# STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing

Dongqing Sun[1], Yihan Xiao[1], Zhaoyang Liu[1], Taiwen Li[3], Qiu Wu[2], Chenfei Wang[1,2]

[1] Translational Medical Center for Stem Cell Therapy, Tongji Hospital, Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China

[2] Clinical and Translational Research Center of Shanghai First Maternity and Infant Hospital, Frontier Science Center for Stem Cells, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[3] State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, Chinese Academy of Medical Sciences Research Unit of Oral Carcinogenesis and Management, West China Hospital of Stomatology, Sichuan University, Chengdu, Sichuan 610041, China

Corresponding authors: qiu_wu@tongji.edu.cn, 08chenfeiwang@tongji.edu.cn

**Abstract**

The recent advances in spatial transcriptomics have brought unprecedented opportunities to understand the cellular heterogeneity in the spatial context. However, the current limitations of spatial technologies hamper the exploration of cellular localizations and interactions at single-cell level. Here, we present spatial transcriptomics deconvolution by topic modeling (STRIDE), a computational method to decompose cell-types from spatial mixtures by leveraging topic profiles trained from single-cell transcriptomics. STRIDE accurately estimated the cell-type proportions and showed balanced specificity and sensitivity compared to existing methods. We demonstrate STRIDE's utility by applying it to different spatial platforms and biological systems. Deconvolution by STRIDE not only mapped rare cell-types to spatial locations but also improved the identification of spatial localized genes and domains. Moreover, topics discovered by STRIDE were associated with cell-type-specific functions, and could be further used to integrate successive sections and reconstruct the three-dimensional architecture of tissues. Taken together, STRIDE is a versatile and extensible tool for integrated analysis of spatial and single-cell transcriptomics and is publicly available at https://github.com/DongqingSun96/STRIDE.

## Introduction

The rapid development of high-throughput single-cell sequencing technologies[1-3] allows the investigation of cellular heterogeneity and the specificity of gene expression at an unprecedented resolution. However, recent reports indicate that cellular heterogeneity is not only modulated by the intracellular regulatory network, but also influenced by the extracellular microenvironment[4,5]. Traditional bulk and single-cell RNA-sequencing (scRNA-seq) require the dissociation of tissues or the isolation of single cells, resulting in the loss of cell positions and their proximities to each other. The recent advances in spatial transcriptomics have enabled the measurement of gene expression while preserving the spatial information[6-10], which provide great opportunities to investigate the cellular heterogeneity[11], cell-cell communication[12], and the interplay between each other[13] in the spatial context.

Despite the existing applications of spatial transcriptomics on different tissues[12,14] and disease models[15,16], there remain many computational challenges. The cell-type assignment is one of the most important issues to be addressed, considering that other downstream analyses such as the detection of spatially restricted gene expression or cell-cell interactions highly rely on correct cell-type annotations. Many *in situ* capturing technologies, such as spatial transcriptomics (ST) and 10X Genomics Visium utilized spots with a diameter of 55~100μm to record mRNA positions, and thus might cover multiple homogeneous or heterogeneous cells (1~20 cells). Other high-resolution methods such as Slide-seq[9] and HDST[17] could improve the resolution to 1~2 cells. However, the fixed beads or wells used in those techniques cannot match the cell boundaries perfectly, resulting in the overlap with several cells. As a consequence, the gene expression measured at one spatial location (i.e., spot or bead) can be regarded as a mixture of multiple cells. The within-spot heterogeneity greatly affects the performance of unsupervised clustering, which was commonly used in the analysis of scRNA-seq. To understand the cell-type distribution from spatial transcriptomics, some studies performed enrichment analysis based on the gene signature derived from scRNA-seq[16,18]. However, these analyses cannot perform the quantitative inference of the cell-type compositions for each spatial location.

Here, we presented STRIDE, a topic-model-based deconvolution method for spatial transcriptomics by integrating with matched scRNA-seq. STRIDE first discovers cell-type-associated topics from annotated single-cell transcriptome by performing topic modeling. Then STRIDE applies the pre-trained topic model to infer the cell-type compositions for each location of spatial transcriptomics. By using simulated spatial transcriptomics data, we validated the power of STRIDE to predict the cell-type proportions in high accuracy and sensitivity. To demonstrate the broad utility of STRIDE, we applied it to three spatial transcriptomics datasets of different tissues including mouse cerebellum, human squamous cell carcinoma (hSCC), and human developing heart. We demonstrated that the topics generated by STRIDE could accurately reflect the spatial signatures of each cell-type, improve the resolution of spatial clustering, and finally help in reconstructing the three-dimensional (3D) structures of tissue based on the integration of

multiple slides. STRIDE will be of great utility to study the cellular organizations in complex biological tissues.

## Results

### STRIDE: a topic-model-based method to deconvolve spatial transcriptomics using scRNA-seq

STRIDE is designed to deconvolve the cell-type composition of spatial transcriptomics locations by integrating scRNA-seq data from the same tissue (Figure 1). We assumed that transcriptional measurements obtained from spatial transcriptomics and scRNA-seq of the same tissue have common cell-types with similar features, and thus can be projected to a common latent space. Here, we characterized the common latent space by topic modeling, a statistical model originally used in natural language processing. The first step of our method is to infer the cell-type-associated topic profiles from scRNA-seq. Topic modelling on cell-by-gene matrix from scRNA-seq data could generate two probability distributions. One is the probability of a gene belonging to a topic (i.e., gene-by-topic distribution). In addition, topic modelling also estimates the contributions of different topics for each cell (i.e., topic-by-cell distribution), which could be summarized to the cell-type-by-topic distribution using Bayesian method. Next, the pre-trained topic model could be employed to infer the topic distributions of each spot in spatial transcriptomics. Finally, STRIDE could estimate the cell-type fractions of each spot by combining cell-type-by-topic distribution and topic-by-spot distribution.

Besides the cell-type composition deconvolution for each spot, STRIDE also provides several downstream analysis functions, including (1) signature (i.e., topic) detection and visualization, (2) spatial clustering and domain identification based on neighborhood cell populations and (3) reconstruction of three-dimensional architecture from sequential ST slides of the same tissue. The methodology details and examples are described in the followed results and method section.

### STRIDE yielded more accurate cell-type composition estimates on simulated ST data than existing tools

To evaluate the performance of STRIDE, we generated synthetic mixtures of cells with known cell-type labels. Specifically, we simulated a spatial transcriptomics dataset with 1,000 locations from a breast cancer (BRCA) scRNA-seq dataset[19]. For each location, single cells from scRNA were randomly selected and their transcriptomic profiles were combined to synthesize a mixture. The synthetic mixtures with known cell-type composition could serve as ground truth to benchmark the performance of STRIDE on decomposing cell-types.

We first validated the ability of topic modelling to discover cell-type-specific topics. We derived 28 different topics which were enriched in distinct cell-types (Figure 2a), indicating the association between topics and specific cell-types. We also performed Gene ontology (GO) function analysis on top enriched genes for each topic. GO biological process (BP) terms related to tumor

metastasis, such as cell-cell adhesion and platelet aggregation, were detected in malignant-cell-associated topic (Topic 12), whereas immune-related functions were enriched in immune-cell specific topics (Supplementary Figure 1a). In addition, the B-cell, T-cell and myeloid cell enriched topics have distinct enriched terms, suggesting that the topics generated by STRIDE could accurately separate different immune cell-types (Supplementary Figure 1a). In addition, when we validated the trained topic model on the same scRNA-seq dataset used for training, STRIDE achieved high cell-type assignment accuracy (87.13%, n = 33043 cells) (Figure 2b). Cell-types such as $CD4^+$ and $CD8^+$ T cells, which are similar in their transcriptome, accounted for most of the errors (79.51%, n = 4251 errors).

Feature selection is usually an important step to achieve a balance between resource consumption and model performance in machine learning algorithms. We then benchmarked the impact of feature gene selection on STRIDE's deconvolution performance using simulated data. We calculated the overall Pearson's correlation coefficients between estimated cell-type proportions and real proportions within each spot for evaluating the performance. Interestingly, model trained using marker genes (defined as DE genes for each cell-type) showed comparably decent performance with model trained with all genes, while model trained using highly variable genes (HVGs) or markers plus HVGs showed poor performance (Figure 2c, Supplementary Figure 1b). Therefore, STRIDE selects marker genes for model training by default, but also allows users to provide more specific cell-type signatures.

Finally, we compared the performance of STRIDE with other published cell-type deconvolution methods developed for spatial transcriptomics, including SPOTlight[20], NMFreg[9], Seurat CCA[21], RCTD[22], and cell2location[23]. STRIDE showed the highest concordance between prediction and ground truth, and RCTD and cell2location showed slightly worse consistency (Figure 2d). To further evaluate the sensitivity and specificity, we divided all simulated locations into two groups according to the presence and absence of each cell-type, and calculated root mean squared error (RMSE) within each group. STRIDE achieved a balance between sensitivity and specificity, while other methods such as CCA, RCTD and cell2location, achieved high specificity at the expense of low sensitivity (Figure 2e). We also assessed the methods' capability to distinguish different cell-types. STRIDE exhibited overall best performance among all methods (Figure 2f, Supplementary Figure 1c). All of these methods showed poor performance in discriminating three T cell subtypes (i.e., $CD4^+$ T cells, $CD8^+$ T cells and proliferating T cells), which shows highly similar expression profiles even in the scRNA-seq data. Collectively, these results demonstrated that STRIDE could estimate the proportions of diverse cell-types with high accuracy.

**Application of STRIDE on mouse cerebellum dataset revealed spatial localization patterns of cell-types and novel spatial signatures**

To demonstrate STRIDE's usability on real spatial transcriptomics data, we applied it on a mouse cerebellum section profiled by Slide-seq2[22]. The mouse cerebellum presents clearly defined cell-type layer structures, and thus can be used to assess the cell-type decomposition performance.

149 We collected a published snRNA-seq dataset as reference[24], and applied STRIDE to map
150 reference cell-type to spatial locations of Slide-seq2 image (Figure 3a). Consistent to the prior
151 literature (Figure 3b)[25], the two types of molecular layer interneurons, MLI1 and MLI2, were
152 mapped to the top and outermost layer of the cerebellar cortex (i.e., molecular layer). Bergmann
153 and Purkinje cells were colocalized to the same middle layer, Purkinje layer, while granule cells
154 were localized to the bottom layer, granular layer. The oligodendrocytes and astrocytes were
155 scattered below the granular layer (Figure 3a). Taken together, STRIDE could accurately
156 deconvolve the cell-types and reconstruct the layered structure of the mouse cerebellum.
157 Although there's no ground truth for real spatial transcriptomics data, we could potentially
158 compare the deconvolution result achieved from STRIDE with the cell-type-specific markers. We
159 found that the spatial distribution of cell-types inferred by STRIDE was generally consistent with
160 the spatial distribution of marker genes (Figure 3c). For example, the marker genes of
161 oligodendrocytes were highly expressed in the white matter underneath the gray matter of the
162 cerebellar cortex, where oligodendrocytes are highly enriched (Figure 3c). In line with
163 expectations, the markers for granule cells, Purkinje cells and molecular layer interneurons were
164 distributed from the inner layer to the middle layer, and thence to the outermost layer, respectively
165 (Figure 3c). The further correlation analysis between gene signature score and cell-type
166 proportion revealed roughly corresponding relationships, despite the confusion between
167 astrocytes and Bergmann cells, and between MLI1 and MLI2 (Supplementary Figure 2a). Since
168 Bergmann cells are one type of astrocytes, it seems reasonable that there was a high correlation
169 between these two cell-types.
170 Additionally, STRIDE could discover cell-type-related topics and underlying functions for these
171 topics (Supplementary Figure 2b). Interestingly, we noticed that the topic signatures showed
172 better spatial localization patterns compared to marker genes for molecular layers (Figure 3c).
173 We also performed GO function analysis on cell-type-specific topics and demonstrated the
174 consistency between the topics and their functions (Supplementary Figure 2c). For instance, GO
175 terms involved in myelinogenesis, such as myelination and axon ensheathment were enriched in
176 oligodendrocyte-specific topics (Topic 8, 18 and 30), while biological processes participating in
177 synapse formation and synaptic signal transmission were identified in topics related to Purkinje
178 cells and molecular layer interneurons (Topic 3 and 29). Taken conjointly, these results suggested
179 that topic signatures generated by STRIDE could better describe the spatial localization pattern
180 of cell-types than their marker genes, which can be used to identify novel spatially expressed
181 genes.
182

183 **STRIDE characterized the spatial heterogeneity of tumor cells in human squamous cell**
184 **carcinoma microenvironment**

185 To further explore the application of STRIDE in complex human tissues, we applied it on a hSCC
186 dataset[15] generated using the ST technology. The scRNA-seq data from matched patients were
187 analyzed to decipher the cell-type populations, and subsequently used as a reference to integrate

with ST data. Deconvolution using STRIDE could explicitly resolve the cell-type compositions for spatial locations, thereby delineating the complex tumor microenvironment (Figure 4a). Interestingly, the deconvolution analysis by STRIDE revealed a fibrovascular niche where endothelial cells and fibroblasts were enriched, and another immune-cell-enriched leading-edge region, which were in accordance with the findings from the original study (Figure 4a). To better define spatial domains within the tissue section, we performed spatial clustering analysis combining both the neighborhood information and the cell-type deconvolution result. Locations with similar cell-type compositions and similar surrounding cell populations were clustered together (see Method for more details). This identified six spatial domains, each of which had different cell-type proportions (Figure 4b-c). Cluster C4 and C2 corresponded to the fibrovascular niche and the immune-related leading-edge mentioned above, respectively, while C3 were dominated by epithelial/malignant cells (Figure 4c). We also observed that malignant C3 domain was surrounded by C4 and C1, consistent with the histological examination of the tissue (Supplementary Figure 3a). Taken together, the cell-type fraction deconvolved by STRIDE could be used to identify the spatial domains by integrating neighborhood information.

The tumor cells usually displayed high heterogeneity, we next asked whether the spatial domains identified by STRIDE could be used to investigate the potential relationship between tumor cells' heterogeneity and their spatial locations. We defined C3 as tumor-core, and its surrounding region (C1, C4 and C5) as tumor-edge, respectively (Figure 4d). To describe the difference between the tumor cells from these two regions, we performed differential gene expression analysis and function enrichment analysis for the up-regulated genes of each region. Surprisingly, the tumor-core and the tumor-edge region showed distinct hallmark pathways (Supplementary Figure 3b). The tumor-core region was characterized by the enrichment of estrogen response and cholesterol homeostasis pathways, which were reported to play an important role in SCC carcinogenesis[26]. By contrast, the edge region specific genes were highly enriched in interferon-involved signaling pathways, in line with the results from the previous study[15]. Interestingly, we also found cancer hallmarks such as epithelial mesenchymal transition (EMT) and angiogenesis were enriched in the edge region, suggesting these tumor cells had the ability to modulate immune response and were likely to metastasize (Supplementary Figure 3b-c). To exclude the possibility that the observation was affected by other cell-types presented in the tumor-core and the tumor-edge region, we dissected the single-cell transcriptome of epithelial cells from the spots (see Method). Epithelial cells from scRNA-seq were mapped to the spatial locations based on the similarity of topic distribution. Again, we found that pathways involved in epidermis development and keratinization were identified in epithelial cells within the tumor-core region, while epithelial cells in the tumor-edge were enriched in pathways related to inflammatory and interferon signaling response, as well as EMT associated pathways such as cell migration and angiogenesis (Figure 4e-f). In summary, the cell-type deconvolution results from STRIDE not only helped to define spatial localized domains, but also revealed the spatial heterogeneity within the same cell-type.

227 **STRIDE detected the location of rare cell-types during the development of human heart**
228 To verify the utility of STRIDE in different biological systems, we also applied it to study spatial
229 organization of organ development. A recent study presented a spatiotemporal atlas of the
230 developing human heart by combining scRNA-seq, spatial transcriptomics and *in situ* sequencing
231 (ISS)[14], which provided great resource for us to validate the application of STRIDE. We trained
232 model using scRNA-seq data collected from heart at 6.5-7 post-conception weeks (PCW)
233 (Supplementary Figure 4a), and performed cell-type deconvolution on all samples from three
234 developmental stages (4.5-5, 6.5 and 9 PCW). As expected, the atrial and ventricular
235 cardiomyocytes were predicted to be located in the upper and lower chambers across all three
236 stages (Figure 5a). The epicardial cells were also correctly mapped to the outer layer of the heart,
237 which is defined as the epicardium. Remarkably, cardiac neural crest cells and Schwann
238 progenitor cells, a rare cell-type identified by scRNA-seq, was mapped to the mesenchyme region
239 with smooth muscle cells and fibroblast-like cells enriched (Figure 5a). This spatial localization
240 was not achieved by the direct clustering of spots based on the gene expression profiles[14]. Finally,
241 the cell-type mapping by STRIDE was highly consistent with the spatial cell-type map (Figure 5b)
242 created through the integration of ISS and scRNA-seq by the original study[14].
243 We further explored the distribution of topics discovered by STRIDE. Reassuringly, most cell-
244 type-specific topics were distributed in concordance with the spatial position of their
245 corresponding cell-types (Figure 5c). The cell-type-specific topics also showed highly cell-type-
246 specific functions. For example, GO processes related to cardiac muscle development and
247 contraction were enriched in cardiomyocyte-associated topics, and processes such as
248 erythrocyte or ion homeostasis and oxidation-reduction process were enriched in erythrocyte-
249 associated topics (Supplementary Figure 4c). Taken together, these results demonstrated that
250 STRIDE could be used to infer the cell-type mixture patterns of tissues from different time points,
251 and accurately detected the position even for rare cell-types.
252
253 **STRIDE enabled 3D architecture construction of developing human heart using topics**
254 In order to further demonstrate the application of STRIDE-derived topics, we set out to explore
255 the integrative analysis across multiple samples leveraging the STRIDE deconvolution results.
256 The human heart samples for each timepoint were sectioned along the dorsal-ventral axis, which
257 can be integrated to reconstruct the 3D spatial architecture. We modified PASTE[27] and performed
258 ST slide integration based on STRIDE-derived topics, which could balance both transcriptional
259 similarity and spatial structure similarity through a balancing parameter (Figure 6a,
260 Supplementary Figure 5a, see Methods for details). Spots with similar topic profiles were mapped
261 pairwise. For example, spots mainly containing ventricular cardiomyocytes were mapped to each
262 other within the ventricle region, and the atrial cardiomyocytes in the right and left atrium were
263 correctly mapped within a local range (Figure 6a).
264 We then constructed a 3D model representation of the developing human heart (Figure 6b) on
265 the basis of sequential pairwise alignment of adjacent samples. Our 3D model could accurately

represent the 3D structure only using the spatial expression information without the need for image-based registration. Interestingly, when we focused on the SMCs and erythrocytes, which are the main components of vessels, we observed that these cells started from the OFT, and clearly formed four branches into right atrium (RA), right ventricle (RV), left ventricle (LV), left atrium (LA) (Figure 6c, Supplementary Figure 5b), which might correspond to the vena cava, pulmonary artery, aorta, and pulmonary veins, respectively. In addition, we could clearly observe the location of epicardial cells in the outermost layer of the heart from the 3D model (Figure 6c). Taken together, the 3D reconstruction from multiple slides based on both the topic distribution similarity and the spatial distance similarity enabled a global view of the developing human heart, and is also applicable for other systems.

## Discussion

The recent development of spatial transcriptomics has brought new insights into the understanding of tissue architectures. However, due to the limitations of current technologies, it is difficult to accurately map cell-types to spatial locations at single-cell resolution through ST data alone. To address this issue, we introduce STRIDE, a deconvolution method which utilizes topic modelling to decipher cell-type compositions of spatial transcriptomics by integrating with scRNA-seq data. Benchmarking STRIDE on different feature genes demonstrated that STRIDE could accurately estimate the cell-type proportions, and outperformed other existing methods in terms of the balance between specificity and sensitivity. By applying STRIDE on different scenarios, we confirmed its versatility in different biological systems and flexibility to various spatial transcriptomics techniques. We also demonstrated the broad application of STRIDE in the derivation of spatially localized cell-type signatures, identification of spatial domains, and final reconstruction of 3D tissue architecture based on cell-type-specific topics.

In spite of the aforementioned merits, STRIDE still has some limitations. First of all, the deconvolution by STRIDE depends on the assumption that the cell populations and their proportions are similar and robust between spatial transcriptomics and scRNA-seq. In fact, different sampling strategies and sites make a great difference to the cell-type compositions. It's almost impossible to estimate the fractions of cell-types which are present in spatial data but absent in scRNA-seq data. The issue can be mitigated by either controlling the consistency of sampling sites, or selecting closely matched reference samples. Secondly, STRIDE showed reduced performance in predicting the fractions of transcriptionally similar cell subtypes due to the lack of sufficient subtype-specific marker genes, although all of the other methods also failed. Future implementation of hierarchical topic modelling[28] might discover hierarchical topic structures, thus enhancing the capability to discriminate similar cell-types from the same lineage. The principal objective of spatial transcriptomics is to understand the cellular heterogeneity and interplay in the spatial context. Many methods have been developed to infer the ligand-receptor interactions from spatial transcriptomics through the co-expression or co-localization analysis

within the physically proximate locations[29]. Future deployment of STRIDE could consider to estimate deconvolved cell-type expression level based on scRNA-seq, which could assist in the effective inference of interactions between different cell-types. The coupling of cell-type distribution and intercellular interactions will facilitate the discovery of principles for the spatial organization of cells.

STRIDE could deconvolve the cell-type compositions of spatial transcriptomics based on latent topics. On the other hand, the shared topics by single-cell and spatial transcriptomics could also be used to map single cells to spatial locations. In this way, single-cell multi-omics data, such as scNOMe-seq[30], scNMT-seq[31], could be integrated with spatial transcriptomics through scRNA-seq to uncover the spatial regulatory mechanism. And with the accumulation of spatio-temporal transcriptomics and regulatory profiles, STRIDE could be further enhanced to elucidate the spatio-temporal multi-omics dynamics during the tissue development or the tumor progression. Besides, most of the current spatial techniques quantify the gene expression and infer cell-type distribution in the two-dimensional space. Though the topic-based integration of multiple spatial transcriptomics samples proved to be helpful in the 3D reconstruction of tissue architectures, we anticipate that with the aid of imaging data and data from other modalities, it will be possible to establish a more comprehensive and multi-scale 3D tissue atlas by STRIDE.

## Method

### Implementation of STRIDE

- Discovery of topics from scRNA-seq by latent Dirichlet allocation (LDA)

Topic modelling is frequently used in the field of natural language processing to discover latent sematic structures, referred as topics, from a mass of collected documents. Latent Dirichlet allocation (LDA), a generative probabilistic model, is one of the most popular topic modelling methods[32]. Here we utilize LDA for the discovery of functional topics from scRNA-seq data. In our model, given that there are $M$ cells and $V$ genes in scRNA-seq data, for the $m$th cell, the occurrences of topics are assumed to follow a multinomial distribution

$$Z_{\cdot m} \sim \mathrm{Multinomial}(\theta_m),$$

where $Z_{\cdot m} = (Z_{1m}, Z_{2m}, \cdots, Z_{Km})$, represents the occurrences of all $K$ topics in the $m$th cell, and $\theta_m$ is a $K$-dimensional parameter vector, which follows a Dirichlet distribution

$$\theta_m \sim \mathrm{Dirichlet}(\alpha),$$

where $\alpha$ is a symmetric $K$-dimensional hyper-parameter, which denotes the prior weights of $K$ topics in the $m$th cell. For the $k$th topic, the occurrences of genes are also multinomial-distributed

$$W_{k\cdot} \sim \mathrm{Multinomial}(\varphi_k),$$

where $W_{k\cdot} = (W_{k1}, W_{k2}, \cdots, W_{kV})$, represents the number of occurrences of all $V$ genes in the $k$th topic, and $\varphi_k$ is a $V$-dimensional parameter vector, which follows a Dirichlet distribution

$$\varphi_k \sim \mathrm{Dirichlet}(\beta),$$

where $\beta$ is a $V$-dimensional hyper-parameter, which denotes the prior weight of $V$ genes in the $k$th topic. In STRIDE, an online variational Bayes (VB) algorithm[33] is applied to estimate the parameters and infer the gene-by-topic distribution and the topic-by-cell distribution, which is implemented by the python library Genism[34]. With the output topic-cell distribution and user-provided cell-type labels, we can calculate the topic-by-cell-type distribution

$$T_{kn} = \frac{Z_{km}A_{mn}}{\sum_{i=1}^{M} A_{in}},$$

where $T_{kn}$ represents the probability that the $n$th cell-type has the $k$th topic, $Z_{km}$ represents the probability that the $m$th cell has the $k$th topic, and $A_{mn} = 1\ or\ 0$ means the $m$th cell belongs to the $n$th cell-type or not. According to Bayes' Theorem, the cell-type-by-topic distribution can be inferred as

$$C_{nk} = \frac{T_{kn}Q_n}{P_k},$$

where $C_{nk}$ represents the probability that the $k$th topic exists in the $n$th cell-type, $Q_n$ represents the probability of being the $n$th cell-type (i.e., the percentage of the $n$th type of cells), and $P_k$ represents the probability of the $k$th topic. Then, $SC_{nm}$, the probability that the $m$th cell belongs to the $n$th cell-type can be calculated as

$$SC_{nm} = \sum_{k=1}^{K} C_{nk}Z_{km}.$$

And each cell would be assigned to the cell-type with the largest probability. Accordingly, the cell-type prediction accuracy of single cells could be used to assess the topic model (Figure 2b).

- Selecting the optimal topic number

The number of topics makes a difference to the trained topic models. Theoretically, the identified topics should be representative and can distinguish different types of cells. To select the optimal topic number, STRIDE will run several models with different topic numbers. And the optimal topic number is selected based on the highest accuracy. If multiple models have the highest accuracy, STRIDE will select the simplest model (i.e., the model with the smallest number of topics).

- ST cell-type deconvolution

The gene expression measurements at a location in ST can be regarded as a mixture of multiple cells of different cell-types. No matter what technology we use to characterize the cells, the latent topic structures (i.e., the gene-by-topic distribution) should be the same. With the gene-topic distribution derived from the first step, LDA could further estimate the topic distributions for each location in ST. That is,

$$Y_{\cdot j} = (Y_{1j}, Y_{1j}, \cdots, Y_{Kj}),$$

where $Y_{\cdot j}$ is the distribution of all $K$ topics in the $j$th location. As in the single cell, the probability that the $j$th location belongs to the $n$th cell-type can be calculated as

$$ST_{nj} = \sum_{k=1}^{K} C_{nk}Y_{kj}.$$

From another perspective, $ST_{nj}$ could be viewed as the proportion of the $n$th type of cells at the $j$th location.

**Evaluating STRIDE's performance using simulated ST data**

- Simulating ST data from BRCA scRNA-seq

To evaluate the deconvolution performance, we generated a ST dataset from a breast cancer (BRCA) scRNA-seq dataset[19]. Each spot in the simulated ST dataset was a mixture of 2-10 cells randomly picked from scRNA-seq. The transcriptome profiles of selected cells were aggregated to represent the spot's expression profile. To better mimic the real capture, if the synthetic spot had over 25,000 UMI counts, it would be down-sampled to 20,000 UMI counts. Since the cell-type labels of selected cells are known, the resulting cell-type proportions can be regarded as the golden standard to evaluate the deconvolution performance.

- Function analysis on topics

LDA could discover latent topics from scRNA-seq, and each topic is composed of different genes with different probabilities. To ensure that the discovered topic profile would be able to characterize the cells, we performed functional enrichment analysis on the topics. For each topic, the top 200 genes with the highest probabilities were selected to perform gene ontology (GO) analysis by using clusterProfiler v3.18.1[35]. For function analysis in mouse cerebellum dataset, biological process (BP) terms were form org.Mm.eg.db v3.12.0[36]. In the human heart dataset, the top 50 genes for each topic were used to characterize the topics' functions by enrichment analysis on the GO BP terms provided by org.Hs.eg.db v3.12.0[37].

- Performance evaluation (metrices)

As simulated data provided known cell-type composition, we used two metrices, correlation and root mean squared error (RMSE), to assess the deconvolution performance. For each spot, Pearson's correlation between the predicted cell-type proportions and the ground truth could be calculated to measure each spot's deconvolution accuracy. We could also calculate Pearson's correlation for each cell-type to measure the performance in distinguishing different cell-types. In addition, we used RMSE to evaluate the sensitivity and specificity. For each given cell-type, according to the presence or absence of this cell-type, all the spots were divided into two groups. RMSE was then calculated for each spot in the two groups separately.

In the case when methods are applied to real ST data, there's no golden standard to compare the prediction results with the truth. Instead, correlation between cell-type fractions and signature score was calculated to evaluate the performance. For each cell-type, the top 50 specific marker genes, which were derived from *FindAllMarkers* function of Seurat v4.0.1, were defined as signature genes. The gene signature score was calculated for each spot, based on cell-type-specific signature genes through *AddModuleScore* function of Seurat.

415

416 • Benchmarking different gene sets

417 To select the appropriate features for model training, we benchmarked the impact of different sets
418 of genes. Since highly variable genes (HVGs) are useful for cell clustering in scRNA-seq analysis,
419 we took HVGs into account. In general, all genes, maker genes, HVGs, and combination of marker
420 genes and HVGs were used as the input for STRIDE separately. For each gene set, the optimal
421 number of topics was selected according to the accuracy of cell-type assignment for single cells
422 (Supplementary Figure 1b). Spot-level Pearson's correlation was utilized to evaluate the
423 performance of models with different gene sets.

424

425 • Benchmarking different deconvolution methods

426 We used the simulated ST data to compare the performance of STRIDE with other ST
427 deconvolution tools, including SPOTlight v0.1.6, NMFreg, Seurat v4.0.1 CCA, RCTD v1.2.0 and
428 Cell2location v0.5. Each published method was run with all the parameters set as default values
429 following their documentations. Methods were evaluated using correlation and RMSE as
430 described in 'Performance evaluation' section.

431

432 **Application of STRIDE on Slide-seq V2 mouse cerebellum data**

433 • Collection of mouse cerebellum single-cell and spatial transcriptomics data

434 Due to the well-studied structure and complex cell-type composition, mouse cerebellum can be
435 used as a model tissue to assess the cell-type deconvolution performance naturally. Slide-seq V2
436 mouse cerebellum data was obtained from a previous study[22]. The single-nucleus mouse
437 cerebellum data collected by snRNA-seq protocol was derived from another transcriptomic atlas
438 study[24], which provided detailed cell-type annotation information.

439

440 • Cell-type deconvolution of mouse cerebellum data

441 The single-cell and spatial gene count matrices were scaled by the UMI counts in each cell or
442 spot to avoid the impact of different cell-types. Then we performed unit-variance normalization by
443 gene to standardize the expression levels of different genes. The top 500 differentially expressed
444 (DE) genes of each cell-type were selected as feature genes to run STRIDE. According to the
445 prediction accuracy of single cells in scRNA-seq, the model with 40 topics was selected to perform
446 deconvolution on spatial data. Given that the resolution of Slide-seq V2 is up to 1~ 2 cells, each
447 pixel in spatial data was assigned the cell-type with the largest probability. To reveal the
448 anatomical features clearly, the granule cell region was cropped out from the whole slide as the
449 original study did. And only the 7 of 19 most common cell-types were displayed in Figure 3.

450

451 • Topic and marker visualization
452 STRIDE could discover latent topics and estimate the topic profile of spatial locations through
453 LDA. The topics were proved to be associated with specific cell-types and have cell-type-related
454 functions by GO analysis. To visualize the spatial distribution of topics, we selected four cell-types,
455 oligodendrocytes, granule cells, Purkinje cells, and molecular layer interneurons. For each cell-
456 type, the probability distribution of representative topics was displayed. If a cell-type had multiple
457 associated topics, the sum of these topics was used to represent the cell-type. For comparison,
458 the distribution of marker genes was visualized. The top 100 DE genes of each cell-type were
459 defined as marker genes. The sum of the expression of marker genes was calculated and used
460 as the cell-type's signature score.
461
462 **Application of STRIDE on human squamous cell carcinoma ST dataset**
463 • Collection and preprocessing of SCC scRNA-seq and ST data
464 The human SCC ST dataset was obtained from a published study on tumor microenvironment.
465 The study also provided matched scRNA-seq data of the 10X Genomics Chromium platform. In
466 our study, we only used the sample, replicate 2 of patient 2. The scRNA-seq data of patient 2 was
467 clustered and annotated using the markers provided by the original publication. As one patient
468 only had limited number of cells, the myeloid cells were classified into two clusters, DC and
469 Mono/Macro (monocytes and macrophages).
470
471 • Cell-type deconvolution of human SCC ST data
472 To run STRIDE, the single-cell and spatial gene count matrices were scaled and normalized in
473 the way described in the mouse cerebellum dataset. The top 500 marker genes were identified
474 for each cell-type in scRNA-seq, and were used to train the topic models. The model with 18
475 topics was selected automatically to be the optimal one. The cell-type deconvolution result of ST
476 was visualized by the scatter pie plot.
477
478 • Deconvolution-based spatial clustering
479 The cell-type deconvolution could provide high-resolution view of cell-type composition. Spots
480 with similar cell-type composition are assumed to belong to the same cluster. Combined with the
481 additional spatial information provided by ST, spatial domains could be further identified. To
482 identify the spatial domains, k-means clustering was performed based on both the spot's own
483 cell-type composition and its surrounding cell populations. In the case of SCC, equal weights were
484 given to the two factors, and k was set to 6 to run k-means clustering. C3 cluster was defined as
485 tumor-core region, and spots surrounding the C3 cluster were defined as the tumor-edge region,
486 which mainly consisted of C1, C4 and C5 spots.
487

488 • Deriving high-resolution gene expressions
489 STRIDE could estimate the topic profiles for both scRNA-seq and spatial transcriptomics. Then
490 the correspondence between single cells in scRNA-seq and spots in ST could be constructed
491 according to the similarity of topic distributions. Here we calculated Pearson's correlation to
492 measure the similarity. For each spot in ST, 25 most similar cells from scRNA-seq were selected
493 to represent the spot's cell composition. With the known cellular identities of single cells, the gene
494 expression in each spot can be dissected into cell-type-specific expression.
495
496 • Function analysis of epithelial cells in different clusters
497 The spot-matched single cells selected from the last steps could be pooled together to study the
498 cellular heterogeneity with high resolution. To compare the difference of epithelial cells in different
499 regions, we extracted epithelial cells predicted to be located in tumor-core and tumor-edge for
500 differential analysis. In order to characterize the functions of region-specific epithelial cells, we
501 performed enrichment analysis on hallmark pathways and GO BP terms collected from the
502 Molecular Signatures Database (MSigDB v7.1)[38] and org.Hs.eg.db v3.12.0, respectively.
503
504 **Application of STRIDE on developing human heart ST dataset**
505 • Collection and preprocessing of human heart scRNA-seq and ST data
506 The developing human heart data were derived from a previous study[14]. The study provided 4, 9
507 and 6 heart sections from three timepoints, 4.5-6, 6.5, and 9 PCW, respectively, which were
508 profiled by ST technology. They also provided a 6.5 PCW cell-type map with single-cell resolution
509 constructed by smFISH. The single-cell transcriptional profiles of the 6.5 PCW tissue sample were
510 generated through 10X Genomics Chromium platform. The scRNA-seq data was clustered and
511 annotated using the markers provided by the original study. Notably, three fibroblast-like cell
512 clusters were combined into a large fibroblast-like cluster in our study.
513
514 • Cell-type deconvolution of human heart ST data
515 To run STRIDE on the human heart dataset, the top 100 DE genes of each cluster compared to
516 all other cells were identified by Wilcoxon test. Because epicardium-derived cells, smooth muscle
517 cells, and fibroblast-like cells are transcriptionally similar, the marker genes of each of the three
518 cell-types were identified by comparing with the other two. Markers genes for atrial and ventricular
519 cardiomyocytes were discovered in the same way. The scaled and normalized count matrices
520 with marker genes were utilized to train the topic models. The model with 28 topics was further
521 employed to deconvolve the ST data. The deconvolution results of sample 3, 9 and 16 were
522 visualized by the scatter pie plot.
523

3D model construction of human heart

- Slide alignment based on STRIDE-derived topics and spatial structure

The human heart samples for each timepoint were sectioned along the dorsal-ventral axis, which can be used to reconstruct the spatial architecture of developing human hearts. Here we modified a published method PASTE[27], which utilized Fused Gromov-Wasserstein Optimal Transport (FGW-OT) algorithm, to integrate multiple tissue layers from a ST experiment. The original PASTE aligns and integrates adjacent layers based on the transcriptional and spatial similarity. Here we made a little change to the goal problem that we expected to find a mapping between spots from two layers which had similar topic profiles and preserved spatial relationship. The problem was formulated by finding a mapping $\Pi \in \mathbb{R}_+^{n\times n'}$ that minimized the following transport cost function:

$$F(\Pi; X, D, X', D', c, \alpha) = (1-\alpha)\sum_{i,j} c\left(X_{\cdot i}, X'_{\cdot j}\right)\Pi_{ij} + \alpha \sum_{i,j,k,l}(D_{ik} - D'_{jl})^2 \Pi_{ij}\Pi_{kl}.$$

$(X, D)$ and $(X', D')$ denote two layers of one ST experiment, where $X \in \mathbb{R}_+^{p\times n}$ and $X' \in \mathbb{R}_+^{p\times n'}$ represent topic profiles of the $p$ topics in $n$ and $n'$ spots, and $D \in \mathbb{R}_+^{n\times n}$ and $D' \in \mathbb{R}_+^{n'\times n'}$ represent the spot distance matrices within each layer. The parameter $\alpha \in [0, 1]$ is used to weigh up the topic distribution similarity and spatial distance similarity. If $\alpha = 0$, only topic distribution similarity is taken into consideration, and vice versa. The problem was solved by Python Optimal Transport (POT) library.

- 3D reconstruction of PCW6.5 human heart

To reconstruct the 3D structure of the 6.5 PCW human heart, we sequentially aligned multiple adjacent ST layers. For a series of sequential layers $\left(X^{(1)}, D^{(1)}\right), \cdots, (X^{(9)}, D^{(9)})$ from the dorsal to the ventral, we found the mapping $\Pi^{(k)}$ $(k = 1, \cdots, 8)$ between the $k$th and $(k+1)$th layer. Based on the mappings, we should project all tissue layers to the same spatial coordinate system. This is a generalized weighted Procrustes problem. That is, to find a translation vector $\hat{v} \in \mathbb{R}^2$ and a rotation matrix $\hat{R} \in \mathbb{R}^{2\times 2}$ $(R^T R = 1)$ that minimize the weighted distances between matched spots

$$\hat{R}, \hat{v} = \min \sum_{i,j} \Pi_{ij}^{(k)} \parallel Z_{\cdot i}^{(k)} - R Z_{\cdot j}^{(k+1)} - v \parallel^2,$$

where $Z^{(k)} \in \mathbb{R}^{2\times n}$ represents the spot coordinate matrix of the $k$th layer. The problem was solved by using SVD (see PASTE[27] for more details). For Figure 6a, only three tissue sections, sample 8, 9 and 10 were displayed to visualize the alignment and integration result. For Figure 6b, all nine ST slides from 6.5PCW were displayed to provide a general view of the reconstructed 3D structure of the human heart.

**Data availability**

The mouse cerebellum dataset including spatial transcriptomics and snRNA-seq is available at https://singlecell.broadinstitute.org/single_cell/study/SCP948. The SCC dataset can be accessed

from Gene Expression Omnibus (GEO) through the accession number GSE144240. The human heart dataset is collected from https://www.spatialresearch.org.

**Code availability**

STRIDE is an open-source python package with source code freely available at https://github.com/DongqingSun96/STRIDE.

**Acknowledgements**

This work was supported by the National Natural Science Foundation of China [31801059, 32000561, 81972551]. Shanghai Rising Star Program [21QA1408200]. Natural Science Foundation of Science [21ZR1467600]. The authors thank the Bioinformatics Supercomputer Center of Tongji University for offering computing resource.

**Author Contributions**

C.W. and Q.W. conceived the project. D.S. developed the algorithm, wrote the workflow and benchmarked the performance. Y.X. and Z. L. helped in data processing. T.L offered suggestions to the workflow. D.S., Q.W. and C.W wrote the manuscript with help from all other authors. C.W supervised the whole project. All authors read and approved the final manuscript.

**Competing interests**

The authors declare no competing financial interests.

580 **Reference**


581 1 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**,
582 377-382, doi:10.1038/nmeth.1315 (2009).
583 2 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells
584 Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).
585 3 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic
586 stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).
587 4 Lin, G. L. & Hankenson, K. D. Integration of BMP, Wnt, and notch signaling pathways in
588 osteoblast differentiation. *J Cell Biochem* **112**, 3491-3501, doi:10.1002/jcb.23287 (2011).
589 5 Junttila, M. R. & de Sauvage, F. J. Influence of tumour micro-environment heterogeneity
590 on therapeutic response. *Nature* **501**, 346-354, doi:10.1038/nature12626 (2013).
591 6 Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial
592 transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).
593 7 Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional
594 states. *Science* **361**, doi:10.1126/science.aat5691 (2018).
595 8 Eng, C. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH.
596 *Nature* **568**, 235-239, doi:10.1038/s41586-019-1049-y (2019).
597 9 Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide
598 expression at high spatial resolution. *Science* **363**, 1463-1467,
599 doi:10.1126/science.aaw1219 (2019).
600 10 Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding
601 in Tissue. *Cell* **183**, 1665-1681 e1618, doi:10.1016/j.cell.2020.10.026 (2020).
602 11 Halpern, K. B. *et al.* Erratum: Single-cell spatial reconstruction reveals global division of
603 labour in the mammalian liver. *Nature* **543**, 742, doi:10.1038/nature21729 (2017).
604 12 Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at
605 single-cell resolution. *Cell* **184**, 810-826 e823, doi:10.1016/j.cell.2020.12.016 (2021).
606 13 Baccin, C. *et al.* Combined single-cell and spatial transcriptomics reveal the molecular,
607 cellular and spatial bone marrow niche organization. *Nat Cell Biol* **22**, 38-48,
608 doi:10.1038/s41556-019-0439-6 (2020).
609 14 Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the
610 Developing Human Heart. *Cell* **179**, 1647-1660 e1619, doi:10.1016/j.cell.2019.11.025
611 (2019).

612 15 Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human
613 Squamous Cell Carcinoma. *Cell* **182**, 497-514 e422, doi:10.1016/j.cell.2020.05.039
614 (2020).
615 16 Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell
616 RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat*
617 *Biotechnol* **38**, 333-342, doi:10.1038/s41587-019-0392-8 (2020).
618 17 Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat*
619 *Methods* **16**, 987-990, doi:10.1038/s41592-019-0548-y (2019).
620 18 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-
621 cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).
622 19 Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment
623 revealed by single-cell profiling. *Cell Res* **30**, 745-762, doi:10.1038/s41422-020-0355-0
624 (2020).
625 20 Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF
626 regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.
627 *Nucleic Acids Res* **49**, e50, doi:10.1093/nar/gkab043 (2021).
628 21 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821,
629 doi:10.1016/j.cell.2019.05.031 (2019).
630 22 Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics.
631 *Nat Biotechnol*, doi:10.1038/s41587-021-00830-w (2021).
632 23 Kleshchevnikov, V. *et al.* Comprehensive mapping of tissue cell architecture via integrated
633 single cell and spatial transcriptomics. *bioRxiv*, 2020.2011.2015.378125,
634 doi:10.1101/2020.11.15.378125 (2020).
635 24 Kozareva, V. *et al.* A transcriptomic atlas of the mouse cerebellum reveals regional
636 specializations and novel cell types. *bioRxiv*, 2020.2003.2004.976407,
637 doi:10.1101/2020.03.04.976407 (2020).
638 25 Consalez, G. G., Goldowitz, D., Casoni, F. & Hawkes, R. Origins, Development, and
639 Compartmentation of the Granule Cells of the Cerebellum. *Front Neural Circuits* **14**,
640 611841, doi:10.3389/fncir.2020.611841 (2020).
641 26 Mancuso, M. *et al.* Modulation of basal and squamous cell carcinoma by endogenous
642 estrogen in mouse models of skin cancer. *Carcinogenesis* **30**, 340-347,
643 doi:10.1093/carcin/bgn243 (2009).
644 27 Zeira, R., Land, M. & Raphael, B. J. Alignment and Integration of Spatial Transcriptomics
645 Data. *bioRxiv*, 2021.2003.2016.435604, doi:10.1101/2021.03.16.435604 (2021).

28 Blei, D. M., Griffiths, T. L., Jor Da N, M. I. & Tenenbaum, J. B. Hierarchical Topic Models and the Nested Chinese Restaurant Process. *Advances in neural information processing systems* **16** (2004).

29 Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* **22**, 78, doi:10.1186/s13059-021-02286-2 (2021).

30 Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6**, doi:10.7554/eLife.23203 (2017).

31 Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**, 781, doi:10.1038/s41467-018-03149-4 (2018).

32 Blei *et al.* Latent Dirichlet Allocation. *Journal of Machine Learning Research* (2003).

33 Hoffman, M., Blei, D. M. & Bach, F. R. in *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*

34 Ehek, R. & Sojka, P. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.*

35 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

36 Carlson, M. org.Mm.eg.db: Genome wide annotation for Mouse. (2020).

37 Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. (2020).

38 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

## Figures

### Figure 1 Schematic representation of the STRIDE workflow

First, STRIDE estimates the gene-by-topic distribution and the topic-by-cell distribution from scRNA-seq. The topic-by-cell distribution is then summarized to the cell-type-by-topic distribution by Bayes' Theorem. Next, the pre-trained topic model is applied to infer the topic distributions of each location in spatial transcriptomics. By combining cell-type-by-topic distribution and topic-by-location distribution, the cell-type fractions of each spatial location could be inferred. STRIDE also provides several downstream analysis functions, including signature detection and visualization, spatial domain identification and reconstruction of spatial architecture from sequential ST slides of the same tissue.

### Figure 2 Benchmarking STRIDE's performance using simulated data

a. The cell-type-by-topic distribution estimated by STRIDE. The color represents the probability that one topic exists in one given cell-type.

b. Validation of the trained topic model on the scRNA-seq used for training. The confusion matrix reflects the consistency between the prediction and the truth for the scRNA-seq data. The value represents the number of cells that belong to one cell-type and are predicted to all different cell-types. The color represents the proportion of cells belonging to the cell-type on the y axis and classified as the cell-type on the x axis.

c. Benchmark of STRIDE's performance on different gene sets. The box plot reflects the distribution of Pearson's correlation calculated between the predicted cell-type proportion and the ground truth for each spot.

d. Benchmark of STRIDE's accuracy against different deconvolution methods. The box plot reflects the overall distribution of Pearson's correlation calculated in each spot for each method.

e. Benchmark of STRIDE's sensitivity and specificity against different deconvolution methods. All the simulated locations were divided into two groups according to the presence (blue) and absence (pink) of each cell-type, and RMSE was calculated within each group. The box plot reflects the distribution of RMSE in different methods.

f. Benchmark of the ability to distinguish diverse cell-types across different deconvolution methods. Pearson's correlation between the predicted proportions and the ground truth was calculated for cell-type. The black line in each column indicates the median of different cell-types' correlation for each method.

### Figure 3 Application of STRIDE on the mouse cerebellum

a. The spatial distribution of the 7 most common cell-types predicted by STRIDE. Each point represents the pixel captured by Slide-seq V2, and colors represent different cell-types.

b. The schematic of the layered structure of the mouse cerebellum. The cerebellum cortex is divided into three layers. At the top lies the molecular layer which contains two types of molecular layer interneurons, MLI1 and MLI2, along with the dendritic trees of Purkinje cells. At the middle

710 lies the Purkinje layer which contains the body of Purkinje cells and Bergmann cells. At the bottom
711 lies the granular layer which contains the granule cells. Below the cortex is the region of white
712 matter enriched with oligodendrocytes and astrocytes.
713 c. The expression patterns of cell-type-specific marker genes for oligodendrocytes, granule cells,
714 Purkinje cells and MLIs. The color represents the summed expression values of top 100 marker
715 genes.
716 d. The distribution of cell-type-associated topics for oligodendrocytes, granule cells, Purkinje cells
717 and MLIs. The color represents the summed probability of associated topics in each pixel.
718

719 **Figure 4 Characterizing the heterogeneity of microenvironment in human squamous cell**
720 **carcinoma**

721 a. The scatter pie plot to show the spatial locations of different cell-types predicted by STRIDE.
722 Each scatter represents a spot in the ST slide. The pie chart is used to reflect the proportions of
723 different cell-types within each spot. Colors represent different cell-types.
724 b. The k-means clustering of spots based on the cell-type compositions and surrounding cell
725 populations. Colors represent the cluster labels.
726 c. The cell-type composition of each cluster. The cell-type proportions of all spots in each cluster
727 were averaged to represent the cluster's cell-type composition.
728 d. Spatial location of the tumor-edge region. Spots in the tumor-edge region are highlighted with
729 red.
730 e. Hallmark (left) and GO (right) enrichment analysis on the up-regulated genes of epithelial cells
731 in each region. The size of the dot represents the number of genes enriched in the term, and the
732 color represents the enrichment significance.
733 f. The expression profile of genes associated with different pathways in epithelial cells mapped to
734 each region. Pathways include EMT, interferon response, epidermis development and apical
735 junction.
736

737 **Figure 5 Application of STRIDE on the developing human heart**

738 a. The deconvolution result of sample 3, 9 and 16 from 4.5-5 PCW, 6.5 PCW and 9 PCW,
739 respectively. Each scatter represents a spot in the ST slide. The pie chart is used to reflect the
740 proportions of different cell-types within each spot. Colors represent different cell-types.
741 b. The spatial cell-type map created through the integration of ISS and scRNA-seq by the original
742 study.
743 c. The distribution of topics associated with atrial and ventricular cardiomyocytes, SMC or
744 fibroblast-like cells, and cardiac neural crest cells & Schwann progenitor cells. The colors
745 represent the probability of topics in each spot.
746

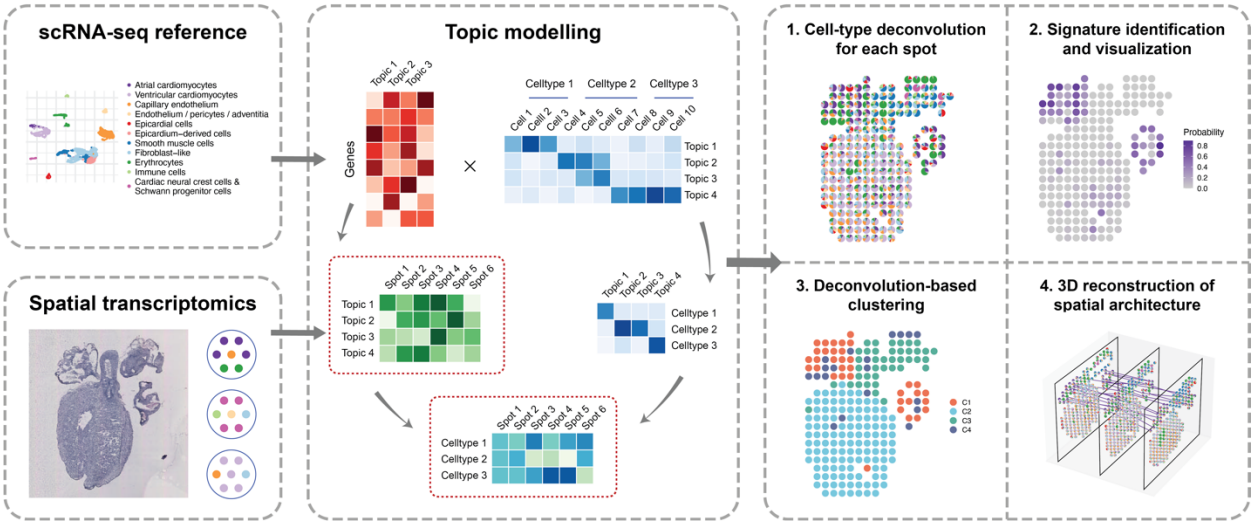
747 **Figure 6 3D model reconstruction of the developing human heart**

a. The alignment between adjacent tissue samples. Each pair of spots is connected by a line if they are matched according to the slide alignment. The left and right show the matching between spots dominated by ventricular and atrial cardiomyocytes, respectively. (The matching of other cell-types are shown in Supplementary Figure 5a.)

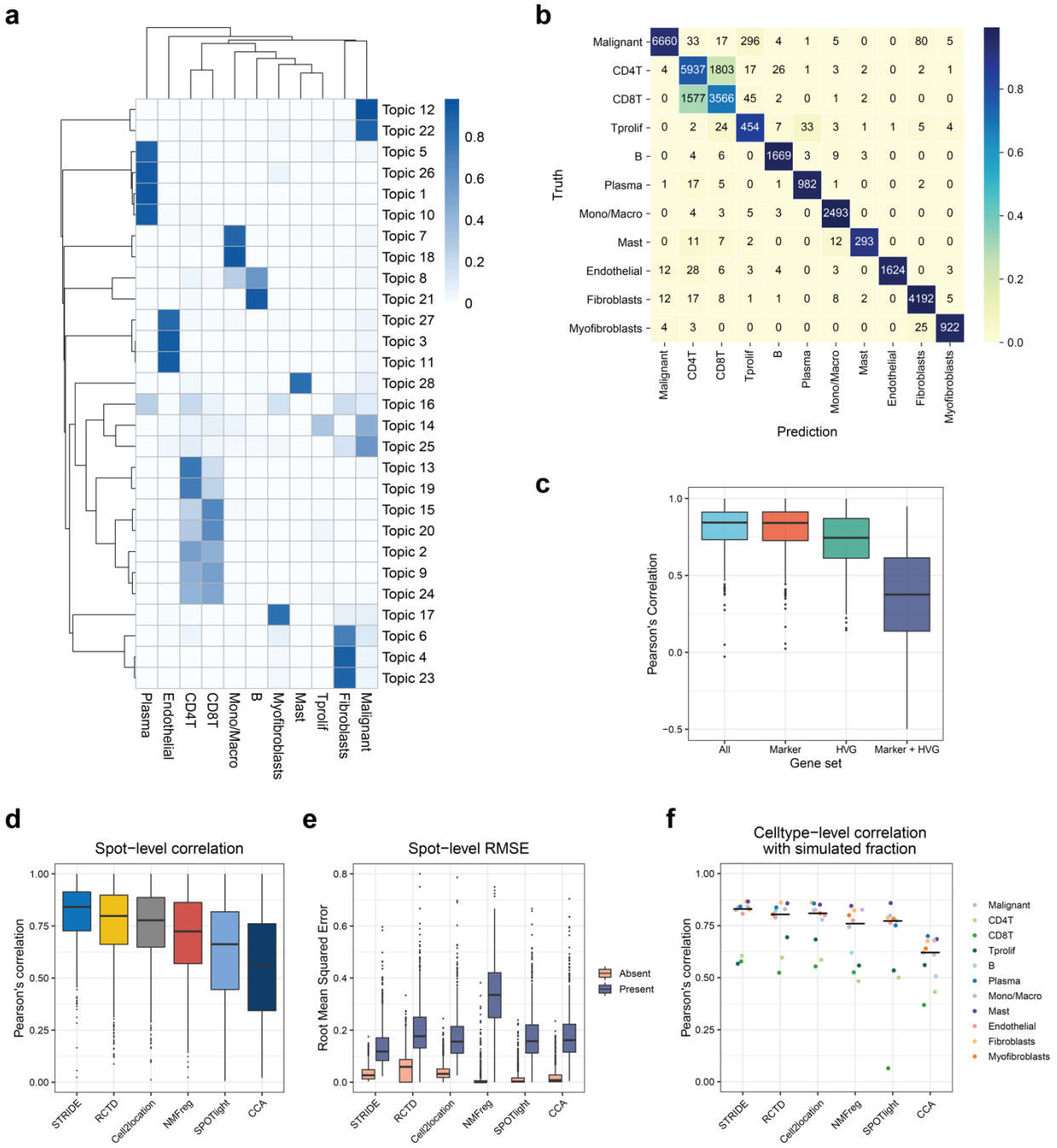
b. 3D model representation of the 6.5 PCW human heart constructed by STRIDE. 9 sequential samples from the 6.5 PCW heart were aligned and integrated together. Each sphere represents a spot in the ST slide, which is colored according to the cell-type with the highest proportion. The translucent outline shows the 3D atlas of the developing human heart (Carnegie stage 18).

c. Left, the spatial distribution of SMCs and erythrocytes. Right, the spatial distribution of epicardial cells.

## Figure 1

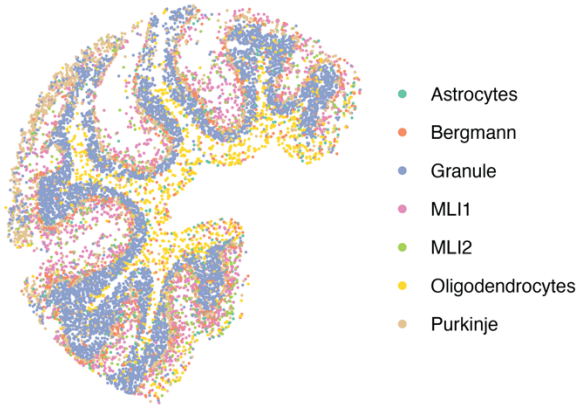

scRNA-seq reference
Atrial cardiomyocytes
Ventricular cardiomyocytes
Capillary endothelium
Endothelium / pericytes / adventitia
Epicardial cells
Epicardium–derived cells
Smooth muscle cells
Fibroblast–like
Erythrocytes
Immune cells
Cardiac neural crest cells & Schwann progenitor cells
Spatial transcriptomics
Topic modelling
Genes
Topic 1
Topic 2
Topic 3
Topic 4
Celltype 1
Celltype 2
Celltype 3
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
Cell 6
Cell 7
Cell 8
Cell 9
Cell 10
Spot 1
Spot 2
Spot 3
Spot 4
Spot 5
Spot 6
1. Cell-type deconvolution for each spot
2. Signature identification and visualization
Probability
0.8
0.6
0.4
0.2
0.0
3. Deconvolution-based clustering
C1
C2
C3
C4
4. 3D reconstruction of spatial architecture

## Figure 2

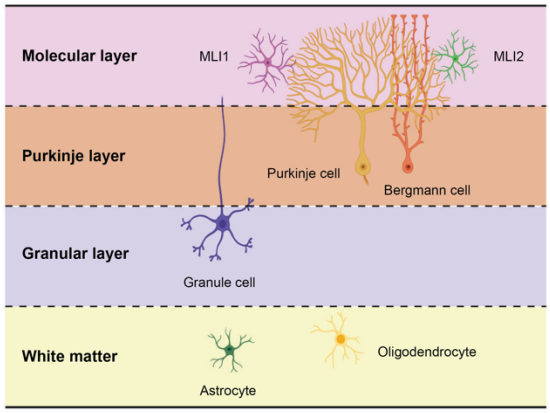

a
Topic 12
Topic 22
Topic 5
Topic 26
Topic 1
Topic 10
Topic 7
Topic 18
Topic 8
Topic 21
Topic 27
Topic 3
Topic 11
Topic 28
Topic 16
Topic 14
Topic 25
Topic 13
Topic 19
Topic 15
Topic 20
Topic 2
Topic 9
Topic 24
Topic 17
Topic 6
Topic 4
Topic 23
Plasma
Endothelial
CD4T
CD8T
Mono/Macro
B
Myofibroblasts
Mast
Tprolif
Fibroblasts
Malignant
b
Truth
Prediction
Malignant
CD4T
CD8T
Tprolif
B
Plasma
Mono/Macro
Mast
Endothelial
Fibroblasts
Myofibroblasts
c
Pearson's Correlation
Gene set
All
Marker
HVG
Marker + HVG
d
Spot−level correlation
Pearson's correlation
STRIDE
RCTD
Cell2location
NMFreg
SPOTlight
CCA
e
Spot−level RMSE
Root Mean Squared Error
Absent
Present
f
Celltype−level correlation
with simulated fraction
Pearson's correlation
Malignant
CD4T
CD8T
Tprolif
B
Plasma
Mono/Macro
Mast
Endothelial
Fibroblasts
Myofibroblasts

## Figure 3

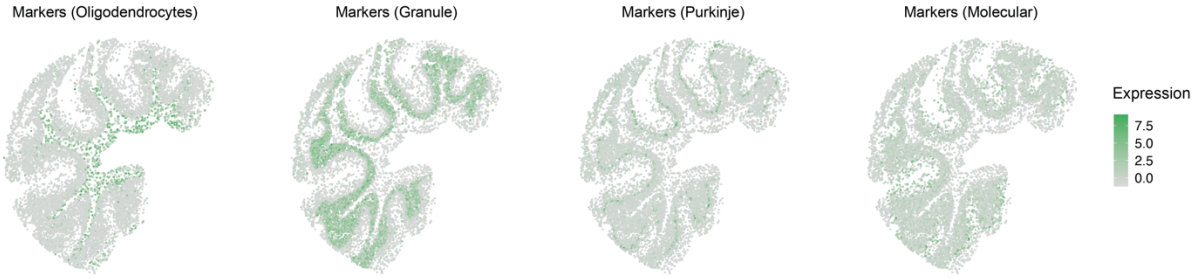
**a**



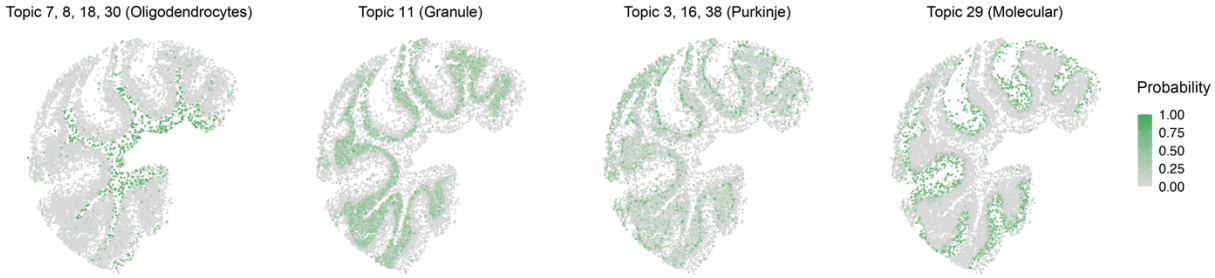

**b**




**c**




**d**



# Figure 4

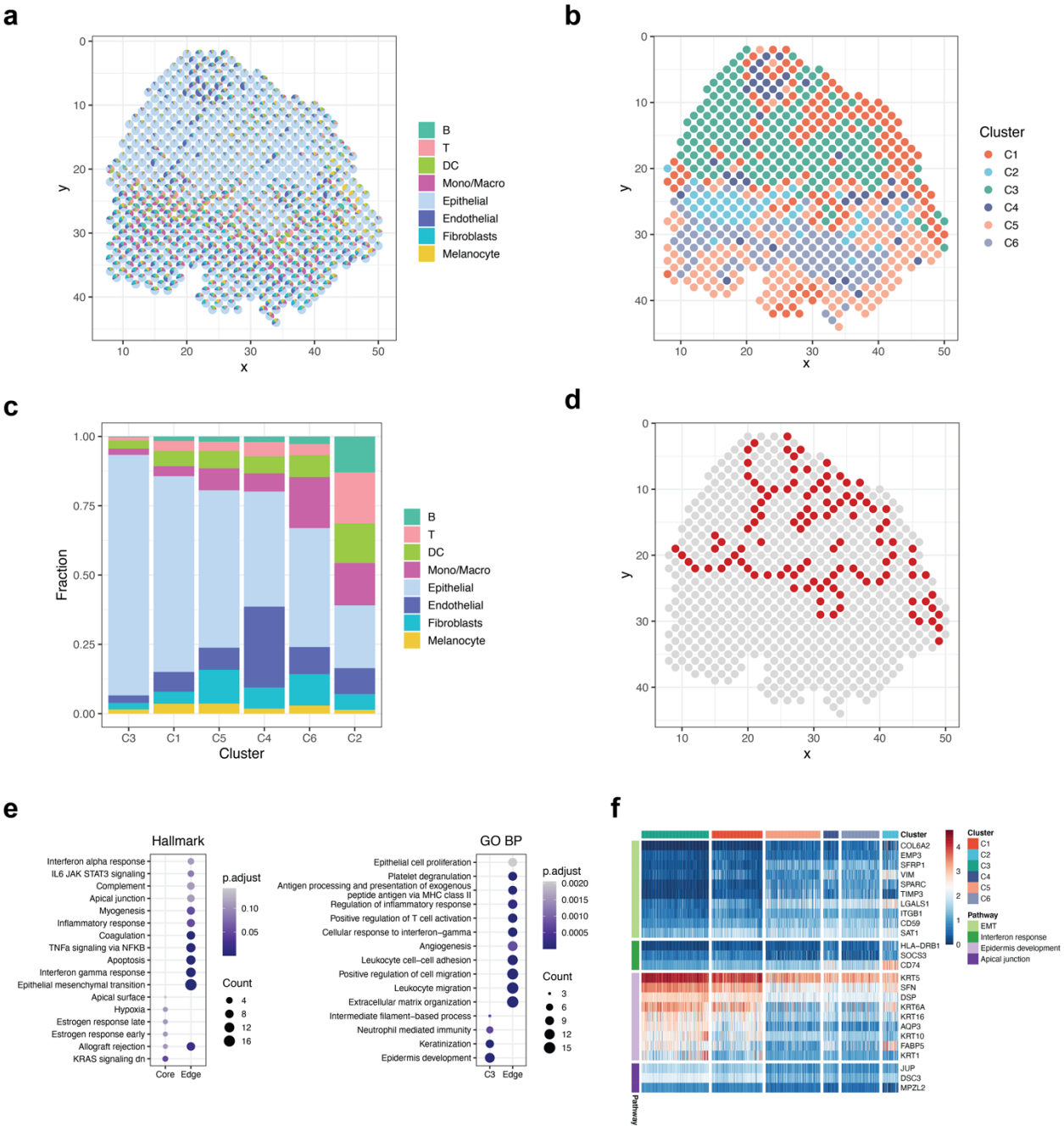
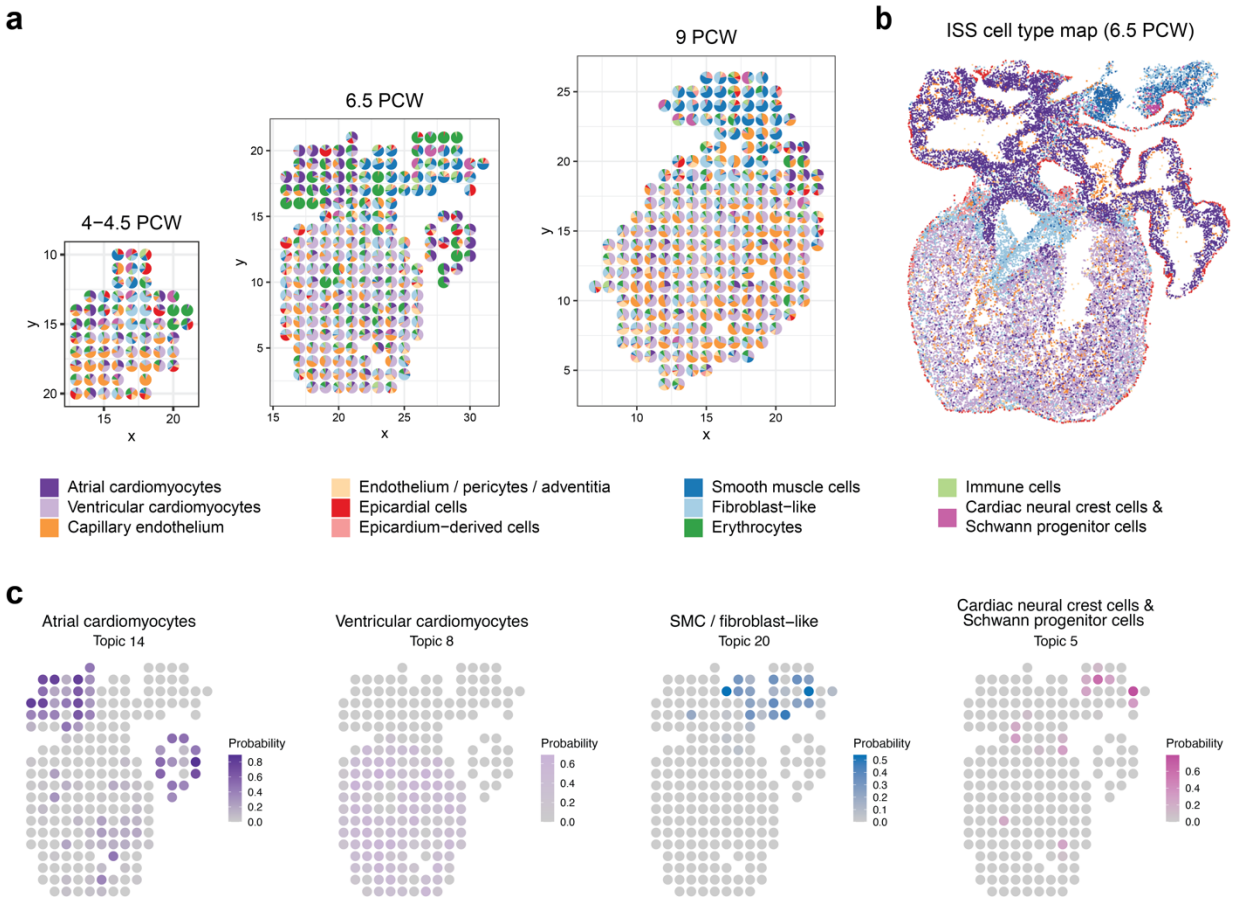

a
b
c
d
e
f
B
T
DC
Mono/Macro
Epithelial
Endothelial
Fibroblasts
Melanocyte
Cluster
C1
C2
C3
C4
C5
C6
Fraction
Hallmark
GO BP
p.adjust
Count
Core
Edge
Pathway
EMT
Interferon response
Epidermis development
Apical junction

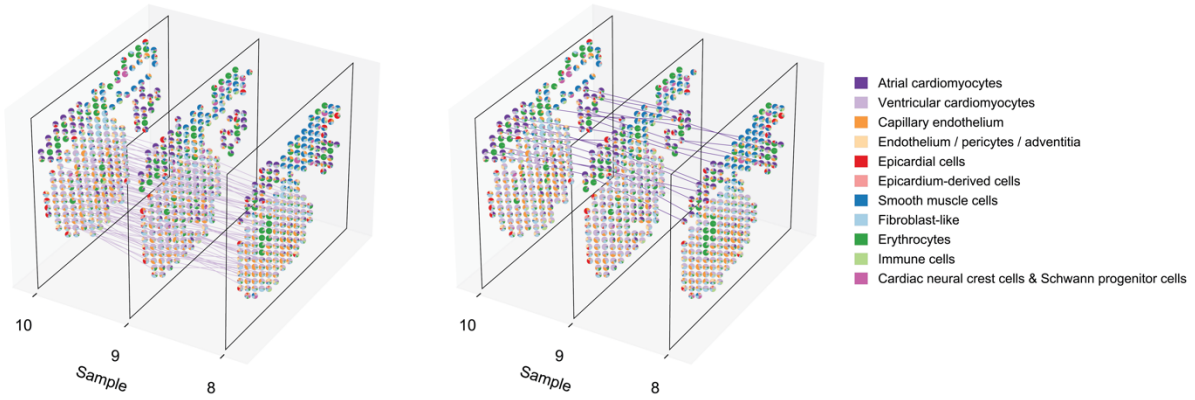
**Figure 5**


a
4−4.5 PCW
6.5 PCW
9 PCW
x
y
b
ISS cell type map (6.5 PCW)
Atrial cardiomyocytes
Ventricular cardiomyocytes
Capillary endothelium
Endothelium / pericytes / adventitia
Epicardial cells
Epicardium−derived cells
Smooth muscle cells
Fibroblast−like
Erythrocytes
Immune cells
Cardiac neural crest cells &
Schwann progenitor cells
c
Atrial cardiomyocytes
Topic 14
Ventricular cardiomyocytes
Topic 8
SMC / fibroblast−like
Topic 20
Cardiac neural crest cells &
Schwann progenitor cells
Topic 5
Probability

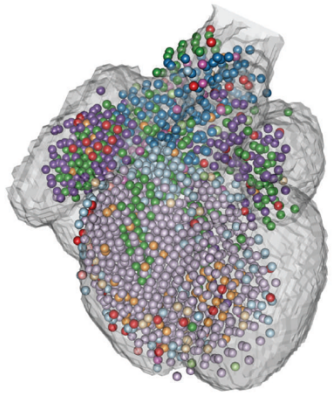
768

**Figure 6**

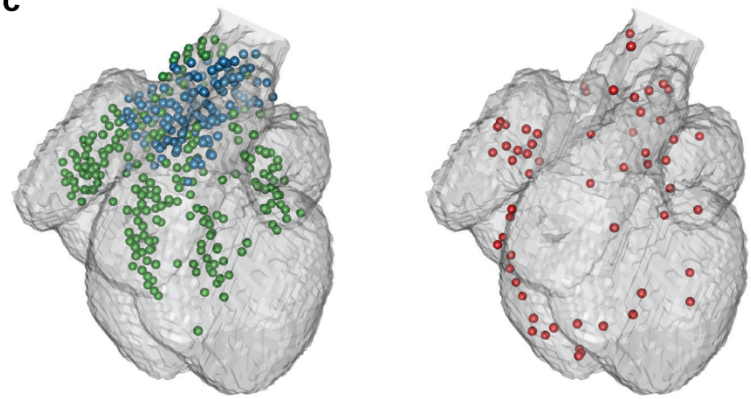
**a**




**b**



**c**



769