iSORT: An Integrative Method for Reconstructing Spatial Organization of Cells using Transfer Learning

Yecheng Tan[1,2†], Ai Wang[3†], Wei Lin[1,4,5*], Yan Yan[3*], and Jifan Shi[1,4,5*]

[1]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200433, China.
[2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China.
[3]Department of Cardiology, Zhongshan Hospital, Fudan University, Shanghai 200032, China.
[4]State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200032, China.
[5]Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.
[*]wlin@fudan.edu.cn
[*]yan.yan@zs-hospital.sh.cn
[*]jfshi@fudan.edu.cn
[†]These authors contributed equally to this work.

ABSTRACT

Understanding the cellular spatial organization is a paramountly important direction of exploring the intricate functionalities of tissues and organs. However, conventional single-cell RNA sequencing (scRNA-seq) technologies often lose cells' spatial information. Recent spatial transcriptomics (ST) techniques aim to preserve locations but easily come with either low resolution or high costs. Here, we introduce iSORT (an **i**ntegrative method for reconstructing **S**patial **OR**ganization of cells via **T**ransfer learning), which can recover the spatial organization of large-scale multi-source scRNA-seq dataset with only one ST reference. iSORT employs a transfer learning approach to assign coordinates for each single cell with scRNA-seq. iSORT also integrates multiple ST tissue slices as references to enhance accuracy and capture multi-slice biological heterogeneity. Furthermore, iSORT conduces to the precise identification of spatially variable genes (SVGs). We demonstrated the effectiveness of iSORT by applying it to four benchmark datasets, including those from *Drosophila* embryo, mouse embryo, mouse brain, and human brain tissues, and by a comparative analysis with five other algorithms as well. Additionally, we collected scRNA-seq data from the patients with atherosclerosis to reconstruct the structure of arteries for the first time, using ST data from other individuals as references. Atherosclerosis-related SVGs were identified and verified by the gene set enrichment analysis. As a promising tool, iSORT provides deep insights into the single-cell omics data analysis and biomedical research.

Introduction

Cells are the fundamental structural and functional units of living organisms. Hosting a diverse array of biological processes ranging from genetics to metabolism, cells are vital for understanding life's complexities. The advent of single-cell omics in the recent decade has opened a new era for the quantitative study of cells from different perspectives[1].

Single-cell RNA sequencing technology (scRNA-seq) is one of the most mature technologies since its first appearance in 2009[2]. Popular protocols such as SMART-seq2[3], 10x Genomics , Drop-seq[4], and CEL-seq2[5], have significantly improved the convenience of obtaining high-resolution and high-throughput gene expression data. scRNA-seq contributes to the discovery of new cell types[4], the analysis of cellular heterogeneity[6], and the tracing of cell differentiation processes[7–9]. However, cells are disassociated during the sequencing process which results in the loss of spatial context.

Spatial transcriptomics (ST)[10] is an emerging technology which aims to simultaneously capture both gene expressions and cell locations. Image-based and sequencing-based methods are the two main approaches for ST. Image-based methods such as MERFISH[11], seqFISH[12], and STARmap[13] provide single-cell or even subcellular spatial resolution through in situ hybridization and fluorescence, but face the limitations in gene/transcript counts and high costs. Sequencing-based methods, such as 10x Visium[14] and Slide-seq[15], offer comprehensive transcriptomic information but lack in the single-cell spatial resolution. One spot in 10x Visium potentially contains 1-10 cells. The latest Stereo-seq[16] achieves a nanoscale resolution with thousands of captured genes/transcripts, but is relatively much more expensive than the scRNA-seq.

Recently, the integration of multi-omics data to complete the missing information in ST has drawn significant attention[17,18]. To complete the undetected gene expressions in the image-based ST, novoSpaRc[19] and spaOTsc[20] applied optimal transport to predict the probabilistic spatial location distribution of cells. Perler[21] employed the linear generative modeling and spatially reconstructed gene expression through weighting. To deconvolve individual cells from spots in the sequencing-based

ST, SPOTlight[22] used the non-negative matrix factorization, and Cell2location[23] built a hierarchical Bayesian framework. Tangram[24] and STEM[25] developed deep neural networks to integrate ST and scRNA-seq data for the spot deconvolution. A survey and benchmarking comparisons of the most common methods are also available[26,27].

The spatial organization of cells within biological tissues and organs is important for understanding their static functions and dynamic mechanisms[28]. Unveiling the spatial arrangement of cells can provide insights into studying the tissue formation[29], the cell-cell communication[30], the identification of spatially variable genes (SVGs)[31,32], and the pathology of diseases[33]. Because scRNA-seq has huge amount of databases and is cheap in practical experiments compared to ST, it is a natural way to study the pseudo-space of scRNA-seq data and reconstruct the cellular spatial organisation with a limited number of ST slices as references. Moreover, assigning spatial coordinates to scRNA-seq data can reveal the spatial origins of cells in high resolution and high throughput[34].

Several algorithms have been proposed to recover the spatial locations for scRNA-seq cells. To name a few, CSOmap[35] performed the spatial reconstruction solely with scRNA-seq data, by estimating cell-cell affinities with ligand-receptor interactions. With a proper targeted space or ST reference, novoSpaRc utilized the optimal transport to assign locations for the scRNA-seq data. Tangram mapped single cells to pre-fixed spatial voxels by designing a probability mapping. scSpace[36] extracted a feature subspace with transfer component analysis and predicted positions using a multi-layer perceptron. CeLEry[37] predicted positions using deep learning and enhanced the spatial transcriptome with data augmentation. CellTrek[38] employed a random forest approach for position prediction but required extensive interpolation of ST data to optimize the reconstruction process. However, most existed algorithms did not consider multi-samples from different sources, and even ignored the difference of cell-type distributions between scRNA-seq and ST data.

In this study, we proposed a framework named iSORT (an **i**ntegrative method for reconstructing **S**patial **OR**ganization of cells using **T**ransfer learning), which integrates high-throughput scRNA-seq with limited ST references and predicts spatial locations for individual cells in the continuous pseudo-space. iSORT can not only handle data from different sources with different cell-type distributions, but also identify SVGs which play a vital role in the tissue organization. We validated the precision and robustness of iSORT using benchmarking datasets, such as *Drosophila* embryo, mouse embryo, mouse brain and human brain tissues. Moreover, we applied iSORT to an experimental dataset of arteries collected from atherosclerotic patients. With iSORT, we successfully reconstructed the spatial organization reconstruction of vascular cells with scRNA-seq, and identified SVGs associated with potential causes of atherosclerosis.

Results

An overview of iSORT

iSORT aims to reconstruct the pseudo-space for the scRNA-seq data using the information from one or multi ST slices. Fig. 1 shows the workflow for the single-slice case. Firstly, iSORT integrates the preprocessed scRNA-seq $\boldsymbol{X}_{\text{sc}}$ and ST data $\boldsymbol{X}_{\text{st}}$ into a common feature space. Batch effects are removed and highly variable genes are selected. The latent features $\boldsymbol{Z}_{\text{sc}}$ and $\boldsymbol{Z}_{\text{st}}$ hold the hypothesis that cells with similar gene expressions have a high probability of being located in close locations, i.e. $p_{\text{sc}}(\boldsymbol{y}|\boldsymbol{z}) = p_{\text{st}}(\boldsymbol{y}|\boldsymbol{z})$, where $\boldsymbol{y}$ represents the spatial coordinate. However, the scRNA-seq data and the ST data are usually obtained from different experiments and have different cell-type distributions, i.e. $p_{\text{sc}}(\boldsymbol{z}) \neq p_{\text{st}}(\boldsymbol{z})$. To deal with the heterogeneity, iSORT utilizes the idea of density ratio originated from the transfer learning in the second step. A weight function $w(\boldsymbol{z})$ is estimated by the Kullback-Leibler importance estimation procedure (KLIEP), and a weight is calculated to give to each sample of ST data. In the third step, iSORT trains a neural network $\boldsymbol{f} : \boldsymbol{z} \mapsto \boldsymbol{y}$ according to a loss function $\mathbb{E}_{(\boldsymbol{z},\boldsymbol{y})\sim p_{\text{st}}}[w(\boldsymbol{z})L(\boldsymbol{z},\boldsymbol{y};\boldsymbol{f})] + \Omega(\boldsymbol{f})$, where L is the mean squared error and $\Omega(\boldsymbol{f})$ is the normalization term. Then, $\boldsymbol{Y}_{\text{sc}} = \boldsymbol{f}(\boldsymbol{Z}_{\text{sc}})$ is the pseudo-space prediction of the scRNA-seq data. Moreover, $I_g = |\nabla_g \boldsymbol{Y}_{\text{sc}}|$ is a measure for ranking the spatial importance of genes. For $I_g \approx 0$, the gene g has little influence on the spatial organization, while, for $I_g \gg 0$, a small perturbation on g leads to a large change of cells' positions. In conclusion, iSORT is applicable to real dataset for reconstructing spatial organization, detecting spatial gene patterns, and identifying SVGs. The detailed derivation and algorithm are presented in the Methods section.

In Supplementary Fig. 1, we also propose the iSORT framework for the case of multiple ST references. The multi-slice ST data obtained from different experiments usually have different shapes, various cell-type distributions, and diverse spatial organizations. To integrate the information gathered from the multiple slices, iSORT designs a slice-specific weight β_i to weight the contribution of the ST slice i. The slice-specific weight can effectively balance the prediction accuracy against model complexity. By combining deep learning techniques and weighting strategies, iSORT makes it possible to estimate an accurate and robust spatial organization for the gene expression data. The detailed description of iSORT for multi-slices is included in the Methods section.

iSORT can reconstruct the spatial organization of cells

iSORT is proposed to reconstruct spatial organization for cells with scRNA-seq. To evaluate the performance of iSORT, we conducted its application using three benchmark datasets, including human dorsolateral prefontal cortex (DLPFC) data, mouse embryo data, and mouse brain data.

Human DLPFC dataset. The DLPFC data comprises three post-mortem brain samples, with each sample containing four tissue slices[39]. 10X Visium technology was conducted to obtain the ST data. We took the slice ID151674 as the scRNA-seq data by removing its spatial coordinates (Fig. 2a). Then, the slice ID151675 was used as the ST reference, which was sampled from the same brain as ID151674 (Fig. 2b). When the scRNA-seq data has the same layer structure and cell-type distribution as the ST reference, iSORT can give an effective reconstruction of the spatial organization (Fig. 2c). Fig. 2c also displays the results from other algorithms, including scSpace, Tangram, novoSpaRc, CeLEry, and CellTrek (Detailed description of these methods could be found in Supplementary Note 1). From the reconstructed results in Fig. 2c, we found that iSORT, scSpace, novoSpaRc, and CeLEry could more distinctly distinguish between layers, while iSORT, Tangram, novoSpaRc, and CeLEry recovered the tissue shape more accurately. We evaluated the six algorithms by four indicators, i.e. the normalized intra-layer distance D_C, the normalized density distribution ρ_C, the normalized volume A_C, and the normalized perimeter P_C (see Methods). The higher values of the metrics, the better quality of the algorithm in clustering different layers. The performance of iSORT in Fig. 2d and Supplementary Fig. 2 supports the effectiveness in remapping scRNA-seq data from the DLPFC to their spatial locations.

Mouse embryo dataset. The mouse embryo data involves sagittal sections from three mouse embryos at the 8-12 somite stage (E8.5-E8.75)[40], which was sequenced by the seqFISH technology[34]. We chose one of these embryos for detailed experimental analysis. As the seqFISH data has the single-cell resolution, scRNA-seq data was taken as the gene expression from each spot by removing its spatial information. To test the performance of iSORT on a low-resolution ST reference, we simulated a low-resolution ST data by a coarse graining method (see Methods). The result of iSORT for reconstructing the spatial organization of the mouse embryo is shown in Fig. 2e. The mouse embryo data are characterized by a rich diversity of cell types, which are displayed in a non-clustered pattern across the simulated spot space. From Fig. 2e, we could see most complex topological structures of the embryo, in particular the presence of intercellular cavities, was recaptured by iSORT from the scRNA-seq, which demonstrates iSORT's capability to reconstruct complex biological tissues. The ground truth, along with a comparison of iSORT with other methods, is presented in Supplementary Fig. 3. We could see that scSpace and CeLEry were generally able to capture the spatial structure, but they had a deficiency in detailing, failing to capture the cavities between embryonic cells. While novoSpaRc and Tangram showed good performance, their capabilities were limited by the sparsity of the simulated spot space.

Mouse brain dataset. The mouse brain dataset is a more challenging case with scRNA-seq data sequenced by Smart-seq2 and ST data from 10X Visium[41]. In this experiment, the scRNA-seq and ST data were from different samples, obtained by different technologies, and had different cell-type distributions. The reconstruction by iSORT from the scRNA-seq is shown in Fig. 2f. We found that iSORT successfully delineated the stratified architecture of the cerebral cortex, reflecting the sequential arrangement of laminar excitatory neuron subgroups: L2/3 intratelencephalic (IT), layer 4, layer 5 intratelencephalic (IT), layer 5 pyramidal tract neurons (PT), layer 6 intratelencephalic (IT), layer 6 corticothalamic (CT), and layer 6b. A comparison with other methods could be found in Supplementary Fig. 4.

iSORT integrating spatial information from multiple slices robustly

ST slices are often obtained from different samples and contain diverse biological information. In Fig. 2c, we showed that when the scRNA-seq (ID151674) and ST data (ID151675) were from the same DLPFC region, iSORT can reconstruct the spatial organization accurately. However, when the single-slice ST reference (ID151507 or ID151671) was from a different tissue and little prior knowledge was available, the reconstruction in scRNA-seq probably shows poor performance (Fig. 3a). To show that iSORT has the advantage of integrating multi-slice ST references to offer a comprehensive spatial organization for the scRNA-seq data, we conducted four tests on the human DLPFC dataset. The original space of all references is provided in Supplementary Fig. 5.

Multiple homologous slices. We first applied iSORT to a set of homologous slices, which we designate as combination 1 (abbreviated as 'C1'), comprising ID151673, ID151675, and ID151676. The target scRNA-seq data is from ID151674 by removing the location information. Fig. 3b shows that, when multiple slices are used as reference, iSORT could get a better spatial organization than the single-slice case (Fig. 2c or Fig. 3a). Since we have the true spatial coordinates of slice ID151674, mean squared error (MSE) are calculated under different conditions in Fig. 3f, which further verified the superiority of multiple slices over single slice in terms of specific metrics.

Multiple rotated homologous slices. To further test the robustness of iSORT, we added rotations to the ST references, which is one of the most common batch effects of experimental data (Supplementary Fig. 6). ST slices ID151673, ID151675, and ID151676 were rotated by 45, 0, -45 degree, respectively. The reconstruction of scRNA-seq ID151674 by iSORT is displayed

in Fig. 3c. The results indicate that iSORT can make accurate spatial reconstruction when there are obvious position differences in different slices. The MSE results presented in Fig. 3f show that, for the rotated version of combination 1, the performance, though slightly inferior to that of the non-rotated combination 1, still surpasses the results obtained with single-slice cases. This finding underscores that using multiple slices, even in a rotated configuration, yields better results than a single slice.

Multiple heterogeneous slices. In practice, scRNA-seq data and ST slices usually originate from different samples with different shapes and cell-type distributions. To assess iSORT's performance, we used three distinct sets of heterogeneous slices as ST references. The first set (Combination 2 or 'C2') comprised slices ID151507, ID151671, and ID151675. The second set (Combination 3 or 'C3') included slices ID151675, ID151507, and ID151508. The third combination (Combination 4 or 'C4') consisted of slices ID151675, ID151670, and ID151671. These sets were used to reconstruct the scRNA-seq from ID151674. The diversity and complexity between these data are high, but iSORT still could get a robust result as shown in Fig. 3d. Fig. 3a demonstrates the reconstruction using a single heterogeneous slice from ID151507/ID151671. Although certain structural layers of the target sample are somewhat recovered, there is a noticeable deviation from the original spatial structure. In contrast, iSORT excellently completes the spatial reconstruction work with multiple slices(as shown in Fig. 3d).

Multiple transformed heterogeneous slices. We also applied rotational transformations to the heterogeneous ST slices. The ST slices ID151507, ID151671, ID151675 (Combination 4 or 'C4') are rotational transformed (Supplementary Fig. 6). The scRNA-seq data were from ID151674. Fig. 3e implies that iSORT can effectively capture and reproduce the original spatial organization of the scRNA-seq data.

The above numerical experiments indicate that iSORT is able to integrate spatial information from multiple ST slices with different situations. With the multi-slice enhancement, iSORT can recover the basic spatial structure of biological tissues robustly.

iSORT revealing the spatial expression patterns of genes

The cell development and tissue organization are determined not only by values of genes but also by their spatial patterns. Gene regulations and cell-cell communications play important roles in complex biological processes. The spatial expression pattern of genes is helpful in uncovering the mechanisms of tissue development and disease progression[42]. iSORT can predict the spatial organizations of scRNA-seq data and reveal the gene patterns at the same time.

We took the *Drosophila* embryonic development dataset as our benchmark[43,44]. This dataset contains a FISH-seq ST slice. We got the scRNA-seq data by removing the spatial information, and by the low-resolution simulation (see Methods) we got a coarse-grained ST data as the reference (Fig. 4b). iSORT reconstructed the pseudo-space and set a spatial coordinate for each scRNA-seq cell (Fig. 4c). For the early development of embryo, several genes, such as *ftz* and *eve*, exhibit unique expression patterns and play decisive roles in the axial establishment and segmentation of the *Drosophila* body[45]. Fig. 4a shows the true spatial pattern for *ftz*, and Fig. 4b is the fuzzy pattern on the low-resolution ST reference. Fig. 4c is the reconstructed pattern for scRNA-seq by iSORT. We can clearly identify the seven stripes, each of which corresponds to a body segment in future. Fig. 4d compare the true location distributions of single cells with the predicted ones. Fig. 4e shows the marginal distributions and reconstruction errors. These results support that the reconstructed gene patterns by iSORT are highly consistent with the original spatial organization. Spatial pattern of 12 special genes such as *eve* can be found in Supplementary Figs. 7, 8 and 9. A comparison with other methods could be found in Supplementary Fig. 10.

iSORT identifying SVGs and discovering disease-related biomarkers

SVGs are essential for understanding cellular heterogeneity within tissues[46]. Mutations or differential expressions in SVGs often result in disorders in organ structures and even diseases. Since iSORT proposes a spatial-organization mapping $\boldsymbol{f}$ from the gene expression to the pseudo-space, the influence of each gene g is quantified by the index I_g (see Methods). Thus, iSORT is able to identify SVGs by the value of I_g and beneficial to discovering disease-related biomarkers[47], which thus provides a bioinformatic approach for the study of the development of organ abnormalities. To validate the effectiveness, we tested iSORT using two datasets.

SVGs in human DLPFC dataset. From the human DLPFC dataset, we chose ID151675 as the ST reference and ID151674 as the scRNA-seq samples. Figure 2c shows the reconstruction of iSORT when taking all the highly-variable genes into consideration. After ranking the genes by the index I_g, we obtained the first 20 important SVGs (Supplementary Table 1). To further test the biological functions of these SVGs, we conducted the knockout validation from the data. The MSE curve is available in Fig. 5 c, where the more genes we knockout, the worse the reconstruction result it is. When the number of genes being knocked out comes to around 300, the MSE curse reaches a tipping point. Figures 5a and b display the reconstruction of spatial organization in the in-silico gene knockout experiment, in which the tissue structure suffered from significant disorders. The results indicate the capability of iSORT in identifying SVGs, the crucial factors in maintaining the spatial organization of cells.

SVGs in human artery dataset. In our analysis, three ST slices were used: one from a patient with atherosclerosis and two from individuals without atherosclerosis. Two independent scRNA-seq data were also used: one sample from a patient with

atherosclerosis and another from an individual without atherosclerosis. Figure 5d illustrates the structure of a healthy artery, while Fig. 5e presents the corresponding hematoxylin and eosin staining of the normal artery in alignment with the spatial transcriptomics results. Endothelial cells, smooth muscle cells, and fibroblasts are respectively marked as representative cells for the intima, media, and adventitia layers within the vascular wall[48]. Since the artery is not a simple connected region, it challenges most conventional algorithms to reconstruct cellular organization. We applied a reversible polar transformation to the ST data before using iSORT to reconstruct spatial organization (see Supplementary Notes 2).

We first reconstructed the scRNA-seq data from the normal individual and the patient with atherosclerosis. The ECs, SMCs and fibroblasts were placed in the correct layers (Fig. 5f). The results with all cells could be found in Supplementary Fig. 14 and Supplementary Fig. 15. Compared with the reconstruction from the normal individual, the result from the patient with atherosclerosis is more concentrated on one side of the vessel, which may be due to the accumulation of lipids and fibrosis within the vessel wall of arteries in patients with atherosclerosis, allowing more cells to be concentrated on one side of the vessel, leading to hardening and narrowing of the vessel.

After the reconstruction of the topology of arteries, we further explored the SVGs (Supplementary Tables 2 and 3). Based on the index I_g, we ranked the genes and picked out the top 20 SVGs in both cases(Supplementary Fig. 11). In the ranking plot of gene scores, we found that both had 9 overlapping genes in the top 10 and 16 overlapping genes in the top 20, which is closely related to the fact that we used both normal and atherosclerotic arteries as references. Nonetheless, the difference in the ordering of their scores, as well as the genes specific to each of them, reflects the variability of the sample and the heterogeneity originating from the data. Among these SVGs, *TNN* achieved the highest score. *TNN*, also known as *tenascin-W*, is implicated in facilitating integrin binding, a quintessential process for cell adhesion and migration[49]. Its association with cell proliferation and potential roles in neuronal generation and osteoblast differentiation highlight its importance in vascular structure and function[50–52]. Meanwhile, a number of genes thought to be strongly associated with atherosclerosis such as *RFLNA*, *SPN*, and *EZH2* achieved high scores. *RFLNA* is hypothesized to promote filamin binding, playing a vital role in organizing actin filaments, bone mineralization, chondrocyte development, and cell motility[53–55]. Such processes are integral to the dynamic nature of vascular cells, which require structural pliability, potentially linking to the atherogenic process. Recent studies propose that *filaminA* expression could emerge as a significant prognostic biomarker for atherosclerosis[56]. *SPN*, widely recognized as *CD43*, is involved in numerous signaling pathways that enhance cell adhesion and migration[57–59]. Notably, it has been shown that anti-*CD43* can significantly reduce monocyte adhesion to endothelial cells during inflammatory responses[60], further corroborating the role of *SPN* in atherosclerotic processes. In both cases, *EZH2* was identified as SVGs. The literatures indicate that *EZH2* is highly conserved from Drosophila to primates and is primarily associated with gene transcriptional repression. *EZH2* is indispensable for regular cellular development, differentiation, maintenance of gene expression, and determination of cell fate[61–63]. Furthermore, its association with cardiovascular diseases, particularly atherosclerosis[64,65], posits a potential role in vascular health maintenance. The concurrent presence of *EZH2* in both conditions may suggest a dual functionality, contributing to vascular homeostasis and pathology. This allows for a comparative observation of *EZH*2's potential bifunctional role in the maintenance of vascular function and in disease manifestation, possibly providing fruitful avenues for future research into vascular disease biomarkers and therapeutic targets.

We then performed GO enrichment analysis on the top 50 genes ranked by SVG scores in our study. The GO analysis reveals that 10 out of the top 30 results in GO Biological Processes (BP) are common to both the AS and normal samples (see Supplementary Table 4, Supplementary Table 5, Supplementary Fig. 12 and Supplementary Fig. 13). These shared processes include the regulation of leukocyte proliferation (GO:0070663), alpha-beta T cell activation (GO:0046631), and T cell activation involved in the immune response (GO:0002286), among others. The presence of these shared processes suggests a commonality in core biological functions between the AS and normal samples. It is noteworthy that unique BP terms appear in the AS samples, including but not limited to tumor necrosis factor superfamily cytokine production (GO:0071706), tumor necrosis factor production (GO:0032640), regulation of tumor necrosis factor superfamily cytokine production (GO:1903555), and regulation of tumor necrosis factor production (GO:0032680). These terms in AS samples indicate their potential involvement in specific biological processes or functions related to tumor necrosis factor during atherosclerosis. This underscores the potential for AS to influence specific biological processes linked to tissue, region, or cell type, particularly those correlated with tumor necrosis factor.

The results indicate that iSORT provides an approach for understanding the cellular organization, and is helpful in exploring SVGs and uncovering disease-related biomarkers.

Discussion

In this study, we proposed an algorithm named iSORT to reconstruct the cellular spatial organization. iSORT utilized the thought of density ratio from the transfer learning to integrate the spatial information in one or multiple ST slices, and gave an accurate pseudo-space estimation for the scRNA-seq cells. Compared to the previous methods, iSORT is able to predict spatial locations in continuous space, fully integrate multi-source information, reveal gene spatial patterns, and identify SVGs.

iSORT showed superior performance in spatial reconstruction using, respectively, human DLPFC data, mouse embryo data, and mouse brain data. The spatial patterns in Drosophila embryo were also recovered. Further, iSORT identified SVGs in the human DLPFC dataset and the human artery dataset. Thereby iSORT can provide insights into gene regulations in the tissue organization and cell interactions during the disease development.

Some of the previous methods could predict the location of a cell by gene expression but only in discrete space, especially the algorithms Tangram originating from the spot deconvolution. Most algorithms cannot deal with multi-slice case. SpatialDE[31] only identifies genes with spatial expression patterns in spatial transcriptomes. Trendsceek[66] requires prior spatial location information of scRNA-seq data without offering prediction methods. iSORT developed the previous approaches and proposed an effective integration way.

While seq-based spatial sequencing technologies like the 10X Visium offer invaluable data to researchers, they come with high cost and spot-level resolution. Single-cell sequencing techniques like Smart-seq2, provide high transcript coverage, but lose cellular spatial positions during sequencing. iSORT is designed to bridge the gap between spatial transcriptomes and scRNA-seq data. Integrative computational methods like iSORT can, to a certain extent, overcome the limitations of both, proving a reliable reconstruction of cellular organization in the single-cell resolution.

The ST references and scRNA-seq data should sampled from the same tissue, though the shape and cell-type distribution could be different. iSORT can not transfer the spatial information between different organs or different species yet. Additional cell-cell interaction would help determine the spatial inference further[35]. In the other hand, we also hope iSORT becomes useful in inferring the cell-cell communication (CCC). By integrating scRNA-seq data with ST, iSORT endows scRNA-seq with precise spatial coordinates, enabling exploration of gene spatial distribution at a higher resolution and thus offering a broader perspective for CCC inference.

Identifying key spatial genes from single-cell data, using spatial transcriptomics as a benchmark, offers new perspectives for unveiling cellular characteristics and deciphering disease pathways. These genes play a pivotal role in disease progression. Traditional methods based on scRNA-seq mainly focus on the functional states and heterogeneity of individual cells but often overlook the spatial context of tissues or organs. For instance, atherosclerosis results from lipid accumulation within vascular walls, an accumulation that is not uniform but localized to specific spatial regions, like areas with low shear stress. Such spatial heterogeneity plays a crucial role in disease onset and progression. Therefore, integrating spatial information to pinpoint genes that can influence tissue structure and function holds immense potential for understanding disease mechanisms and devising therapeutic strategies. In this study, iSORT derived a measure I_g for identify SVGs, and showed the ability in detecting potential factors associated with atherosclerosis. We hope iSORT could be helpful in investigating other diseases.

Methods

Data preprocessing

First, the raw gene expression for each scRNA-seq cell or ST spot is log-transformed and normalized by

$$x_{ij} = \ln\left(s \cdot \frac{d_{ij}}{\sum_{k=1}^{n} d_{kj}} + 1\right), \tag{1}$$

where d_{ij} is the raw expression of gene i in cell/spot j, n is the total number of genes, s is a scaling factor with default value 10000, and x_{ij} is the normalized gene expression. Then, x_{ij} is scaled by z-score

$$x_{ij} \leftarrow \frac{x_{ij} - \mu_i}{\sigma_i}, \tag{2}$$

where μ_i and σ_i are the mean value and standard variation of gene i across cells/spots, respectively. Highly variable genes are selected for further studies. All these preprocessing procedures can be carried out by Seurat[67] or Scanpy[68].

iSORT for the single-slice case

Spatial organization mapping

In the single-slice case, we have scRNA-seq data with only one ST slice as the reference (Fig. 1a). Denote $\boldsymbol{X}_{\text{sc}} = \{\boldsymbol{x}_{\text{sc}_1}, \boldsymbol{x}_{\text{sc}_2}, ..., \boldsymbol{x}_{\text{sc}_M}\}$ as the preprocessed expressions for the scRNA-seq data, and $\boldsymbol{X}_{\text{st}} = \{\boldsymbol{x}_{\text{st}_1}, \boldsymbol{x}_{\text{st}_2}, ..., \boldsymbol{x}_{\text{st}_m}\}$ for the ST data, where M and m are the sample sizes. The location of ST data is $\boldsymbol{Y}_{\text{st}} = \{\boldsymbol{y}_{\text{st}_1}, \boldsymbol{y}_{\text{st}_2}, \ldots, \boldsymbol{y}_{\text{st}_m}\}$. Suppose that there are H highly variable genes after preprocessing, then we have $\boldsymbol{x}_{\text{sc}_i} \in \mathbb{R}^H$, $\boldsymbol{x}_{\text{st}_j} \in \mathbb{R}^H$, $\boldsymbol{y}_{\text{st}_j} \in \mathbb{R}^2$, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, m$. Let $p_{\text{sc}}(\boldsymbol{x}, \boldsymbol{y})$ and $p_{\text{st}}(\boldsymbol{x}, \boldsymbol{y})$ be the joint probability density functions (pdfs) of the scRNA-seq data and the ST data, respectively. Due to the variety in samples and discrepancies in technologies, the marginal cell-type distributions and expression scales are usually different between $\boldsymbol{X}_{\text{sc}}$ and $\boldsymbol{X}_{\text{st}}$, i.e.

$$p_{\text{st}}(\boldsymbol{x}) \neq p_{\text{sc}}(\boldsymbol{x}), \quad \text{and} \quad p_{\text{st}}(\boldsymbol{y}|\boldsymbol{x}) \neq p_{\text{sc}}(\boldsymbol{y}|\boldsymbol{x}). \tag{3}$$

Integration is needed to consider the data in a shared latent space. In iSORT, we employ the reference-based co-embedding approach from the Seurat package. $\boldsymbol{X}_{\text{sc}}$ is set as the query and $\boldsymbol{X}_{\text{st}}$ is as the reference to pass the functions `FindIntegrationAnchors()` and `IntegrateData()` sequentially. Denote the variables after the integration as $\boldsymbol{Z}_{\text{sc}}$ and $\boldsymbol{Z}_{\text{st}}$, respectively. Under the postulation that a cell's spatial location is intrinsically related to its latent expression, we obtain

$$p_{\text{st}}(\boldsymbol{z}) \neq p_{\text{sc}}(\boldsymbol{z}), \quad \text{and} \quad p_{\text{st}}(\boldsymbol{y}|\boldsymbol{z}) = p_{\text{sc}}(\boldsymbol{y}|\boldsymbol{z}), \tag{4}$$

which implies that the samples with similar features $\boldsymbol{z}$ have close pseudo-space $\boldsymbol{y}$[69]. We represent the mapping between latent expressions and locations as $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{f} : \mathbb{R}^H \to \mathbb{R}^2$ and $\boldsymbol{\varepsilon}$ is a random vector standing for the deviation. The task is to give an estimation of $\boldsymbol{f}$ and determine the pseudo-space of $\boldsymbol{X}_{\text{sc}}$ by $\boldsymbol{f}(\boldsymbol{Z}_{\text{sc}})$. Several studies have discussed the construction of $\boldsymbol{f}$ considering the covariance bias[70,71]. iSORT addresses the estimation as a learning task to minimize a loss function

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{z},\boldsymbol{y})\sim p_{\text{sc}}}\left[L(\boldsymbol{z},\boldsymbol{y};\boldsymbol{f})\right] &= \mathbb{E}_{(\boldsymbol{z},\boldsymbol{y})\sim p_{\text{st}}}\left[\frac{p_{\text{sc}}(\boldsymbol{z},\boldsymbol{y})}{p_{\text{st}}(\boldsymbol{z},\boldsymbol{y})}L(\boldsymbol{z},\boldsymbol{y};\boldsymbol{f})\right] = \mathbb{E}_{(\boldsymbol{z},\boldsymbol{y})\sim p_{\text{st}}}\left[\frac{p_{\text{sc}}(\boldsymbol{z})p_{\text{sc}}(\boldsymbol{y}|\boldsymbol{z})}{p_{\text{st}}(\boldsymbol{z})p_{\text{st}}(\boldsymbol{y}|\boldsymbol{z})}L(\boldsymbol{z},\boldsymbol{y};\boldsymbol{f})\right] \\ &= \mathbb{E}_{(\boldsymbol{z},\boldsymbol{y})\sim p_{\text{st}}}\left[\frac{p_{\text{sc}}(\boldsymbol{z})}{p_{\text{st}}(\boldsymbol{z})}L(\boldsymbol{z},\boldsymbol{y};\boldsymbol{f})\right],\end{aligned} \tag{5}$$

where Eq. (4) is applied in the last step, and $L(\cdot,\cdot;\boldsymbol{f})$ is the loss function. For the spatial reconstruction, L is usually selected as the mean squared error loss, i.e. $L(\boldsymbol{x},\boldsymbol{y};\boldsymbol{f}) = ||\boldsymbol{f}(\boldsymbol{z}) - \boldsymbol{y}||_2$. The density ratio $w(\boldsymbol{z}) = p_{\text{sc}}(\boldsymbol{z})/p_{\text{st}}(\boldsymbol{z})$ is the key to address the different cell-type distributions between the scRNA-seq and ST data.

For the m samples in the ST slice, the minimization of (5) is discretized as

$$\underset{\boldsymbol{f}}{\text{argmin}} \frac{1}{m}\sum_{j=1}^{m} w_j L(\boldsymbol{z}_{\text{st}_j}, \boldsymbol{y}_{\text{st}_j}; \boldsymbol{f}) + \Omega(\boldsymbol{f}), \tag{6}$$

where $w_j = w(\boldsymbol{z}_{\text{st}_j})$ is the sample-specific density ratio and $\Omega(\boldsymbol{f})$ is the normalization term. Once w_js are known, we can obtain the spatial organization mapping by optimizing Eq. (6) and then apply $\boldsymbol{f}$ on $\boldsymbol{Z}_{\text{sc}}$.

Estimation of the density ratio

iSORT estimates the density ratio $w(\boldsymbol{z}) = p_{\text{sc}}(\boldsymbol{z})/p_{\text{st}}(\boldsymbol{z})$ by the method of KLIEP[71]. KLIEP demonstrates computational efficiency, stability in performance, and effective mitigation of overfitting. Specifically, $w(\boldsymbol{z})$ is represented linearly by a set of basis functions, i.e.

$$\hat{w}(\boldsymbol{z}) = \sum_{i=1}^{k} a_i \phi_i(\boldsymbol{z}), \tag{7}$$

where $\hat{w}$ is the approximation of w, $\{\phi_i\}_{i=1}^k$ are k given basis functions satisfying $\phi_i(\boldsymbol{z}) \geq 0$, and a_is are the coefficients to be determined. By default, Gaussian basis functions (GBFs) are selected as ϕ_i, offering qualities of smoothness, locality, and universal approximation capability[72,73].

KLIEP estimates $\{a_i\}_{i=1}^k$ by minimizing the KL divergence between $p_{\text{sc}}(\boldsymbol{z})$ and $\hat{p}_{\text{sc}}(\boldsymbol{z}) = \hat{w}(\boldsymbol{z})p_{\text{st}}(\boldsymbol{z})$, i.e.

$$\text{KL}[p_{\text{sc}}(\boldsymbol{z})||\hat{p}_{\text{sc}}(\boldsymbol{z})] = \int p_{\text{sc}}(\boldsymbol{z}) \ln\left(\frac{p_{\text{sc}}(\boldsymbol{z})}{\hat{w}(\boldsymbol{z})p_{\text{st}}(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} = \int p_{\text{sc}}(\boldsymbol{z}) \ln\left(\frac{p_{\text{sc}}(\boldsymbol{z})}{p_{\text{st}}(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} - \int p_{\text{sc}}(\boldsymbol{z}) \ln \hat{w}(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}. \tag{8}$$

Only the second term in Eq. (8) contains a_i, and the optimization is equivalent to maximize

$$J := \int p_{\text{sc}}(\boldsymbol{z}) \ln \hat{w}(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} \approx \frac{1}{M}\sum_{j=1}^{M} \ln\left(\sum_{i=1}^{k} a_i \phi_i(\boldsymbol{z}_{\text{sc}_j})\right), \tag{9}$$

where $\boldsymbol{z}_{\text{sc}_j}$s are the latent expressions of scRNA-seq data, and M is the number of scRNA-seq samples. Considering the constrain of pdf as

$$\int \hat{p}_{\text{sc}}(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} = \int \hat{w}(\boldsymbol{z}) p_{\text{st}}(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} = 1, \tag{10}$$

$\{a_i\}_{i=1}^k$ is finally estimated by solving

$$\begin{aligned} \underset{\{a_i\}}{\text{argmax}} \quad & \sum_{j=1}^{M} \ln\left(\sum_{i=1}^{k} a_i \phi_i(\boldsymbol{z}_{\text{sc}_j})\right) \\ \text{s.t.} \quad & \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{k} a_i \phi_i(\boldsymbol{z}_{\text{st}_j}) = 1, \\ & a_i \geq 0, \; i = 1, 2, \ldots, k. \end{aligned} \tag{11}$$

Optimizing the reconstruction mapping

Neural network technology can uncover the nonlinear mapping between variables with just a few hidden layers[74–76]. To solve the optimization problem (6), we approximate $\boldsymbol{f}$ by a multi-layer BP neural network (Fig. 1a). The architecture is briefly described as follows:

- **Network Architecture:** The default setting includes five fully connected layers. Each hidden layer is followed by a dropout layer to prevent overfitting. Users can adjust the depth and width of the network according to the complexity of the input data.
- **Input Layer:** The input layer contains H neurons, which is the number of genes. Latent expressions in $\boldsymbol{Z}_{\text{sc}}$ or $\boldsymbol{Z}_{\text{st}}$ are used as inputs.
- **Hidden Layers:** The model contains several hidden layers, and ReLU activation function is used in each layer to enhance the network's non-linear mapping capability.
- **Dropout Layer:** Dropout layers are inserted after each hidden layer with a dropout rate of 0.1 during the training process.
- **Output Layer:** The output layer has a dimension of 2, representing the spatial coordinates on the x-axis and y-axis. If a z-axis is required, this can be extended to 3D.
- **Loss Function:** Typically, the mean squared error loss is chosen and the training process employs the Adam optimizer with a learning rate of 0.001.

iSORT for the multi-slice case

Spatial organization mapping with multiple slices

For the multi-slice case, we have scRNA-seq with l ST slices as references (Fig. 1b). Denote $\boldsymbol{X}_{\text{sc}} = \{\boldsymbol{x}_{\text{sc}_1}, \boldsymbol{x}_{\text{sc}_2}, \ldots, \boldsymbol{x}_{\text{sc}_M}\}$ as the scRNA-seq expression, $\boldsymbol{X}_{\text{st}}^{(i)} = \{\boldsymbol{x}_{\text{st}_1}^{(i)}, \boldsymbol{x}_{\text{st}_2}^{(i)}, \ldots, \boldsymbol{x}_{\text{st}_{m_i}}^{(i)}\}$ as the expression of the ith ST slice and $\boldsymbol{Y}_{\text{st}}^{(i)} = \{\boldsymbol{y}_{\text{st}_1}^{(i)}, \boldsymbol{y}_{\text{st}_2}^{(i)}, \ldots, \boldsymbol{y}_{\text{st}_{m_i}}^{(i)}\}$ as the locations, where $i = 1, 2, \ldots, l$ is the number of ST slices, and M and m_i are the sample sizes. The latent expressions after integration are $\boldsymbol{Z}_{\text{sc}} = \{\boldsymbol{z}_{\text{sc}_1}, \boldsymbol{z}_{\text{sc}_2}, \ldots, \boldsymbol{z}_{\text{sc}_M}\}$ and $\boldsymbol{Z}_{\text{st}}^{(i)} = \{\boldsymbol{z}_{\text{st}_1}^{(i)}, \boldsymbol{z}_{\text{st}_2}^{(i)}, \ldots, \boldsymbol{z}_{\text{st}_{m_i}}^{(i)}\}$, respectively. The challenge is that the scRNA-seq and ST slices are usually sampled from different experiments with different tissue shapes and cell-type distributions, i.e.

$$p_{\text{sc}}(\boldsymbol{z}) \neq p_{\text{st}^{(1)}}(\boldsymbol{z}) \neq p_{\text{st}^{(2)}}(\boldsymbol{z}) \neq \ldots \neq p_{\text{st}^{(l)}}(\boldsymbol{z}), \quad \text{and} \quad p_{\text{st}^{(1)}}(\boldsymbol{y}|\boldsymbol{z}) \neq p_{\text{st}^{(2)}}(\boldsymbol{y}|\boldsymbol{z}) \neq \ldots \neq p_{\text{st}^{(l)}}(\boldsymbol{y}|\boldsymbol{z}), \tag{12}$$

where p_{sc} and $p_{\text{st}^{(i)}}$ are the corresponding pdfs. To address this issue, iSORT generalizes the single-slice case (6) to the minimization problem

$$\underset{\boldsymbol{f}}{\text{argmin}} \sum_{i=1}^{l} \frac{\beta_i}{m_i} \sum_{j=1}^{m_i} w_j^{(i)} L(\boldsymbol{z}_{\text{st}_j}^{(i)}, \boldsymbol{y}_{\text{st}_j}^{(i)}; \boldsymbol{f}) + \Omega(\boldsymbol{f}), \tag{13}$$

where β_j is the slice-specific weight integrating the heterogeneity of different slices, and $w_j^{(i)}$ is the sample-specific density ratio. Similar to the single-slice case, L is set as the mean squared error loss, and $w_j^{(i)}$ is estimated by the KLIEP. Once β_is are obtained, the mapping $\boldsymbol{f}$ can be approximated by a multi-layer BP neural network (Fig. 1b). The pseudo-space estimation is given by $\boldsymbol{f}(\boldsymbol{Z}_{\text{sc}})$. When $l = 1$, the procedure naturally degenerates to the single-slice case.

Estimation of the slice-specific weight

The slice-specific weight $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_l)^T$ is used to balance the importance of multiple slices. We estimate $\boldsymbol{\beta}$ by

$$\underset{\boldsymbol{\beta}:\boldsymbol{\beta}^T\mathbf{1}=1,\ \boldsymbol{\beta}\geq 0}{\text{argmin}} \sum_{i,j=1}^{M} \rho_{ij} ||\boldsymbol{\beta}^T \boldsymbol{Y}_{\text{sc}}^{(i)} - \boldsymbol{\beta}^T \boldsymbol{Y}_{\text{sc}}^{(j)}||_2^2, \tag{14}$$

where $\boldsymbol{Y}_{\text{sc}}^{(i)} = (\boldsymbol{f}^{(1)}(\boldsymbol{z}_{\text{sc}_i}), \boldsymbol{f}^{(2)}(\boldsymbol{z}_{\text{sc}_i}), \ldots, \boldsymbol{f}^{(l)}(\boldsymbol{z}_{\text{sc}_i}))^T$, $\boldsymbol{f}^{(i)}$ is the single-slice mapping estimated by scRNA-seq and the ith ST slice, and ρ_{ij} is the value of the correlation coefficient between $\boldsymbol{z}_{\text{sc}_i}$ and $\boldsymbol{z}_{\text{sc}_j}$. Eq. (14) implies that with the help of $\boldsymbol{\beta}$, the predictions using different slices between cells with high similarity become smaller and the differences between cells with low similarity become larger.

Eq. (14) is a convex quadratic optimization problem with an equivalent form:

$$\underset{\boldsymbol{\beta}:\boldsymbol{\beta}^T\mathbf{1}=1,\ \boldsymbol{\beta}\geq 0}{\text{argmin}} \sum_{k=1}^{2} \boldsymbol{\beta}^T \hat{\boldsymbol{Y}}_{\text{sc}}^{[k]} \boldsymbol{L} (\hat{\boldsymbol{Y}}_{\text{sc}}^{[k]})^T \boldsymbol{\beta}, \tag{15}$$

where $\hat{\boldsymbol{Y}}_{\text{sc}}^{[k]} = (\boldsymbol{Y}_{\text{sc}}^{(1)}(k), \boldsymbol{Y}_{\text{sc}}^{(2)}(k), \ldots, \boldsymbol{Y}_{\text{sc}}^{(M)}(k))$ is the axis-k of the single-slice mappings, and $\boldsymbol{L}$ is a $M \times M$ matrix with

$$\boldsymbol{L}_{ij} = \begin{cases} -\rho_{ij}, & i \neq j, \\ \sum_{k=1}^{M} \rho_{ik}, & i = j. \end{cases} \tag{16}$$

Identification of spatially variable genes

SVGs are critical genes underlying spatial gene expression patterns and their associated cellular functionalities. To forge a coherent understanding of how specific genes influence spatial configurations within cells, we engaged a gradient-based analytical approach, which subtly interweaves concepts from the domain of Saliency Maps[77]—a renowned technique in deep learning for unraveling model decision rationale.

Within the framework of iSORT, the mapping $\boldsymbol{f}$ maps the gene expression data $\boldsymbol{x} \in \mathbb{R}^n$ to a two-dimensional spatial coordinate $\boldsymbol{y} \in \mathbb{R}^2$, i.e. $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. We determined the importance of gene g according to

$$I_g = \mathbb{E}\left\{ \frac{1}{2} \sum_{p=1}^{2} \left| \frac{\partial y_p}{\partial x_g} \right| \right\}, \tag{17}$$

where y_1, y_2 are the components of $\boldsymbol{y}$, x_g is the gene g expression in $\boldsymbol{x}$, and the expectation $\mathbb{E}$ is taken over all cells/samples. I_g takes both positive and negative influences into the consideration. SVGs are identified as the genes with top I_g scores, reflecting their significant contributions to the spatial configuration and biological functionality.

Comparative metrics

To evaluate the performance of spatial reconstruction algorithms, we use four indices to assess the capability in preserving the layered structure of a tissue.

Normalized Intra-layer Distance

For a given layer/cluster C, the intra-layer average distance $\hat{D}_C$ is defined as

$$\hat{D}_C = \frac{1}{|C|} \sum_{\boldsymbol{x}_i \in C} \|\boldsymbol{x}_i - \boldsymbol{\mu}_C\|_2, \tag{18}$$

where $\boldsymbol{x}_i$ are samples with in C, $\boldsymbol{\mu}_C$ is the centroid of C, and $|C|$ is the total number of samples in C. $\hat{D}_C$ measures the average distance of all points within a specific layer to the centroid. The normalized intra-layer distance D_C is defined by

$$D_C = \psi(\hat{D}_C) := 1 - \frac{\hat{D}_C - \hat{D}_{\min}}{\hat{D}_{\max} - \hat{D}_{\min} + \varepsilon} \tag{19}$$

where ψ represents the normalization function, ε is a small constant to prevent extremes, and $\hat{D}_{\min}$ and $\hat{D}_{\max}$ are the minimum and maximum values of $\hat{D}_C$ among all layers, respectively. D_C is between 0 and 1, and a larger value implies a better clustering result.

Normalized Density Distribution

For a layer/cluster C, its density distribution ρ_C is defined as

$$\rho_C = \frac{\hat{\rho}_C}{\max_C \hat{\rho}_C}, \quad \text{and} \quad \hat{\rho}_C = \frac{|C|}{V(C)}, \tag{20}$$

where $|C|$ is the number of samples in C, and $V(C)$ is the volume of the convex hull occupied by C. ρ_C is between 0 and 1, and evaluates the spatial concentration of data points within a given layer or cluster C. A larger value of ρ_C indicates a better clustering.

Volume and Perimeter

For a layer/cluster C, the normalized volume A_C and normalized perimeter P_C are defined as

$$A_C = \psi(V_C), \quad P_C = \psi(S_C), \tag{21}$$

Where $V(C)$ is the volume of the convex hull of C, S_C is the perimeter, and ψ is the normalized function used in Eq. (19). A_C and P_C are between 0 and 1. Generally, larger values of A_C and P_C (smaller values of V_C and S_C) indicate a more compact and well-defined layer.

Simulation of the low-resolution ST data

FISH technology can provide high-resolution ST data[34] in the single-cell scale, while most sequence-based approaches such as 10X Visium can only achieve spot-resolution. The diameter of a spot in 10X Visum is 55μm, while the distance between the centers of two adjacent spots is 100μm. One spot may contain 1-10 cells.

To test the effectiveness of iSORT when using low-resolution ST data as the reference, we simulate coarse ST from the seqFISH data in mouse embryo and FISH data in *Drosophila* embryo experiments. In simulation, we first divided the area into a uniform grid and the intersections of the lines serve as spots. The radius of each coarse spot is set to be one quarter of the adjacent spot spacing, and the gene expression is set as the accumulated value of all cells within the radius. We only considered spots with gene expression greater than zero.

Collection of human artery samples and ethics statement

The human artery data were collected from patients undergoing coronary artery bypass grafting (CABG) or heart transplantation at Zhongshan Hospital, Fudan University. Written informed consent was obtained from each participant before surgery, with the study approved by the Ethics Committee of Zhongshan Hospital, Fudan University (ethical approval number: B2022-031R), and conducted in strict accordance with the principles outlined in the Declaration of Helsinki.

For scRNA-seq, artery samples were processed to generate single-cell suspensions. The process involved cell separation, labeling, and library construction, followed by sequencing on the Illumina NovaSeq platform. The sequencing data were processed with Cell Ranger (10x Genomics, version cellranger-6.0.0), aligning them to the human reference genome (GRCh38) to create a matrix of gene expression barcodes.

For ST, the samples that passed quality inspection were re-sectioned for permeabilization experiments. The Visium Spatial Tissue Optimization Slide & Reagent Kit (PN-1000193, 10X Genomics) was used to release mRNA from cells and bind it to spatially barcoded oligonucleotides on the slides. Imaging to determine the appropriate permeabilization time was performed using Leica Aperio CS2 and Leica THUNDER Imager Tissue. The libraries were then prepared using the Visium Spatial Gene Expression kit (PN-1000184, 10X Genomics).

Data availability

The original public data used in this paper can be accessed through the following links: (1) FISH data from the Berkeley Drosophila Transcription Network Project (BDTNP) https://www.fruitfly.org/; (2) 10X Visium data of the human dorsolateral prefrontal cortex (DLPFC) http://spatial.libd.org/spatialLIBD/; (3) Adult mouse cortical cell datasets: GEO accession GSE71585; (4) The mouse embryo data https://marionilab.cruk.cam.ac.uk/SpatialMouseAtlas; (5) The Seurat objects ST data (10X Genomics Visium) of mouse brain https://satijalab.org/seurat/articles/spatial_vignette.html. The human artery data are available from the corresponding author on reasonable request.

Code availability

The iSORT algorithm is available at GitHub: https://github.com/xiaojierzi/iSORT.

References

1. Teichmann, S. & Efremova, M. Method of the Year 2019: single-cell multimodal omics. *Nat. Methods* **17**, 2020 (2020).
2. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
3. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
4. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
5. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 1–7 (2016).
6. Ulrich, N. D. *et al.* Cellular heterogeneity of human fallopian tubes in normal and hydrosalpinx disease states identified using scRNA-seq. *Dev. Cell* **57**, 914–929 (2022).
7. Shi, J., Li, T., Chen, L. & Aihara, K. Quantifying pluripotency landscape of cell differentiation from scRNA-seq data by continuous birth-death process. *PLoS Comput. Biol.* **15**, e1007488 (2019).
8. Shi, J., Teschendorff, A. E., Chen, W., Chen, L. & Li, T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Briefings Bioinform.* **21**, 248–261 (2020).
9. Fernandez-Garcia, J. *et al.* CD8+ T cell metabolic rewiring defined by scRNA-seq identifies a critical role of ASNS expression dynamics in T cell differentiation. *Cell Rep.* **41** (2022).

10. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

11. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).

12. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).

13. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).

14. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

15. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).

16. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).

17. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).

18. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).

19. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).

20. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).

21. Okochi, Y., Sakaguchi, S., Nakae, K., Kondo, T. & Naoki, H. Model-based prediction of spatial gene expression via generative linear mapping. *Nat. Commun.* **12**, 3731 (2021).

22. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50–e50 (2021).

23. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).

24. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).

25. Hao, M. *et al.* Stem enables mapping of single-cell and spatial transcriptomics data with transfer learning. *Commun. Biol.* **7**, 56 (2024).

26. Li, B. *et al.* Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).

27. Li, H. *et al.* A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nat. Commun.* **14**, 1548 (2023).

28. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).

29. Scheele, C. & Wolfrum, C. Functional diversity of human adipose tissue revealed by spatial mapping. *Nat. Rev. Endocrinol.* **17**, 713–714 (2021).

30. Cang, Z. *et al.* Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nat. Methods* **20**, 218–228 (2023).

31. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. methods* **15**, 343–346 (2018).

32. Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. methods* **18**, 1342–1351 (2021).

33. Jung, S.-H. *et al.* Spatiotemporal dynamics of macrophage heterogeneity and a potential function of Trem2hi macrophages in infarcted hearts. *Nat. Commun.* **13**, 4580 (2022).

34. Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023).

35. Ren, X. *et al.* Reconstruction of cell spatial organization from single-cell rna sequencing data based on ligand-receptor mediated self-assembly. *Cell Res.* **30**, 763–778 (2020).

36. Qian, J. *et al.* Reconstruction of the cell pseudo-space from single-cell RNA sequencing data with scSpace. *Nat. Commun.* **14**, 2484 (2023).

37. Zhang, Q. *et al.* Leveraging spatial transcriptomics data to recover cell locations in single-cell RNA-seq with CeLEry. *Nat. Commun.* **14**, 4050 (2023).

38. Wei, R. *et al.* Spatial charting of single-cell transcriptomes in tissues. *Nat. Biotechnol.* **40**, 1190–1199 (2022).

39. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

40. Lohoff, T. *et al.* Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.* **40**, 74–85 (2022).

41. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).

42. Gilmour, D., Rembold, M. & Leptin, M. From morphogen to morphogenesis and back. *Nature* **541**, 311–320 (2017).

43. Berkeley Drosophila Transcription Network Project. Available at: http://bdtnp.lbl.gov/.

44. Luengo Hendriks, C. L. *et al.* Three-dimensional morphology and gene expression in the Drosophilablastoderm at cellular resolution I: data acquisition pipeline. *Genome Biol.* **7**, 1–21 (2006).

45. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801 (1980).

46. Zuo, C. *et al.* Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. *Nat. Commun.* **13**, 5962 (2022).

47. Montaner, J. *et al.* Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat. Rev. Neurol.* **16**, 247–264 (2020).

48. Libby, P. The changing landscape of atherosclerosis. *Nature* **592**, 524–533 (2021).

49. Scherberich, A. *et al.* Tenascin-W is found in malignant mammary tumors, promotes alpha8 integrin-dependent motility and requires p38MAPK activity for BMP-2 and TNF-alpha induced expression in vitro. *Oncogene* **24**, 1525–1532 (2005).

50. Tucker, R. P., Ferralli, J., Schittny, J. C. & Chiquet-Ehrismann, R. Tenascin-C and tenascin-W in whisker follicle stem cell niches: possible roles in regulating stem cell proliferation and migration. *J. Cell Sci.* **126**, 5111–5115 (2013).

51. Neidhardt, J., Fehr, S., Kutsche, M., Löhler, J. & Schachner, M. Tenascin-N: characterization of a novel member of the tenascin family that mediates neurite repulsion from hippocampal explants. *Mol. Cell. Neurosci.* **23**, 193–209 (2003).

52. Meloty-Kapella, C. V., Degen, M., Chiquet-Ehrismann, R. & Tucker, R. P. Effects of tenascin-W on osteoblasts in vitro. *Cell Tissue Res.* **334**, 445–455 (2008).

53. Zhou, A.-X., Hartwig, J. H. & Akyürek, L. M. Filamins in cell signaling, transcription and organ development. *Trends Cell Biol.* **20**, 113–123 (2010).

54. Nakamura, F., Stossel, T. P. & Hartwig, J. H. The filamins: organizers of cell structure and function. *Cell Adhesion & Migr.* **5**, 160–169 (2011).

55. Campbell, I. D. Studies of focal adhesion assembly. *Biochem. Soc. Transactions* **36**, 263–266 (2008).

56. Bandaru, S. *et al.* Targeting filamin A reduces macrophage activity and atherosclerosis. *Circulation* **140**, 67–79 (2019).

57. Seo, W. & Ziltener, H. J. CD43 processing and nuclear translocation of CD43 cytoplasmic tail are required for cell homeostasis. *Blood, The J. Am. Soc. Hematol.* **114**, 3567–3577 (2009).

58. Mody, P. D. *et al.* Signaling through CD43 regulates CD4 T-cell trafficking. *Blood, The J. Am. Soc. Hematol.* **110**, 2974–2982 (2007).

59. Serrador, J. M. *et al.* CD43 interacts with moesin and ezrin and regulates its redistribution to the uropods of T lymphocytes at the cell-cell contacts. *Blood, The J. Am. Soc. Hematol.* **91**, 4632–4644 (1998).

60. McEvoy, L. M., Jutila, M. A., Tsao, P. S., Cooke, J. P. & Butcher, E. C. Anti-CD43 inhibits monocyte-endothelial adhesion in inflammation and atherogenesis. *Blood, The J. Am. Soc. Hematol.* **90**, 3587–3594 (1997).

61. Chittock, E. C., Latwiel, S., Miller, T. C. & Müller, C. W. Molecular architecture of polycomb repressive complexes. *Biochem. Soc. Transactions* **45**, 193–205 (2017).

62. Schmitges, F. W. *et al.* Histone methylation by PRC2 is inhibited by active chromatin marks. *Mol. Cell* **42**, 330–341 (2011).

63. Schuettengruber, B. & Cavalli, G. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136**, 3531–3542 (2009).

64. Greißel, A. *et al.* Histone acetylation and methylation significantly change with severity of atherosclerosis in human carotid plaques. *Cardiovasc. Pathol.* **25**, 79–86 (2016).

65. Meng, X.-D. *et al.* Knockdown of GAS5 inhibits atherosclerosis progression via reducing EZH2-mediated ABCA1 transcription in ApoE-/- mice. *Mol. Ther. Acids* **19**, 84–96 (2020).

66. Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342 (2018).

67. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

68. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).

69. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* **90**, 227–244 (2000).

70. Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. & Smola, A. Correcting sample selection bias by unlabeled data. *Adv. Neural Inf. Process. Syst.* **19** (2006).

71. Sugiyama, M. *et al.* Direct importance estimation for covariate shift adaptation. *Ann. Inst. Stat. Math.* **60**, 699–746 (2008).

72. Poggio, T. & Girosi, F. Networks for approximation and learning. *Proc. IEEE* **78**, 1481–1497 (1990).

73. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245 (2000).

74. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

75. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).

76. Eldan, R. & Shamir, O. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, 907–940 (PMLR, 2016).

77. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

Acknowledgements

This work is supported by National Key R&D Program (2022YFC2704604 to J.S.), National Science Foundation of China (No. 12301620 to J.S., No. 11925103 to W.L., No. 82070463 to Y.Y.), and Science and Technology Commission of Shanghai Municipality (No. 21DZ1201402 to W.L.).

Author contributions statement

J.S. designed the research; Y.T. and J.S. conceived the method; Y.T. and J.S. conducted the numerical results; A.W. and Y.Y. conducted the experiments; Y.T., A.W., Y.Y. and J.S. interpreted the results; W.L., Y.Y. and J.S. supervised the work. All the authors wrote and reviewed the manuscript.

Additional information

Competing interests: The authors declare no competing interests.

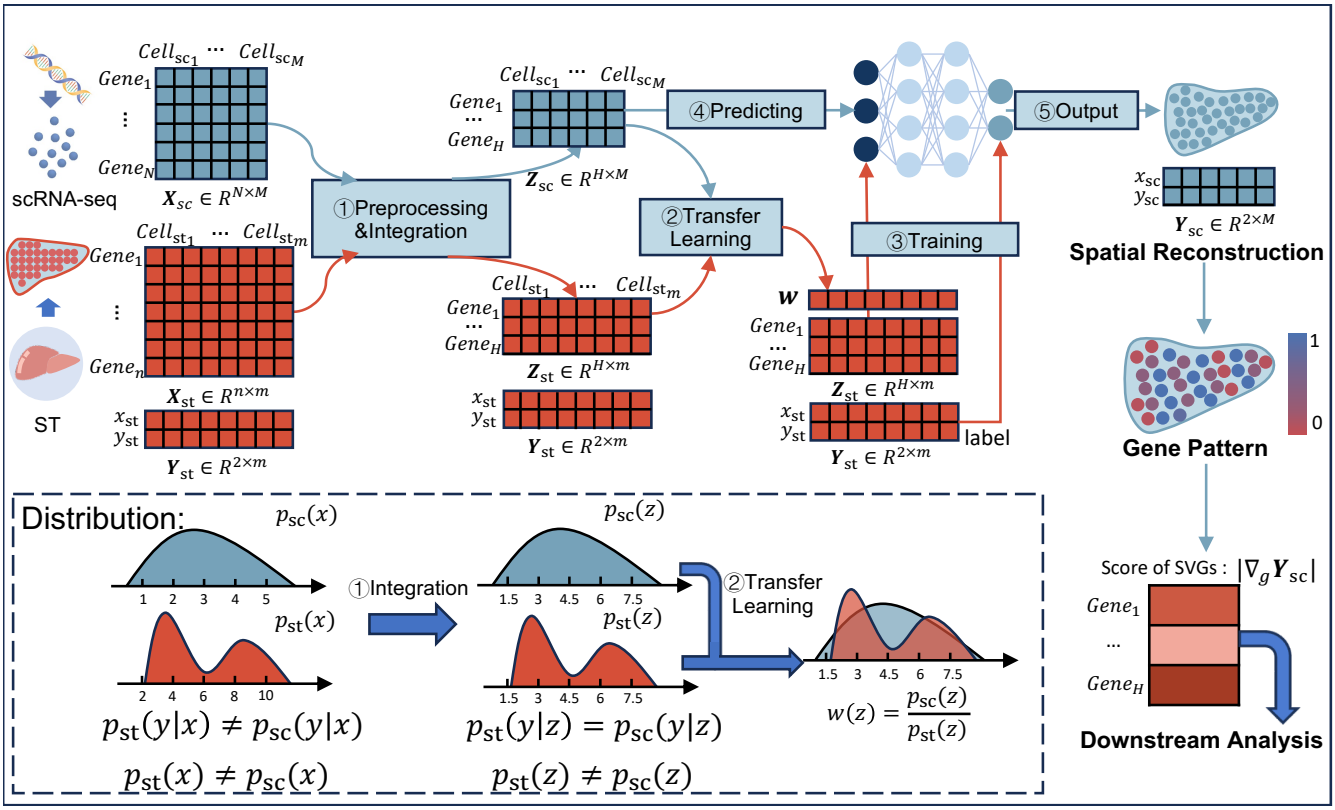

Figure 1. Pipeline of iSORT. The iSORT algorithm workflow, designed to integrate scRNA-seq data and ST data and uncover spatial gene expression patterns along with spatially variable genes (SVGs) consists of five steps. More precisely, 1. Preprocessing and Integration: In the first step, both the scRNA-seq data $\boldsymbol{X}_{\text{sc}}$ and the ST data $\boldsymbol{X}_{\text{st}}$ are preprocessed and integrated onto a shared latent common space $\boldsymbol{Z} \in R^H$. Batch effects are removed and data share the same conditional distributions $p_{\text{st}}(\boldsymbol{y}|\boldsymbol{z}) = p_{\text{sc}}(\boldsymbol{y}|\boldsymbol{z})$, where y represents the spatial coordinate. 2. Transfer Learning: Transfer learning is employed to estimate the density ratio between the processed matrices $\boldsymbol{Z}_{\text{sc}}$ and $\boldsymbol{Z}_{\text{st}}$. This step assigns weight $\boldsymbol{w}$ to each sample in the ST data. 3. Model Training: Using the weighted gene matrix derived from the ST data as input, a multi-layer backpropagation neural network (BP neural network) is trained. The network is designed to predict spatial coordinates as its output. The training of this model aims to capture the relationship between gene expression and spatial coordinates. 4. Prediction: The trained model is applied to the scRNA-seq data to reconstruct its spatial organization. 5. Output: The final output is the spatial coordinates of scRNA-seq, which are used to reconstruct spatial organization, reveal gene spatial expression patterns, and identify SVGs.

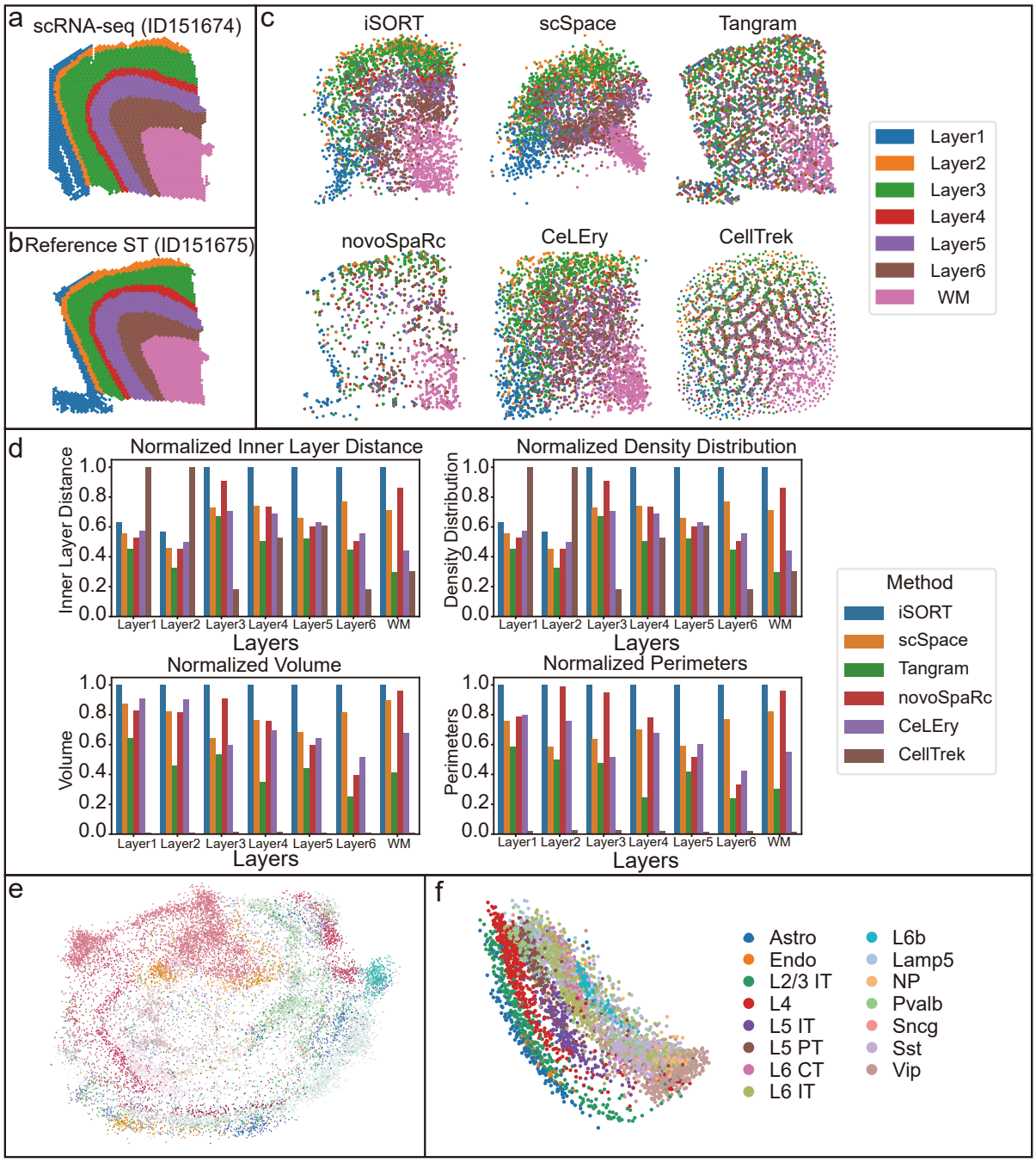

Figure 2. Spatial organization reconstruction by iSORT on the DLPFC, mouse embryo, and mouse brain datasets. (**a**) The spatial representation of ST slice ID151674 in human DLPFC layers, which is taken as the scRNA-seq data without spatial coordinates in the experiment. (**b**) The spatial representation of ST slice ID151675, which is used as the ST reference slice. (**c**) Spatial reconstruction results for ID151674 by six algorithms: iSORT, scSpace, Tangram, novoSpaRc, CeLEry, and CellTrek. The different colors represent different cortical regions. (**d**) Bar charts to compare the performance of the six algorithms based on four indicators: normalized intra-layer distance, normalized density distribution, normalized volume of each layer, and normalized perimeter of each layer. iSORT (blue bar) outperforms the others in most of cases. (**e**) Spatial reconstruction of the mouse embryo dataset with iSORT. The different colors represent different cell types. (**f**) Spatial reconstruction of the mouse brain dataset with iSORT. The different colors represent different cell types.

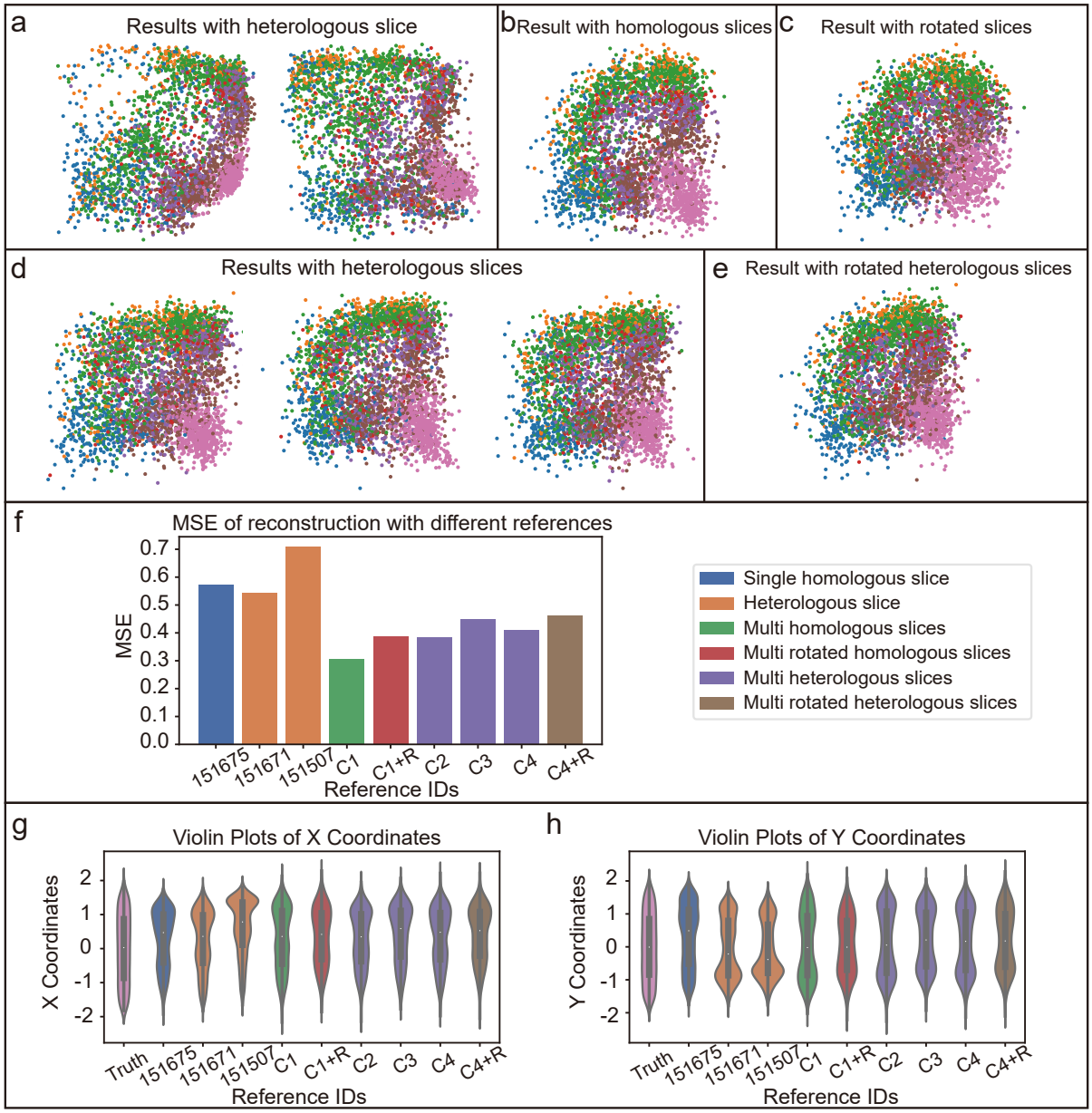

Figure 3. Results for reconstructing human DLPFC slice ID151674 using multiple ST references. (**a**) Reconstruction results using individual reference slices ID151507 (Left) and ID151671 (Right). (**b**) Reconstruction results with combination 1 (C1) using three homologous slices (Sample ID151673, ID151675, and ID151676) as references via iSORT. (**c**) Reconstruction results using three rotated homologous slices (Sample ID151673, ID151675, and ID151676) as references via iSORT. (**d**) Reconstruction results obtained using iSORT with three sets of heterogeneous slices as references: the first combination (C2) includes sample IDs 151675, 151507, and 151508 (left); the second combination (C3) comprises sample IDs 151675, 151670, and 151671 (middle); and the third combination (C4) consists of sample IDs 151675, 151607, and 151671 (right). (**e**) Reconstruction results using three rotated heterogeneous slices (Sample ID151675, ID151671, and ID151507) as references via iSORT. (**f**) Mean square error (MSE) across different reconstruction scenarios (R represents 'Rotation'). (**g**) Violin plots of X coordinates across different reconstruction scenarios. (**h**) Violin plots of Y coordinates across different reconstruction scenarios.

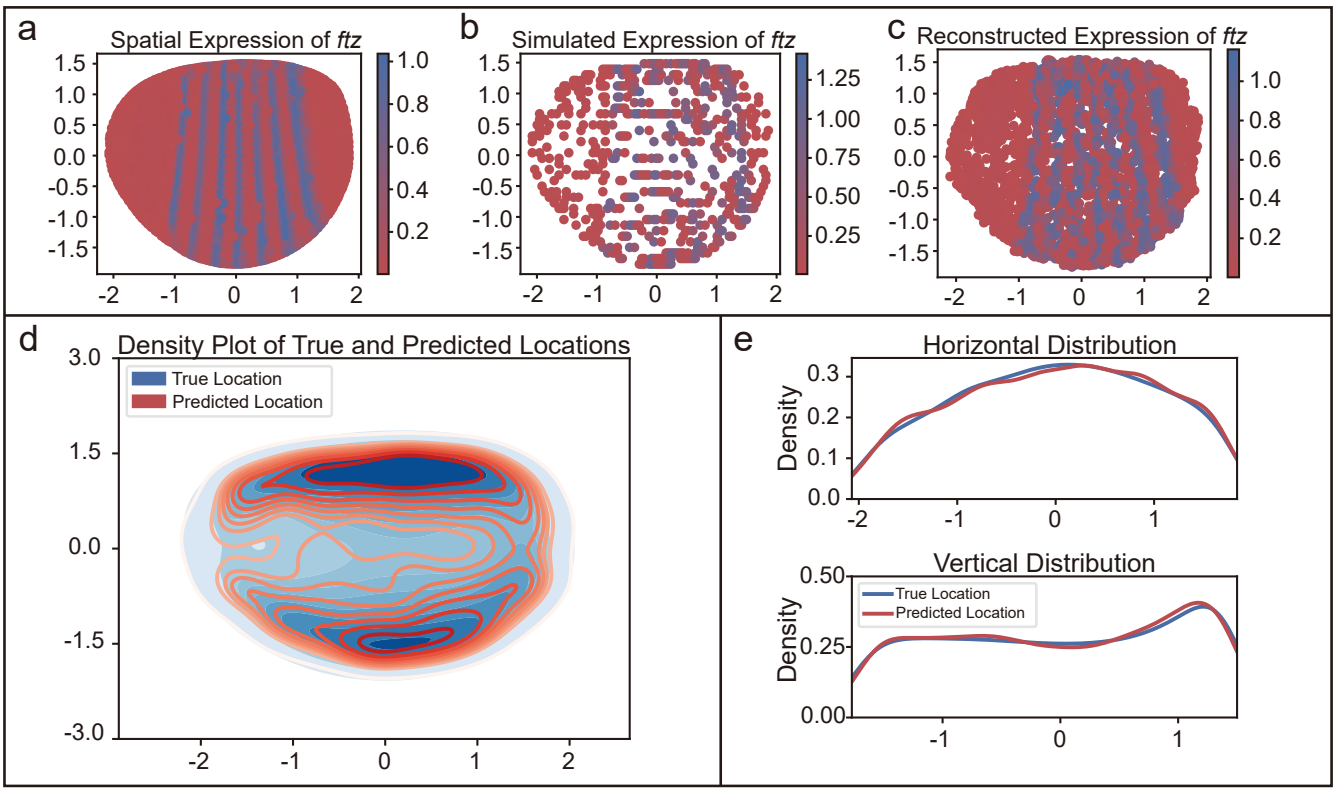

Figure 4. Pattern reconstruction of the *ftz* gene from *Drosophila* embryo data using iSORT. (**a**) Visualization of *ftz*'s distribution in the original space. (**b**) Visualization of *ftz*'s distribution in the simulated coarse-grained space, demonstrating the disruption of gene patterns in the low-resolution ST spots. (**c**) Visualization of *ftz*'s distribution reconstructed from the scRNA-seq data by iSORT. (**d**) Density plot contrasting the true (Blue) and predicted (Red) spatial distribution of the *ftz* gene. (**e**) Kernel density estimation plot contrasting the true (Blue) and predicted (Red) spatial distributions of the *ftz* gene. The horizontal distribution corresponds to the distribution of cells along the horizontal axis in figures a and c, while the vertical distribution corresponds to the distribution along the vertical axis in these figures.

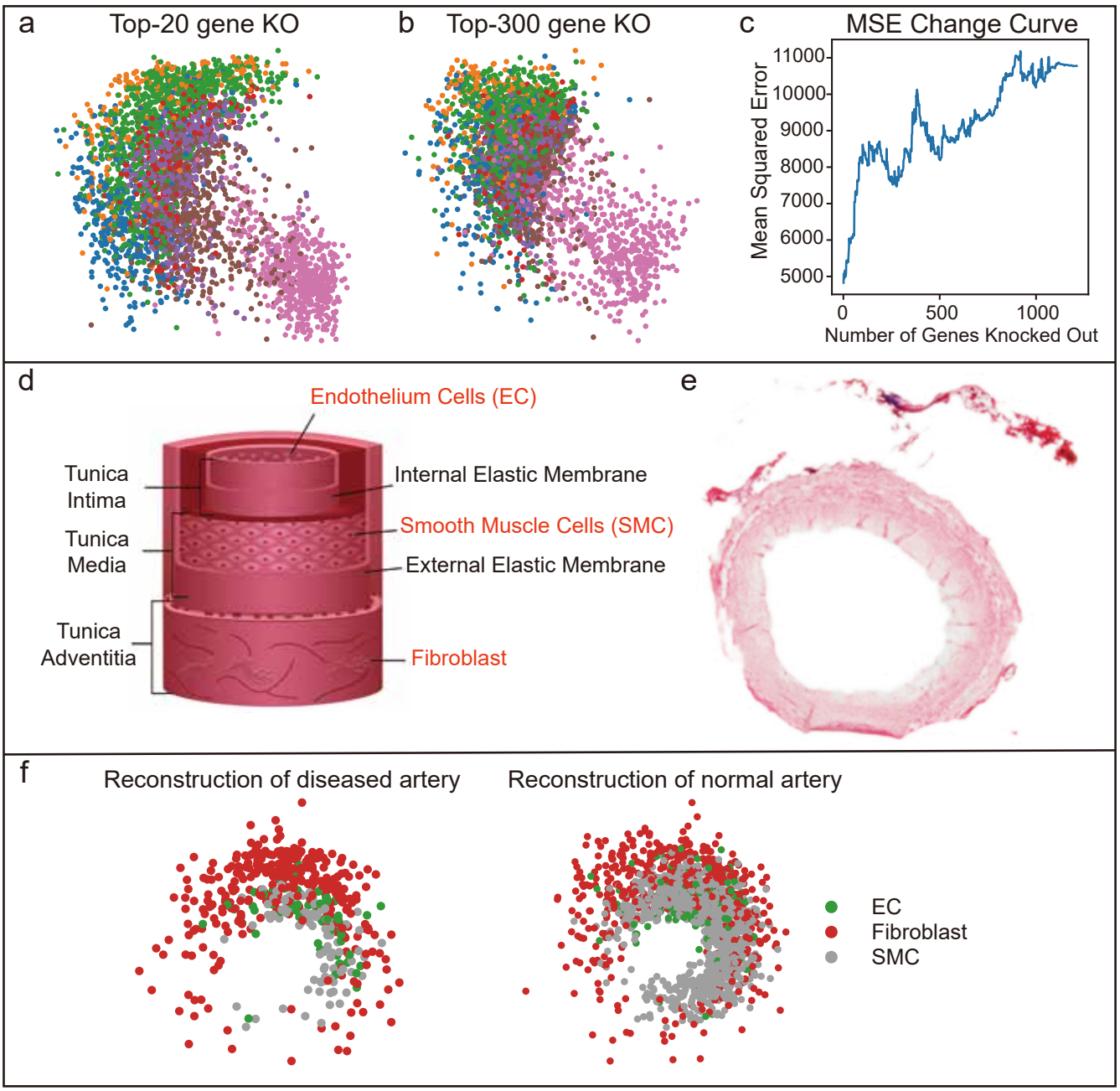

Figure 5. Simulated gene knockout experiments on DLPFC dataset and analysis of the human artery dataset. (**a**) Reconstruction result of DLPFC with the top 20 SVGs knocked out. (**b**) Reconstruction result of DLPFC with the top 300 SVGs knocked out. (**c**) MSE change curve showing stepwise knockout of all genes according to gradually varied SVG scores. (**d**) Schematic diagram of artery structure illustrating layered composition: the innermost layer is lined with Endothelial Cells (EC), followed by Smooth Muscle Cells (SMC), and the outermost layer composed of Fibroblasts. (**e**) Hematoxylin and Eosin staining of the non-atherosclerotic (normal) artery aligned with spatial transcriptomics results. (**f**) Reconstruction results for an artery with atherosclerosis (left) and a normal artery (right), showcasing selected cell types from (**d**): Endothelial Cells (EC), Smooth Muscle Cells (SMC), and Fibroblasts.