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 single- and sub-cellular resolution profiling of gene expression across the whole transcriptome. However, the transition to high-definition ST significantly increased data sparsity and dimensionality, posing computational challenges in identifying cell types, deciphering neighborhood structure, and detecting 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 the current binning-aggregation-deconvolution strategy, STHD directly models gene expression at single-spot level to infer cell type identities without cell segmentation or spot aggregation. STHD addresses sparsity by modeling count statistics, incorporating neighbor similarities, and leveraging reference single-cell RNA-seq data. We show in VisiumHD data that STHD accurately predicts cell type identities at single-spot level, which achieves precise segmentation of both global tissue architecture and local multicellular neighborhoods. The high-resolution labels facilitate various downstream analyses, including cell type-stratified bin aggregation, spatial compositional comparisons, and cell type-specific differential expression analyses. Moreover, STHD labels further reveal

frontlines of inter-cell type interactions at immune hubs in cancer samples. STHD is scalable and generalizable across diverse samples, tissues, and diseases, facilitating genome-wide analyses in various spatial organization contexts. Overall, computational modeling of individual spots with STHD facilitates discoveries in cellular interactions and molecular mechanisms in whole-genome spatial technologies with high resolution. STHD is available at https://github.com/yi-zhang/STHD.

Keywords

Machine learning, spatial transcriptomics, high-resolution, tumor microenvironment, cell-cell interactions

Background

Spatial transcriptomics (ST) technologies have enabled gene expression profiling in the native spatial context of tissues. These technologies provide a systematic approach to profile cell type distribution, regional variations, and cellular neighborhoods across 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 cover towards 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 measurement of hundreds of genes in a designed panel (e.g. 10X Xenium, STARmap(4), MERFISH(5)). The technical tradeoff between resolution and throughput determined that seq-ST is generally suitable for hypothesis generation while image-ST for mechanism validation(6).

Emerging ST technologies have rapidly advanced with image-ST increasing gene coverage and seq-ST enhancing resolution to sub-cellular scale. Notable technologies that scale to whole-

52 transcriptome and provide single- or sub-cellular resolution include HDST(7) (2um spot), Seq-
53 Scope(8) (0.5-0.8um resolution), Stereo-seq(3) (spot size about 0.22um and center-to-center
54 distance 0.5 or 0.715um), and Open-ST(3) (center-to-center distance 0.6um), *etc*. Recently,
55 VisiumHD from 10X Genomics is further available for formalin-fixed paraffin-embedded (FFPE)
56 samples to profile whole transcriptome in spots of size 2x2um, a sub-cellular size for most cell
57 types and tissue contexts(9). Despite these technological advancements, the analysis of this
58 data type presents significant computational challenges. The read coverage in each high-
59 resolution spot is often highly sparse, and it is usually required to bin adjacent spots at a chosen
60 size. For example, aggregation of 4x4 spots into 8x8um bins is usually recommended in
61 VisiumHD sample processing. Similarly, typical Stereo-seq analyses aggregate data by binning
62 50 spots(3,10). The binning process computationally merges spots for sufficient read depth in
63 downstream analytical tasks(11). However, this approach creates bins spanning multiple single
64 cells, potentially mixing distinct cell types and thus compromising resolution. To address this,
65 computational deconvolution of cell proportions is further needed, similar to the scenario in low-
66 resolution seq-ST using tools like RCTD(12), CARD(13), Cell2Location(14), *etc*. Furthermore,
67 the vast number of million spots or bins in high-resolution ST data poses additional challenges
68 in computational efficiency.
69
70 We here developed STHD (probabilistic cell typing of single Spots in whole Transcriptome
71 spatial data with High Definition), a machine learning method designed for cell typing of
72 individual spots in high-resolution, whole-transcriptome spatial data. STHD reverses the
73 traditional approach by first inferring cell type identities directly at the raw spot level, without
74 requiring cell segmentation or bin aggregation. This is followed by cell type-specific bin
75 aggregation as needed for downstream genome-wide analyses. To tackle high dimensionality
76 and data sparsity, the STHD model employs a neighbor-augmented loss. It leverages cell type-
77 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 output includes cell type labels and probabilities for each individual spot. The STHD labels further facilitate multiple downstream analyses, including cell type-specific compositional comparison, cell type-specific differential expression, global tissue architecture segmentation, local cellular niche identification, cell type-augmented binning, and cell-cell interactions analyses. We demonstrated application of STHD to a colon cancer VisiumHD sample to uncover spatially heterogenous cancer cell expression programs and identifying immune infiltrate hubs with diverse immune functions and interactions. Moreover, we demonstrated that STHD is scalable and generalizable across different tissues, providing both local subcellular classification and global tissue segmentation.

Results

STHD efficiently models high-resolution spots using a machine learning model with neighbor regularization.

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, Supplementary File 1**). The first part optimizes log likelihood of spot gene counts following Poisson distribution with the latent cell type variable, 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 the neighbor parameter β, which can be tuned with automatic stopping criterion to achieve balance between statistical likelihood and neighborhood similarity. 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 gradient derivation, numba(15), and Adam(16) optimizer (full derivation of loss gradient in **Supplementary File 1**). STHD also includes local low-count detection to filter cell and tissue gaps (**Methods, Supplementary Fig.1**). To be compatible with the millions of spots in one sample, we built the STHD pipeline that parallelizes patch-level inference followed by whole-sample integration (**Supplementary Fig.2**). The pipeline is thus adaptive to various computing resources without requiring GPU for optimization. Finally, to enable close examination of cellular neighborhoods, we developed STHDviewer for interactive and scalable visualization of spots in the entire sample (**Supplementary Fig.2**).

Different from current standards that start from binning, we propose in STHD that the analyses of whole-transcriptome, high-resolution ST data can begin with modeling the sub-cellular HD spots. 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 nucleus boundaries, but whole-cell segmentation has remained a challenging task(17) (**Fig.1b**, right).

STHD infers cell type identities at the spot level, outperforms other inference tools, and guides cell type-specific binning.

We first demonstrated STHD using the publicly available human VisiumHD sample from 10X Genomics, where a 6.5x6.5mm area of human colorectal tumor was profiled by 8,726,600 HD spots covering 18,085 genes. We constructed the normalized gene expression reference from

130 an independent human colon cancer atlas study that profiled 43,113 genes from 370,115 single
131 cells annotated into 98 fine-grained cell types from GSE178341(18) (**Supplementary Fig.3**).
132 The reference gene profile construction step of STHD selected 4618 cell type-informative genes
133 (**Methods, Supplementary Fig.3**). **Fig.2a** illustrated STHD label for a tumor-containing patch in
134 the colorectal cancer sample containing 22,336 spots in a 300x300um area or 1100x1100 pixels
135 in the full-resolution H&E image. First, we observed that STHD-predicted cell type labels
136 collectively captured the compartmentalization of colon crypt-like structure with tumor-like
137 epithelial cell type signatures (**Fig.2a**). In detail, stem-like proliferating cells were positioned
138 inside, circled by the abundant enterocyte cells with scattered tumor stem and transit amplifying
139 (TA)-like cells(18,19). The intercrypt regions were composed of abundant endothelial and
140 fibroblast cells, with diverse immune cell populations such as plasma cells, monocytes,
141 macrophages, dendritic cells, and T cells (**Fig.2a**). Interestingly, the cell type probabilities
142 displayed high confidence for abundant epithelial spots as well as for rarer cells like
143 macrophages and plasma IgG B cells (**Fig.2b**, **Supplementary Fig.4**). By default, spots with
144 maximum posterior probability below 0.8 were assigned as ambiguous. The ambiguous spots
145 mainly located at boundaries of different cell types and tissue structures and could represent
146 cell type mixtures or cell gaps (**Fig.2a, b**). In comparison, firstly, clustering of spot-level gene
147 expression vaguely reflected colon crypt structure with mingled cell identities (**Fig.2c,**
148 **Supplementary Fig.5a**). This contradicts the expectation of consistent cell type labels within
149 each cell, given that the raw spots (2x2um) are much smaller than a single cell (about 10um in
150 diameter). Secondly, clustering of gene expression of bins aggregated from 4x4 spots (8x8um
151 bins) within the patch resulted in less noise in crypt center and intercrypt regions (**Fig.2d** left**,**
152 **Supplementary Fig.5b**). Moreover, similar bin grouping from the whole-sample clustering result
153 could clearly distinguish the stemness center, enterocytes, and intercrypt regions, but the
154 boundaries are assigned as one separated cluster that mixed in intercrypt immune cells (**Fig.2d**

155 right). Further deconvolution on the aggregated bins using RCTD separated epithelial structure
156 and immune cells but demonstrated a loss of spot-level resolution (**Fig.2e**).
157
158 Since high-resolution ST data lacks ground truth of spot labels, we evaluated the performance
159 from aspects of histopathology, comparison to other methods, simulation, and marker gene
160 expression. First, the spatial distribution of STHD cell type labels aligns with cell organization
161 morphology based on H&E staining (**Fig.2f**). Second, we compared to cell type deconvolution
162 on 8x8um bins by RCTD doublet mode or full mode using the same reference human colon
163 cancer scRNA-seq reference (**Methods**). RCTD also resolved cell type proportions that match
164 the colon crypt organization and the distribution of STHD labels for enterocytes and Stem/TA-
165 like proliferating cells (**Fig.2e**, **Supplementary Fig.6a, b**). At the raw 2x2um spot level, we next
166 tested computational methods that were able to label cell types by reference-based
167 deconvolution like RCTD(12) and CARD(13), Cell2Location(14), Celloscope(20),
168 Stereoscope(21), or by neighborhood-augmented clustering like Banksy(22). We noted that the
169 common gene count filters in these tools need to be removed to avoid filtering of most spots; for
170 example, 20436 of the 22336 spots are removed by RCTD default UMI cutoff 100
171 (**Supplementary Fig.6c**). At single-spot level, RCTD, CARD, Cell2Location, and Stereoscope
172 can vaguely reveal the crypt structure while with highly inter-mingled cell types from different
173 lineages (**Fig.2g**, **Supplementary Fig.7a-e**). The unsupervised Banksy efficiently separated
174 epithelial crypt and intercrypt fibroblast/endothelial structures, but mixed the rarer immune cells
175 inside the fibroblast/endothelial cluster and called the boundary as a separate cluster
176 (**Supplementary Fig.7f**). Moreover, STHD has highest computational efficiency among all
177 methods, which enables efficient inference on the entire sample (**Supplementary Fig.8**).
178
179 Since binned outputs are commonly used in VisiumHD, we noted that the STHD labels at the
180 subcellular 2x2um resolution can further guide cell type-specific binning in cell-size resolution,

181 by aggregating adjacent spots of the same cell type, thereby improving gene coverage while
182 preventing cell type mixing. With spot-level cell types, an appropriate bin size can be chosen,
183 allowing separation of distinct cell types within the same bin areas for cell type-specific gene
184 expression profiles. **Fig.2h** showcased STHD-guided binning with two commonly used bin
185 sizes, 8x8um (single-cell size) and 16x16um; each cell type-specific bin was positioned at
186 centroid of the covered spots of corresponding cell type. Thus, tumor epithelial cells and
187 intercrypt endothelial, fibroblasts and immune cells were clearly separated on the STHD-guided
188 binning spatial maps (**Fig.2h**, **Supplementary Fig.5c, d**).
189

190 **STHD achieves high performance in benchmark simulations and preserves cell type-**
191 **specific expression profiles.**

192 To quantitatively assess performance of STHD, we simulated comprehensive whole-
193 transcriptome, high-resolution ST data *in silico* (**Methods, Supplementary Fig.9**). As shown in
194 **Fig.2i**, we simulated 500 cells with heterogeneous cell sizes and cell types on 22,500 spots,
195 each with a true cell type label and gene counts following expression levels from the same
196 human colon cancer reference. STHD achieved average area under the curve (AUC) 0.9977
197 with overall accuracy 93.97% for this multi-class classification task with 98 categories (**Fig.2i,**
198 **Supplementary Fig.9**). STHD outperformed other popular spot-level cell type annotation
199 approaches when tested in repeated simulations, achieving highest accuracy, AUC, and
200 percentage of retained spots (**Methods, Fig.2j**). Next, we systematically investigated cell type-
201 specific expression of marker genes based on STHD prediction across the entire colon cancer
202 sample. Despite spot sparsity (**Supplementary Fig.5e, Supplementary Fig.3d**), the STHD
203 labels displayed similar cell type-specific expression patterns of marker genes as in scRNA-seq
204 reference data in normal epithelial cells (**Fig.2k**, left two panels, **Supplementary Fig.10a**) and
205 all other cell lineages: tumor epithelial cells (**Supplementary Fig.10b**), myeloid cells
206 (**Supplementary Fig.11a**), B and plasma cells (**Supplementary Fig.11b**), CD8T, CD4T, and

207 other T cells (**Supplementary Fig.12, Supplementary Fig.13a**), fibroblasts (**Supplementary**
208 **Fig.13b**), and endothelial cells and pericytes (**Supplementary Fig.14**). STHD-guided binning
209 effectively increased marker gene coverage and meanwhile retained cell-type specificity, shown
210 in two example bin sizes: 8x8um bins (**Fig.2k**, third panel, **Supplementary Figs.10-14**) and
211 16x16um bins (**Fig.2k**, fourth panel, **Supplementary Figs.10-14**). We also observed that the
212 traditional binning approach that aggregated 8x8um bins yielded 311,396 bins with single cell
213 type and 169,188 bins with more than one cell type, creating one third of 8x8um grid bins with
214 mixed cell type signals.
215
216 **STHD enables scalable spatial inference and interactive exploration of cellular niches**
217 **and immune hubs, facilitating genome-wide and cell type-specific spatial comparisons.**
218 The STHD pipeline enabled scalable inference across the entire ST sample through parallelized
219 inference of spatial patches followed by whole-sample integration (**Supplementary Fig.2c**). The
220 parameter tuning was thus performed for the colon cancer VisiumHD sample to balance
221 between statistical likelihood and neighborhood similarity (**Supplementary Fig.15**). To facilitate
222 close examination of spot identities and scalable investigation of cellular neighborhoods, we
223 developed STHDviewer, an interactive and scalable visualization functionality capable of
224 rendering millions of spots in a webpage (**Fig.3a, Supplementary Fig.16**). As demonstrated in
225 **Fig.3a**, STHD prediction automatically achieved precise segmentation of global tissue
226 architecture and local multicellular neighborhoods. Specifically, fibroblasts separated regions of
227 normal-like area with colon gland morphology (named Epi-Normal-like, **Supplementary Fig.16-**
228 **17**) and tumor epithelial areas at the margins (Epi-Tumor-left, Epi-Tumor-top, Epi-Tumor-right)
229 compared to tumor in deposit (Epi-Tumor-inside)(23,24) (**Fig.3b**). Across tumor regions of
230 interest (ROI), STHD labels delineated compositional heterogeneity of cell lineages and types
231 (**Fig.3c**). Notably, spatial heterogeneity of immune cell populations was also observed; for
232 instance, the Epi-Tumor-inside region contained macrophages(25,26) but few T cells compared

233 to tumor regions at the margin, indicating distinct “hot” verses “cold” T cell infiltration spatial
234 patterns within the same sample (**Fig.3c** top, **Supplementary Fig.18**).
235
236 Besides cell type compositional comparisons, STHD facilitates genome-wide and cell type-
237 specific differential expression (DE) analyses across spatial regions. Comparing different tumor
238 ROI, epithelial cells in Epi-Tumor-inside expressed highest stemness gene *LGR5*(27), whereas
239 epithelial cells in Epi-Tumor-top highly express *CCL20*(28) and enrich pathways like TNF-alpha
240 signaling (**Fig.3d,e, Supplementary Fig.19-20**). Since the Epi-Tumor-inside region enrich
241 macrophages, we compared differentially expressed genes specific to macrophage spots and
242 observed that macrophage infiltrating inside tumor expressed higher extracellular matrix gene
243 *SPP1* and lower metalloelastase *MMP12*, cytokine *CXCL8*, and antigen presentation gene *B2M*,
244 indicating that macrophage states segregated spatially in the same tumor(26,29,30)(**Fig.3f**,
245 **Supplementary Fig.21**). The cell type-specific DE analyses can also be performed on the level
246 of STHD-guided bins with higher depth, which resulted in a higher number of significant DE
247 genes with well correlated effects (**Fig.3f**, **Supplementary Fig.20-21**).
248
249 We next demonstrated STHD in analyzing the less abundant immune cells after using
250 STHDviewer to identify immune cell hubs at different locations in the colon cancer VisiumHD
251 sample. An advantage of STHD is that rarer cell types and fine-grained cellular neighborhoods
252 are maintained on top of the denoising effect for those abundant cell types forming continuous
253 tissue architecture; an example of immune-endothelial cell interactions was demonstrated in
254 **Supplementary Fig.22**. **Fig.4a** highlighted two T cell-rich areas marked in **Fig.3b**. The
255 Immune-Trich-left region contained CD4T and CD8T cells with abundant macrophages and
256 C1Q-high dendritic cells (DC)(31), while the Immune-Trich-right area contained CD4T, CD8T
257 cells, mature DCs enriched in immunoregulatory molecules (mregDC)(32), and Plasma IgG B
258 cells (**Fig.4b**). In neighborhood enrichment analyses, each area displayed unique co-occurrence

259 patterns among immune cell types; for example, T cells in Immune-Trich-left area frequently
260 interacted with macrophage and DCs, while T cells in Immune-Trich-right area co-occurred with
261 DCs (**Fig.4a,c**). Utilizing cell type-specific DE analysis, we observed that T cells in the Immune-
262 Trich-left region expresses higher cytotoxicity genes like GZMB comparing to T cells in the
263 Immune-Trich-right region, which can be associated with the immunosuppressive function of
264 mregDCs in tumor(33–35) (**Supplementary Fig.23**).
265
266 **STHD identifies precise cell-cell interactions frontlines in immune hubs and reveals**
267 **direct ligand-receptor communication activities.**
268 Cell-cell communications between cell types in disease microenvironment could drive disease
269 progression like tumorigenesis(36,37). Molecular interactions between cells types in single-cell
270 transcriptomics analyses are traditionally inferred by evaluating averaged expression of
271 communicating gene pairs, which lacks support on direct and physical interactions(38). With the
272 sub-cellular resolution spots labeled, STHD can pinpoint the precise frontlines of inter-cell type
273 interactions. As shown in **Fig.4d**, we highlighted the frontlines of T cell-macrophage interaction
274 and T cell-DC cell interaction in each of the T cell-rich regions. We hypothesized that cell-cell
275 communications mediated by ligand-receptor signaling near-frontline could be stronger than that
276 in far-frontline spots. For a given cell type pair, we identified spots at boundary of two cell types
277 based on posterior probabilities and considered STHD-labeled spots within 4 spots (comparable
278 to single cell size) in distance as near-frontline spots (**Methods, Fig.4d**). We calculated ligand-
279 receptor interaction scores (LRIS)[21,22] between the near-frontline spots and compared with
280 background interaction scores from far-frontline spots, conditioning on the same cell type pair.
281 As shown in **Fig.4e**, near-frontline spots (grouped as 3,4) harbored higher LRIS activities
282 compared to the far-frontline spots (grouped as 1,2) in both cell type pairs of T-Macrophage
283 (**Fig.4d,e** top) and T-DC interactions (**Fig.4d,e** bottom). Specifically, the T-Macrophage
284 communications in Immune-Trich-left were largely driven by *CD74-IFNG* interaction, while the

285 T-DC communications in Immune-Trich-right were driven by *CCL17*/*CCL19*-*CCR7* interactions
286 (**Fig.3k**), which indicated the immunoregulatory effect of mature DCs on T cells(39).
287
288 **Multi-sample analyses of human colon cancer identified sample-specific immune**
289 **enrichment patterns.**
290 As data accumulate for whole-genome spatial transcriptomics with high-definition, we next
291 extended STHD to additional VisiumHD samples. Oliveira et al. recently profiled multiple
292 colorectal cancer samples and adjacent normal colon tissues using VisiumHD(9). We applied
293 STHD to analyze two cancer samples (P1, P5) and one adjacent normal colon sample (P3), in
294 addition to the previously released sample (P2) included in this study. **Fig.5a** illustrated STHD-
295 inferred cell type identities for colon cancer sample P1 containing 8,115,201 spots at 2x2um
296 spot resolution. The colon cancer sample P1 exhibits a similar tumor organization to the colon
297 cancer sample P2, with segmented tumor regions containing colon cancer epithelial cells
298 interspersed with fibroblasts, and adjacent colon tissue enriched with goblet cells. Notably, the
299 colon cancer sample P1 features a prominent tertiary lymphoid structure (TLS), characterized
300 by abundant germinal center B cells and surrounding T cells (**Fig.5b** top). The TLS-like regions
301 was also validated by B cell markers (*CD19*, *MS4A1*) and the TLS-associated T cell cytokine
302 *CXCL13* (**Fig.5b** bottom)(40). In contrast, other colon cancer samples P2 and P5 lack TLS-like
303 structures (**Fig.2, Fig.5c** left). STHD also provided detailed insights into the normal colon tissue
304 from sample P3, revealing goblet cells, an adjacent lymphoid structure, and a myofibroblast
305 region (**Fig.5c** right). Overall, STHD captures heterogenous cellular compositions across
306 multiple samples highlighting distinct immune cell profiles (**Fig.5d**).
307
308 **STHD is generalizable to diverse samples across tissues, diseases, and organisms.**
309 We then extended the application of STHD to normal tissues from both human and mouse
310 model systems. Using the STHD pipeline with default parameters, we analyzed three VisiumHD

311 samples from 10X Genomics including human pancreas, mouse small intestine, and mouse
312 brain (**Methods**). For the adult human pancreas sample, we constructed cell type-specific gene
313 expression estimation using reference single-cell RNA-seq data from Baron et al.(41), which
314 included 12,000 pancreatic cells from 4 adult donors profiled by inDrop and annotated 14 cell
315 types. STHD inference, based on this reference, demonstrated the pancreas's abundant acinar
316 cells and embedded pancreatic islets - organized cell structure harboring endocrine cells
317 (**Supplementary Fig.24a**, left). We observed that marker gene expression levels matched
318 expected patterns for various cell types, such as beta cells (*IAPP*, *INS*), alpha cells (*GCG*,
319 *TTR*), epsilon cells (*GHRL*), and pancreatic stellate cells (*RGS5*) (**Supplementary Fig.24a**,
320 right)(41). For the mouse small intestine, we utilized a reference scRNA-seq profile from the
321 Mouse Cell Atlas 2.0(42) (**Methods**). STHD inference of spot identities collectively reveals the
322 tissue structure of the small intestine (**Fig.5e**). Specifically, we highlighted the mesenteric lymph
323 node with immune cells and the layered intestine-epithelium organizing enterocytes, epithelial
324 cells, goblet cells, Paneth cells, and fibroblasts (**Fig.5f,g** top row), with matched expression of
325 cell type-specific gene markers(42) observed at 8x8um bin-aggregated level (**Fig.5f,g** bottom
326 row). Finally, we mapped the mouse brain VisiumHD data to scRNA-seq data based on Allen
327 Brain Atlas(43) (**Methods**). STHD revealed brain layer-specific cell types for each high-
328 resolution spot and differentiated known cortical layers and hippocampus regions
329 (**Supplementary Fig.24b**). Overall, we observe a high degree of spatial correspondence to
330 tissue structures collectively by STHD spot-level labels, demonstrating STHD's generalizability
331 across diverse tissues and disease samples.
332

333 ## Discussion

334 As spatial transcriptomics technologies advance towards whole-genome coverage at single-
335 and sub-cellular resolution, the development of computational methods that offer both local
336 precision and global scalability is crucial for enabling comprehensive genome-wide analyses

337 across spatial regions, samples, and conditions. However, the enhanced dimension and
338 resolution introduces challenges such as data sparsity. To address this issue, existing analytical
339 approaches often rely on hard aggregation of adjacent spots, which compromises resolution
340 and leads to the mixing of distinct cell types. We presented STHD, a machine learning method
341 designed to model cell types at the single-spot level for high-definition, genome-wide spatial
342 transcriptomics data. Unlike traditional methods, STHD achieves soft aggregation of cell type-
343 specific gene expression information contributed by neighboring spots, effectively addressing
344 the computational difficulties posed by sparse data through a neighbor-augmented Expectation
345 Maximization (EM) model. Our benchmarking simulations and analyses of recently available
346 public VisiumHD samples demonstrate STHD's accuracy, precision, and efficiency. With
347 genome-wide gene expression profiles, STHD's spot-level cell type labels enabled a range of
348 downstream analyses, including the study of global tissue architecture, local cellular
349 neighborhoods, cell type-specific differential expression, cell type-augmented binning, and cell-
350 cell interactions. The STHD pipeline and viewer are implemented to provid efficient modeling
351 and interactive examination, scalable to the entire sample containing millions of spots.
352 Moreover, we have shown STHD's generalizability across different samples, tissues, and
353 disease contexts.
354
355 A tradeoff has existed in ST technologies between resolution and throughput. For seq-ST
356 technologies, which typically cover a high number of genes, the limitation in spot resolution
357 often results in the mixing of gene expression signals from different cell types within a single
358 spot (e.g. Visium's 55um spot diameter). Previous computational efforts have aimed to estimate
359 cell type proportions using spot-level gene expression decomposition, such as RCTD(12),
360 Cell2Location(14), CARD(13), *etc*. As seq-ST technology progresses towards subcellular spot
361 size with genome-wide coverage, we adopted a different assumption: most spots could be
362 attributed to a single cell type. This assumption reduces model complexity and allows for

363 estimation based on sparse readouts. Notably, STHD's efficient optimization enabled analyses
364 of an entire VisiumHD sample with millions of spots using around ten CPU hours without
365 requirement GPU resource. In both normal and tumor tissue samples, we observed that this
366 assumption held for most spots, except for a subset classified as ambiguous. STHD detected
367 these ambiguous spots based on the multinomial cell type probability vector, where lack of a
368 dominant cell type (default posterior cutoff 0.8) indicates ambiguity. Comparison to histology
369 reveal that these ambiguous spots often correspond to regions like cell gaps, certain non-
370 nucleus cytoplasm areas, and mostly inter-cell type boundaries. Consequently, these cell type
371 boundary spots can be utilized in identifying inter-cell type transitions at the highest possible
372 resolution; also, STHD's estimated cell type probabilities could implicate spot-level
373 compositions.
374
375 The data sparsity in gene counts presents a significant challenge in subcellular spot modeling.
376 For instance, in the colon cancer sample P2, the average count depth of a 10X VisiumHD spot
377 was 1/262 of that of a single cell from 10X scRNA-seq. Consequently, this sparsity often results
378 in the exclusion of most spots in existing spot-level methods or generation of noisy cell type
379 labels when each HD spot was treated independently. To tackle sparsity, STHD incorporates a
380 soft regularization of the loss function based on cross-entropy among neighboring spots' cell
381 type probability vectors. This approach leverages the neighbor similarities in spatial data, as
382 seen in previous methods like Banksy, which utilized Gabor filter to incorporate structured
383 neighborhood contribution in unsupervised tasks like tissue segmentation and spatial
384 clustering(22). Our results show that STHD labels not only captures the continuous zonation of
385 those abundant cell types but also identifies rarer cell types, such as immune cells in tissue and
386 tumor samples. The recovered continuity in the spot identities further enabled cell type-guided
387 binning, which maintains cell type-specific marker expression and enhances gene coverage for
388 downstream differential expression analyses. In a recent study, Bin2Cell also avoids traditional

binning methods but operates on the raw VisiumHD spots by pooling spots overlapping the same nucleus identified from H&E image(44). In another recent study, TopACT also models neighbor aggregation, where the range of multiscale aggregation was optimized for cell type classification(45). Similar to the neighbor parameters in Banksy and multiscale design in TopACT, STHD allows for tunning of neighborhood aggregation effect: the β parameter and the iteration step parameter control the loss contribution from neighbors, and the neighbor size parameter can control the range of spatial similarity aggregation, making it adaptable to various cell sizes and tissue types.

Despite the advancements, several challenges remain in analyzing high-dimensional data with subcellular resolution. First, the lack of group truth for spot-level cell identities in tissue ST samples complicates the evaluation of method performance. Our spatial simulations, which created spots with true cell type labels matching scRNA atlases, provide a starting point for such evaluations, but future integration with other spatial technologies (e.g., image-ST or proteomics) could offer additional ground truth data. The STHD model is also generalizable to other high-resolution, high-dimensional spatial data platforms like Stereo-seq and Xenium 5K, *etc*. Second, the availability and quality of the reference scRNA-seq data can limit the applicability of STHD, as it relies on these references for cell type identification but does not discover new cell types, a common feature also existed for spot deconvolution methods. In cases where detailed and cell type annotation are lacking or cell type profiles are highly correlated, STHD label can provide an initial labeling, STHD can still provide initial labeling, allowing for further discovery of transcriptional variations within each cell type through clustering of STHD-guided bins. We anticipate expanding of STHD's capabilities for semi-supervised cell type identification. Third, the issue of molecular diffusion has been found to exist across ST methods(46), which may interfere with STHD prediction particularly in spots with weak marker gene signals. Computational tools to correct spot diffusion, such as SpotClean(47) for low-

resolution data, have been developed, but there remains a need for methods compatible with millions of sparse spots. Additionally, molecular diffusion could also blur the tissue boundaries, as observed in some regions outside H&E images containing gene counts. This also imply that integrating histopathology properties can be informative in modeling the high-resolution spatial transcriptomics sample, as done in recent studies like iStar, which infers spatial gene expression from H&E images(48).

Looking ahead, we believe that the combination of individual subcellular spot modeling, neighborhood information sharing, and multi-modality neighbor information incorporation, including H&E images, will be key to overcoming the resolution-throughput tradeoff in emerging spatial transcriptomics technologies. This rationale fully harnesses the potential of recent advancements in spatial technology and enables the discovery of spatially organized cellular heterogeneity in samples. The algorithm design of STHD is thus generalizable in other platforms of high-dimensional spatial transcriptomics profiling with subcellular resolution. In summary, STHD offers probabilistic subcellular spot-level predictions that facilitate precise and scalable analyses, with a software toolkit that is generalizable across various tissues and diseases, providing new insights into cellular and molecular landscapes.

Conclusions

We have presented STHD, a novel machine learning method for probabilistic cell typing of single spots in whole-transcriptome, high-resolution ST data. Unlike the current binning-aggregation-deconvolution strategy, STHD directly models gene expression at single-spot level to infer cell type identities without cell segmentation or spot aggregation. STHD addresses sparsity by modeling count statistics, incorporating neighbor similarities, and leveraging reference single-cell RNA-seq data. We show that STHD accurately predicts cell type identities at single-spot level in simulated benchmarking and biological spatial samples, where STHD

441 achieves precise segmentation of both global tissue architecture and local multicellular
442 neighborhoods. The high-resolution STHD labels facilitate various downstream analyses,
443 including cell type-stratified bin aggregation, spatial compositional comparisons, and cell type-
444 specific differential expression analyses. Moreover, STHD labels further reveal frontlines of
445 inter-cell type interactions at immune hubs in cancer samples. Finally, STHD is efficient,
446 scalable, and generalizable across diverse samples, tissues, and diseases. Overall, STHD will
447 facilitate biomarker and mechanistic discoveries in spatial heterogeneity, cellular interactions,
448 and molecular mechanisms in various tissue and disease contexts.
449

450 **Methods**

451 **STHD algorithm.**

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

459 $$\mathcal{L} = -LL + \beta CE$$

460 For the negative log likelihood part, we model counts of each gene at each spot to follow
461 Poisson distribution. Let the normalized average gene expression level for gene g within the
462 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
463 type t is $P_a(t)$. Thus, the count of gene g at spot a: $n_g^a \sim Poission(d_a \lambda_g^t)$.

464 $$LL = log \prod_a \prod_g P(n_g^a)$$

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

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

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

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

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.

$$CE = -\sum_a \sum_{a' \in NN(a)} \sum_t \left(P_a(t) logP_{a'}(t)\right)$$

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:

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

More details of the model and derivation of explicit form of gradient for fast optimization is available separately in **Supplementary File 1**. In short, we derived that the explicit form of gradient of the loss function $\mathcal{L}$ at $w_{\tilde{a}}^{\tilde{t}}$ for spot $\tilde{a}$ and cell type $\tilde{t}$ as:

$$\frac{\partial \mathcal{L}}{\partial w_{\tilde{a}}^{\tilde{t}}} = -\frac{\partial LL}{\partial w_{\tilde{a}}^{\tilde{t}}} + \beta \frac{\partial CE}{\partial w_{\tilde{a}}^{\tilde{t}}}$$

$$= -P_{\tilde{a}}(\tilde{t}) \mathcal{F}(\tilde{a}, \tilde{t}) + \sum_t P_{\tilde{a}}(t) P_{\tilde{a}}(\tilde{t}) \mathcal{F}(\tilde{a}, t)$$

$$-\beta \left[P_{\tilde{a}}(\tilde{t}) \sum_{a' \in NN(\tilde{a})} \left(logP_{a'}(\tilde{t}) - \sum_t \left(P_{\tilde{a}}(t) logP_{a'}(t)\right) \right) - \sum_{a \neq \tilde{a}} \mathbb{I}_{a' \in NN(a)} \left(P_a(\tilde{t}) - P_{\tilde{a}}(\tilde{t})\right) \right]$$

483 Where

484 $$\mathcal{F}(a,t) = \sum_{g} \left(n_g^a log\left(d_a \lambda_g^t\right) - d_a \lambda_g^t \right)$$

485 $$P_a(t) = \frac{e^{w_a^t}}{\sum_{s \in T} e^{w_a^s}}$$

486 STHD thus optimizes $w_{\tilde{a}}^{\tilde{t}}$ and the probability of spot $\tilde{a}$ having the identity of cell type $\tilde{t}$: $P_{\tilde{a}}(\tilde{t})$,

487 using Adam(16). We implemented iterative updates using numba(15) on sparse matrix to

488 achieve high computational efficiency and scalable optimization.

489

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

491 The average normalized gene expression λ_g^t can be estimated from a pre-defined single-cell

492 RNA-seq reference dataset. A reference single-cell dataset shall match the tissue or disease

493 origin of the high resolution, whole-transcriptome spatial transcriptomics data. It shall also

494 contain cell type annotation that is usually based on multi-sample harmonization and

495 unsupervised clustering of single cells into various cell type or groups. The granularity of cell

496 type annotation directly influence the number of predicted cell type and downstream analyses.

497 For each cell type, average normalized gene expression profiles are estimated with $\lambda_g^t = \frac{n_g^t}{\sum_g n_g^t}$

498 so that $\sum_g \lambda_g^t = 1$. Specifically for human colon cancer, we downloaded the raw counts file of the

499 single-cell RNA-seq data from the human colon cancer atlas from Pelka et.al(18) through

500 GSE178341. The data contains 370115 tumor and adjacent normal single cells from 62 primary

501 treatment-naïve colorectal cancer patients, with 43113 genes profiled. Quality control of single

502 cells was performed by filtering cells with minimum 100 counts and 100 genes, filtering genes

503 with minimum 3 cells, and removing cells with mitochondrial read ratio larger than 27.7%, which

504 is mean plus one standard deviation of the mitochondrial read ratio across all cells. After the

505 above quality control, 303358 cells and 31,873 genes remain. The cell annotations were

506 obtained from the study that provided 98 fine-grained cell types and 7 cell lineages. We used
507 the 98 high-granularity cell type annotation throughout this study. To select genes informative in
508 distinguishing the 98 cell types, we followed the calculation of cell type expression profile as in
509 RCTD(12). First, we calculated the estimated cell type expression profiles λ_g^t for all genes and
510 cell types. We then selected genes with average expression profile above 0.000125, which is
511 0.0625 counts per 500, and at least 0.5 log-fold-change compared to the average expression
512 across all cell types, using log2 scale. Moreover, we removed the mitochondrial genes and
513 ribosome genes in the selection. The selection resulted in 4618 genes that are cell type-
514 informative, for each we calculated λ_g^t in all cell types. Note that STHD uses the 4618 genes to
515 infer cell type identity of spots, and later can consider whole-transcriptome for downstream
516 analyses.
517
518 **STHD hyperparameter tuning.**
519 To select hyperparameters that balance between statistical likelihood and neighborhood
520 similarity, we comprehensively tested hyperparameter tuning and provided an automatic
521 stopping criterion and optional tuning tutorial. The main parameters in STHD includes the
522 neighborhood parameter β, which controls the contribution of cross entropy loss among local
523 neighbors to the total loss, the optimization steps n_iteration, and the learning rate that can be
524 fixed in practice. We provided a tutorial on parameter tuning based on performance in random
525 sampled patches; for the human colon cancer sample, the optimized parameters are 23 iteration
526 steps and β=0.1. In detail, for the colon cancer sample, we randomly selected six different
527 patches in size of 1100x1100 pixels (in full resolution, around 22500 spots). For the optimization
528 of each patch, we tested a range of β from 0, 0.03, 0.1, 0.3, 1, and 3, each outputting optimal
529 iteration steps based on the early stopping criterion, where the drop of loss in the past 10 steps
530 falls within 1%. Since the regularization term from neighborhood contribution β will lower log

531 likelihood, as β increases, we select the optimal parameter pairs where drop in cross entropy
532 slows down to around 10% while a high log likelihood is maintained. The optimal iteration steps
533 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
534 is robustly 23.17 on average. Therefore, we chose β=0.1 and n_iteration=23 as optimal
535 parameters for the whole sample modeling that balance the count-based likelihood and the
536 contribution from neighbors.
537
538 **STHD low-count region detection and filtering.**
539 Throughout the spatial transcriptomics data, regions of low gene counts exist, usually arising
540 from tissue boundaries, tissue gaps, and regions low in DNA/RNA such as dead cells, necrosis,
541 vessel enriching red blood cells, and collagen rich area. We included an automatic local low-
542 count region detection and filtering functionality in STHD. By default, STHD computed neighbor
543 graphs within 2 rings of spots (13 spots for the grid system), aggregated local connected spots,
544 and detected low-count local regions not exceeding 50 counts, where the model will skip
545 modeling due to low confidence in cell type inference. Other customized tissue gap and
546 boundary masks can also be provided to further filter spots.
547
548 **Assessing model accuracy with simulated spatial transcriptomics data.**
549 To assess the accuracy of STHD model prediction, we simulated the high-dimensional, high-
550 resolution spatial transcriptomics data *in silico*. We simulated a region with 22500 (150x150)
551 grid spots and position 500 simulated cells overlapping the region with cell sizes spanning 1-20
552 spots. Each cell is randomly assigned a cell type from the 98 cell types annotated in colon
553 cancer atlas, and spots overlapping each whole cell are assigned the corresponding cell type.
554 To mimic the real-world scenario in gene expression profiles and spot sparsity, we adapted the
555 same normalized gene expression profile from the same colon cancer atlas data with 98 cell
556 types. The depth of each spot was simulated following the distribution of exponential decay with

557 mean read depth 47 matching the colon cancer VisiumHD sample. After simulating counts of
558 each gene at each spot in the region, we applied STHD with default optimal parameters (β=0.1)
559 and also STHD no-neighbor model (β=0) to the simulated spatial data, and measured prediction
560 accuracy, standard error, and area under the receiver operating characteristic curve (ROC), for
561 this task of multi-class classification of 98 categories at each spot. To compared with other
562 methods that enable spot-level cell type annotation, we examined six different settings: STHD,
563 RCTD with default count filter, RCTD with no count filter, CARD with default count filter, CARD
564 with no count filter, and Cell2Location. The inclusion of both default parameter and adjusted
565 count filtering for RCTD and CARD is due to the fact that most spots will be removed if following
566 default parameters, where accuracy measurements were based on non-filtered spots. We
567 evaluated the methods in ten simulation replicates and assessed performance metrics in mean
568 accuracy, AUC, and percentage of retained spots.
569
570 **Comparison with other spot cell typing methods.**
571 To benchmark the STHD results at 2x2 μm resolution, we compared STHD to six methods that
572 can output cell typing, including the deconvolution-based cell typing tools RCTD(12), CARD(13),
573 Cell2Location(14), Celloscope(20), Stereoscope(21), and the unsupervised method Banksy(22).
574 Because of computational efficiency and data size limit of the existing methods, we used the colon
575 cancer tumor region as in Fig.2a to perform benchmarking, which contains 22,336 cells and
576 18,085 genes. For methods requiring the reference scRNA-seq data as input, we used the same
577 human colorectal cancer scRNA-seq datasets (GSE178341) with 98 cell types. Since many
578 methods in comparison requires a maximum number of cells from reference scRNAseq dataset
579 – for example, CARD requires below 50000 cells - we down-sampled the reference data to include
580 maximum 600 cells for each cell type, resulting in a total of 49,351 cells and 41,195 genes. For
581 running RCTD on single spots, we used RCTD doublet mode, testing both the default minimum
582 UMI 100 and also by removing the minimum UMI requirement due to loss of most raw spots. For

583 testing CARD, we followed guidelines from https://github.com/YMa-lab/CARD. Similarly, to avoid
584 most spots being filtered, the minimum spot count and minimum gene numbers were set to 0.
585 Cell2location and Stereoscope were ran with default parameters according to the tutorials
586 https://cell2location.readthedocs.io/en/latest/notebooks/cell2location_tutorial.html, and
587 https://github.com/almaan/stereoscope. In Cell2location, by default, the reference data was
588 filtered using the following parameters: cell_count_cutoff = 5, cell_percentage_cutoff = 0.03, and
589 nonz_mean_cutoff = 1.12. The reference single-cell regression model was trained with default
590 parameter max_epochs = 250, lr = 0.002. The spatial cell2location model was obtained with
591 default parameters max_epoch = 30,000. In Stereoscope, by default, the reference data was
592 filtered using a minimum count threshold of min_counts = 10, and all mitochondrial genes were
593 removed. The reference single-cell model was trained with parameters max_epochs = 100; the
594 spatial stereoscope model was obtained using parameters max_epochs = 10,000. Finally, for
595 Celloscope, a binary matrix with prior knowledge about marker genes is required among input,
596 for which we selected top 200 genes in normalized expression from our informative gene matrix
597 for each cell type. Another required input is estimated number of cells in each spot, for which we
598 set to 1 due to sub-cellular-resolution nature of the data. Celloscope was then tested following
599 guidelines at https://github.com/szczurek-lab/Celloscope. Finally, Banksy was tested with default
600 parameters following the tutorial at
601 https://github.com/prabhakarlab/Banksy_py/blob/main/slideseqv2_analysis.ipynb. For all toolkits,
602 the weights or cell type abundance for each spot were extracted and normalized as cell
603 proportions for subsequent visualizations.
604
605 **Cell type deconvolution in aggregated bins.**
606 Following current standards of binning-deconvolution, we binned the spatial data and aggregated
607 the bins containing 4x4 spots. We applied RCTD deconvolution of doublet mode on the same
608 spatial patch as in Fig.2a, following the default filtering and parameters of RCTD as on spacexr

609 (https://github.com/dmcable/spacexr), and using the same subsampled colon cancer reference
610 scRNA-seq data as in comparison with other methods. To compare cell proportions across bins,
611 we counted STHD cell type labels and obtained cell proportions for each 4x4 bin after excluding
612 filtered and ambiguous spots. We also computed cosine similarity between STHD's spot count-
613 based proportions and RCTD's deconvoluted proportions and visualize the proportions and
614 cosine similarity on a spatial plot.
615
616 **Estimating running time in computational efficiency.**
617 To assess computational efficiencies, different cell typing tools are applied onto the tumor
618 epithelial patch using a server with 8 CPU cores and maximum 128GB memory, running time
619 recorded. For tools supporting GPU acceleration (cell2location and stereoscope) the time cost
620 was evaluated on a single GPU, and the spatial training time was recorded without including
621 scRNA training time. For computational efficacy, we also estimated the time cost for the entire
622 sample by linearly scaling running time based on number of spots in patch (22,336 spots) and in
623 the entire colon cancer sample P2 (8,726,224 spots).
624
625 **STHD pipeline on an entire VisiumHD sample.**
626 The STHD software and pipeline contains several components to flexibly split the whole sample
627 into patches for inferences and parallelization. The input/output functionalities (STHDio)
628 expanded from Squidpy Visium interface to enable cropping of full-resolution image. Second, for
629 the full-size sample, STHD will automatically patchify the full sample into multiple regions with
630 overlaps, followed by low-count filtering, per-patch inference, and result aggregation where
631 probabilities of overlapping spots are averaged. The STHD cell type label results are
632 determined by Maximum A Posterior (MAP) with a preset cutoff (default 0.8). An optional fast
633 implementation for STHD-guided binning with a predefined bin size is provided, which outputs
634 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(49). 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(50) 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.

660 **Neighborhood interaction enrichment analysis for immune-rich regions in colon cancer**
661 **VisiumHD sample.**
662 To investigate the neighborhood interaction patterns across various cell types in the two immune-
663 infiltrated regions, local neighborhoods were calculated using the 4 nearest cells in the grid and
664 considering 6 rings of adjacent spots to get a obtain a local neighborhood structure using
665 squidpy.gr.spatial_neighbors in Squidpy. We used STHD predicted cell type labels and excluded
666 spots in the filtered and ambiguous categories. The neighborhood enrichment Z-scores were
667 calculated using squidpy.gr.nhood_enrichment, which follows CellChat(38) with ligand-receptor
668 databases and permutation of 1000 times. Heatmaps of neighborhood interaction patterns were
669 plotted based on the co-occurrence score of pairwise cell types in each region of interest.
670
671 **Inter-cell type frontline and cell-cell communication analysis.**
672 For a given pair of cell types, we first defined their precise frontlines to be spots where those two
673 cell types are adjacent to each other. We further restrict frontline spots to be those whose neighbor
674 only contains that pair of cell types based on maximum posterior to eliminate potential noises
675 from other cell types. Considering a single cell is comparable to the size of 4 spots, we assumed
676 spots within 4 spots from the frontlines represented one cell with direct interaction with another
677 cell type. Thus, spots were categorized based on their cell type and proximity to frontline. In the
678 colon cancer sample P2, we examined interactions between T cells and dendritic cells, and T
679 cells and macrophages separately for the two immune regions of interest, Immune-Trich-left and
680 Immune-Trich-right. In the chord plots, for a pair of cell types A and B, the labels were assigned
681 as follows: '1' for far-frontline cell type A, '2' for far-frontline cell type B, '3' for near-frontline cell
682 type A, and '4' for near-frontline cell type B. We then hypothesized that cell-cell communications
683 like ligand-receptor expression levels were higher between group 3 and 4 compared to group 1
684 and 2. To perform the ligand-receptor interaction analysis, normalized data was applied the
685 squidpy.gr.ligrec function which permuted 1000 times to detect significant ligand-receptor pairs.

686 Top 10 ligand-receptor pairs different between near-frontline and far-frontline communications
687 were visualized in the chord diagram showing the mean expression of each ligand-receptor pair
688 and sender-receiver directions.
689
690 **STHD on additional colon samples and other human and mouse tissue samples.**
691 We performed STHD analyses on the extended list of human colon samples using the same colon
692 cancer reference gene estimation, yielding cell type inference on 8,115,201 spots in colon cancer
693 P1, 8,665,510 spots in colon cancer P5 sample, and 7,791,374 spots in the normal colon sample
694 P3 using normal cell types. We further extended analyses to the human pancreas sample
695 containing 6,680,357 spots for 18,085 genes, the mouse intestine sample containing 5,479,447
696 spots with 19,059 genes, and the mouse brain sample containing 6,296,637 spots with 19,059
697 genes. All analyses were run using default parameters. To match the tissue system, we used the
698 non-cancerous colon scRNA for the normal colon sample, normal human adult pancreas scRNA
699 dataset for the human pancreas sample, the adult Mouse Cell Atlas intestine scRNA dataset for
700 the mouse small intestine sample, and the Allen Brain Atlas reference for the mouse brain sample.
701
702 For constructing reference gene expression profile for human pancreas, we downloaded raw
703 counts file of single-cell RNA-seq data from the human pancreas atlas by Baron et al.(41) from
704 GSE84133. The dataset contains 8,569 normal single cells from four donors, profiling 20,124
705 genes. Quality control involved filtering cells with a minimum of 100 counts and 100 genes,
706 filtering genes with a minimum of 3 cells, and removing cells with a mitochondrial read ratio
707 larger than 5%. After quality control, 8,569 cells and 16,358 genes remained. The study
708 provided 14 cell type annotations. Using STHD default setting, we selected 4,436 cell type-
709 informative genes based on average expression profile above 0.000125 (0.0625 counts per
710 500) and at least 0.5 log-fold change, excluding mitochondrial and ribosomal genes. For mouse
711 intestine tissue, the raw counts file of single-cell RNA-seq data from the mouse cell atlas by Fei

et al.(42) was downloaded from GSE176063. The dataset contains 6,684 normal single cells from three adult mouse intestine samples, profiling 16,416 genes. Quality control involved filtering cells with a minimum of 100 counts and 100 genes, filtering genes with a minimum of 3 cells, and removing cells with a mitochondrial read ratio larger than 7.15%. After quality control, 6,102 cells and 15,745 genes remained, with 28 cell types annotated. Similarly, we selected 4,083 cell type-informative genes based on average expression profile above 0.0002 (0.1 counts per 500) and at least 0.5 log-fold change, excluding mitochondrial and ribosomal genes. For mouse brain reference gene expression construction, the processed single-cell RNA-seq data from the Allen Brain Atlas was downloaded from the Seurat tutorial website (https://satijalab.org/seurat/articles/visiumhd_analysis_vignette), which contained 199,993 single cells and 31,053 genes. Quality control involved filtering cells with a minimum of 100 counts and 100 genes, filtering genes with a minimum of 3 cells, and removing cells with a mitochondrial read ratio larger than 4.8%. After quality control, 181,250 cells and 264,484 genes remained, with 41 subclass cell types annotated. Similarly, we selected 5,171 cell type-informative genes based on average expression profile above 0.000125 (0.0625 counts per 500) and at least 0.5 log-fold change, excluding mitochondrial and ribosomal genes.

Declarations

Availability of data and materials

Software availability.

The current STHD software source code is 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 P2 is available at https://yi-zhang-compbio-lab.github.io/STHDviewer_colon_cancer_hd/STHDviewer_crchd.html.

Data Availability.

All analyses in this paper are based on publicly available data.

Colon cancer reference scRNA-seq atlas: data and cell type annotation from Human Colon Cancer Atlas is available from Pelka et. al through GSE178341: (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178341).

Colon cancer VisiumHD data (patient P2) is available at 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).

Colon cancer VisiumHD data (patient P1, P3, P5) from Oliveira et al.(9) include two colon cancer and one normal colon, available at https://www.10xgenomics.com/products/visium-hd-spatial-gene-expression/dataset-human-crc.

The human pancreas VisiumHD data is available at https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-human-pancreas.

The mouse small intestine VisiumHD data is available at https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-intestine.

The mouse brain VisiumHD data is available at https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he.

Competing interests

The authors declare no competing interests.

Funding

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.

Authors' contributions

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.

Acknowledgements

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.

Figures

Fig.1.

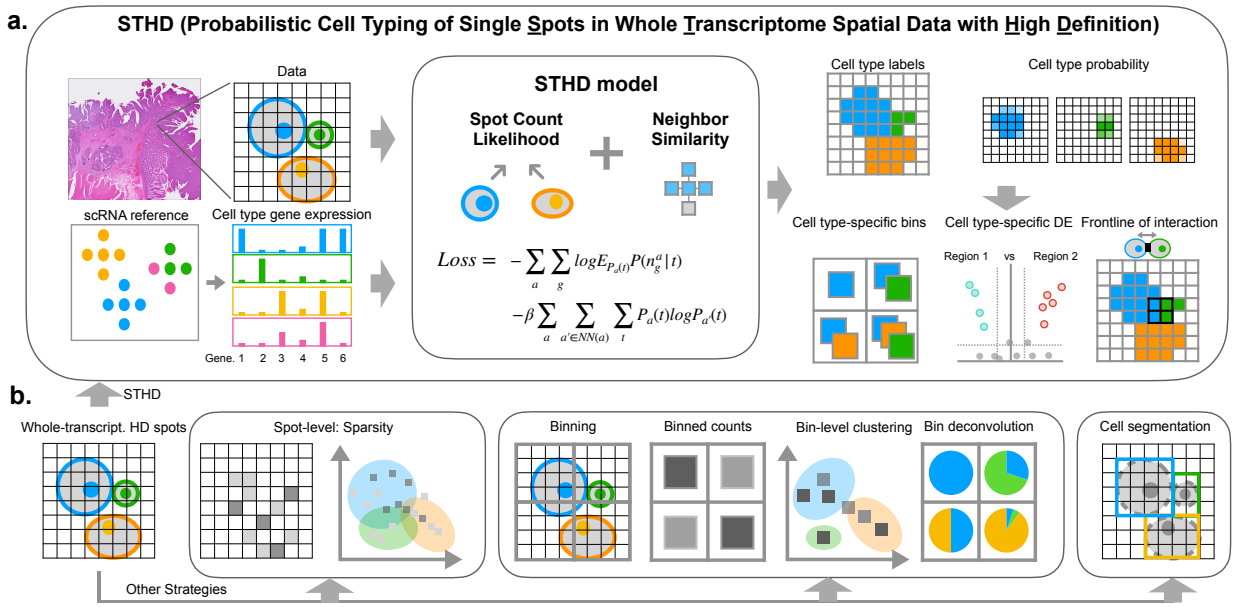

Fig.1. Overview of STHD model. a. STHD (probabilistic Cell Typing of Single Spots in Whole Transcriptome Spatial Data with High Definition) takes two inputs: whole-transcriptome spatial data of high resolution, and gene expression profiles by cell type from reference single-cell dataset. The STHD model 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.

Fig.2.

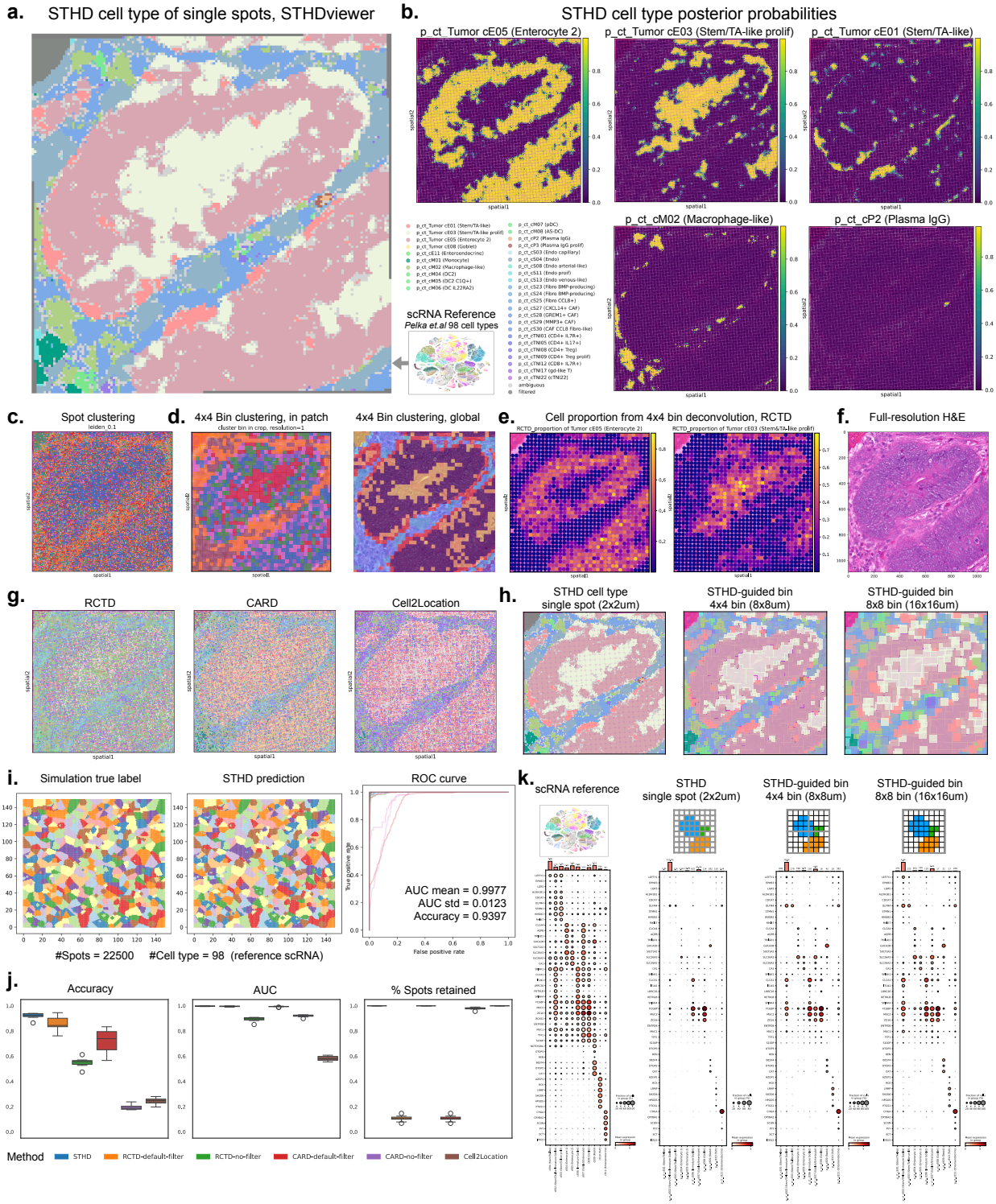

Fig.2. Spot-level predictions and benchmarking of a zoomed patch in a colon cancer VisiumHD sample. a. Example of STHD results of a patch from the human colon cancer

808 VisiumHD sample, revealing tumor-like epithelial cells and intercryptic immune population,
809 visualized using the interactive STHDviewer. **b.** Cell type posterior probabilities at spot-level for
810 the same patch, visualized using Squidpy. Top row, three epithelial cell types; Bottom row, two
811 immune cell types. **c.** A spatial plot of unsupervised leiden clustering on spot gene expression.
812 **d.** Spatial plots of unsupervised leiden clustering of gene expression aggregated by bins in size
813 of 4x4 spots. Left, clustering bins within patch, right, global cluster group of bins. **e.** Cell type
814 proportion decomposed by RCTD on bins of size of 4x4 spots, demonstrating two tumor
815 epithelial classes, enterocytes, and stem-like cells. **f.** The H&E histopathology image in full
816 resolution of the same crop. **g.** Comparison to other cell typing and clustering methods adaptive
817 to high-resolution spots. **h.** The STHD predicted labels at spot-level and STHD-guided binning
818 in bins of 4x4 spots and 8x8 spots, visualized using Squidpy using the same colon cancer cell
819 type colormap. **i.** Benchmarking with simulated high-resolution spatial data. Left, ground truth
820 cell type labels; middle, STHD predicted cell type labels; right, receiver operating characteristic
821 curve (ROC) curve for all cell types. **j.** Benchmarking against other methods for spot annotation
822 in repeated spatial transcriptomics simulations. **k.** Expression dot plots for marker genes of
823 normal epithelial cells, at spot level or STHD-guided bin level. Left to right: normal epithelial cell
824 types in reference human colon cancer sample, normal epithelial cell types based on STHD
825 predicted spots, normal epithelial cell types from STHD-guided bins of size 4x4 spots, normal
826 epithelial cell types in STHD-guided bins of size 8x8 spots. The same cell type-specific gene
827 markers from the original colon cancer study were used. ROC, Receiver operating
828 characteristic. AUC, area under the curve.
829
830
831
832
833

Fig.3.

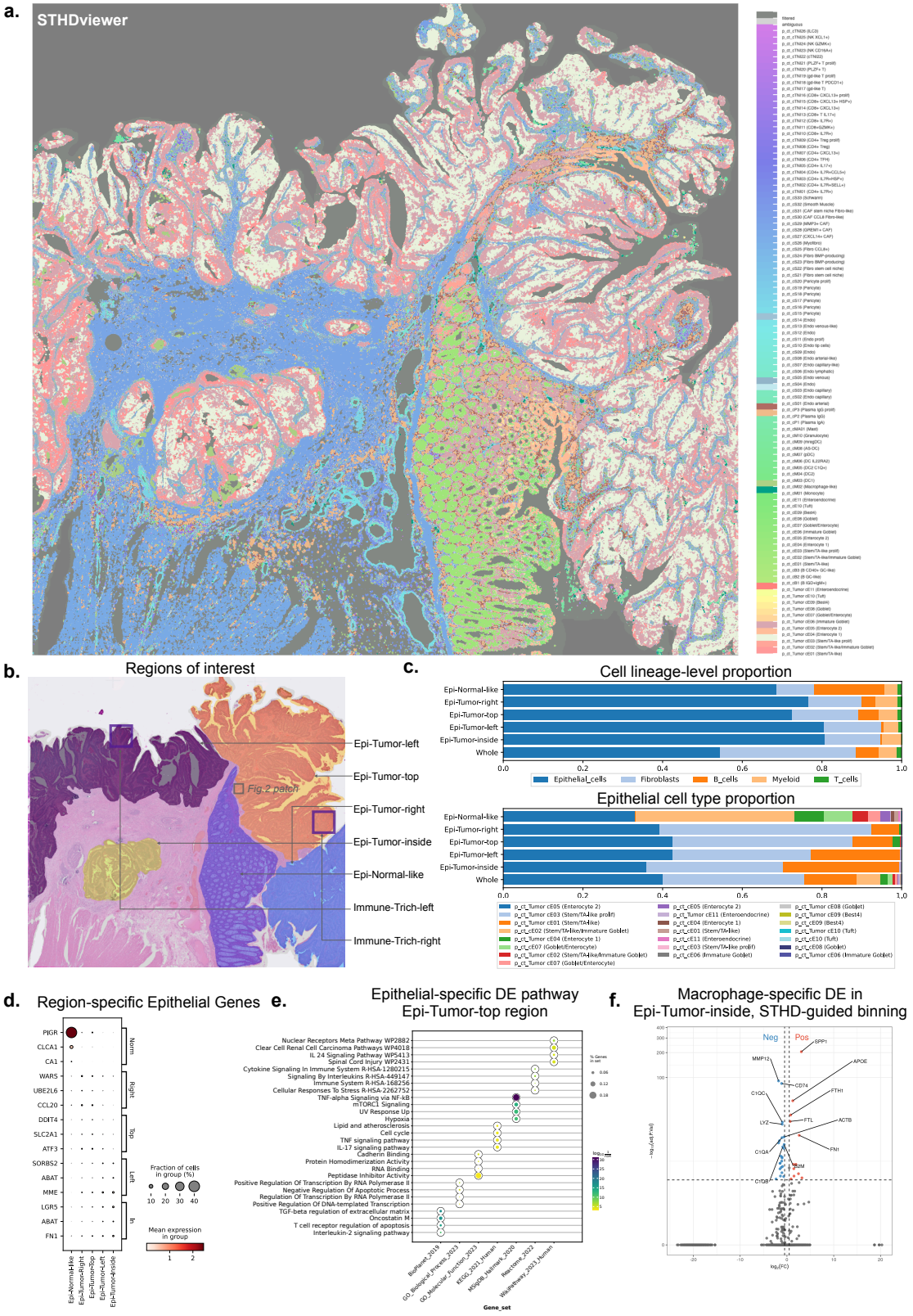
a.
STHDviewer
b.
Regions of interest
Fig.2 patch
Epi-Tumor-left
Epi-Tumor-top
Epi-Tumor-right
Epi-Tumor-inside
Epi-Normal-like
Immune-Trich-left
Immune-Trich-right
c.
Cell lineage-level proportion
Epi-Normal-like
Epi-Tumor-right
Epi-Tumor-top
Epi-Tumor-left
Epi-Tumor-inside
Whole
Epithelial_cells
Fibroblasts
B_cells
Myeloid
T_cells
Epithelial cell type proportion
d.
Region-specific Epithelial Genes
PIGR
CLCA1
CA1
WARS
UBE2L6
CCL20
DDIT4
SLC2A1
ATF3
SORBS2
ABAT
MME
LGR5
ABAT
FN1
Norm
Right
Top
Left
In
Fraction of cells in group (%)
Mean expression in group
e.
Epithelial-specific DE pathway
Epi-Tumor-top region
Nuclear Receptors Meta Pathway WP2882
Clear Cell Renal Cell Carcinoma Pathways WP4018
IL 24 Signaling Pathway WP5413
Spinal Cord Injury WP2431
Cytokine Signaling In Immune System R-HSA-1280215
Signaling By Interleukins R-HSA-449147
Immune System R-HSA-168256
Cellular Responses To Stress R-HSA-2262752
TNF-alpha Signaling via NF-kB
mTORC1 Signaling
UV Response Up
Hypoxia
Lipid and atherosclerosis
Cell cycle
TNF signaling pathway
IL-17 signaling pathway
Cadherin Binding
Protein Homodimerization Activity
RNA Binding
Peptidase Inhibitor Activity
Positive Regulation Of Transcription By RNA Polymerase II
Negative Regulation Of Apoptotic Process
Regulation Of Transcription By RNA Polymerase II
Positive Regulation Of DNA-templated Transcription
TGF-beta regulation of extracellular matrix
Oncostatin M
T cell receptor regulation of apoptosis
Interleukin-2 signaling pathway
Gene_set
f.
Macrophage-specific DE in
Epi-Tumor-inside, STHD-guided binning
Neg
Pos
SPP1
MMP12
APOE
CD74
C1QC
FTH1
LYZ
FTL
ACTB
C1QA
FN1
C1QB

Fig.3. STHD prediction facilitates segmentation of global structures and cell type-specific differential analyses across spatial regions. 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. DE, differential expression.

Fig.4.

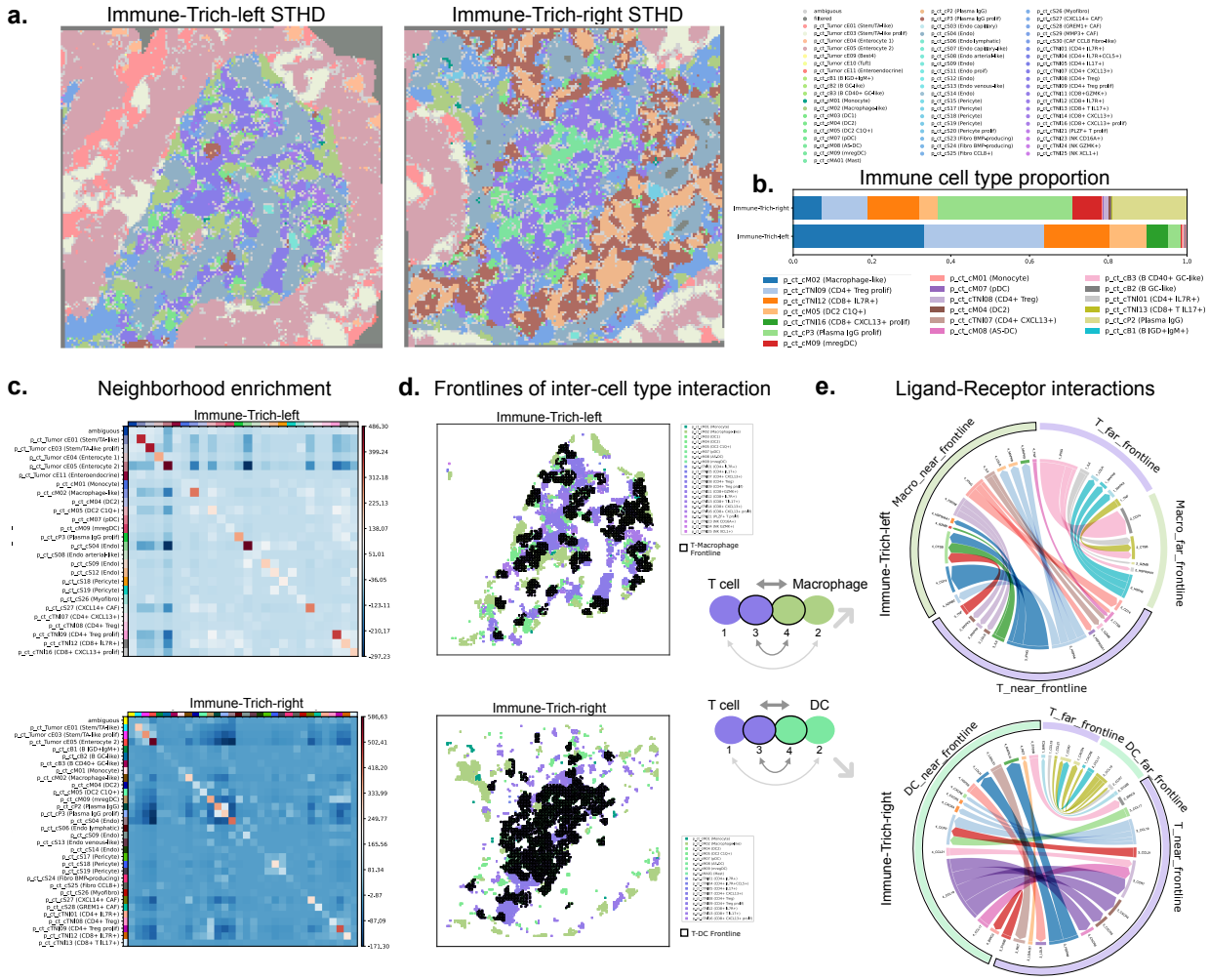

Fig.4. STHD enabled analyses of local neighborhoods for cellular interactions in tumor-infiltrating immune hubs. a. STHD cell type labels for two T cell-rich regions in the colon cancer sample. **b.** Proportions of immune cell types in the two immune regions. **c.** Cell neighborhood interaction patterns of the two T cell-rich regions showing neighborhood enrichment z-scores where spots within 6 spots are considered neighbors. **d.** 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. **e.** 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.

Fig.5.

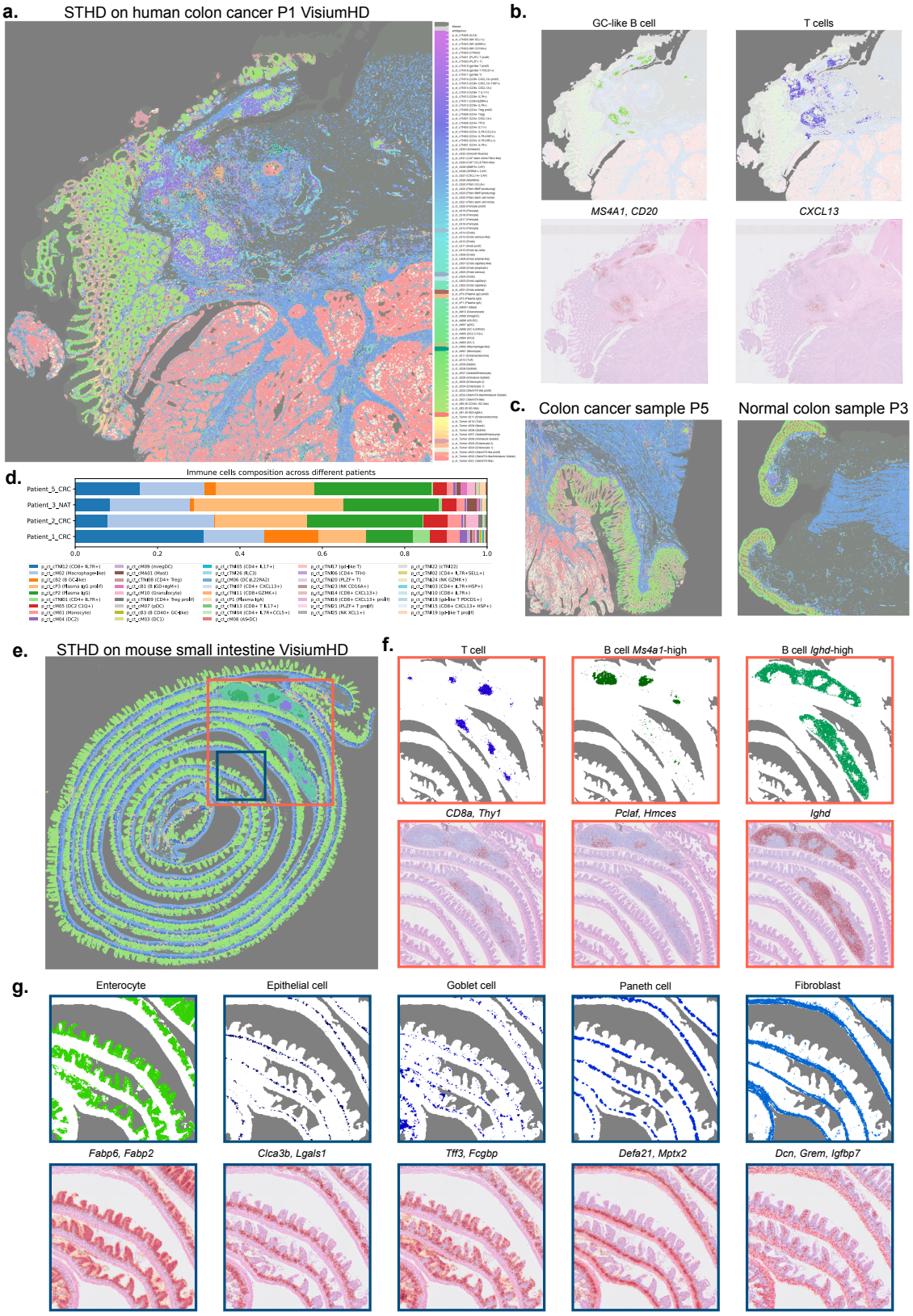
a.
STHD on human colon cancer P1 VisiumHD
b.
GC-like B cell
T cells
MS4A1, CD20
CXCL13
c.
Colon cancer sample P5
Normal colon sample P3
d.
Immune cells composition across different patients
Patient_5_CRC
Patient_3_NAT
Patient_2_CRC
Patient_1_CRC
0.0
0.2
0.4
0.6
0.8
1.0
e.
STHD on mouse small intestine VisiumHD
f.
T cell
B cell Ms4a1-high
B cell Ighd-high
CD8a, Thy1
Pclaf, Hmces
Ighd
g.
Enterocyte
Epithelial cell
Goblet cell
Paneth cell
Fibroblast
Fabp6, Fabp2
Clca3b, Lgals1
Tff3, Fcgbp
Defa21, Mptx2
Dcn, Grem, Igfbp7

Fig.5. STHD is scalable to other samples, revealing subcellular cell type identities and tissue structures. a. STHD spot-level inference of a human colon cancer Visium HD sample showing tumor, colon tissue, and enriched lymphoid structure. **b.** Left, spots (2x2um) labeled as germinal center B cells, with B cell marker expression aggregated per 8x8um bins. Right, spots (2x2um) labeled as T cells, and expression of TLS-specific T cell cytokine *CXCL13* by 8x8um bins. Marker expression plots are illustrated by loupe browser. **c.** Spot-level inference of human colon VisiumHD sample P5 (cancer) and P3 (normal adjacent tissue). **d.** Immune cell-specific composition across all four human colon Visium HD samples. **e.** STHD results for the mouse small intestine VisiumHD sample, with two zoomed-in regions highlighted. **f.** Spots with immune cell identities in the lymph node-like region, showing T cell, *Ms4a1*+ B cell and *Ighd*+ B cells. Top, STHD labels in 2x2um spots, Bottom, marker gene expression in 8x8um bins. **g.** Spots of intestine cells forming layered structure, including intestine enterocyte cell, epithelial cell, goblet cell, Paneth cell, and fibroblasts. Top, STHD labels in 2x2um spots, Bottom, marker gene expression in 8x8um bins. GC, germinal center. CRC, colorectal cancer. NAT, normal adjacent tissue.

References

1. Tian L, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. Nat Biotechnol. 2023 Jun;41(6):773–82.

2. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat Biotechnol. 2021 Mar;39(3):313–9.

3. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell. 2022 May 12;185(10):1777-1792.e21.

4. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018 Jul 27;361(6400):eaat5691.

5. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science. 2018 Nov 16;362(6416):eaau5324.

6. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021 Aug;596(7871):211–20.

7. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods. 2019 Oct;16(10):987–90.

8. Cho CS, Xi J, Si Y, Park SR, Hsu JE, Kim M, et al. Microscopic examination of spatial transcriptome using Seq-Scope. Cell. 2021 Jun 24;184(13):3559-3572.e22.

9. Oliveira MF, Romero JP, Chung M, Williams S, Gottscho AD, Gupta A, et al. Characterization of immune cell populations in the tumor microenvironment of colorectal cancer using high definition spatial profiling [Internet]. bioRxiv; 2024 [cited 2024 Jul 28]. p. 2024.06.04.597233. Available from: https://www.biorxiv.org/content/10.1101/2024.06.04.597233v1

10. Wu L, Yan J, Bai Y, Chen F, Zou X, Xu J, et al. An invasive zone in human liver cancer identified by Stereo-seq promotes hepatocyte–tumor cell crosstalk, local immunosuppression and tumor progression. Cell Res. 2023 Aug;33(8):585–603.

11. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024 Feb;42(2):293–304.

12. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022 Apr;40(4):517–26.

13. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. Nat Biotechnol. 2022 Sep;40(9):1349–59.

14. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022 May;40(5):661–71.

15. Lam SK, Pitrou A, Seibert S. Numba: a LLVM-based Python JIT compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC [Internet]. New York, NY, USA: Association for Computing Machinery; 2015 [cited 2024 May 14]. p. 1–6. (LLVM '15). Available from: https://dl.acm.org/doi/10.1145/2833157.2833162

969 16. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization [Internet]. arXiv; 2017 [cited
970 2024 May 12]. Available from: http://arxiv.org/abs/1412.6980

971 17. Hartman A, Satija R. Comparative analysis of multiplexed in situ gene expression profiling
972 technologies [Internet]. bioRxiv; 2024 [cited 2024 May 12]. p. 2024.01.11.575135. Available
973 from: https://www.biorxiv.org/content/10.1101/2024.01.11.575135v2

974 18. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, et al. Spatially organized
975 multicellular immune hubs in human colorectal cancer. Cell. 2021 Sep 2;184(18):4734-
976 4752.e20.

977 19. Merlos-Suárez A, Barriga FM, Jung P, Iglesias M, Céspedes MV, Rossell D, et al. The
978 intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease
979 relapse. Cell Stem Cell. 2011 May 6;8(5):511–24.

980 20. Geras A, Darvish Shafighi S, Domżał K, Filipiuk I, Rączkowska A, Szymczak P, et al.
981 Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial
982 transcriptomics data. Genome Biology. 2023 May 17;24(1):120.

983 21. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, et
984 al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type
985 topography. Commun Biol. 2020 Oct 9;3(1):1–8.

986 22. Singhal V, Chou N, Lee J, Yue Y, Liu J, Chock WK, et al. BANKSY unifies cell typing and
987 tissue domain segmentation for scalable spatial omics data analysis. Nat Genet. 2024 Feb
988 27;1–11.

989 23. Höppener DJ, Stook JLPL, Galjart B, Nierop PMH, Nagtegaal ID, Vermeulen PB, et al. The
990 relationship between primary colorectal cancer histology and the histopathological growth
991 patterns of corresponding liver metastases. BMC Cancer. 2022 Aug 22;22(1):911.

992 24. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng PA, Guo YB, et al. Gland
993 segmentation in colon histology images: The glas challenge contest. Med Image Anal. 2017
994 Jan;35:489–502.

995 25. Ozato Y, Kojima Y, Kobayashi Y, Hisamatsu Y, Toshima T, Yonemura Y, et al. Spatial and
996 single-cell transcriptomics decipher the cellular environment containing HLA-G+ cancer cells
997 and SPP1+ macrophages in colorectal cancer. Cell Rep. 2023 Jan 31;42(1):111929.

998 26. Zhang Y, Xiang G, Jiang AY, Lynch A, Zeng Z, Wang C, et al. MetaTiME integrates single-
999 cell gene expression to characterize the meta-components of the tumor immune
1000 microenvironment. Nature Communications. 2023 May 6;14(1):2634.

1001 27. Xu L, Lin W, Wen L, Li G. Lgr5 in cancer biology: functional identification of Lgr5 in cancer
1002 progression and potential opportunities for novel therapy. Stem Cell Res Ther. 2019 Jul
1003 29;10(1):219.

1004 28. Kapur N, Mir H, Clark III CE, Krishnamurti U, Beech DJ, Lillard JW, et al. CCR6 expression
1005 in colon cancer is associated with advanced disease and supports epithelial-to-
1006 mesenchymal transition. Br J Cancer. 2016 Jun;114(12):1343–51.

1007 29. Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional
1008 atlas of tumor infiltrating myeloid cells. Cell. 2021 Feb 4;184(3):792-809.e23.

1009 30. Matusiak M, Hickey JW, van IJzendoorn DGP, Lu G, Kidzinski L, Zhu S, et al. Spatially
1010 Segregated Macrophage Populations Predict Distinct Outcomes In Colon Cancer. Cancer

1011 Discovery [Internet]. 2024 Mar 28 [cited 2024 Jun 8]; Available from:
1012 https://doi.org/10.1158/2159-8290.CD-23-1300

1013 31. Baruah P, Dumitriu IE, Malik TH, Cook HT, Dyson J, Scott D, et al. C1q enhances IFN-
1014 gamma production by antigen-specific T cells via the CD40 costimulatory pathway on
1015 dendritic cells. Blood. 2009 Apr 9;113(15):3485–93.

1016 32. Maier B, Leader AM, Chen ST, Tung N, Chang C, LeBerichel J, et al. A conserved dendritic-
1017 cell regulatory program limits antitumour immunity. Nature. 2020 Apr;580(7802):257–62.

1018 33. Kvedaraite E, Ginhoux F. Human dendritic cells in cancer. Science Immunology. 2022
1019 Apr;7(70):eabm9409.

1020 34. Puleston DJ, Baixauli F, Sanin DE, Edwards-Hicks J, Villa M, Kabat AM, et al. Polyamine
1021 metabolism is a central determinant of helper T cell lineage fidelity. Cell. 2021 Aug
1022 5;184(16):4186-4202.e20.

1023 35. You S, Li S, Zeng L, Song J, Li Z, Li W, et al. Lymphatic-localized Treg-mregDC crosstalk
1024 limits antigen trafficking and restrains anti-tumor immunity. Cancer Cell [Internet]. 2024 Jul
1025 18 [cited 2024 Jul 31];0(0). Available from: https://www.cell.com/cancer-cell/abstract/S1535-
1026 6108(24)00239-3

1027 36. Wang X, Tokheim C, Gu SS, Wang B, Tang Q, Li Y, et al. In vivo CRISPR screens identify
1028 the E3 ligase Cop1 as a modulator of macrophage infiltration and cancer immunotherapy
1029 target. Cell. 2021 Oct 14;184(21):5357-5374.e22.

1030 37. Burdziak C, Alonso-Curbelo D, Walle T, Reyes J, Barriga FM, Haviv D, et al. Epigenetic
1031 plasticity cooperates with cell-cell interactions to direct pancreatic tumorigenesis. Science.
1032 2023 May 12;380(6645):eadd5327.

1033 38. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and
1034 analysis of cell-cell communication using CellChat. Nat Commun. 2021 Feb 17;12(1):1088.

1035 39. Li J, Zhou J, Huang H, Jiang J, Zhang T, Ni C. Mature dendritic cells enriched in
1036 immunoregulatory molecules (mregDCs): A novel population in the tumour
1037 microenvironment and immunotherapy target. Clin Transl Med. 2023 Feb 17;13(2):e1199.

1038 40. Chen Z, Wang X, Jin Z, Li B, Jiang D, Wang Y, et al. Deep learning on tertiary lymphoid
1039 structures in hematoxylin-eosin predicts cancer prognosis and immunotherapy response.
1040 NPJ Precis Oncol. 2024 Mar 22;8(1):73.

1041 41. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A Single-Cell
1042 Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell
1043 Population Structure. Cell Syst. 2016 Oct 26;3(4):346-360.e4.

1044 42. Fei L, Chen H, Ma L, E W, Wang R, Fang X, et al. Systematic identification of cell-fate
1045 regulatory programs using a single-cell atlas of mouse development. Nat Genet. 2022
1046 Jul;54(7):1051–61.

1047 43. Yao Z, Velthoven CTJ van, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution
1048 transcriptomic and spatial atlas of cell types in the whole mouse brain [Internet]. bioRxiv;
1049 2023 [cited 2023 Sep 8]. p. 2023.03.06.531121. Available from:
1050 https://www.biorxiv.org/content/10.1101/2023.03.06.531121v1

1051 44. Polański K, Bartolomé-Casado R, Sarropoulos I, Xu C, England N, Jahnsen FL, et al.
1052 Bin2cell reconstructs cells from high resolution Visium HD data [Internet]. bioRxiv; 2024
1053 [cited 2024 Jul 31]. p. 2024.06.19.599766. Available from:
1054 https://www.biorxiv.org/content/10.1101/2024.06.19.599766v1

1055 45. Multiscale topology classifies cells in subcellular spatial transcriptomics | Nature [Internet].
1056 [cited 2024 Aug 12]. Available from: https://www.nature.com/articles/s41586-024-07563-1

1057 46. You Y, Fu Y, Li L, Zhang Z, Jia S, Lu S, et al. Systematic comparison of sequencing-based
1058 spatial transcriptomic methods. Nat Methods. 2024 Jul 4;1–12.

1059 47. Ni Z, Prasad A, Chen S, Halberg RB, Arkin LM, Drolet BA, et al. SpotClean adjusts for spot
1060 swapping in spatial transcriptomics data. Nat Commun. 2022 May 27;13(1):2971.

1061 48. Zhang D, Schroeder A, Yan H, Yang H, Hu J, Lee MYY, et al. Inferring super-resolution
1062 tissue architecture by integrating spatial transcriptomics with histology. Nat Biotechnol. 2024
1063 Jan 2;1–6.

1064 49. Bokeh Development Team. Bokeh: Python library for interactive visualization [Internet].
1065 2018. Available from: https://bokeh.pydata.org/en/latest/

1066 50. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set
1067 enrichment analysis in Python. Bioinformatics. 2023 Jan 1;39(1):btac757.

1068