SOTIP: a Unified Framework for Microenvironment Modelling with Spatial Omics Data

Zhiyuan Yuan

Tsinghua University

Yisi Li

Tsinghua University

Minglei Shi

Tsinghua University

Fan Yang

Tencent

Juntao Gao

Tsinghua University

Jianhua Yao

Tencent

Michael Zhang (✉ michael.zhang@utdallas.edu)

University of Texas at Dallas <https://orcid.org/0000-0002-7022-6115>

Article

Keywords:

Posted Date: January 19th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1263025/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Title: SOTIP: a Unified Framework for Microenvironment Modelling with

Spatial Omics Data

Zhiyuan Yuan^{1,2,5}, Yisi Li^{1,5}, Minglei Shi³, Fan Yang², Juntao Gao¹, Jianhua Yao^{2*}, Michael Q. Zhang^{1,3,4*}

1, MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRist; Department of Automation, Tsinghua University, Beijing 100084, China.

2, Tencent AI Lab, Shenzhen, China.

3, MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRist; Department of Basic Medical Sciences, School of Medicine, Tsinghua University, Beijing 100084, China.

4, Department of Biological Sciences, Center for Systems Biology, The University of Texas, Richardson, TX 75080-3021, USA.

5, These authors contributed equally: Zhiyuan Yuan, Yisi Li

*These authors jointly supervised this work: Jianhua Yao, Michael Q. Zhang.

J. Y. (jianhuayao@tencent.com)

M. Q. Z. (michael.zhang@utdallas.edu)

26 **Abstract**

27 The rapidly developing spatial omics techniques generate datasets with
28 diverse scales and modalities. However, most existing methods focus on
29 modeling dynamics of single cells while ignore microenvironments (MEs)
30 which bridge single cells to tissues. Here we present SOTIP (Spatial Omics
31 mulTIPIe-task analysis), a scalable framework incorporating MEs and their
32 interrelationships into a unified graph. Based on this graph, three tasks can be
33 performed, including spatial heterogeneity (SHN) quantification, spatial
34 domain (SDM) identification and differential microenvironment (DME)
35 analysis. We validate SOTIP's accuracy, robustness, interpretation, and
36 scalability on various datasets by comparing with state-of-art methods. In
37 mammalian cerebral cortex, we reveal a striking gradient SHN pattern with
38 strong correlations with the cortical depth. In human triple-negative breast
39 cancer (TNBC), we identify previously unreported molecular polarizations
40 around SOTIP-detected tumor-immune boundaries. Most importantly, we
41 discover TNBC subtype-specific MEs, which exhibit interesting compositional
42 and spatial properties, and could be powerful clinical indicators. Overall, by
43 modeling biologically explainable microenvironments, SOTIP outperforms
44 state-of-art methods and provides new perspectives for data interpretation,
45 which facilitates further understanding of spatial information on a variety of
46 biological issues.

47

Background

Tissue and cell state are jointly modulated by the intracellular gene regulatory network and extracellular ecosystem (i.e., the microenvironment). Recent studies highlighted the importance of such microenvironment in tissue homeostasis¹⁻⁴, disease occurrence⁵⁻⁷, and tumor progression⁸⁻¹⁰. For instance, in healthy vertebrate liver, the zonation patterns of hepatocytes perform mutual effect on the microenvironment within liver lobules⁴. While during liver fibrosis, there occurs metabolic alteration around the liver-fibrotic microenvironment⁷. In cancer, researchers also reported that different patterns of immune infiltration were associated with patient survival¹¹⁻¹³.

Recent years, various spatial omics technologies have been developed, including spatial transcriptomics (e.g. in situ sequencing (ISS)¹⁴, single-molecule fluorescence in situ hybridization (smFISH)¹⁵, multiplexed error-robust fluorescence in situ hybridization (MERFISH)¹⁶, sequential fluorescence in situ hybridization (seqFISH)¹⁷, spatial transcriptomics (ST)¹⁸, slide-seq^{19,20}, High-definition spatial transcriptomics (HDST)²¹, stereo-seq²²), spatial proteomics (e.g. iterative indirect immunofluorescence imaging (4i)²³, co-detection by indexing (CODEX)²⁴, multiplexed ion beam imaging by time-of-flight (MIBI-TOF)^{12,25}, imaging mass cytometry (IMC)^{13,26}), as well as spatial metabolomics (e.g. airflow-assisted desorption electrospray ionization mass spectrometry imaging (AFADESI-MSI)²⁷, SpaceM²⁸, spatial single nuclear metabolomics (SEAM)⁷). These advances provide fertile biomolecular profiles of each observation unit (viz. spots, pixel, or single cells) as well as the corresponding spatial coordination, thus facilitating deep and systematic investigation of tissue microenvironment.

Although above technologies have achieved unprecedented profiling coverage and/or spatial resolution, only a few studies focused on the quantitative description of microenvironment (ME)²⁹⁻³¹. Existing studies typically represent an ME as a vector of cell type frequency within cellular neighborhood, and use Euclidean distance among ME representations to construct a graph (e.g. KNN graph³²), which is further exploited to perform downstream analysis such as clustering (termed classical cell neighborhood representation, CCNR, Supplementary Table 1)²⁹⁻³¹. While easily implemented, this practice fails to consider the mutual relationship of cells within the ME, which may undermine the reliability of the subsequently constructed ME graph. This defect would easily hinder the downstream analytical performance especially in cases containing continuous cell states, such as liver² and cortex³³.

The complexity of the cell composition within the spatial context is reported to be associated with disease progression³⁴ and tissue development^{16,33}. Classical methods quantify the heterogeneity of an ME by calculating the number of unique cell clusters within it (termed NUCC, Supplementary Table 1)^{16,33,35}. Such practice assumes equal importance of different clusters, potentially leading to the inaccuracy of heterogeneity quantification. For example, they may fail to distinguish the heterogeneity difference between two MEs (one is composed of tumor cells and immune cells, and another is composed of two subsets of immune cells).

Partitioning tissues into spatial domains with the integration of both gene expression and spatial information is an important task in tissue biology. Existing methods can be divided into three categories based on their computational principles: Zhu et al.³⁶ and BayesSpace³⁷ (Supplementary Table 1) applied Hidden-Markov random field (HMRF) to model the spatial dependency of hidden variables associated with spots. SpaGCN³⁸ (Supplementary Table 1) utilized existing structure in the field of graph convolutional network (GCN) to aggregate gene expression features according to the spatial graph. StLearn³⁹ (Supplementary Table 1) normalized the spatial information, tissue histology and gene expression features to perform clustering. While these methods achieved good performance in applications, they may lack proper interpretation for the parameters (e.g. weight trade-off between spatial prior and conditional distributions for HMRF-based methods) or features (e.g. features learned from deep learning based methods) of biological entities. In addition, they may not be flexible enough to be compatible with different spatial omics data type.

Recent researches emphasize that tissue homeostasis is hierarchically modulated by at least three scales: tissue regions, single cells, and microenvironment as the intermediate of them^{29,31}. A single cell is easily affected by the external changes, while the microenvironment provides a more robust buffer. When the local microenvironment is destroyed, it will spread to a wide range of tissue disorder and gradually develop into disease. Algorithms for identification of changes in response to biological insult (e.g. disease, ontogeny, or experimental perturbations) have been developed based on single-cell RNA sequencing (scRNA-seq) data⁴⁰⁻⁴³. The advances of spatial omics technologies have revealed the phenomenon that tissues at different disease state exhibit similar cell type composition but distinct spatial organizations^{1,29,34}, thus urging for developing algorithms to identify differential microenvironment among different conditions.

We presented SOTIP (Fig. 1a), a scalable method to jointly perform spatial heterogeneity (SHN) quantification, spatial domain (SDM) identification, and differential microenvironment (DME) analysis within a unified and biologically

interpretable framework. With SOTIP, we proposed to use MECN (molecular-expression-aware cellular neighborhood) as an agent to abstract and mathematically represent the microenvironment. SOTIP's algorithm modules performed these three tasks based on an MECN graph (MEG), in which each node represented an MECN and the edge weight between two nodes was the function of cell states discrepancy and cell state composition within MEs. In this manner, gene expressions, single cells, microenvironments, and tissue regions could be linked in a biological meaningful way. Then SHN was calculated by measuring the gene expression variation within each node, SDM was identified by hierarchical merging nodes guided by the connectivity derived from edges, and DME analysis was performed by assessing nodes' relative densities between conditions on the graph. More details about the algorithm principles and how SOTIP works could be found in Methods.

In this manuscript, we firstly demonstrated the main idea of SOTIP and assessed the quality of MEG with simulated spatial transcriptomics data. Then we conducted exhaustive experiments on various spatial omics modalities with ground truth, to show the superior performance of SHN quantification and SDM identification compared with the state-of-art methods. With SOTIP-quantified SHN, we confirmed a gradient SHN pattern towards deeper cerebral cortical layer. We found a strong correlation between SHN values and the cortical depth on different mouse samples of independent experiment protocols. With the SOTIP-detected SDMs, we found previously unreported protein polarization regarding to the tumor-immune boundary in triple negative breast cancer (TNBC) samples. With DME analysis, a novel task not presented by current methods, we firstly validated our results by identifying known fibrotic-liver-specific MECNs with the H&E staining as positive controls, then utilized SOTIP to identify highly specific MECNs in two subtypes of TNBC, which might be associated with tumor-immune interactions according to previous research⁴⁴. This discovery showed high consistency across a 41-human patient cohort¹². Unlike most methods, SOTIP is highly scalable to a wide range of technologies, our experiments covered various types of spatial omics data⁴⁵⁻⁴⁸, including FISH-based⁴⁹ and sequencing-based⁵⁰ spatial transcriptomics, TOF mass spectrometry based spatial proteomics^{12,51}, and secondary ion mass spectrometry (SIMS) based spatial metabolomics⁷.

Results

SOTIP framework demonstration with *in silico* spatial transcriptomics data.

To demonstrate SOTIP's utility and to conduct adequate comparisons with other methods, we generated three sets of *in silico* spatial transcriptomics data by simulating scRNA-seq data, followed by arranging cell spatial coordinates (see Methods). It is worth noting that since each MECN is generated by searching the spatially nearest neighbors of each cell, in the following sections, every MECN (consisted with multiple cells within the neighborhood of a center cell) is associated with its center cell.

In the first simulation, we aimed to compare SOTIP with other methods in SHN quantification. To achieve this goal, we simulated three clusters of single cells with inter-cluster discrepancy (Fig. 1b left). Then we randomly positioned these three cell clusters as a sequential tissue band on a two dimensional plane (Fig. 1b right). With this simulation, we expected higher spatial heterogeneity around tissue boundaries than inner tissue, which corresponded to higher SHN values between C2 (middle band in Fig. 1b right) and C3 (right band in Fig. 1b right) than between C1 (left band in Fig. 1b right) and C2. We next separately performed SOTIP, IGD and NUCC^{16,33,35} (Supplementary Table 1) to quantify the spatial heterogeneity across the tissue sample. The result showed that all three methods highlighted two tissue boundaries (Fig. 1c), while only SOTIP successfully quantified the difference between two boundaries as expected (Fig. 1c right).

Since SOTIP relies on the MECN graph (MEG) to represent MECN relationships, the quality of MEG determines the performance of downstream tasks. In the second simulation, we aimed to evaluate the quality of the MECN graph constructed by SOTIP. To do this, we generated 5 single cell clusters along a continuous manifold (Fig. 1d left), and regularly mixed these 5 clusters in 5 sequential bands on a two-dimensional plane (Fig. 1d right). In this data, each cell type occupies a major component in a band, and other cell types occupy minor components in its band. The arrangement of major cell types along these tissue bands (R1 to R5) is in accordance with the order of cell types along the gene expression manifold (C1 to C5) (Fig. 1d middle). With this simulation, we expected that R1 to R5 microenvironments should display a continuous pattern when embedded into low dimensional space. We separately performed SOTIP and CCNR²⁹⁻³¹ (Supplementary Table 1) on this simulation data, and embedded their generated graphs with three popular embedding algorithms (diffusion maps⁵², UMAP⁵³ and PHATE⁵⁴, Fig. 1e). The

208 results showed that both methods can separate these 5 MECNs in the
209 embedded space (Fig. 1e), while only SOTIP well preserved the continuous
210 order of R1 to R5 with different embedding methods (Fig. 1e bottom row).

211
212 In the third simulation, we want to ask whether SOTIP can perform differential
213 microenvironment (DME) analysis to identify specific microenvironments
214 between samples, even if these samples shared exactly the same cellular
215 composition. Since there is currently no other method for this task, it was not
216 possible to compare different methods on this type of data. We firstly
217 generated three discrete single cell clusters (Fig. 1f left), then positioned them
218 in two different orders, thus generating two different in silico tissue samples
219 with exactly the same cell composition but different spatial information (Fig. 1f
220 middle and right). With this simulation, differential abundance analysis based
221 on non-spatial scRNA-seq^{40,42,43} failed to detect any difference between the
222 two samples, but SOTIP successfully highlighted (Fig. 1g) the major
223 differential MECNs (C1-C2 boundary in sample 1, and C1-C3 boundary in
224 sample 2), by estimating the relative likelihood of observing each
225 microenvironment in the two tissue samples⁴³ (see Methods).

226
227 All these simulation results supported that SOTIP accurately described the
228 microenvironments and corresponding properties, and detected finer
229 biological differences, complementing with the existing methods.

SOTIP quantifies accurate SHNs in accordance with subcellular and tissue structures.

In eukaryotic cells, the nuclear envelope compartmentalizes the DNA in nucleus and the protein synthesis machine in the cytoplasm, thus separating transcription from translation⁵⁵. The spatial separation of two fundamentally different compositions makes the nuclear envelope to be most heterogeneous region within a eukaryotic cell⁵⁶ (Fig. 2a). Iterative indirect immunofluorescence imaging (4i)²³ is a robust protocol that can achieve multiplexed protein staining at nanometer resolution, thus providing opportunities to characterize the covariance among molecular, spatial, and morphological properties of subcellular structures. Squidpy⁵⁷ provided a subset of original data, consisting of 270,876 pixels from 13 HeLa cells (~20,836 pixels per cell) with 40 protein assayed²³. The data also provided pixel-level nucleus/cytoplasm annotations. Within each cell, since the nuclear envelope is expected to display the highest heterogeneity, if we take the scaled heterogeneity as the probability of being the nuclear envelope, the obtained area under curve (AUC) given ground truth should be high for a good heterogeneity estimation method (see Methods).

For all 13 cells, we compared SOTIP with NUCC^{16,33,35} and IGD (Supplementary Table 1), the quantitative results (Fig. 2b) showed that SOTIP (median AUC 0.85) significantly outperformed NUCC (median AUC is 0.68, $p = 7.34 \times 10^{-6}$, two-sided Wilcoxon rank-sum test, $n=13$ cells), and IGD (median AUC is 0.71, $p = 1.16 \times 10^{-5}$, two-sided Wilcoxon rank-sum test, $n=13$ cells). To dive into details of above methods, we unbiasedly selected two cells as examples to qualitatively compare the computationally delineated nuclear envelopes with the ground truth (Fig. 2c,d). In cell 118, both IGD (AUC=0.66, Fig. 2c third column) and SOTIP (AUC=0.77, Fig. 2c fourth column) delineated clear curves with closed pattern, while NUCC (AUC=0.64, Fig. 2c second column) tended to output blurred curves with irregular and unsmooth pattern. When further comparing IGD and SOTIP, we found that SOTIP exhibited the highest SHN value around the nuclear envelope (white arrows in Fig. 2c fourth column) with strong specificity, but IGD failed to differentiate the relative values of SHN around different regions (white and orange arrows in Fig. 2c third column). Similar results have also been found in cell 119 (Fig. 2d).

To further confirm these results, we inspected the relationship between the spatial localization of SOTIP-computed SHN values and protein enrichment. We implemented this comparison by encoding SHN values into red channel, and targeted protein intensities into green channel, then plotting the encoded colors for each cell. Again we used cell 118 as an example, and found that

SOTIP-identified SHN values aligned well with proliferating cell nuclear antigen (PCNA), a known nuclear marker⁵⁸ (Fig. 2e). Although SOTIP successfully highlighted the region of the highest heterogeneity within a cell, we additionally asked whether SOTIP also delineated other cellular structures (e.g. orange arrow in Fig. 2c fourth column) with relatively lower heterogeneity. To answer this and to check whether SOTIP could reflect different levels of heterogeneity, we first clipped the highest SHN values, then encoded targeted protein intensities and the clipped SHN values in colors as before (Fig. 2f). It turned out that SOTIP-identified structure aligned well with ribosomal protein S6 (RPS6) (Fig. 2f), which is reported as a marker of endoplasmic reticulum (ER)⁵⁸.

To test whether SOTIP can quantify SHN values on tissue samples, we used a spatial transcriptomics dataset containing sections from *BRAF*^{V600E} melanoma zebrafish model⁵⁹. We firstly applied NUCC, IGD and SOTIP on the sample containing a clear tumor-muscle interface (Fig. 2g). The results showed that SOTIP-computed SHN values delineated the tumor-muscle interface, which was expected to be the most heterogeneous region⁵⁹, in accordance with histological annotation (Fig. 2g,h white arrow) with high specificity. On the contrary, both NUCC-computed and IGD-computed SHN values exhibited indistinguishable pattern within tissue bulk and the tumor-muscle interface (Supplementary Fig. 2g,h).

Since SOTIP involved a clustering step (i.e. Leiden⁶⁰) before MEG construction (see Methods), we wanted to ask whether it could output stable results for different clustering resolutions. To evaluate the robustness of SOTIP regarding to the clustering resolution, we firstly quantified the SHN values of the zebrafish sample (Fig. 2g,h) by SOTIP with different Leiden clustering resolutions ranging from 1 to 10. Then we computed the Pearson correlation coefficient (Pearson's *r*) between SHN values on resolution 1 with SHN values on different resolutions. Finally we plotted the computed correlation as the function of Leiden resolution (Fig. 2i). With this procedure, we compared the robustness of SOTIP, NUCC, and IGD, and showed that SOTIP have a substantially stronger robustness than the existing methods, and could output stable SHN values even with extreme over-clustering (Fig. 2i). The other two samples of the zebrafish⁵⁹ study also supported this claim (Supplementary Fig. 2i,j).

SOTIP reveals gradient spatial heterogeneity along cerebral cortical layer axis.

The laminar structure (Fig. 2j) of mammalian cerebral cortex is reported to exhibit layer-specific properties on gene expression, morphology and connectivity in single cell level⁶¹⁻⁶³. One study³³ of the recent BRAIN Initiative Cell Census Network (BICCN) program⁶³ proposed an observation that the composition complexity of cell neighborhoods increased towards deeper layers. They defined the neighborhood complexity of a cell as the number of different cell clusters present within a radius of 100 μm from the given cell (the same manner as NUCC). As our aforementioned analysis, the reliability of their statement could be compromised on different clustering resolutions (Fig. 2i, Supplementary Fig. 2i,j). To confirm and refine their conclusion, we applied SOTIP, which was proved robust and accurate in SHN quantification, on other mouse cortex datasets of independent experiment protocols.

To this end, we firstly adopted a single cell spatial transcriptomics dataset (cyclic-ouroboros single molecule fluorescence in situ hybridization, osmFISH⁴⁹) which covered the main layers of mouse somatosensory cortex with layer annotation⁴⁹. This dataset contains 5328 single cells across cell-wise annotated cortical layers (Fig. 2k) ranging from pia layer 1 (n=159), layer 2-3 medial (n=549), layer 2-3 lateral (n=254), layer 3-4 (n=131), layer 4 (n=1002), layer 5 (n=295), layer 6 (n=1015), and other regions. We applied SOTIP on this dataset to quantify each ME's SHN value and plotted the distribution of SHN values as violin-plots to display the trends along the cortical depth axis (Fig. 2m). Specifically, each violin-plot is a distribution of SOTIP-computed SHN values within a specific cortical layer (x-axis), and the x-axis conforms to the order of the cortex from shallow (viz. pia layer 1) to deep (viz. Layer 6) layer. The result visually showed a striking gradient pattern of SHN values from shallow layers to deep (Fig. 2m). To further quantify the ordinal agreement between the SHN values and layer depth, we used both ME-wise and layer-wise Spearman's rank correlation coefficients (Spearman's ρ) between them (see Methods). The results showed a strong correlation between the SHN value and the cortical depth, in both manner of ME-wise (Spearman's $\rho = 0.674, p < 1 \times 10^{-15}$) and layer-wise (Spearman's $\rho = 0.847, p = 0.016$).

To strengthen our conclusion, we additionally applied SOTIP on another independent single cell spatial transcriptomics dataset (evolution of sequential fluorescence in situ hybridization, seqFISH+¹⁷) of mouse cortex covering different layers¹⁷. This dataset contains 913 single cells from 5 spatially continuous field of views (FOVs) ranging from layer 2 to layer 6 (Supplementary Fig. 2c). We conducted the same SOTIP analysis as with

osmFISH dataset, the result also confirmed a continuous gradient pattern of SHN values along the cortical axis (Supplementary Fig. 2d), and a strong and positive correlation between the SHN value and the layer depth (ME-wise Spearman's $\rho = 0.719, p < 1 \times 10^{-15}$ and layer-wise Spearman's $\rho = 1, p < 1 \times 10^{-15}$). For osmFISH and seqFISH+, the performance of SOTIP were not disrupted by neither different over-clustering levels, nor different neighborhood sizes, being more robust than other methods, e.g., NUCC, IGD (Supplementary Fig. 2a,b,e,f).

374 SOTIP stratifies known layers in brain tissues.

375 To test the capability of SOTIP in the task of spatial domain identification, we
376 took the SpatialLIBD dataset, which provided 10x Visium (Fig. 3a for
377 illustration) spatial transcriptomics data on 12 tissue sections of human
378 dorsolateral prefrontal cortex (DLPFC)⁵⁰. For each tissue section, the original
379 research⁵⁰ also provided spot-level annotations of cortical layers as well as
380 white matter regions, which can be used as ground truth to numerally
381 evaluate performance. Since other methods³⁷⁻³⁹ have already demonstrated
382 the improved performance of integrating spatial information by comparing with
383 non-spatial clustering algorithms, e.g. k-means, Louvain⁶⁰, and SC3⁶⁴, on this
384 dataset, we compared SOTIP only with the spatial methods, including
385 stLearn³⁹, BayesSpace³⁷, and SpaGCN³⁸. Following the comparison
386 methodology of previous studies^{37,38,50}, we input to the algorithms the number
387 of clusters, and used adjusted rand index (ARI) to quantify the similarity
388 between the partition results and manual annotations.

389
390 We took one of the representative tissue sample, sample 151673 (n=1639
391 spots, shown in Figure 3b), to demonstrate the performance of above
392 methods. Visually, SOTIP identified SDM exhibited a layered pattern
393 consistent with the ground truth (Fig. 3c). Since it's difficult to reproduce the
394 original ARI (also reported by other group⁶⁵), especially for deep learning
395 based methods^{38,39}, we directly used the ARI and clustering results as
396 claimed in the original paper. SOTIP achieved the best performance with
397 ARI=0.593, followed by BayesSpace (ARI=0.550), SpaGCN (ARI=0.522) and
398 stLearn (ARI=0.333). When we applied these algorithms to all the 12 tissue
399 sections, we found that SOTIP outperformed all other methods (Fig. 3d,
400 Supplementary Fig. 3a, n=12) in terms of mean, median, bad case, and
401 consistency. Specifically, the mean ARI of SpaGCN (0.436) was very close to
402 BayesSpace (0.434), and substantially better than stLearn (0.363), while
403 SOTIP reached an even higher step than all of them (0.456). The median ARI
404 of SOTIP (0.459) was also the highest, followed by SpaGCN (0.445),
405 BayesSpace (0.418), and stLearn (0.363). Even the worst case of SOTIP
406 achieved ARI as high as 0.328, still better than BayesSpace (0.251), SpaGCN
407 (0.245), and stLearn (0.186), confirming the robustness of SOTIP. As for the
408 performance consistency, SOTIP performed best among all compared
409 methods in most cases (Supplementary Fig. 3a). SOTIP also showed
410 improved performance in 8 out of 12 (Supplementary Fig. 3b) samples
411 compared with second best method SpaGCN, which required extra
412 histological information and spatial label refinement³⁸.

413
414 Apart from the quantitative comparison above, we next discussed about
415 clustering procedure, extra information usage, and compatibility. For
416 clustering procedure, SOTIP followed a hierarchical merging procedure, and

the intermediate clustering result could be maintained and evaluated in just one-time running. While other three methods do not enjoy this benefit due to their different principles. Specifically, SpaGCN needs to initialize the cluster labels with predefined number of clusters, then applying the iterative clustering procedure to stochastically optimize parameters and cluster labels. As a probabilistic graphical model, BayesSpace also needs predefined number of clusters to formulize the model. It regarded the labels as hidden variable, then building a fully Bayesian model with Gaussian as conditional distribution and Markov random field as smooth prior. As to stLearn, it adopts either k-means or Louvain to perform clustering after generating the normalized graph adjacency matrix. Because the true number of clusters is not accessible, which is the common situation in real applications, all these three methods need to constantly re-run entire program to test different number clusters or resolution parameters.

For the extra information usage, as with BayesSpace, SOTIP does not require incorporate histological information (e.g., H&E) to obtain better performance than SpaGCN and stLearn. In the common situation, producing spatial omics data together with paired histological image needs additional efforts on experimental design and finer operations, so that many mainstream spatial omics researches would not actually provide paired or vertically adjacent histological images^{12,17,19,20,24,33,51}. What's more, the discontinuity of adjacent sections and the uncontrollable noise further complicate the utility of histological information.

For the compatibility, BayesSpace was restricted to the lattice-shaped spatial data like 10x Visium (Fig. 3a), while SOTIP and SpaGCN does not rely on the spatial arrangement of spots/cells. To demonstrate whether SOTIP surpass other methods on non-lattice-shaped spatial transcriptomics data (Fig. 3e), we examined the two algorithms on a single cell spatial transcriptomics dataset (osmFISH⁴⁹) of the mouse brain somatosensory cortex. We evaluated the performance with ARI taking the known cell-wise region annotations as the ground truth (Fig. 3f). The result showed that SOTIP (ARI=0.566, Fig. 3h) substantially outperformed SpaGCN (ARI=0.257, Fig. 3g). In addition to all the above advantages, SOTIP is suitable for spatial proteomics, which is hardly manageable for any existing methods. We will further demonstrate SOTIP utility and performance with spatial proteomics technologies in the next section.

SOTIP characterizes spatial and molecular tumor-immune organization.

After demonstrating the superior performance of SOTIP on various spatial transcriptomics datasets, we next ask whether SOTIP is also compatible with spatial proteomics datasets. Since there are currently few methods that was explicitly applied to perform spatial domain identification on spatial proteomics data, we only compared SpaGCN with SOTIP. We found two spatial proteomics datasets, of which the first was produced by MIBI-TOF technique on 41 triple negative breast cancer (TNBC) patients¹², and the second was produced by scMEP technique on 2 colorectal carcinoma (CRC) patients and 2 healthy controls⁵¹. With these datasets, we attempted to identify the tumor-immune interplay (Fig. 4a for illustration), which has been widely regarded as a prognostic indicator in tumor progression⁶⁶⁻⁶⁸.

As the original papers provided tumor/immune partition annotations for 2 samples in CRC (Fig. 4d), and 2 samples in TNBC (Fig. 4g), we used these 4 samples to evaluate the performance of different methods on stratifying tumor and immune regions. For better visualization and qualitative comparison²⁹, we plotted the detected spatial domains using Voronoi diagrams (Fig. 4b for illustration). The results (Fig. 4e,h) showed that in all datasets, SOTIP achieved better results in terms of visual coherence compared with SpaGCN, and almost perfectly matched with the originally reported region annotations (Fig. 4d,g).

Tumor and immune cells are reported to perform mutual modulation by frequent cross-talk within the tumor-immune ecosystem⁶⁹. Their intercellular communications and intracellular regulation networks generate patterned phenotypic cell identities and molecular distribution⁷⁰. Molecular polarization (a form of spatial molecular distribution pattern that certain molecule tends to differentially enriched distal/proximal to a tissue landmark) was recently reported to appear nearby pathological tissue boundaries^{7,12,59} (Fig. 4c for illustration). To show that SOTIP-detected tumor/immune regions better fit for downstream analysis, we next inspected the relationship between molecular polarization and tissue architecture. To be clear, in the following statements, when we say that a protein is polarized to/away a region, we mean that the protein tends to distributed proximally/distally to that region.

In the CRC datasets, we first classified immune cells as distal and proximal to tumor-immune boundary (see Methods, Supplementary Fig. 4a-b) using SOTIP and SpaGCN results, respectively. Next, Wilcoxon rank sum test was adopted to estimate the protein enrichment/depletion with respect to the boundary, and polarized proteins with statistical significance were detected. We measured the detection power on the basis of false-discovery rate (FDR) in both datasets to conduct a fair comparison across methods. The results

showed that SOTIP was more powerful than SpaGCN across a range of FDR cutoffs (Fig. 4f left). We also performed similar analysis on TNBC datasets (see Methods, Supplementary Fig. 4c-d), to identify molecular polarization to/away tumor (Fig. 4i left) and immune regions (Supplementary Fig. 4e-g), confirming the superior detection power of SOTIP.

Last, we showed representative (see Methods) polarized proteins detected by SOTIP in CRC (Fig. 4f right, Supplementary Fig. 4e-g), and in TNBC (Fig. 4i right) samples. Apart from consistency with the original paper (e.g. CD98, CD11c polarization in CRC immune region⁵¹, and CD45RO, CD11c polarization in TNBC immune region¹², and Keratin6, Keratin17, HLA-DR polarization in TNBC tumor region¹²), we also detected novel polarizations which had not been reported yet. For example, in TNBC patient 9, we found that CD45 was significantly polarized to the tumor region ($p = 6.88 \times 10^{-44}$, two-sided Wilcoxon rank sum test, Fig. 4i bottom right), which was also uncovered in patient 4 ($p = 1.30 \times 10^{-9}$, two-sided Wilcoxon rank sum test). CD45 plays a significant role in T cell receptor signaling pathway⁷¹, while CD45RO, an isoform of CD45, is of vital importance for the production of TNF- α and IFN- γ ⁷². TNF- α and IFN- γ are powerful in monitoring tumor proliferation and trigger cell death⁷³. TNF- α involves in complex immune response to induce tumor cell apoptosis and destroy the blood vessels which supports the tumor growth by transporting essential nutrients⁷⁴. IFN- γ , a member of type II interferon group, possibly contributes to the tumor cell apoptosis via JAK-STAT pathway in a microenvironment-dependent manner^{75,76}. The enrichment of CD45 and CD45RO around the tumor-immune boundaries (Fig 4i), which was also described in the original paper, indicated that they may function during the battle of immune tissue against the tumor tissue via multiple apoptosis-associated immune signaling pathways.

In both TNBC patient 4 and 9, CD138 was significantly polarized away the tumor-immune boundary ($p = 5.12 \times 10^{-27}$ in patient 4 and $p = 7.17 \times 10^{-52}$ in patient 9, two-sided Wilcoxon rank sum test, Fig. 4i top right and bottom right). CD138, also known as heparan sulfate (HS) proteoglycan Syndecan-1, is a key regulator responsible for inflammatory cytokines modulation^{77,78}. Cytokines have been found to play anti-tumor effects by binding to the receptors to mediate cell-cell communications, supported by both animal studies and clinical researches⁷⁹. Specifically, CD138 could upgrade the survival of mature plasma cells, which is critical for long-term humoral immune⁸⁰. This function is in accordance with our observation that CD138 expression is highly enriched in the core of immune cell cluster (Fig 4i). It indicates that activated B cells may play an active anti-cancer role in a persistent way with recruitment of other immune cells.

SOTIP recovers known differential microenvironments in cirrhotic liver.

The last task of SOTIP is differential microenvironment analysis (DMA), which aims to identify microenvironments between conditions (e.g. disease status, drug treatment, and experimental perturbation) with high specificity. This task is in analogy with the differential abundance analysis task^{40,42,43} in single cell analysis, but unavailable in spatial omics analysis. To demonstrate the utility of SOTIP on this task, and as a sanity check, we firstly applied DMA in a spatial metabolomics dataset including 2 samples from healthy and fibrotic liver (Fig. 5a for illustration). They provided paired H&E staining, which could be taken as a positive control for our detection result (Fig. 5b). The healthy liver sample (Fig. 5b left) was a section from a standard liver lobule, which is a repeating hexagonal-shaped units of liver. The fibrotic sample (Fig. 5b right) was a liver section containing both regions of fibrotic niche (red arrow) and hepatic tissue bulk. In this case, we expected that SOTIP could spot the fibrotic-specific microenvironments, e.g. the fibrotic region and tissue boundary.

To investigate on that, we firstly built an MEG of the joint MECNs from both samples, resulting in total 1,608 nodes and 237,604 edges (Supplementary Fig. 5a). The complete high-resolution graph was too complicated to be read and interpreted, so we next detected communities in this graph and analyzed the abstracted topological structure (see Methods). The resulted MEG abstraction (a simplified version of MEG, Fig. 5d) consisted of 16 nodes, each of which is an MECN cluster (a group of similar MECNs). The edge width can be considered as similarity between nodes. To get a global view of the spatial organization of various microenvironments within the tissue field, we plotted the clustered MECN groups on physical space with Voronoi diagram (Fig. 5e), so that MECNs from the same cluster share the same color according to the MEG abstraction (Fig. 5d). From the MECN cluster map (Fig. 5e), we observed that certain microenvironments showed consistency with the histological images (Fig. 5b right).

To select the MECN clusters with high sample-specificity and cell type complexity, we plotted each MECN cluster on an EH plot (abbreviation for “entropy-of-ME-cluster (EMC)” versus “spatial heterogeneity (SHN)” plot, see Methods). On the EH plot, MECN clusters with higher sample-specificity corresponded to lower EMC (y-axis), and MECN clusters with higher compositional complexity corresponded to higher SHN value(x-axis), so that MECN clusters at the bottom right corner should be of special interest (Fig. 5f). Based on this criterion, we selected 7 MECN clusters, and all of them were specifically occurred in the fibrotic liver sample. We first looked at MECNs with highest cell composition complexity, i.e., the MECN cluster 15 (Fig. 5f,i). The hepatocytes account for more than half of the cell counts,

587 followed by all other major cell types, fibroblast, Kupffer cell, immune cell, and
588 endothelial cell. This high level of SHN has been expected, since the
589 metabolomic profile of hepatocytes is reported to exhibit substantial difference
590 with other non-hepatic parenchymal cells (NPCs)^{2,3,7}. Given that MECN
591 cluster 0,9,11,12,14 were densely connected in the MEG abstraction (Fig. 5d),
592 we jointly plotted them in a single MECN cluster map (Fig. 5g left). The result
593 indicated that these MECNs displayed fibrotic-niche phenotype in cell type
594 composition (Fig. 5g right) and also exhibited high consistency with the
595 histological information (right arrows in Fig. 5g, Fig. 5b right) in spatial
596 localization. We also investigate MECN cluster 3 (Fig. 5h), whose cell type
597 composition (Fig. 5h bottom) showed a balanced mixture of the fibrotic niche
598 and the hepatic tissue bulk. Its spatial distribution confirmed this composition
599 by approximately delineating the hepatic-fibrotic tissue boundary (green
600 dashed line in Fig. 5b right and Fig. 5h). In summary, these analyses
601 demonstrate the unique capability of SOTIP to identify samples associated
602 microenvironments with high specificity, and interrogate the relationship
603 between the phenotypic and spatial features of them.

SOTIP identifies highly specific MECNs associated with tumor-immune interactions in subtypes of TNBC.

Triple-negative breast cancer (TNBC), defined by lack of therapeutic targets (estrogen receptor, progesterone receptor, and Her2), is considered to be a form of breast cancer with more aggression and poorer prognosis⁸¹. Keren, et al¹². applied multiplexed ion beam imaging by time-of-flight (MIBI-TOF) to generate a 36-plex spatial single cell proteomics dataset from 41 TNBC patients. In that paper, the researchers classified TNBC patients into 3 archetypical subtypes (viz. *cold*, *mixed*, and *compartmentalized*), by quantifying the mixture degree of tumor and immune cells¹². The dataset hasn't been fully mined since the original paper mainly concentrated their analysis on protein co-expression and basic spatial statistics¹², we here focused on the microenvironment to draw a spatial organization of the tumor-immune landscape and to inspect interesting microenvironment with strong specificity.

Since the “cold” samples, lack of immune infiltration, can be easily differentiated from other two subtypes. To demonstrate SOTIP's capability, we choose the more challenging case to identify differential MECNs between *mixed* (immune cells mixed with tumor cells)¹² and *compartmentalized* (immune cells spatially separated from tumor cells)¹² samples. The challenge to distinguish the *mixed* from the *compartmentalized* came from the fact that these two subtypes of TNBC samples typically contained similar cell type compositions¹², such that they could not be easily identified by current single-cell based analysis^{40,42}, or protein co-expression analysis⁶⁶. To this end, we chose the most representative samples, suggested by the original research¹², from the two subtypes of TNBC, i.e., patient 4 for *compartmentalized* and patient 12 for *mixed* (Fig. 6a,b). As with previous section, we firstly built an MEG of the joint MECNs from both samples (Supplementary Fig. 6a), then generated an MEG abstraction (Fig. 6c), consisting of 29 MECN clusters as nodes and the corresponding edges as similarities between MECN clusters. We further plotted the MECN cluster map for the two samples (Fig. 6d for *compartmentalized* and Fig. 6e for *mixed*), and selected the MECN clusters with high SHN and low EMC on EH plot (Fig. 6f). Interestingly, we identified several pronounced MECN clusters, of which MECN cluster 9 and MECN cluster 24 were strictly enriched in specific samples with disappeared EMC.

MECN cluster 9 is specifically restricted in the *compartmentalized* sample. To inspect the cellular complexity of MECN cluster 9, we found that its SHN value was fairly low, since it only consisted of two subtypes of cells, tumor cell (Tumor cluster) and its Keratin (pan-Keratin, Keratin6, Keratin17, Supplementary Fig. 6b) positive counterpart (Keratin+ tumor cluster). Since

there lacks cells of Tumor cluster in *mixed* sample (Supplementary Fig. 6c, Fig. 6g,h), it's not surprising to detect this MECN differential between samples. More interestingly, MECN cluster 24 is specifically restricted in the *mixed* sample (Fig. 6i), and is composed of a high percentage of keratin+ tumor cells, as well as a mixture of CD8 T cells and macrophages (Fig. 6i bottom). We noticed that all the three cell types occupied a substantial proportion within both *mixed* and *compartmentalized* samples (Supplementary Fig. 6c), so we reasoned that these three cell types may distribute exclusively in the *compartmentalized* sample but closely in the *mixed* sample, which was confirmed by the spatial relationships of them (Fig. 6j).

We also looked at other differential MECNs (with non-zero EMCs, but still differentiated between samples by occurrence (with EMC cutoff of 0.5)), and found that they were mostly derivative from MECN cluster 9, and 24 (Supplementary Fig. 6d-g). Specifically, for the MECN clusters mainly occurred in the *compartmentalized* sample (Supplementary Fig. 6d,e), they were dominated by the same two cell types (i.e. Tumor and Keratin+ tumor) as MECN cluster 9. Correspondingly, for those MECN clusters mainly occurring in the *mixed* samples (Supplementary Fig. 6f,g), they were dominated by the same three cell types (i.e., CD8 T, macrophage, and Keratin+ tumor) as MECN cluster 24, except for a rather minor proportion of CD4 T cells. This indicated that MECN cluster 9 (Tumor and Keratin+ tumor) and/or 24 (CD8 T, macrophage, and Keratin+ tumor) may be the main driver for the differentiation of *mixed* and *compartmentalized* forms of TNBC. To examine this statement, we evaluated the occurrence of MECN cluster 9 and 24 in all *compartmentalized* (n=15) and *mixed* (n=19) patients, and compared the occurrence score of these MECNs between two patient groups (see Methods). The results showed that the occurrence score of MECN cluster 9 (Tumor and Keratin+ tumor) displayed more consistency in *mixed* samples than in *compartmentalized* samples, with lower variance (Supplementary Fig. 6h), but there is no significant difference of the occurrence of MECN cluster 9 between the *mixed* and *compartmentalized* groups ($p = 0.986$, two-sided Wilcoxon rank-sum test, Supplementary Fig. 6h). On the contrary, MECN cluster 24 (CD8 T, macrophage, and Keratin+ tumor) significantly showed more occurrence in *mixed* samples than in *compartmentalized* samples ($p = 9.20 \times 10^{-6}$, one-sided Wilcoxon rank-sum test, Supplementary Fig. 6i).

Above analysis implied that MECN cluster 24 (CD8 T, macrophage, and Keratin+ tumor) may drive the difference between *compartmentalized* and *mixed* subtypes of TNBC. Tumor-infiltrating lymphocytes (TILs) have aroused wide interest because of its prognostic significance⁸². As an important part of TILs, CD8 T cells have been proved to be associated with lots of immune checkpoint molecules as well as the survival rate of patients^{83,84}. However, a pooled analysis⁸⁵ containing 3,771 patients showed that increased TILs were

not always a positive prognostic factor for breast cancer patients' survival. How TILs related to the prognosis of patients differed in breast cancer subtypes. For example, high-level TILs contradicted with the survival benefit in patients of luminal–HER2-negative breast cancer. On the other hand, a recent study revealed that macrophages played a role in excluding CD8 T cells from reaching the tumor cells, thus causing a poor clinical outcome⁴⁴. Combining these studies, an assumption came up that TILs should help to fight against tumor but it may be hindered by other cells in the microenvironment. Once the hindering happened, it would indicate a condition where immunocompetent hosts failed to defend against the antigenic tumor. For example, the mixture of CD8 T cells and macrophages could be regarded as a malignant signal indicating bad survival chance. This assumption was supported by our finding that MECN cluster 24 specifically occurred more in the *mixed* group (reported to be more malignancy and output worse survival¹²) than in the *compartmentalized* group across the patient cohort. Altogether, by fully integrating the spatial information and quantitatively measuring the microenvironment difference between TNBC samples, our analysis demonstrated a solid case to show the power of our ME-based computational framework. Our findings suggested a potential theory to explain why compartmentalized-pattern TNBC patients would response better than the mix-pattern ones and exhibited the potential of MECN as a powerful clinical indicator in a biologically meaningful way.

Discussion

In this study, we presented SOTIP, a unified framework to perform multiple important tasks with various spatial omics technologies. The core of SOTIP is the construction of MECN graph (MEG), of which each node represents an MECN and each edge encodes the relationship between MEs. Based on MEG, two existing tasks, namely, spatial heterogeneity (SHN) quantification and spatial domain (SDM) identification, and one novel task, namely, differential microenvironment (DME) analysis can be performed.

We conducted simulation experiments to demonstrate the utility of SOTIP, and compared the performance with different methods. We also performed exhaustive comparisons between SOTIP and state-of-art methods on different spatially resolved datasets. In the task of SHN quantification, we benchmarked SOTIP against NUCC and IGD by delineating the nuclear envelope in HeLa cell line (spatial proteomics, 4i²³) and recovering the gradient heterogeneity pattern in mouse brain cortex (spatial transcriptomics, osmFISH⁴⁹). In the task of SDM identification, we firstly showed SOTIP's superior performance compared with other spatial clustering algorithms on the commonly used SpatialLIBD dataset of human brain cortex⁵⁰ (spatial transcriptomics, 10x Visium). On another spatial transcriptomics dataset of mouse brain cortex⁴⁹, SOTIP also outperformed others in terms of ARI. We validated the compatibility of SOTIP on other spatial omics data by collecting samples from CRC (spatial proteomics, scMEP⁵¹) and TNBC (spatial proteomics, MIBI¹²). Although SpaGCN is also extendable to spatial proteomics dataset, SOTIP showed better performance when comparing with the ground truth in multiple cases (Fig. 4). Based on SOTIP detected tumor-immune boundary, we also detected proteins with previously unreported spatial polarization. In the novel task of DME analysis, we firstly compared a fibrotic liver sample with healthy liver sample⁷, then identified MECNs specifically enriched in the fibrotic sample. The spatial localization and cellular composition of the detected MECN is consistent with the histology image. We finally applied SOTIP to compare the MECN differences between *compartmentalized* and *mixed* subtypes of triple negative breast cancer (TNBC), consistent with previously reported tumor-immune interactions and revealing possible theory to explain the driven factor for the differentiation of these two subtypes.

Apart from the aforementioned performance advantages and novel discoveries, there are other several points that make SOTIP a distinct method from others. As for interpretability, deep learning-based algorithms are typically considered as black boxes and hard to interpret the result and parameters. On the contrary, each component of SOTIP can be easily

mapped back to a biological entity, which is beneficial for model diagnosis and biological interpretation. As for clustering procedure, current methods³⁷⁻³⁹ need multiple runs to search for the optimal result when the true number of clusters is not accessible, while SOTIP maintains the intermediate result for every level during the one-time hierarchical merging process. As for scalability, unlike methods whose applications may be limited to certain range of modalities, SOTIP can be freely generalized to different spatial techniques.

The core of SOTIP, MEG, was designed to only require distance/similarity matrix, without full accessibility of the coordination of the two spaces (i.e., physical space and molecular expression space). Several variants of SOTIP can be quickly implemented with different configurations of the two spaces. For example, SOTIP can be used to characterize microenvironment of other spatial coordination, for example 3D (given 3D spatially resolved techniques), and virtually inferred spatial affinity^{86,87}. For another example, SOTIP can use the physical space of the high resolution histological images and the molecule expression space of the high throughput scRNA-seq, to integrate their benefits. With the most advanced multi-modal biotechnologies^{88,89}, which can obtain multi-omics measurements within single cells, the physical space of SOTIP can be even replaced with protein expression, chromatin accessibility or metabolites profiles.

One limitation is that SOTIP defines MECNs with predefined shape and size, which might lead to false negative discoveries. The high complexity and diversity of microenvironments in different tissue states and disease progressions make it intricate to fully understand the mechanistic properties, for example, the shapes and sizes of MEs, while SOTIP's rigid manner of MECN definition reduces the searching space of MEs, since those MEs falling out of the beforehand definition (e.g. MEs with non-spherical shapes) could be overlooked. This problem could especially stand out when the real applications need investigations of finer resolutions, for example when comparing the MECN differences between an early stage of disease (with fairly subtle alterations) and the healthy control. To alleviate this problem, two lines of methods can be considered. One is to formalize the geometric and connectomic properties as priors to be incorporated into MECN modelling. This approach might be hindered by the divergence of MEs. Another one is to expend the searching space by considering as many shapes and sizes and combinations of MEs. This approach might be troubled by computational efficiency, and supervisions need to be properly used to guide the searching process. We believe that methods belonging to this niche is essential for future research.

Methods

MEG construction

SOTIP takes spatial omics data (gene expression/protein/metabolomic profiles of spots/cells/pixels) as input. For ease of explanation, we will use spatial single cell transcriptomics data (formed as gene expression matrix and 2D spatial coordinates for each cell) to illustrate the method, which can be extended to other spatial omics modalities without loss of generality.

The core of SOTIP is the definition of the MECN graph (MEG). With the MEG, the relationships between every pair of MECNs are encoded by considering both the occurrence, and the gene expression dissimilarities of constituted single cell identities. In this way, gene expressions, single cells, microenvironments, and tissue regions could be linked in a biological meaningful way.

To describe MEG, its nodes and edges should be defined. Specifically, each node is defined as an MECN, which is a bag of cells encoded by cell clusters within the spatial neighborhood. Then each MECN is represented by counting the frequency of each cell cluster (generated by user-defined clustering algorithm, e.g. Leiden) within the MECN, resulting a histogram. In this way, for a spatial omics dataset of a tissue sample, the number of MECNs is the same as the number of cells, and every MECN are represented by a vector with the length of number of cell clusters.

To define the edge between two nodes in MEG, SOTIP computes earth mover's distance (EMD)⁹⁰ between them. Considering the cell discrepancy within high-dimensional gene expression space, we derive a connectivity guided minimum graph distance (CGMGD) as the ground distance which approximates the geodesic distance along the low-dimensional manifold. We elaborate CGMGD 's better performance compared with other choices in "Connectivity guided minimum graph distance (CGMGD)" of Methods section. With simulation data, we also demonstrate the better quality of MEG constructed based on CGMGD than based on an isotropic ground distance (IGD), which assumes equal distances among cell groups (Fig. 1d). After computing the pairwise distance, SOTIP estimates the connectivity between nodes using an efficient neighbor search similar with UMAP^{32,91}, and the MEG is then constructed.

We next formally define MEG and associated terms. Different with general application of earth mover's distance, in the same biological sample, SOTIP

represents every MECNs using single cells with the same configuration of clusters. In a spatial omics sample with n cells and t cell clusters, we can formulize the definition of $MECN$ as a histogram: $MECN_p = \{(T_i, w_{T_i}^p)\}, T_i \in [1, t], p \in [1, n]$. T_i enumerates all possible cell cluster labels in the tissue sample, and $w_{T_i}^p$ is the total-normalized count of cluster T_i in $MECN_p$. The distance between two MECNs, e.g. $MECN_p$ and $MECN_q$ is defined as:

$$EMD(MECN_p, MECN_q) = \frac{\sum_{i=1}^t \sum_{j=1}^t f_{ij} CGMGD(i, j)}{\sum_{i=1}^t \sum_{j=1}^t f_{ij}}$$

Where $CGMGD(i, j)$ is precomputed between all pairs of cell clusters according to Methods section “Connectivity guided minimum graph distance (CGMGD)”. And $\sum_{i=1}^t \sum_{j=1}^t f_{ij} CGMGD(i, j)$ is minimized subject to

$$\begin{aligned} f_{ij} &\geq 0, i, j \in [1, t], \\ \sum_{i=1}^t f_{ij} &\leq w_j^{MECN_q}, i, j \in [1, t], q \in [1, n], \\ \sum_{j=1}^t f_{ij} &\leq w_i^{MECN_p}, i, j \in [1, t], p \in [1, n]. \end{aligned}$$

Spatial heterogeneity (SHN) quantification

We define the spatial heterogeneity (SHN) as a numerical property of a node in MEG, to assess the total gene expression variation within the represented MECN. For a $MECN$, suppose it consists of k single cells, that is:

$$CONSIST(MECN) = (c_1 \dots c_i \dots c_k), i \in [1, k]$$

c_i is vectorized by a high-dimensional gene expression feature. Each single cell c_i can be encoded with a categorical variable T_i , representing the belonging cell group, which is precomputed by clustering with the gene expression feature vector.

$$GROUP(c_i) = T_i, i \in [1, k], T_i \in [1, t]$$

Since we have also precomputed the pairwise CGMGD among cells, and the CGMGD between two single cells is a function of their belonging cell groups. We denote the pairwise gene expression variation (PGEV) as:

$$PGEV(c_i, c_j) = CGMGD(GROUP(c_i), GROUP(c_j)), i, j \in [1, k]$$

The SHN of a $MECN$ is computed as the total gene expression variation by summing over i and j :

$$SHN(MECN) = \sum_{c_i, c_j \in CONSIST(MECN)} PGEV(c_i, c_j)$$

When estimating SHN with an isotropic ground distance (IGD), the pairwise gene expression variation is simply replaced with:

$$PGEV_{IGD}(c_i, c_j) = IDENTITY(GROUP(c_i), GROUP(c_j)), i, j \in [1, k]$$

Where IDENTITY is an indicator function defined as:

$$IDENTITY(p, q) = \mathbf{1}_{p=q}(p, q)$$

Spatial domain (SDM) identification

The relationship among MECNs encoded by MEG considers both the cellular composition by representing each MECN as a histogram, and the discrepancy among cells described by gene expression profiles. To connect the local microenvironment with the tissue domain at a relatively broader scale, SOTIP performs the task of spatial domain identification based on the pre-constructed MEG.

Leiden community detection⁶⁰ is firstly performed on the MEG to partition MECNs into clusters, to this end the clustered result already reflect meaningful tissue organization. To reach the pre-defined number of clusters, SOTIP next performs cluster merging with a hierarchical scheme based on the connectivity between MECN clusters. Specifically, similar to modularity^{92,93}, SOTIP measures the degree of connectivity of two MECN clusters by considering the ratio between the number of inter-edges and the expected number of inter-edges under random assignment. SOTIP then iterates between two sub-steps: assess the pairwise connectivity between MECN clusters, and update MECN cluster assignment by merging MECN clusters with largest connectivity, until the number clusters reached the predefined.

Since the process is guided by the finely designed MEG, in each step the algorithm merges two most similar sub regions of tissue. After the hierarchical merging, users can get a multi-level tissue partitioning hierarchy, from which tissue regions with different resolutions can be investigated. This is particularly an advantage over other SDM identification algorithms, e.g. stLearn³⁹, bayesSpace³⁷, and SpaGCN³⁸, especially in real cases that the ground truth of number of tissue regions is not known in advance.

Differential microenvironment (DME) analysis

The aim of DME analysis is to identify specifically enriched microenvironments between samples (or conditions, disease status, drug treatment, etc.). This task is also based on the MEG construction, but from

joining MECNs of two samples. After that, Leiden⁶⁰ is used for community detection based on MEG, resulting a cluster assignment for MEs. Then an abstract version of MEG is constructed with partition-based graph abstraction (PAGA)⁹³. Each node of the resulted MEG abstraction is an MECN cluster, and the connectivity between nodes is computed as the ratio between the number of inter-edges and the expected number of inter-edges under random assignment, similar to modularity⁹².

To identify interesting MECN clusters with specific sample associated enrichment and more complex cellular composition, an EH plot (abbreviation for “entropy-of-MECN-cluster (EMC)” versus “spatial heterogeneity (SHN)” plot) is drawn for each DME analysis application. EH plot is a scatter plot, in which each point is an MECN cluster, the x-axis is defined as average spatial heterogeneity (SHN) across MECNs of the MECN cluster, and the y-axis is defined as the entropy of the MECN cluster (EMC) between the compared two samples. Suppose we want to compare the MECN differences of 2 samples: sample 0 and sample 1, the EMC (EMC_C) of an MECN cluster C (C is an MECN cluster, viz. a set of MECNs) is defined on the Bernoulli distribution (parameterized by p) modelling the probability of observing an MECN from either samples. That is:

$$EMC_C(p) = -p_C \log_2 p_C - (1 - p_C) \log_2 (1 - p_C), \text{ where } C \text{ is a MECN set}$$

The parameter of the distribution is estimated as the proportion of normalized count of MECNs ($NCME$) from sample 0, that is:

$$p_C = \frac{NCME(C, \text{sample } 0)}{NCME(C, \text{sample } 0) + NCME(C, \text{sample } 1)},$$

where $NCME$ of sample k is defined as:

$$NCME(C, \text{sample } k) = \frac{|\{MECN | MECN \in C \text{ and } SAMPLE(MECN) = \text{sample } k\}|}{|\{MECN | SAMPLE(MECN) = \text{sample } k\}|},$$

where $SAMPLE(MECN)$ is defined as which sample the MECN comes from. Interesting MECNs (strong specificity and high SHN) lies in the bottom right corner of the EH plot. Above analysis can be extended to multiple samples/conditions with multinomial distribution.

Connectivity guided minimum graph distance (CGMGD)

Computing the earth mover’s distance between microenvironments needs to define the ground distance between cell clusters. Directly computing Euclidean distance in the high dimensional gene expression space would be cursed by the dimensionality. Early single cell analysis⁹⁴ attempted to compute distance in the embedded space, e.g. t-SNE⁹⁵, or UMAP⁹¹, to perform downstream analysis such as clustering, but was proved problematic since the principle of these algorithms made it poorly preserving global

topological data structure⁹⁶⁻⁹⁸. Diffusion pseudo-time (DPT)⁹⁹ computes the geodesic distance between each data point and a given root point in a diffusion map embedded space, which seems consistent with our objective. However, since we need to compute the ground distance between cell clusters, we need to run DPT (e.g. SCANPY implementation³²) for multiple times by setting the root point to the center cell of every cell cluster. So the computational efficiency is a problem, not to mention the choice of the cluster centers. More recently, PAGA⁹³ is reported to preserve the topological data structure of the high-dimensional scRNA-seq data, and has been widely used as an official module of SCANPY³². It is also not suitable for us since it assesses the connectivity between clusters, instead of the distance we need. PHATE⁵⁴ is reported to preserve better distance than t-SNE and UMAP, there are also researches applying Euclidean distance within PHATE embedding to assess the distance between single cells, we found it produced errors with our simulated data.

To more accurately approximate the geodesic distance along the single cell manifold, and also accounting for computational efficiency, we propose connectivity guided minimum graph distance (CGMGD). Specifically, since we only need the cluster level distance instead of single cell level, CGMGD firstly assesses the connectivity between cell clusters with PAGA to produce a binary connectivity matrix (BCM) justifying whether there are edges (1) or not (0) between two cell clusters. In the second step, CGMGD embeds the single cells into UMAP space, and calculates the pairwise distance of cluster centers within the embedding space to produce a UMAP distance matrix (UDM). In the third step, a connectivity guided cluster graph (CGG) is constructed, of which each node is defined as a cell cluster. The edge of the graph is defined with the adjacent matrix computed by element-wise multiplication between BCM and UDM. Finally, the CGMGD is computed by searching pairwise minimum distance of the CGG. With this procedure, CGMGD approximates the geodesic distance along the single cell manifold by integrating the global topological preservation of PAGA, and the local data structure preservation of UMAP. We used simulation data 4 (see Methods) to demonstrate the advantage of CGMGD over 6 manifold learning algorithms, PCA, UMAP⁹¹, DPT⁹⁹, ForceAtlas¹⁰⁰, PHATE⁵⁴ and PAGA⁹³ (Supplementary Fig. 7).

Correlation analysis between SHN and cortical depth

To quantify the ordinal agreement between the SHN values and the cortical layer, we used both ME-wise and layer-wise Spearman's rank correlation coefficients (Spearman's ρ) between them. For ME-wise manner, the SHN was computed for each MECN, and the layer depth of each MECN was assigned by setting a number for each layer, Pia Layer 1 (1), Layer 2-3 medial

(2.5), Layer 2-3 lateral (2.5), Layer 3-4 (3.5), Layer 4 (4), Layer 5 (5), Layer 6 (6). The Spearman's ρ was computed between the SHNs and layer depths of MEs. For layer-wise manner, the SHN was computed as median SHN of MECNs for each layer, and the layer depth was assigned as with ME-wise manner. The Spearman's ρ was computed between the SHNs and layer depths of layers.

Validation of identified DMEs in all TNBC patients

To validate that our identified differential MEs, i.e., MECN cluster 9 (Tumor and Keratin+ tumor) and 24 (CD8 T, macrophage, and Keratin+ tumor) consistently differentiated between compartmentalized and mixed samples from other patients, we performed MECN occurrence analysis as follows. For ease of explanation, we used MECN cluster 9 as an example.

For each patient sample, we firstly counted the number of target MECNs (MECNs which contained cells from both Tumor and Keratin+ tumor clusters) as k_{obs} . In order to account for difference in number of cells, the expected number of target MECNs (k_{exp}) under background distribution was computed by random permutation of cell cluster labels, followed by averaging across their number of target MEs. Formally, $\bar{k}_{exp} = \frac{1}{N} \times \sum_{i=1}^N \bar{k}_i$, where N is the number of permutations, and $\bar{k}_i$ is the count of target MECNs in each permutation. The occurrence score of MECN cluster 9 is computed as $k_{obs}/\bar{k}_{exp}$. Following this procedure, we computed the occurrence score for every patient, and compare the occurrence scores between compartmentalized (N=15) and mixed (N=19) patients. The significance is estimated using one-sided Wilcoxon rank-sum test.

Simulation datasets

There are 4 simulation datasets used in this study. Simulation 1~3 were firstly generated with Splatter¹⁰¹, a R package for scRNA-seq data simulation, then arrange the spatial coordination for the generated cells. Simulation 4 is generated by Splatter as single cell data, without spatial coordination. The full information for these datasets is summarized in Supplementary Table 2. In the analysis of simulation 3 (Fig. 1f,g), the relative likelihood of observing each microenvironment in specific sample is performed by MELD⁴³, with the precomputed MECN distance matrix as input.

Public datasets

This study involved multiple datasets from 7 different spatial omics technologies. All datasets are publicly available. Human HeLa cell line 4i dataset: <https://squidpy.readthedocs.io/en/stable/>²³. Mouse brain cortex osmFISH dataset: <http://linnarssonlab.org/osmFISH>⁴⁹. Mouse brain cortex seqFISH+ dataset: <https://github.com/CaiGroup/seqFISH-PLUS>¹⁷. Zebrafish melanoma Visium dataset: GSE159709⁵⁹. Human brain cortex Visium dataset: <http://research.libd.org/g/spatialLIBD/>⁵⁰. Human colorectal carcinoma scMEP dataset: <https://zenodo.org/record/3951613>⁵¹. Human triple negative breast cancer MIBI dataset: <https://mibi-share.ionpath.com/>¹². TOF-SIMS liver dataset: https://github.com/yuanzhiyuan/SEAM/tree/master/SEAM/data/raw_tar⁷. The full information of public datasets used in this study can be found in Supplementary Table 3.

Method comparison

To evaluate the performance in spatial heterogeneity (SHN) quantification, NUCC and IGD was compared with SOTIP. As with SOTIP, NUCC, IGD both need to perform clustering and define spatial neighborhood. For a fair comparison, NUCC, IGD and SOTIP utilized exactly the same parameters (MECN size and cluster resolution). We also compared the relative performance under different combinations of these parameters, the parameters were also consistent in every comparison. With 4i dataset (Fig. 2a-f), since the original data (Supplementary Table 3) provided ground truth of nucleus and cytoplasm, we defined their boundary as width of 2 pixels. We used area under curve (AUC) to evaluate and compare the performance. With osmFISH dataset (Fig. 2g-i), the original data (Supplementary Table 3) provided cell-wise annotations of different cortical layers. We used Spearman's rank correlation coefficient to evaluate the ordinal agreement between the layer order and the computed SHN. To digitalize the layer order, we set a number for each layer, Pia Layer 1 (1), Layer 2-3 medial (2.5), Layer 2-3 lateral (2.5), Layer 3-4 (3.5), Layer 4 (4), Layer 5 (5), Layer 6 (6). To test the impact of over-clustering on SHN quantification (Supplementary Fig. 2a,e), on osmFISH data and seqFISH data, we set the clustering resolution to *moderate* resolution, over-clustering ($3 \times moderate$), and extreme over-clustering ($5 \times moderate$). The moderate resolution was determined by resolution researching given true number of clusters (provided by original paper^{17,49}).

To evaluate the performance in spatial domain (SDM) identification, for both brain cortex datasets (Fig. 3), we set the true number of clusters to run compared algorithms, and used adjusted Rand index (ARI) to compare the similarity between clustering result and ground truth. Note that for human sample, since original ground truth labels contain “N/A” (Fig. 3b), we filtered out these spots before the comparison. The ARI of the human sample was directly adopted from the ARI as they reported in the original paper^{37,38}. We also applied the same gene filtering and data normalization procedures as with code provided by SpaGCN. For the mouse cortex osmFISH dataset, the original ground truth labels contain “excluded” (Fig. 3e), we filtered out these single cells before the comparison, and the parameter of SpaGCN was set as default. As for the spatial proteomics datasets (Fig. 4), the original paper provided ground truth for tumor-immune boundaries, based on which we compared SOTIP and SpaGCN qualitatively. For better visualization²⁹, we plotted the results by Voronoi diagrams, adopted from <https://stackoverflow.com/questions/20515554/colorize-voronoi-diagram/20678647#20678647%203.18.2019>.

Parameter settings

There are three parameters for user to choose according to their applications. These parameters are set with a consistent manner except when evaluating the robustness to different combinations of parameters.

Across the manuscript, the resolution of Leiden clustering is set by searching according to the true number of cell clusters if available, otherwise 2. The parameter of MECN size (k-NN parameter) is set to 10. The number of spatial domains is set according to the ground truth. More discussions of parameter settings can be found in Supplementary Notes.

Molecular polarization with spatial proteomics data

In the CRC datasets, following the consistent manner with Hartmann et al.⁵¹, we first classified immune cells within a 51 pixels (20 μ m) radius to tumor region as proximal, and other immune cells as distal (Supplementary Fig. 4a,b). In the TNBC datasets, and also consistent with Keren et al.¹², the definition of proximity of both “tumor cells to immune region” and “immune cells to tumor region” is defined with radius of 100 pixels (39 μ m) (Supplementary Fig. 4c,d). The criterion for selection of representative polarized proteins (Fig 4.f,i, Supplementary Fig. 4g) is by ascendingly sorting proteins by FDR, and the representative proteins are overlay and plotted by the interactive plotting webserver Ionpath (<https://www.ionpath.com/>).

Boxplot

All boxplots in the manuscript share the same settings: the lower and upper hinges show the first and third quartiles (the 25th and 75th percentiles); the center lines correspond to the median; the upper whisker extends from the upper hinge to the largest value, which should be less than $1.5\times$ the interquartile range (or distance between the first and third quartiles) and the lower whisker extends from the lower hinge to the smallest value, which is at most the $1.5\times$ interquartile range. Data beyond the end of the whiskers are 'outlying' points and are plotted individually.

Quantitative and statistical analysis.

All statistical tests used in this study are described in detail in the corresponding figure legends. Spearman's rank correlation coefficients, Pearson correlation coefficient, and Wilcoxon rank-sum test are performed using `scipy`¹⁰². To account for multiple hypotheses testing, we applied the Benjamini–Hochberg (BH) procedure to report the associated FDR¹⁰³, which is performed using `pingouin`¹⁰⁴. ARI and AUC are performed using `scikit-learn`¹⁰⁵.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

An open-source Python implementation of SOTIP and reproduction code are available at <https://github.com/TencentAILabHealthcare/SOTIP>.

Acknowledgements

This research was supported by National Natural Science Foundation of China 81890994 to M.L.S., 81890991 to J.G., National Key Research and Development Program of China (2018YFA0801402) to M.L.S., Beijing Municipal Natural Science Foundation (Z200021), CAS Interdisciplinary Innovation Team (JCTD-2020-04), and the State Key Research Development

1146 Program of China (2021YFE0201100) to J.G., M.Q.Z acknowledges the
1147 support by the Cecil H. and Ida Green Distinguished Chair.

1148 **Author contributions**

1149 M.Q.Z. and Y.L. conceived and designed the project. Z.Y. and Y.L. developed
1150 and implemented the algorithms under the guidance of M.Q.Z. and J.Y.. Z.Y.
1151 and Y.L. collected and processed public datasets. Z.Y. and Y.L. conducted
1152 the data analysis and methods comparisons. Y.L. and M.S. did the biological
1153 interpretation. M.S. and J.G. gave suggestions on the applications of the
1154 method. Z.Y. and Y.L. completed the figures and manuscript with the
1155 guidance of J.Y., M.S., and M.Q.Z.. F.Y. helped with the figure generation. All
1156 authors approved the manuscript.

1158 **Competing interests**

1159 The authors declare no competing interests.

1160
1161
1162

1163 **References**

- 1164 1 Cho, C. S. *et al.* Microscopic examination of spatial transcriptome using Seq-Scope. *Cell*
1165 **184**, 3559-3572 e3522, doi:10.1016/j.cell.2021.05.010 (2021).
- 1166 2 Ben-Moshe, S. & Itzkovitz, S. Spatial heterogeneity in the mammalian liver. *Nat Rev*
1167 *Gastroenterol Hepatol*, doi:10.1038/s41575-019-0134-x (2019).
- 1168 3 Ben-Moshe, S., Shapira, Y., Moor, A. E., Bahar Halpern, K. & Itzkovitz, S. Spatial sorting
1169 enables comprehensive characterization of liver zonation. doi:10.1101/529784 (2019).
- 1170 4 Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in
1171 the mammalian liver. *Nature* **542**, 352-356, doi:10.1038/nature21065 (2017).
- 1172 5 Iannacone, M. & Guidotti, L. G. Immunobiology and pathogenesis of hepatitis B virus
1173 infection. *Nat Rev Immunol*, doi:10.1038/s41577-021-00549-4 (2021).
- 1174 6 Kisseleva, T. & Brenner, D. Molecular and cellular mechanisms of liver fibrosis and its
1175 regression. *Nat Rev Gastroenterol Hepatol* **18**, 151-166, doi:10.1038/s41575-020-00372-
1176 7 (2021).
- 1177 7 Yuan, Z. *et al.* SEAM is a spatial single nuclear metabolomics method for dissecting tissue
1178 microenvironment. *Nat. Methods* **18**, 1223-1232, doi:10.1038/s41592-021-01276-3
1179 (2021).
- 1180 8 Bergers, G. & Fendt, S. M. The metabolism of cancer cells during metastasis. *Nat Rev*
1181 *Cancer* **21**, 162-180, doi:10.1038/s41568-020-00320-2 (2021).
- 1182 9 Elia, I. *et al.* Breast cancer cells rely on environmental pyruvate to shape the metastatic
1183 niche. *Nature* **568**, 117-+, doi:10.1038/s41586-019-0977-x (2019).
- 1184 10 Li, X. *et al.* The immunological and metabolic landscape in primary and metastatic liver
1185 cancer. *Nature Reviews Cancer*, doi:10.1038/s41568-021-00383-9 (2021).
- 1186 11 Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human
1187 Squamous Cell Carcinoma. *Cell* **182**, 497-514 e422, doi:10.1016/j.cell.2020.05.039 (2020).
- 1188 12 Keren, L. *et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast
1189 Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* **174**, 1373-+,
1190 doi:10.1016/j.cell.2018.08.039 (2018).
- 1191 13 Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature*,
1192 doi:10.1038/s41586-019-1876-x (2020).
- 1193 14 Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods*
1194 **10**, 857-860, doi:10.1038/nmeth.2563 (2013).
- 1195 15 Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. Y. & Zhuang, X. W. Spatially resolved,
1196 highly multiplexed RNA profiling in single cells. *Science* **348**, doi:ARTN aaa6090
1197 10.1126/science.aaa6090 (2015).
- 1198 16 Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the
1199 hypothalamic preoptic region. *Science* **362**, doi:10.1126/science.aau5324 (2018).
- 1200 17 Eng, C. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH.
1201 *Nature* **568**, 235-239, doi:10.1038/s41586-019-1049-y (2019).
- 1202 18 Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial
1203 transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).
- 1204 19 Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide

1205 expression at high spatial resolution. *Science* **363**, 1463-+, doi:10.1126/science.aaw1219
 1206 (2019).
 1207 20 Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with
 1208 Slide-seqV2. *Nat Biotechnol*, doi:10.1038/s41587-020-0739-1 (2020).
 1209 21 Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat*
 1210 *Methods* **16**, 987-990, doi:10.1038/s41592-019-0548-y (2019).
 1211 22 Chen, A. *et al.* Stereo Seq. doi:10.1101/2021.01.17.427004 (2021).
 1212 23 Gut, G., Herrmann, M. D. & Pelkmans, L. Multiplexed protein maps link subcellular
 1213 organization to cellular states. *Science* **361**, doi:10.1126/science.aar7042 (2018).
 1214 24 Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed
 1215 Imaging. *Cell* **174**, 968-981 e915, doi:10.1016/j.cell.2018.07.010 (2018).
 1216 25 Keren, L. *et al.* MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and
 1217 tissue structure. *Sci Adv* **5**, doi:10.1126/sciadv.aax5851 (2019).
 1218 26 Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution
 1219 by mass cytometry. *Nat. Methods* **11**, 417-+, doi:10.1038/Nmeth.2869 (2014).
 1220 27 Sun, C. *et al.* Spatially resolved metabolomics to discover tumor-associated metabolic
 1221 alterations. *Proc Natl Acad Sci U S A* **116**, 52-57, doi:10.1073/pnas.1808950116 (2019).
 1222 28 Rappez, L. *et al.* SpaceM reveals metabolic states of single cells. *Nat Methods* **18**, 799-
 1223 805, doi:10.1038/s41592-021-01198-0 (2021).
 1224 29 Schürch, C. M. *et al.* Coordinated Cellular Neighborhoods Orchestrate Antitumoral
 1225 Immunity at the Colorectal Cancer Invasive Front. *Cell*, doi:10.1016/j.cell.2020.07.005
 1226 (2020).
 1227 30 Rovira-Clave, X. *et al.* Subcellular localization of biomolecules and drug distribution by
 1228 high-definition ion beam imaging. *Nat Commun* **12**, 4628, doi:10.1038/s41467-021-
 1229 24822-1 (2021).
 1230 31 Bhate, S. S., Barlow, G. L., Schürch, C. M. & Nolan, G. P. Tissue schematics map the
 1231 specialization of immune tissue motifs and their appropriation by tumors. *Cell Syst*,
 1232 doi:10.1016/j.cels.2021.09.012 (2021).
 1233 32 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data
 1234 analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).
 1235 33 Zhang, M. *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by
 1236 MERFISH. *Nature* **598**, 137-143, doi:10.1038/s41586-021-03705-x (2021).
 1237 34 Chen, H. *et al.* Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell*
 1238 *Rep* **37**, 109915, doi:10.1016/j.celrep.2021.109915 (2021).
 1239 35 Chen, H. *et al.* Dissecting Mammalian Spermatogenesis Using Spatial Transcriptomics.
 1240 *bioRxiv*, doi:10.1101/2020.10.17.343335 (2020).
 1241 36 Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G. C. Identification of spatially associated
 1242 subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization
 1243 data. *Nat Biotechnol* **36**, 1183-+, doi:10.1038/nbt.4260 (2018).
 1244 37 Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat*
 1245 *Biotechnol*, doi:10.1038/s41587-021-00935-2 (2021).
 1246 38 Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify
 1247 spatial domains and spatially variable genes by graph convolutional network. *Nat*
 1248 *Methods* **18**, 1342-1351, doi:10.1038/s41592-021-01255-8 (2021).

1249 39 Pham, D. *et al.* stLearn: integrating spatial location, tissue morphology and gene
1250 expression to find cell types, cell-cell interactions and spatial trajectories within
1251 undissociated tissues. *bioRxiv* (2020).

1252 40 Zhao, J. *et al.* Detection of differentially abundant cell subpopulations in scRNA-seq data.
1253 *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2100293118 (2021).

1254 41 Reshef, Y. A. *et al.* Co-varying neighborhood analysis identifies cell populations associated
1255 with phenotypes of interest from single-cell transcriptomics. *Nat Biotechnol*,
1256 doi:10.1038/s41587-021-01066-4 (2021).

1257 42 Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential
1258 abundance testing on single-cell data using k-nearest neighbor graphs. *Nat Biotechnol*,
1259 doi:10.1038/s41587-021-01033-z (2021).

1260 43 Burkhardt, D. B. *et al.* Quantifying the effect of experimental perturbations at single-cell
1261 resolution. *Nat Biotechnol*, doi:10.1038/s41587-020-00803-5 (2021).

1262 44 Peranzoni, E. *et al.* Macrophages impede CD8 T cells from reaching tumor cells and limit
1263 the efficacy of anti-PD-1 treatment. *Proceedings of the National Academy of Sciences*
1264 **115**, E4041-E4050 (2018).

1265 45 Lewis, S. M. *et al.* Spatial omics and multiplexed imaging to explore cancer biology. *Nat*
1266 *Methods*, doi:10.1038/s41592-021-01203-6 (2021).

1267 46 Rao, A., Barkley, D., Franca, G. S. & Yanai, I. Exploring tissue architecture using spatial
1268 transcriptomics. *Nature* **596**, 211-220, doi:10.1038/s41586-021-03634-9 (2021).

1269 47 Hickey, J. W. *et al.* Spatial mapping of protein composition and tissue organization: a
1270 primer for multiplexed antibody-based imaging. *Nat. Methods*, doi:10.1038/s41592-021-
1271 01316-y (2021).

1272 48 Nemes, P. Mass spectrometry comes of age for subcellular organelles. *Nat. Methods* **18**,
1273 1157-1158 (2021).

1274 49 Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH.
1275 *Nat. Methods* **15**, 932-+, doi:10.1038/s41592-018-0175-z (2018).

1276 50 Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human
1277 dorsolateral prefrontal cortex. *Nat Neurosci* **24**, 425-436, doi:10.1038/s41593-020-
1278 00787-0 (2021).

1279 51 Hartmann, F. J. *et al.* Single-cell metabolic profiling of human cytotoxic T cells. *Nat*
1280 *Biotechnol*, doi:10.1038/s41587-020-0651-8 (2020).

1281 52 Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell
1282 analysis of differentiation data. *Bioinformatics* **31**, 2989-2998,
1283 doi:10.1093/bioinformatics/btv325 (2015).

1284 53 Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat*
1285 *Biotechnol* **37**, 38-+, doi:10.1038/nbt.4314 (2019).

1286 54 Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data.
1287 *Nat Biotechnol* **37**, 1482-1492, doi:10.1038/s41587-019-0336-3 (2019).

1288 55 Stewart, M. Molecular mechanism of the nuclear protein import cycle. *Nat Rev Mol Cell*
1289 *Bio* **8**, 195-208 (2007).

1290 56 Rubin, R., Goldstein, L. & Ko, C. Differences between nucleus and cytoplasm in the degree
1291 of actin polymerization. *The Journal of cell biology* **77**, 698-701 (1978).

1292 57 Palla, G. *et al.* Squidpy: a scalable framework for spatial single cell analysis. *bioRxiv* (2021).

1293 58 Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356** (2017).

1294 59 Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I. & White, R. M. Spatially resolved
1295 transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat*
1296 *Commun* **12**, 6278, doi:10.1038/s41467-021-26614-z (2021).

1297 60 Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-
1298 connected communities. *Sci Rep* **9**, 1-12 (2019).

1299 61 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse
1300 cortex. *Nature* **573**, 61-68 (2019).

1301 62 Narayanan, R. T., Udvary, D. & Oberlaender, M. Cell type-specific structural organization
1302 of the six layers in rat barrel cortex. *Frontiers in neuroanatomy* **11**, 91 (2017).

1303 63 Network, B. I. C. C. A multimodal cell census and atlas of the mammalian primary motor
1304 cortex. *Nature* **598**, 86-102, doi:10.1038/s41586-021-03950-0 (2021).

1305 64 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*
1306 **14**, 483-+, doi:10.1038/Nmeth.4236 (2017).

1307 65 Fu, H. *et al.* Unsupervised Spatial Embedded Deep Representation of Spatial
1308 Transcriptomics. *bioRxiv*, doi:10.1101/2021.06.15.448542 (2021).

1309 66 Patwa, A. *et al.* Multiplexed imaging analysis of the tumor-immune microenvironment
1310 reveals predictors of outcome in triple-negative breast cancer. *Commun Biol* **4**, 852,
1311 doi:10.1038/s42003-021-02361-1 (2021).

1312 67 Leone, R. D. & Powell, J. D. Metabolism of immune cells in cancer. *Nat Rev Cancer* **20**,
1313 516-531, doi:10.1038/s41568-020-0273-y (2020).

1314 68 Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast
1315 Cancer. *Cell*, doi:10.1016/j.cell.2019.03.005 (2019).

1316 69 Martinez-Reyes, I. & Chandel, N. S. Cancer metabolism: looking forward. *Nat Rev Cancer*,
1317 doi:10.1038/s41568-021-00378-6 (2021).

1318 70 Craig, A. J., von Felden, J., Garcia-Lezana, T., Sarcognato, S. & Villanueva, A. Tumour
1319 evolution in hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol* **17**, 139-152,
1320 doi:10.1038/s41575-019-0229-4 (2020).

1321 71 Williams, A. Glycoprotein antigens of the lymphocyte surface and their purification by
1322 antibody affinity chromatography. *Handbook of experimental immunology* **22** (1986).

1323 72 Dawes, R. *et al.* Combinations of CD45 isoforms are crucial for immune function and
1324 disease. *The Journal of Immunology* **176**, 3417-3425 (2006).

1325 73 Shen, J. *et al.* Anti-cancer therapy with TNF α and IFN γ : A comprehensive review. *Cell*
1326 *proliferation* **51**, e12441 (2018).

1327 74 Carswell, E. A. *et al.* An endotoxin-induced serum factor that causes necrosis of tumors.
1328 *Proceedings of the National Academy of Sciences* **72**, 3666-3670 (1975).

1329 75 Wheelock, E. F. Interferon-like virus-inhibitor induced in human leukocytes by
1330 phytohemagglutinin. *Science* **149**, 310-311 (1965).

1331 76 Ikeda, H., Old, L. J. & Schreiber, R. D. The roles of IFN γ in protection against tumor
1332 development and cancer immunoediting. *Cytokine & growth factor reviews* **13**, 95-109
1333 (2002).

1334 77 Ibrahim, S. A. *et al.* Syndecan-1 is a novel molecular marker for triple negative
1335 inflammatory breast cancer and modulates the cancer stem cell phenotype via the IL-
1336 6/STAT3, Notch and EGFR signaling pathways. *Molecular cancer* **16**, 1-19 (2017).

1337 78 Götte, M. Syndecans in inflammation. *The FASEB Journal* **17**, 575-591 (2003).
 1338 79 Berraondo, P. *et al.* Cytokines in clinical cancer immunotherapy. *British journal of cancer*
 1339 **120**, 6-15 (2019).
 1340 80 McCarron, M. J., Park, P. W. & Fooksman, D. R. CD138 mediates selection of mature
 1341 plasma cells by regulating their survival. *Blood, The Journal of the American Society of*
 1342 *Hematology* **129**, 2749-2759 (2017).
 1343 81 Denkert, C., Liedtke, C., Tutt, A. & von Minckwitz, G. Molecular alterations in triple-
 1344 negative breast cancer—the road to new treatment strategies. *The Lancet* **389**, 2430-
 1345 2442 (2017).
 1346 82 Yu, P. & Fu, Y.-X. Tumor-infiltrating T lymphocytes: friends or foes? *Laboratory*
 1347 *investigation* **86**, 231-245 (2006).
 1348 83 Ikeda, Y. *et al.* Clinical significance of T cell clonality and expression levels of immune-
 1349 related genes in endometrial cancer. *Oncology reports* **37**, 2603-2610 (2017).
 1350 84 Li, F. *et al.* The association between CD8+ tumor-infiltrating lymphocytes and the clinical
 1351 outcome of cancer immunotherapy: A systematic review and meta-analysis.
 1352 *EClinicalMedicine* **41**, 101134 (2021).
 1353 85 Denkert, C. *et al.* Tumour-infiltrating lymphocytes and prognosis in different subtypes of
 1354 breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *The*
 1355 *lancet oncology* **19**, 40-50 (2018).
 1356 86 Ren, X. *et al.* Reconstruction of cell spatial organization from single-cell RNA sequencing
 1357 data based on ligand-receptor mediated self-assembly. *Cell Research*,
 1358 doi:10.1038/s41422-020-0353-2 (2020).
 1359 87 Nitzan, M., Karaikos, N., Friedman, N. & Rajewsky, N. Gene expression cartography.
 1360 *Nature* **576**, 132-137, doi:10.1038/s41586-019-1773-3 (2019).
 1361 88 Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene
 1362 expression and protein levels in single cells. *Nat Biotechnol* **39**, 1246-1258,
 1363 doi:10.1038/s41587-021-00927-2 (2021).
 1364 89 Chung, H. *et al.* Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat.*
 1365 *Methods* **18**, 1204-1212, doi:10.1038/s41592-021-01278-1 (2021).
 1366 90 Rubner, Y., Tomasi, C. & Guibas, L. J. The earth mover's distance as a metric for image
 1367 retrieval. *Int J Comput Vision* **40**, 99-121 (2000).
 1368 91 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection
 1369 for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
 1370 92 Newman, M. E. Modularity and community structure in networks. *Proceedings of the*
 1371 *national academy of sciences* **103**, 8577-8582 (2006).
 1372 93 Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference
 1373 through a topology preserving map of single cells. *Genome Biol* **20**, 59,
 1374 doi:10.1186/s13059-019-1663-x (2019).
 1375 94 Grun, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types.
 1376 *Nature* **525**, 251-+, doi:10.1038/nature14966 (2015).
 1377 95 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *J Mach Learn Res* **9**, 2579-2605
 1378 (2008).
 1379 96 Wattenberg, M., Viégas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1**, e2 (2016).
 1380 97 Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun*

1381 **10**, 5416, doi:10.1038/s41467-019-13056-x (2019).

1382 98 Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in
1383 both t-SNE and UMAP. *Nat Biotechnol* **39**, 156-157, doi:10.1038/s41587-020-00809-z
1384 (2021).

1385 99 Haghverdi, L., Buttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime
1386 robustly reconstructs lineage branching. *Nat Methods* **13**, 845-848,
1387 doi:10.1038/nmeth.3971 (2016).

1388 100 Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph
1389 Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *Plos*
1390 *One* **9**, doi:ARTN e98679
1391 10.1371/journal.pone.0098679 (2014).

1392 101 Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing
1393 data. *Genome Biol* **18**, 174, doi:10.1186/s13059-017-1305-0 (2017).

1394 102 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python.
1395 *Nat. Methods* **17**, 261-272, doi:10.1038/s41592-019-0686-2 (2020).

1396 103 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful
1397 approach to multiple testing. *Journal of the Royal statistical society: series B*
1398 *(Methodological)* **57**, 289-300 (1995).

1399 104 Vallat, R. Pingouin: statistics in Python. *Journal of Open Source Software* **3**, 1026 (2018).

1400 105 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine*
1401 *Learning research* **12**, 2825-2830 (2011).

Figures

Figure 2

Figure 2

SOTIP accurately quantifies spatial heterogeneity.

Figure 3

Figure 3

SOTIP stratifies known layers in brain tissues.

Figure 4

Figure 4

SOTIP applications on two spatial proteomics datasets.

Figure 5

Figure 5

SOTIP recovers differential microenvironments between liver samples.

Figure 6

Figure 6

SOTIP identifies differential microenvironments between subtypes of TNBC.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SOTIPsuppFigure.pdf](#)
- [SOTIPsuppTablesAndNotes.docx](#)