through GSE178341: (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178341). The reference data is also available through https://portals.broadinstitute.org/crc-immune-hubs/ and Broad Institute Single Cell Portal (Human Colon Cancer Atlas c295) at https://singlecell.broadinstitute.org/single_cell/study/SCP1162/human-colon-cancer-atlas-c295. We downloaded the publicly available VisiumHD data demonstrated at https://www.10xgenomics.com/support/software/loupe-browser/latest/tutorials/assay-analysis/lb-hd-spatial-gene-expression from https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-human-crc, named Visium HD Spatial Gene Expression Library, Human Colorectal Cancer (FFPE), by Space Ranger 3.0.0, 10x Genomics (2024, Mar 25).

Code Availability

The STHD software was implemented in Python and available on GitHub at https://github.com/yi-zhang/STHD. An interactive STHDviewer of the high-resolution cell type

result for the human colon cancer sample is available at https://yi-zhang-compbio-lab.github.io/STHDviewer_colon_cancer_hd/STHDviewer_crchd.html.

Acknowledgements

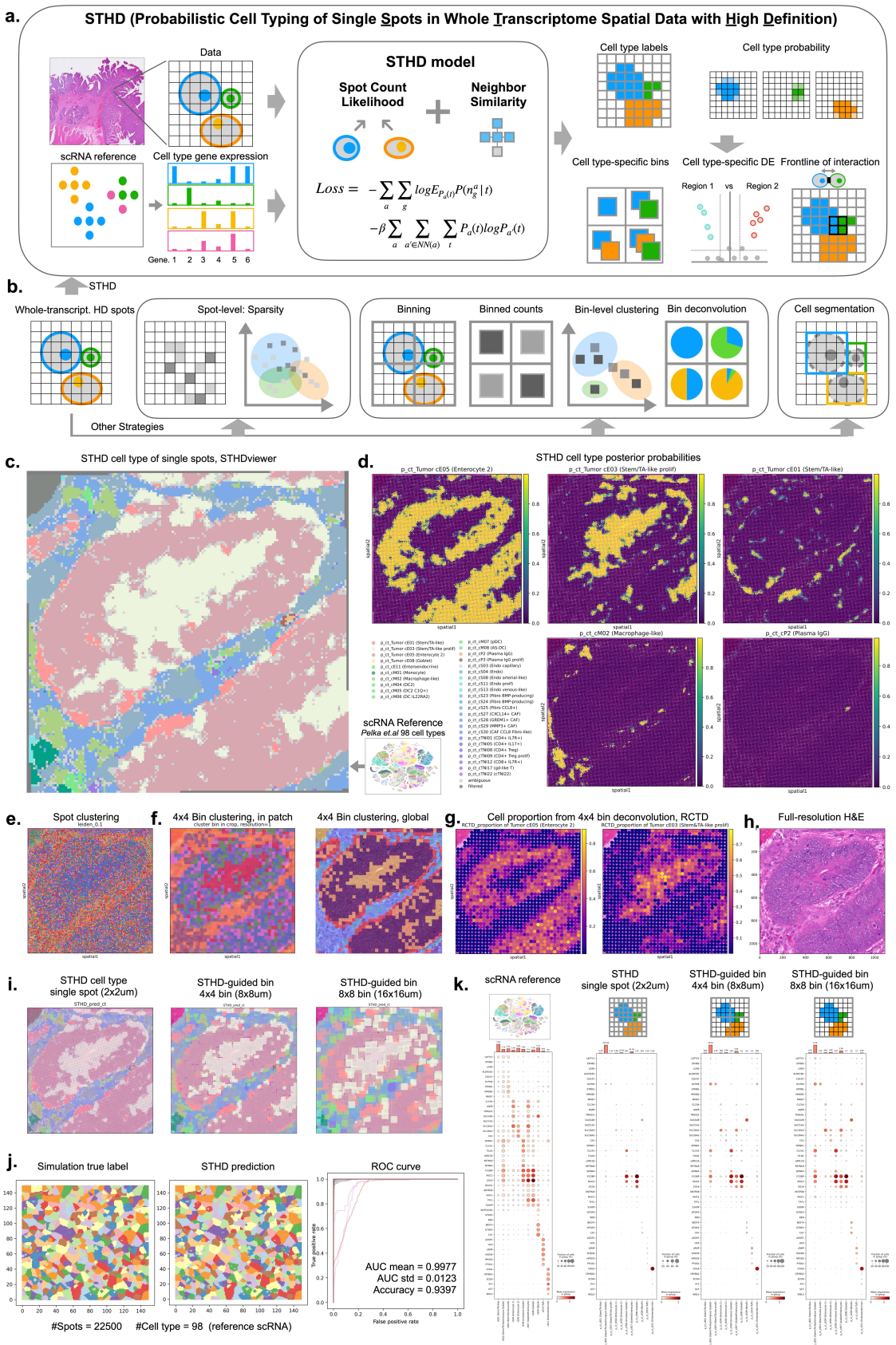
C.S. and Y.Z. was supported by start-up funds to Y.Z. from Brain Tumor Omics Program, Preston Robert Tisch Brain Tumor Center, Department of Neurosurgery and Department of Biostatistics and Bioinformatics at Duke University School of Medicine. We acknowledge 10X Genomics to share VisiumHD datasets publicly available. We acknowledge authors of the Human Colon Cancer Atlas to make scRNAseq datasets publicly available.

Author Information

This study was conceived and led by Y.Z. Y.Z. designed the model and algorithm, implemented the STHD software, led the data analyses, and wrote the manuscript. C.S. analyzed the data and helped with manuscript writing.

Figures

276 **Fig.1**

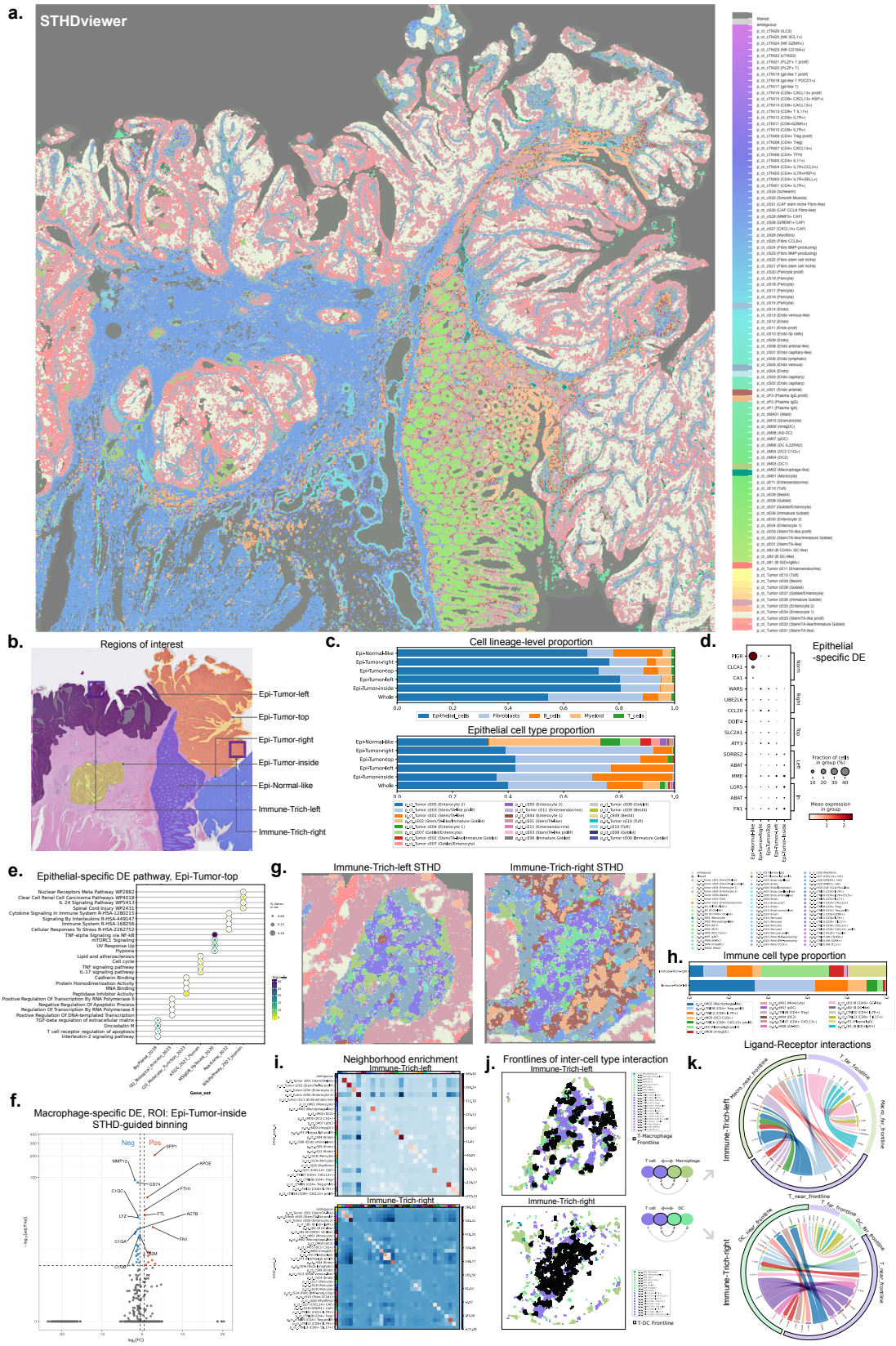


277

Fig.1. Summary of STHD model. a. STHD, (Probabilistic Cell Typing of Single Spots in Whole Transcriptome Spatial Data with High Definition) takes two inputs: whole-transcriptome spatial data of high resolution, and gene expression profiles by cell type from reference single-cell dataset. The STHD model is a machine learning method that infers latent cell type identity of individual spots by simultaneously modeling gene counts and neighbor similarity. STHD outputs cell type identities for each single spot in formats of cell type labels and posterior probabilities. The spot-level cell identities can then be used for multiple downstream analyses, including STHD-guided binning with cell type stratification, spatial comparison of cell type-specific compositional and differential expression analyses, and cell-cell interaction frontline analyses. **b.** Alternative approaches to analyze the HD datatype. Left, direct unsupervised clustering of spot-level gene expression is affected by sparsity. Middle, current binning-deconvolution approach increases coverage but compromise resolution. Right, ideal cell-level aggregation based on segmentation. **c.** Example of STHD results of a patch from the human colon cancer VisiumHD sample, revealing tumor-like epithelial cells and intercryptic immune population, visualized using the interactive STHDviewer. **d.** Cell type posterior probabilities at spot-level for the same patch, visualized using Squidpy. Top row, three epithelial cell types; Bottom row, two immune cell types. **e.** A spatial plot of unsupervised leiden clustering on spot gene expression. **f.** Spatial plots of unsupervised leiden clustering of gene expression aggregated by bins in size of 4x4 spots. Left, clustering bins within patch, right, global cluster group of bins. **g.** Cell type proportion decomposed by RCTD on bins of size of 4x4 spots, demonstrating two tumor epithelial classes, enterocytes, and stem-like cells. **h.** The H&E histopathology image in full resolution of the same crop. **i.** The STHD predicted labels at spot-level and STHD-guided binning in bins of 4x4 spots and 8x8 spots, visualized using Squidpy. **j.** Simulated high-resolution spatial data. Left, ground truth cell type labels; middle, STHD predicted cell type labels; right, receiver operating characteristic curve (ROC) curve for all cell types. **k.** Expression dot plots for marker genes for normal epithelial cells. Left to right: normal epithelial cell types in reference human colon cancer

304 sample, normal epithelial cell types based on STHD predicted spots, normal epithelial cell types
305 from STHD-guided bins of size 4x4 spots, normal epithelial cell types in STHD-guided bins of
306 size 8x8 spots. AUC, area under the curve.
307

308 **Fig.2.**



309

Fig.2. STHD prediction facilitates segmentation of global structures and local neighborhoods. a. Interactive spatial visualization by STHDviewer of predicted cell type labels for spots in the entire human colon cancer VisiumHD sample. **b.** Annotation of regions of interest, separating epithelial regions and two T cell-infiltrated regions. **c.** Cell type proportion bar plots at cell lineage and cell type level for epithelial regions and whole map. Upper: Proportions of major cell lineages including epithelial cells, fibroblasts, B cells, myeloid cells, and T cells. Bottom: Proportions of finer cell types of epithelial lineage. **d.** Spot-level gene expression for epithelial-specific differential genes across regions. Top three genes by log fold change were shown. **e.** Pathway enrichment of epithelial cell-specific genes differentially expressed in the Epi-Tumor-top region. **f.** Macrophage-specific differential genes for the Epi-Tumor-inside region, where differential analyses were performed at bin level of size 4x4 spots with cell type stratification from STHD. **g.** STHD cell type labels for two T cell-rich regions marked in panel b. **h.** Proportions of immune cell types in the two immune regions. **i.** Cell neighborhood interaction patterns of the two T cell-rich regions showing neighborhood enrichment z-scores where spots within 6 spots are considered neighbors. **j.** T cell and myeloid cells in the two immune-rich regions, with frontlines of inter-cell type interactions. Top, T cell-macrophage frontline in Immune-Trich-left region; Bottom, T cell-dendritic cell frontline in Immune-Trich-right region. Considering a pair of cell types, spots within 4 spots (8um) from the inter-cell type interaction location are highlighted as near-frontline spots, with the rest as far-frontline spots. **k.** The ligand-receptor communications at near-frontline spots compared to far-frontline communication scores of the same cell type pair as background, where the ligand-receptor interaction scores were calculated by squidpy.ligrec. Top 10 ligand-receptor pairs different between near-frontline and far-frontline communications were visualized where width represents average expression level of ligand receptor pairs. Spots are classified into four groups 1,2,3,4 based on their cell type and near- or far-frontline categories.

Methods

STHD algorithm

The STHD model is a machine learning method that combine modeling of per-spot gene counts based on latent cell type identities with regularizing similarity among neighboring spots in terms of cell type probability distribution. The overall loss function $\mathcal{L}$ composes of negative log likelihood $-LL$ that models gene count of all spots and the cross entropy loss CE measuring similarity of neighboring spots. The amount of contribution from neighbors is controlled by the parameter β. The cell type identities at every single spot are thus estimated by Expectation Maximization.

$$\mathcal{L} = -LL + \beta CE$$

For the negative log likelihood part, we model counts of each gene at each spot to follow Poisson distribution. Let the normalized average gene expression level for gene g within the latent cell type t is λ_g^t, read depth of spot a is d_a, and spot a's probability of belonging to cell type t is $P_a(t)$. Thus, the count of gene g at spot a: $n_g^a \sim Poission(d_a\lambda_g^t)$.

$$LL = log\prod_a\prod_g P(n_g^a)$$

$$= \sum_a\sum_g logP(n_g^a)$$

$$= \sum_a\sum_g log\, E_{P_a(t)}P(n_g^a|t)$$

$$\geq \sum_a\sum_g E_{P_a(t)}\, logP(n_g^a|t)$$

$$= \sum_a\sum_t P_a(t)\sum_g (n_g^a log(d_a\lambda_g^t) - d_a\lambda_g^t) + const$$

Considering high sparsity of single spots, the second part of the loss function is related to cross entropy loss between spot a and nearest neighboring spots $a' \in NN(a)$ in terms of distribution of cell type probability.

$$CE = -\sum_{a}\sum_{a' \in NN(a)}\sum_{t}\left(P_a(t) log P_{a'}(t)\right)$$

Note that one constraint is $\sum_t P_a(t) = 1$. In calculation, we use the following for inherent incorporation of constraint: $P_a(t) := \frac{e^{w_a^t}}{\sum_{s \in T} e^{w_a^s}}$, where T is the full set of cell types. We thus optimize $P_a(t)$ or w_a^t to minimize the total loss:

$$Loss = -\sum_{a}\sum_{t} P_a(t) \sum_{g}\left(n_g^a log\left(d_a \lambda_g^t\right) - d_a \lambda_g^t\right) - \beta \sum_{a}\sum_{a' \in NN(a)}\sum_{t}\left(P_a(t) log P_{a'}(t)\right)$$

Construct normalized gene expression profile from reference single-cell atlas dataset

The average normalized gene expression λ_g^t can be estimated from a pre-defined single-cell RNA-seq reference dataset. A reference single-cell dataset shall match the tissue or disease origin of the high resolution, whole-transcriptome spatial transcriptomics data. It shall also contain cell type annotation that is usually based on multi-sample harmonization and unsupervised clustering of single cells into various cell type or groups. The granularity of cell type annotation directly influence the number of predicted cell type and downstream analyses. For each cell type, average normalized gene expression profiles are estimated with $\lambda_g^t = \frac{n_g^t}{\sum_g n_g^t}$ so that $\sum_g \lambda_g^t = 1$. Specifically for human colon cancer, we downloaded the raw counts file of the single-cell RNA-seq data from the human colon cancer atlas from Pelka et.al[13] through GSE178341. The data contains 370115 tumor and adjacent normal single cells from 62 primary treatment-naïve colorectal cancer patients, with 43113 genes profiled. Quality control of single cells was performed by filtering cells with minimum 100 counts and 100 genes, filtering genes with minimum 3 cells, and removing cells with mitochondrial read ratio larger than 27.7%, which is mean plus one standard deviation of the mitochondrial read ratio across all cells. After the above quality control, 303358 cells and 31,873 genes remain. The cell annotations were obtained from the study that provided 98 fine-grained cell types and 7 cell lineages. We used

the 98 high-granularity cell type annotation throughout this study. To select genes informative in distinguishing the 98 cell types, we followed the calculation of cell type expression profile as in RCTD[7]. First, we calculated the estimated cell type expression profiles λ_g^t for all genes and cell types. We then selected genes with average expression profile above 0.000125, which is 0.0625 counts per 500, and at least 0.5 log-fold-change compared to the average expression across all cell types, using log2 scale. Moreover, we removed the mitochondrial genes and ribosome genes in the selection. The selection resulted in 4618 genes that are cell type-informative, for each we calculated λ_g^t in all cell types. Note that STHD uses the 4618 genes to infer cell type identity of spots, and later can consider whole-transcriptome for downstream analyses.

STHD hyperparameter tuning

To select hyperparameters that balance between statistical likelihood and neighborhood similarity, we comprehensively tested hyperparameter tuning and provided an automatic stopping criterion and optional tuning tutorial. The main parameters in STHD includes the neighborhood parameter β, which controls the contribution of cross entropy loss among local neighbors to the total loss, the optimization steps n_iteration, and the learning rate that can be fixed in practice. We provided a tutorial on parameter tuning based on performance in random sampled patches; for the human colon cancer sample, the optimized parameters are 23 iteration steps and β=0.1. In detail, for the colon cancer sample, we randomly selected six different patches in size of 1100x1100 pixels (in full resolution, around 22500 spots). For the optimization of each patch, we tested a range of β from 0, 0.03, 0.1, 0.3, 1, and 3, each outputting optimal iteration steps based on the early stopping criterion, where the drop of loss in the past 10 steps falls within 1%. Since the regularization term from neighborhood contribution β will lower log likelihood, as β increases, we select the optimal parameter pairs where drop in cross entropy

406 slows down to around 10% while a high log likelihood is maintained. The optimal iteration steps
407 when fixing β of 0, 0.03, 0.1, 0.3, 1, 3, is each 24, 23, 22, 23, 23, 24. The optimal iteration steps
408 is robustly 23.17 on average. Therefore, we chose β=0.1 and n_iteration=23 as optimal
409 parameters for the whole sample modeling that balance the count-based likelihood and the
410 contribution from neighbors.
411
412 **STHD low-count region detection and filter**
413 Throughout the spatial transcriptomics data, regions of low gene counts exist, usually arising
414 from tissue boundaries, tissue gaps, and regions low in DNA/RNA such as dead cells, necrosis,
415 vessel enriching red blood cells, and collagen rich area. We included an automatic local low-
416 count region detection and filtering functionality in STHD. By default, STHD computed neighbor
417 graphs within 2 rings of spots (13 spots for the grid system), aggregated local connected spots,
418 and detected low-count local regions not exceeding 50 counts, where the model will skip
419 modeling due to low confidence in cell type inference. Other customized tissue gap and
420 boundary masks can also be provided to further filter spots.
421
422 **Assessing model accuracy with simulated spatial transcriptomics data**
423 To assess the accuracy of STHD model prediction, we simulated the high-dimensional, high-
424 resolution spatial transcriptomics data *in silico*. We simulated a region with 22500 (150x150)
425 grid spots and position 500 simulated cells overlapping the region with cell sizes spanning 1-20
426 spots. Each cell is randomly assigned a cell type from the 98 cell types annotated in colon
427 cancer atlas, and spots overlapping each whole cell are assigned the corresponding cell type.
428 To mimic the real-world scenario in gene expression profiles and spot sparsity, we adapted the
429 same normalized gene expression profile from the same colon cancer atlas data with 98 cell
430 types. The depth of each spot was simulated following the distribution of exponential decay with
431 mean read depth 47 matching the colon cancer VisiumHD sample. After simulating counts of

each gene at each spot in the region, we applied STHD with default optimal parameters (β=0.1) and also STHD no-neighbor model (β=0) to the simulated spatial data, and measured prediction accuracy, standard error, and area under the receiver operating characteristic curve (ROC), for this task of multi-class classification of 98 categories at each spot.

Comparison with other cell typing methods

To benchmark the STHD results at 2x2 μm resolution, we compared STHD to six methods that can output cell typing, including the deconvolution-based cell typing tools RCTD[7], CARD[8], Cell2Location[9], Celloscope[15], Stereoscope[16], and the unsupervised method Banksy[17]. Because of computational efficiency and data size limit of the existing methods, we used the tumor region as in Fig.1 to perform benchmarking, which contains 22,336 cells and 18,085 genes. For methods requiring the reference scRNA-seq data as input, we used the same human colorectal cancer scRNA-seq datasets (GSE178341) with 98 cell types. Since many methods in comparison requires a maximum number of cells from reference scRNAseq dataset – for example, CARD requires below 50000 cells - we down-sampled the reference data to include maximum 600 cells for each cell type, resulting in a total of 49,351 cells and 41,195 genes. For running RCTD on single spots, we used RCTD doublet mode, testing both the default minimum UMI 100 and also by removing the minimum UMI requirement due to loss of most raw spots. For testing CARD, we followed guidelines from https://github.com/YMa-lab/CARD. Similarly, to avoid most spots being filtered, the minimum spot count and minimum gene numbers were set to 0. Cell2location and Stereoscope were ran with default parameters according to the tutorials https://cell2location.readthedocs.io/en/latest/notebooks/cell2location_tutorial.html, and https://github.com/almaan/stereoscope. In Cell2location, by default, the reference data was filtered using the following parameters: cell_count_cutoff = 5, cell_percentage_cutoff = 0.03, and nonz_mean_cutoff = 1.12. The reference single-cell regression model was trained with default parameter max_epochs = 250, lr = 0.002. The spatial cell2location model was obtained with

default parameters max_epoch = 30,000. In Stereoscope, by default, the reference data was filtered using a minimum count threshold of min_counts = 10, and all mitochondrial genes were removed. The reference single-cell model was trained with parameters max_epochs = 100; the spatial stereoscope model was obtained using parameters max_epochs = 10,000. Finally, for Celloscope, a binary matrix with prior knowledge about marker genes is required among input, for which we selected top 200 genes in normalized expression from our informative gene matrix for each cell type. Another required input is estimated number of cells in each spot, for which we set to 1 due to sub-cellular-resolution nature of the data. Celloscope was then tested following guidelines at https://github.com/szczurek-lab/Celloscope. Finally, Banksy was tested with default parameters following the tutorial at https://github.com/prabhakarlab/Banksy_py/blob/main/slideseqv2_analysis.ipynb. For all toolkits, the weights or cell type abundance for each spot were extracted and normalized as cell proportions for subsequent visualizations.

Comparison with deconvoluted cell proportions in aggregated bins

Following current standards of binning-deconvolution, we binned the spatial data and aggregated the bins containing 4x4 spots. We applied RCTD deconvolution of doublet mode on the same spatial patch as in Fig.1, following the default filtering and parameters of RCTD as on spacexr (https://github.com/dmcable/spacexr), and using the same subsampled colon cancer reference scRNA-seq data as in comparison with other methods. To compare cell proportions across bins, we counted STHD cell type labels and obtained cell proportions for each 4x4 bin after excluding filtered and ambiguous spots. We also computed cosine similarity between STHD's spot count-based proportions and RCTD's deconvoluted proportions and visualize the proportions and cosine similarity on a spatial plot.

Timing computational efficiency

To assess computational efficiencies, different cell typing tools are applied onto the tumor epithelial patch using a server with 8 CPU cores and maximum 128GB memory, running time recorded. For tools supporting GPU acceleration (cell2location and stereoscope) the time cost was evaluated on a single GPU, and the spatial training time was recorded without including scRNA training time. For computational efficacy, we also estimated the time cost for the entire sample by linearly scaling running time based on number of spots in patch (22,336 spots) and in the entire sample (8,726,224 spots).

STHD for full-size ST sample

The STHD software contains several components to flexibly split the whole sample into patches for inferences and parallelization. The input/output functionalities (STHDio) expanded from Squidpy Visium interface to enable cropping of full-resolution image. Second, for the full-size sample, STHD will patchify the full sample into multiple regions with overlaps, followed by low-count filtering, per-patch inference, and result aggregation where probabilities of overlapping spots are averaged. The STHD cell type label results are determined by Maximum A Posterior (MAP) with a preset cutoff (default 0.8). An optional fast implementation for STHD-guided binning with a predefined bin size is provided, which outputs per-gene count for each bin stratified by STHD cell type labels. Finally, the predicted cell labels are exported to STHDviewer for a rasterized and interactive scatter plot implemented using Bokeh[34]. In STHDviewer, the spot grid of array_row and array_column were used for x-axis and y-axis for visualization efficiency.

Cell type-specific differential gene expression analysis across spatial regions of interest

To study regional differences within the cancer sample, we first focused on regions enriching epithelial cell types and manually separated four different regions and extracted spot barcodes. Based on STHD predicted labels, we performed differential expression analyses specific for

epithelial cell types. Gene counts were normalized towards the spot count depth (target total counts 10000) and log-transformed followed by Wilcoxon test in Scanpy. Genes passing adjusted p-value cutoff 10^{-6} and absolute log2 fold change 0.5 were selected as significant, and top genes passing log2 fold change 0.5 are exported to pathway enrichment analyses by GSEApy[35] using EnrichR and a few pathway databases, including GO Molecular Function 2023, GO Biological Process 2023, MSigDB Hallmark 2020, Reactome 2022, WikiPathway 2023 Human, KEGG 2021 Human, and BioPlanet 2019. The FDR cutoff of 0.1 was used to filter the pathways and top four pathways were visualized for tumor epithelial pathway comparison. We also tested cell type-specific differential expression using both STHD-labeled spot and STHD-guided bin aggregates of size 4x4 spots. At STHD-guided bin level, we followed the same step of normalization and performed differential expression with Wilcoxon with the same cutoff for epithelial and macrophage comparison. For T cell-specific differential expression analyses between the two T cell rich regions, the cutoff was set log2 fold change 0.5 and adjusted p-value cutoff 0.05 due to small number of spots from patches. For differential gene dotplots, significant genes are sorted by log fold change and the top three genes for the tumor comparison and top ten genes for the immune comparison were plotted.

Neighborhood interaction enrichment analysis for immune-rich regions

To investigate the neighborhood interaction patterns across various cell types in the two immune-infiltrated regions, local neighborhoods were calculated using the 4 nearest cells in the grid and considering 6 rings of adjacent spots to obtain a local neighborhood structure using squidpy.gr.spatial_neighbors in Squidpy. We used STHD predicted cell type labels and excluded spots in the filtered and ambiguous categories. The neighborhood enrichment Z-scores were calculated using squidpy.gr.nhood_enrichment, which follows CellChat[32] with ligand-receptor

databases and permutation of 1000 times. Heatmaps of neighborhood interaction patterns were plotted based on the co-occurrence score of pairwise cell types in each region of interest.

Inter-cell type frontline and cell-cell communication analysis frontline computation

For a given pair of cell types, we first defined their precise frontlines to be spots where those two cell types are adjacent to each other. We further restrict frontline spots to be those whose neighbor only contains that pair of cell types based on maximum posterior to eliminate potential noises from other cell types. Considering a single cell is comparable to the size of 4 spots, we assumed spots within 4 spots from the frontlines represented one cell with direct interaction with another cell type. Thus, spots were categorized based on their cell type and proximity to frontline. We examined interactions between T cells and Dendritic cells, and T cells and Macrophages separately for the two immune regions of interest, Immune-Trich-left and Immune-Trich-right. In the chord plots, For a pair of cell types A and B, the labels were assigned as follows: '1' for far-frontline cell type A, '2' for far-frontline cell type B, '3' for near-frontline cell type A, and '4' for near-frontline cell type B. We then hypothesized that cell-cell communications like ligand-receptor expression levels were higher between group 3 and 4 compared to group 1 and 2. To perform the ligand-receptor interaction analysis, normalized data was applied the squidpy.gr.ligrec function which permuted 1000 times to detect significant ligand-receptor pairs. Top 10 ligand-receptor pairs different between near-frontline and far-frontline communications were visualized in the chord diagram showing the mean expression of each ligand-receptor pair and sender-receiver directions.

References

1. Tian, L., Chen, F. & Macosko, E. Z. The expanding vistas of spatial transcriptomics. *Nat Biotechnol* **41**, 773–782 (2023).

2. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* **39**, 313–319 (2021).
3. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).
4. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
5. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
6. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
7. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40**, 517–526 (2022).
8. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol* **40**, 1349–1359 (2022).
9. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* **40**, 661–671 (2022).
10. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2017).
11. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* 1–6 (Association for Computing Machinery, New York, NY, USA, 2015). doi:10.1145/2833157.2833162.
12. Hartman, A. & Satija, R. Comparative analysis of multiplexed in situ gene expression profiling technologies. 2024.01.11.575135 Preprint at https://doi.org/10.1101/2024.01.11.575135 (2024).

583 13. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal
584 cancer. *Cell* **184**, 4734-4752.e20 (2021).
585 14. Merlos-Suárez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer
586 stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).
587 15. Geras, A. *et al.* Celloscope: a probabilistic model for marker-gene-driven cell type
588 deconvolution in spatial transcriptomics data. *Genome Biology* **24**, 120 (2023).
589 16. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic
590 inference of cell type topography. *Commun Biol* **3**, 1–8 (2020).
591 17. Singhal, V. *et al.* BANKSY unifies cell typing and tissue domain segmentation for
592 scalable spatial omics data analysis. *Nat Genet* 1–11 (2024) doi:10.1038/s41588-024-01664-
593 3.
594 18. Höppener, D. J. *et al.* The relationship between primary colorectal cancer histology and
595 the histopathological growth patterns of corresponding liver metastases. *BMC Cancer* **22**,
596 911 (2022).
597 19. Sirinukunwattana, K. *et al.* Gland segmentation in colon histology images: The glas
598 challenge contest. *Med Image Anal* **35**, 489–502 (2017).
599 20. Ozato, Y. *et al.* Spatial and single-cell transcriptomics decipher the cellular environment
600 containing HLA-G+ cancer cells and SPP1+ macrophages in colorectal cancer. *Cell Rep* **42**,
601 111929 (2023).
602 21. Zhang, Y. *et al.* MetaTiME integrates single-cell gene expression to characterize the
603 meta-components of the tumor immune microenvironment. *Nature Communications* **14**, 2634
604 (2023).
605 22. Xu, L., Lin, W., Wen, L. & Li, G. Lgr5 in cancer biology: functional identification of Lgr5 in
606 cancer progression and potential opportunities for novel therapy. *Stem Cell Res Ther* **10**, 219
607 (2019).

608 23. Kapur, N. *et al.* CCR6 expression in colon cancer is associated with advanced disease
609 and supports epithelial-to-mesenchymal transition. *Br J Cancer* **114**, 1343–1351 (2016).
610 24. Cheng, S. *et al.* A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid
611 cells. *Cell* **184**, 792-809.e23 (2021).
612 25. Matusiak, M. *et al.* Spatially Segregated Macrophage Populations Predict Distinct
613 Outcomes In Colon Cancer. *Cancer Discovery* (2024) doi:10.1158/2159-8290.CD-23-1300.
614 26. Baruah, P. *et al.* C1q enhances IFN-gamma production by antigen-specific T cells via
615 the CD40 costimulatory pathway on dendritic cells. *Blood* **113**, 3485–3493 (2009).
616 27. Maier, B. *et al.* A conserved dendritic-cell regulatory program limits antitumour immunity.
617 *Nature* **580**, 257–262 (2020).
618 28. Kvedaraite, E. & Ginhoux, F. Human dendritic cells in cancer. *Science Immunology* **7**,
619 eabm9409 (2022).
620 29. Puleston, D. J. *et al.* Polyamine metabolism is a central determinant of helper T cell
621 lineage fidelity. *Cell* **184**, 4186-4202.e20 (2021).
622 30. Wang, X. *et al.* In vivo CRISPR screens identify the E3 ligase Cop1 as a modulator of
623 macrophage infiltration and cancer immunotherapy target. *Cell* **184**, 5357-5374.e22 (2021).
624 31. Burdziak, C. *et al.* Epigenetic plasticity cooperates with cell-cell interactions to direct
625 pancreatic tumorigenesis. *Science* **380**, eadd5327 (2023).
626 32. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat*
627 *Commun* **12**, 1088 (2021).
628 33. Li, J. *et al.* Mature dendritic cells enriched in immunoregulatory molecules (mregDCs): A
629 novel population in the tumour microenvironment and immunotherapy target. *Clin Transl Med*
630 **13**, e1199 (2023).
631 34. Bokeh Development Team. *Bokeh: Python Library for Interactive Visualization*. (2018).
632 35. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set
633 enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).