SSLAR: Deconvoluting spatial transcriptomics data with single-cell transcriptomes through semi-supervised NMF and least angle regression

January 25, 2025

Contents

1 Supplementary Methods

1.1 Data preparation

As shown in Table S1, we used 22 real sc/snRNA-seq datasets with available and purified cell types to evaluate the SSLAR performance on different tissues. The first 13 datasets were from PBMC and were obtained based on 13 different sequence platforms including Cel-Seq2, ChromiumV2, Chromium V2 single nucleus, C1HT-medium, C1HT-Small, ddSeq, Drop-Seq, gmcSCRB-Seq, ICELL8, inDrop, MARS-Seq, QUARTZSeq2, and SMART-Seq2, respectively. They were downloaded from the Gene Expression Omnibus (GSE133549)[1] and taken as benchmarks. The other 9 scRNA-seq datasets were from HNSCC [2], melanoma [3], human pancreas [4], human kidney tumors [5], mouse primary visual cortex [6], mouse brain [7], PDAC-A [8], PDAC-B [8], and pancreatic immune [9], respectively.

Among the 9 datasets, the former 4 scRNA-seq datasets from HNSCC, human melanoma, pancreas, and kidney tumors were used to construct the benchmarking synthetic data. To investigate the SSLAR deconvolution ability under different conditions, we generated 10 ST datasets for each of the above 17 scRNA-seq datasets, which included 13 PBMC datasets, and the former 4 datasets from HNSCC, human melanoma, pancreas, and kidney tumors, respectively. The mouse primary visual cortex were then analyzed in conjunction with two gridded spatial transcriptome datasets[10, 11]. In addition, the remaining 4 scRNA-seq datasets adding to 4 ST datasets were used to evaluate the SSLAR performance on real datasets. The 4 ST datasets were from mouse brain anterior (10X), mouse brain posterior (10X), PDAC-A [8], and PDAC-B [8], respectively.

Table S1: Real data information

tissue	platform	cell types	cells / spots	gene zero percentage
PBMC [1]	C1HT-medium	8	2,210	67.70%
	C1HT-small	8	1,603	75.57%
	CEL-Seq2	8	1,082	78.82%
	Chromium(sn)	8	1,512	93.39%
	Chromium	8	1,604	82.65%
	ddSEQ	8	2,104	91.16%
	Drop-Seq	8	2,260	94.31%
	ICELL8	9	1,927	90.38%
	inDrop	8	685	94.33%
	MARS-Seq	7	1,458	95.11%
	mcSCRB-Seq	7	1,645	91.96%
	Quartz-Seq2	7	1,314	75.76%
	Smart-Seq2	8	732	90.19%
HNSCC [2]	Smart-Seq2	9	5,578	81.68%
melanoma [3]	Smart-Seq2	7	4,092	81.32%
pancreas [4]	Cel-Seq2	9	2,122	73.02%
kidney tumor [5]	10X	11	5,685	94.70%
mouse primary visual cortex [6]	Smart-Seq	15	14,249	88.49%
mouse brain [7]	Smart-Seq2	42	73,363	77.20%
PDAC-A [8]	inDrop	19	1,926	91.54%
PDAC-B [8]	inDrop	13	1,733	92.85%
pancreatic immune [9]	10X	10	57,530	89.62%
mouse cortex (ST) [10]	SeqFish+	13	524 (grids: 72)	66.62%
mouse visual cortex (ST) [11]	STARmap	15	1,523 (grids: 189)	78.63%
mouse brain anterior (ST)	10X	undefined	2,696	81.70%
mouse brain posterior (ST)	10X		3,353	85.37%
PDAC-A (ST) [8]	indrop		428	94.09%
PDAC-B (ST) [8]	indrop		224	95.16%

1.2 Synthetic mixtures

To measure the cell type annotation performance of different ST analysis methods on various data from different sequencing platforms and different tissues, we generated 1,000 cellular synthetic mixtures for each scRNA-seq dataset from PBMC, HNSCC, melanoma, pancreas, and kidney tumors, respectively. The 1,000 cellular synthetic mixtures represented 1,000 synthetic spots and were taken as an ST dataset. Similar to the synthetic mixture generation method in SPOTlight [12], we randomly selected 2-10 cells from each scRNA-seq dataset and then combined their transcriptomic profiles to generate a synthetic mixture using the $test\_spot\_fun$ function in SPOTlight. If the resulting mixture had $> 25,000$ UMI counts, we down-sampled it to $20,000$ UMI counts to better simulate real capture information. Finally, we generated 10 ST datasets for each scRNA-seq dataset.

1.3 Notation

The variables and matrices in the SSLAR framework are defined as follows:

- N - A set of all cells from an scRNA-seq dataset.
- M - A set of all spots from an ST dataset.
- G - Intersection between sets of all genes from ST and scRNA-seq data.
- C - Number of cell types in an scRNA-seq dataset.
- T - Number of topics used to dimensionality reduction, equal to C.
- X - Matrix with dimension of $G \times N$ containing an scRNA-seq dataset.
- A - Matrix with dimension of $G \times T$ containing the gene distribution of each topic.
- S - Matrix with dimension of $T \times N$ containing the topic distribution of each cell.
- X' - Matrix with dimension of $G \times M$ containing an ST dataset.
- S' - Matrix with dimension of $T \times M$ containing the topic distributions of each spot.
- Q - Matrix with dimension of $T \times C$ containing the topic distributions for each cell type.
- P - Matrix with dimension of $C \times M$ containing the cell type weights of each spot.

1.4 Topics identification

To identify topics from scRNA-seq data, first, a unit-variance normalization method [13, 14], implemented by the SCTransform function in the Seurat package, is used to normalize paired and unmatched scRNA-seq and ST raw count matrices in X and X', respectively. Next, the basis matrix A and coefficient matrix S are initialized to reduce the SSLAR variability and promote its consistency through marker genes: each topic and each column in A are initialized based on unique marker genes that are selected using the *FindAllMarkers* function in the Seurat package. Each topic in S is initialized as 1 using the corresponding belongance of each cell. As a result, we seed the SSLAR method with prior information to guide it towards biologically significant results. Lastly, the initial gene expression matrix X is factorized into two low-dimensional non-negative matrices based on ssNMF. Consequently, we obtain two sparser matrices to further decode cell-type-specific topic profiles.

1.4.1 Non-negative matrix factorization

To factorize an scRNA-seq data into two non-negative matrices, we fully utilize two matrix similarity measurement terms: the first term is the standard Frobenius norm $\|U - V\|_F$, and the second one is the information divergence [15] defined by Eq. (1):

$$D(U\|V) = \sum_{i,j} \left(U_{ij} \log \frac{U_{ij}}{V_{ij}} - U_{ij} + V_{ij} \right) \tag{1}$$

where $D(U\|V) \geq 0$ when $U = V$. The information divergence can be reduced and represented using the Kullback-Leibler divergence when both U and V denote probability distributions (i.e., $\sum U_{ij} = \sum V_{ij} = 1$),

thus, it is usually replaced using the generalized Kullback-Leibler divergence [16]. Consequently, for an scRNA-seq dataset $X \in \mathbb{R}^{G\times N} \geq 0$ with T dimensions, NMF decomposes it into a product of two low-dimensional non-negative matrices A and S by Eq. (2):

$$X \approx A * S \tag{2}$$

where $A \in \mathbb{R}^{G\times T} \geq 0$ and $S \in \mathbb{R}^{T\times N} \geq 0$ denote the dictionary matrix and representation matrix, respectively. And $T < min\{||G||, ||N||\}$ is used to reduce the dimension of scRNA-seq data. Notably, each column in A usually denotes one topic. Each column in S is an approximate representation corresponding to one column in X. Consequently, scRNA-seq data are well approximated based on a linear combination of multiple latent topics.

1.4.2 Semi-supervised NMF

ssNMF is a modified version of NMF and can conduct maximum likelihood estimation for a given uncertainty model based on reconstruction and supervision errors. Inspired by the ssNMF model proposed by Haddock et al. [17], we build an objective function to jointly incorporates scRNA-seq dataset X and its class labels by Eq. (3):

$$\begin{aligned} &F(A, B, S; X, Y) \\ &= \underset{A,B,S\geq 0}{\operatorname{argmin}} D(W \odot X \| W \odot AS) + \lambda \| L \odot (Y - BS) \|_F^2 \end{aligned} \tag{3}$$

where Y denotes target matrix, $A \in \mathbb{R}^{G\times T} \geq 0$, $B = E^{T\times T}$, $S \in \mathbb{R}^{T\times N} \geq 0$, $Y \in \mathbb{R}^{T\times N} \geq 0$, W and L are two matrices with all element values of 1, $\odot$ and AS indicate element-wise multiplication and standard matrix multiplication, respectively. And the parameter $\lambda = \|X\|_F = \sqrt{\operatorname{tr}(X^\top X)} = \sqrt{\sum_{i,j} x_{ij}^2}$ is used to balance the relative importance between the reconstruction error (the first term) penalized by the information divergence and the supervision term (the second term) penalized by the Frobenius norm.

To solve model (3), we iteratively update A, B, and S for n iterations by Eq. (4):

$$\begin{aligned} A &\leftarrow \frac{A}{WS^\top} \odot \left[\frac{(W \odot X)}{(W \odot AS)} \odot W \right] S^\top \\ B &\leftarrow B \odot \frac{(L \odot Y)S^\top}{(L \odot BS)S^\top} \\ S &\leftarrow S \odot \frac{A^\top \left[\frac{(W\odot X)}{(W\odot A)} \odot W \right] + 2\lambda B^\top (L \odot Y)}{A^\top W + 2\lambda B^\top (L \odot BS)} \end{aligned} \tag{4}$$

where Eq. (4) is an entrywise gradient descent model and can individually select each entry step by step to ensure nonnegativity. Consequently, we compute the gradient of the objective function3 with respect to A, B, and S by Eq. (5):

$$\begin{aligned} \nabla_A F &= WS^\top - \left[\frac{W \odot X}{W \odot AS} \odot W \right] S^\top \\ \nabla_B F &= -2[L \odot (Y - BS)]S^\top \\ \nabla_S F &= \lambda A^\top W - \lambda A^\top \left[\frac{W \odot X}{W \odot AS} \odot W \right] - 2B^\top [L \odot (Y - BS)] \end{aligned} \tag{5}$$

Next, A, B, and S is solved by Eq. (6):

$$\begin{aligned} A &\to A - \Gamma \odot \nabla_A F \quad when \quad \Gamma = \frac{A}{WS^\top} \\ B &\to B - \Gamma \odot \nabla_B F \quad when \quad \Gamma = \frac{B}{2(L \odot BS)S^\top} \\ S &\to S - \Gamma \odot \nabla_S F \quad when \quad \Gamma = \frac{S}{2B^\top (L \odot BS) + \lambda A^\top W} \end{aligned} \tag{6}$$

1.5 Topic distribution analysis

LARS [18] is a novel optimal stepwise selection model. It elucidates very strong data modeling ability and thus obtains wide applications. To populate topic distributions, we use the obtained basis matrix A as the basis and design an LARS-based topic distribution analysis method to map each spot's ST data into topic distribution. Given three non-negative matrices $X' \in \mathbb{R}^{G\times M} \geq 0$, $A \in \mathbb{R}^{G\times T} \geq 0$, and $S' \in \mathbb{R}^{T\times M} \geq 0$, and T predictors, the LARS algorithm builds a linear regression model by selecting $k(k \leq T)$ predictors in turn.

First, T regression coefficients are set to zero. Next, predictor a_1 with the closest correlation with A is selected from the T predictors by calculating the cosine of A and x'_j, and computing the regression coefficient $\hat{\beta}_{j1}$. Third, a linear regression model between x'_j and a_2 is built based on a_1. And a_2 is selected from the remaining $(T-1)$ predictors along the direction of a_1 until there is the smallest residual between $(x'_j - a_1\hat{\beta}_{j1})$ and a_2. Similar to the a_2 selection, we obtain the following predictors. Finally, the k predictors with the strongest linkages with A are selected from the T predictors. Particularly, we use a simple Mallows's C_p statistic method to obtain the optimal k value by Eq. (7):

$$C_p = \frac{SSE_k}{MSE_T} - G + 2(T+1) \tag{7}$$

where $MSE_T = \frac{SSE_T}{T}$ and $SSE_k = \sum_{i=1}^{T}\left(x'_{ji} - \hat{x}'_{jki}\right)^2$.

The matrix $X' \in \mathbb{R}^{G\times M} \geq 0$ represents the distribution of G genes in M spots and $X' = \{x_1, x_2, ..., x_M\}$. Taking $\{x_1, x_2, ..., x_M\}$ as the input of LARS, we obtain M regression coefficients $\{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_M\}$ corresponding to X by Eq. (8):

$$\widehat{x}_i = A\widehat{\beta}_i = \sum_{j=1}^{T} a_j\widehat{\beta}_{ij} \tag{8}$$

where $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_M\right)^T = S'$.

Finally, each spot's transcriptome in X' is mapped into S', and the topic profile distribution over each spot is used to decode its composition by Eq. (9):

$$X' \approx A * S' \tag{9}$$

Cell type annotation

To annotate cell types, first, we capture all cells from the same cell type and compute the median of each topic. Consequently, we obtain a consensus cell-type-specific topic signature and cell-type-specific topic profiles Q from S. Next, we utilize NNLS to compute the weight of each cell type by Eq. (10):

$$S' \approx Q * P \tag{10}$$

In particular, we calculate the total sum of squares (TSS) and the residual sum of squares (RSS) for each spot by Eq. (11):

$$\begin{aligned} TSS &= \sum_{i=1}^{C}(p_i - 0)^2 \\ RSS &= \sum_{i=1}^{C}(p_i - \hat{p}_i)^2 \end{aligned} \tag{11}$$

where p and $\hat{p}$ denote a cell type distribution gold standard and the predicted cell type distribution within a spot, respectively. And the ratio of unexplained residuals R_u for each spot is computed to evaluate the quality of a predicted spot composition by Eq. (12):

$$R_u = RSS/TSS \tag{12}$$

1.6 Label refinement

Due to the presence of spatial heterogeneity in spatial transcriptomics, the expression patterns of cells are influenced by their spatial location within the tissue[19]. This means that the gene expression patterns or cellular compositions of a spot are affected by its position, thereby resulting in a certain degree of continuity in the transcriptional profiles of adjacent spots.

To leverage this theory and optimize cell type prediction using the spatial positions of spots, we employed spatial proximity and cell type similarity analysis to refine the proportions of cell types. Firstly, spatial proximity and cell type similarity are computed using Euclidean distance and Manhattan distance, respectively. Subsequently, two kd-trees are employed to find two sets of k nearest neighbors of each spot in the above two similarity matrices, respectively. The intersection between the two kd-trees is used to simulate mixtures from neighboring spots.

In the subsequent step, It is assumed that spatial information is advantageous only when the composite mixtures derived from adjacent locations exhibit cell type compositions similar to the predicted values. Locations that exceed the mean distance are excluded from the optimization process. Thereafter, for those within the specified range, optimization is achieved through the computation of a weighted mean, combining the synthetic mixtures with the initially predicted proportions. Finally, the distances that remain are normalized within the interval of 0 to 0.5 to establish the weighting factors, which are then utilized to adjust the initial predictions according to the synthesized data.

This approach ensures that only relevant spatial information is used to improve the prediction accuracy, focusing on those spots that are closely related spatially. By doing so, we can enhance the reliability of predictions while mitigating the influence of distant or less relevant spots.

Furthermore, the minimum weight θ is applied to determine which cell types a spot belongs to. For b cell types $\{c_1, c_2, \cdots, c_b\}$ in a spot, their weights are $\{w_1, w_2, \cdots, w_b\}$, the spot belongs to c_i and c_j when $w_i \geq \theta$ and $w_j \geq \theta$.

2 Supplementary Results

2.1 Performance evaluation

We used 4 evaluation metrics, i.e., F1 Score, Jensen-Shannon divergence (JSD), Pearson's correlation coefficient (PCC), and root mean square error (RMSE), to assess the SSLAR cell type annotation performance on real and synthetic datasets. We evaluated: (i) if it can accurately predict cell types within a mixture (F1 Score). (ii) if the predicted cell type proportions accurately represent the true composition in a spot (JSD, PCC, and RMSE).

The F1 Score denote correctly predicted cell type presence or absence. The PCC index is used to measure the similarity between cell type distributions P on synthetic ST data and the predicted cell type distributions Q. The RMSE and JSD indices are used to evaluate the differences between the above two distributions P and Q.

The PCC value is defined by Eq. (13):

$$\rho_{P,Q} = \frac{cov(P,Q)}{\sigma_P \sigma_Q} \tag{13}$$

where cov denotes the covariance between P and Q, σ_P and σ_Q indicate the standard deviations corresponding to P and Q, respectively.

The RMSE value is denoted by Eq. (14):

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (P_t - Q_t)^2}{n}} \tag{14}$$

The JSD value can be computed by Eq. (15):

$$\text{JSD}(P||Q) = \frac{1}{2} D_{kl}(P\|M) + \frac{1}{2} D_{kl}(Q\|M) \tag{15}$$

where

$$\begin{aligned} D_{kl}(P\|Q) &= \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \\ M &= \frac{1}{2}(P+Q) \end{aligned} \tag{16}$$

Higher F1 Score and PCC value and lower JSD and RMSE values indicate better cell type annotation performance. The 2.3.2, 2.3.1, and 2.3.3 experiments were iteratively performed for 10 times. The average performance of experimental results for the 10 times was taken as the final performance.

2.2 Performance comparison

To compare the cell type identification performance of our proposed SSLAR framework and other 8 state-of-the-art ST analysis tools, (i.e., STRIDE [20], DSTG [21], RCTD [22], SpatialDDLS [23], SCDC [24], SPOTlight v0.1.7 [12], POLARIS [25], and CARD [26]) on scRNA-seq and ST data, we conducted a series of experiments on 19 datasets. Particularly, synthetic mixtures can't generate the spatial correlation information, but CARD must model spatial correlation when deconvoluting cell-type compositions, thus, CARD was not compared in Section 2.3.2 and 2.3.1.

2.2.1 Performance on predicting spatial distribution

To assess the effectiveness of SSLAR and other 8 spatial deconvolution methods when predicting cell type composition of spots, similar to Ref. [27], we simulated 'multi-cell spot problem' generated in 10X Visium ST datasets by 'gridding' a dataset without this problem (acquired using seqFISH+; Smart-seq; mouse cortex). The cell type composition of each spot has been annotated and is taken as the ground truth when simulating an ST dataset with potentially unconfirmed cell type compositions in each spot (Figure S1A). The original dataset contains 524 cells and has been annotated as 13 cell types. After gridding, the simulated dataset comprised of 72 spots where each spot captured 1-18 cells. The locations of microglia cells were plotted. The results demonstrated that SSLAR performed better than other 8 methods in terms of PCC and JSD (0.7790 and 3120), followed by CARD (0.7631 and 4640), STRIDE (0.6482 and 0.5567) (Figure S1B).

The F1 score and RMSE were used to quantify the performance of all 9 deconvolution methods when inferring the cell type composition of spots in mouse cortex (Figure S2A). Figure S3 and S4 elucidated locations of other cell types on the mouse cortex dataset. SSLAR computed the highest average F1 score and PCC along with the lowest average JSD and RMSE for all spots. Additionally, computational time of SSLAR and the 8 tools is shown in S11. The results show that SSLAR can generate predictions within a relatively short period of time.

SSLAR was also applied to analysis on mouse visual cortex containing 15 cell types composed of 1,523 cells. After gridding, the simulated dataset only had 189 spots (Figure S1C). SSLAR computed the best average F1 score, PCC, RMSE, and JSD when annotating all cell types on mouse visual cortex. (Figure S2B). Moreover, through the ground truth of the locations for the L6 excitatory neurons, SSLAR computed PCC, RMSE and JSD values of 0.8684, 0.1484 and 0.1651 for the assignations of the L6 excitatory neurons, surpassing other 8 deconvolution methods (Figure S1D). Figure S5, S6, S7, S8, S9 and S10 delineate locations of other cell types on mouse visual cortex.

2.3 Benchmarking

To analyze the affects of scRNA-seq datasets from different sequencing depths and different sequencing platforms on the deconvolution performance, we constructed benchmarking datasets. First, to assess the performance of SSLAR on different sequencing depths, we generated 1,000 spots by aggregating 2-10 randomly selected cells and separately down-sampling to 30,000, 20,000, 10,000, 5,000 and 1,000 counts. The 1,000 spots constituted an ST dataset. The cell type labels of the selected 2-10 cells were known in the original scRNA-seq dataset, the obtained cell type proportions in a spot can be taken as a golden standard to measure the deconvolution performance.

Additionally, to detect cell type identification accuracy of SSLAR and the 7 other methods (not including CARD), we adopted a down-sampled technique with the same sequencing depth (20,000 reads/cell) on 13 peripheral blood mononuclear cells (PBMC) datasets based on the specifications provided by Ref. [1].

2.3.1 Benchmarking SSLAR

On the real ST data, since some spots might be partially captured, sequencing depth might influence the captured gene numbers and further affect the deconvolution performance. In this section, we evaluated the performance of SSLAR and other 7 comparison methods on different sequencing depths. Figure S12A illustrates F1 Score, PCC, JSD, and RMSE computed by the above 8 methods on sequencing depths of 1,000, 5,000, 10,000, 20,000, and 30,000, respectively. Table S2 show F1 Score, JSD, PCC, and RMSE computed by the 8 methods under the above different sequencing depths. SSLAR computed the highest F1 Score, and PCC as well as the lowest

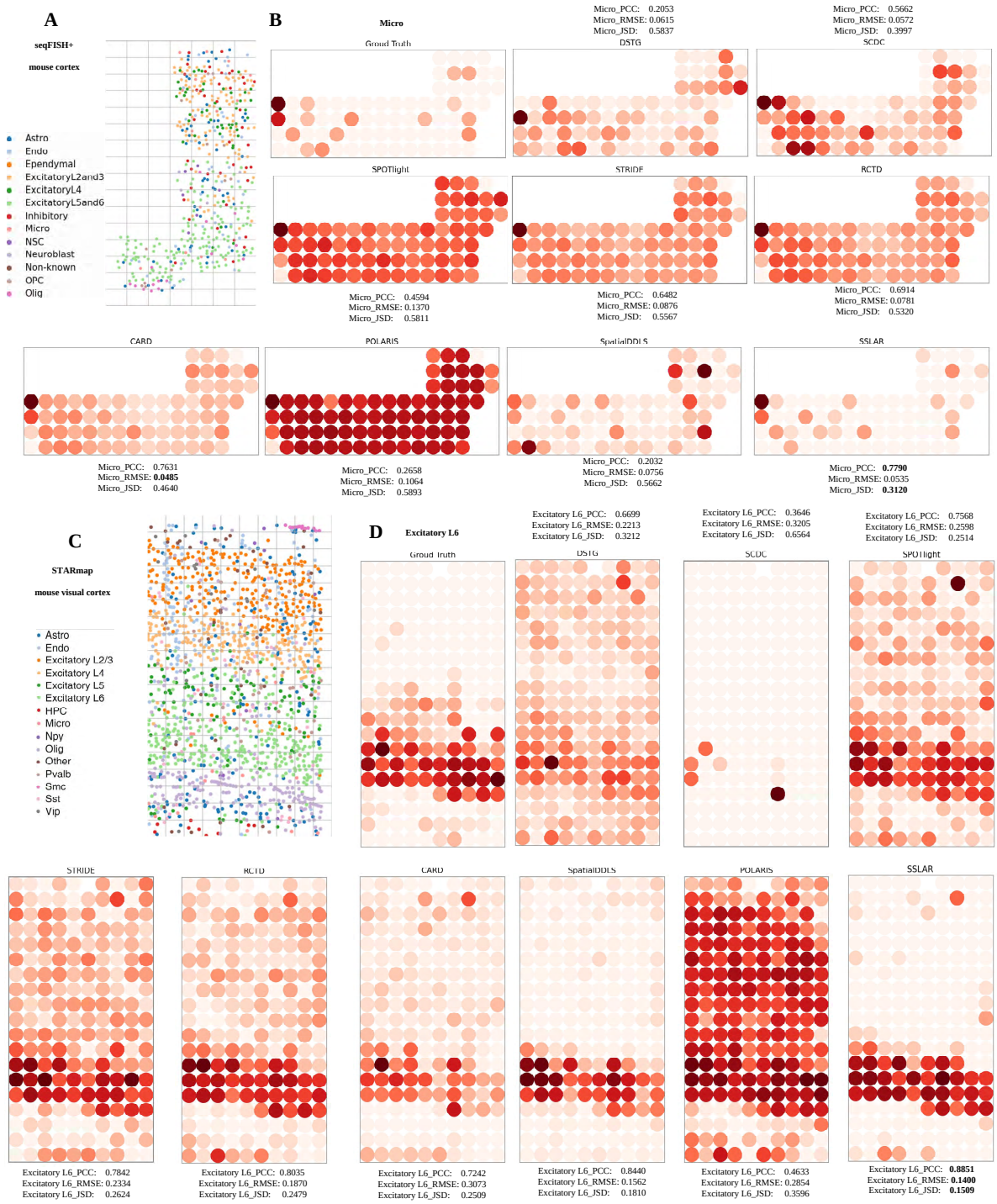


Figure S1: Performance of 9 deconvolution methods on one cell type. (A) A seqFISH+ slide of mouse cortex with cell type annotation. Each grid denotes a simulated spot composed of multiple cells. (B) The proportion of microglia cells in spots simulated from mouse cortex, containing the ground truth and predictions. (C) A STARmap slide of mouse visual cortex. (D) The ground truth and the predicted proportion of L6 excitatory neurons in mouse visual cortex.

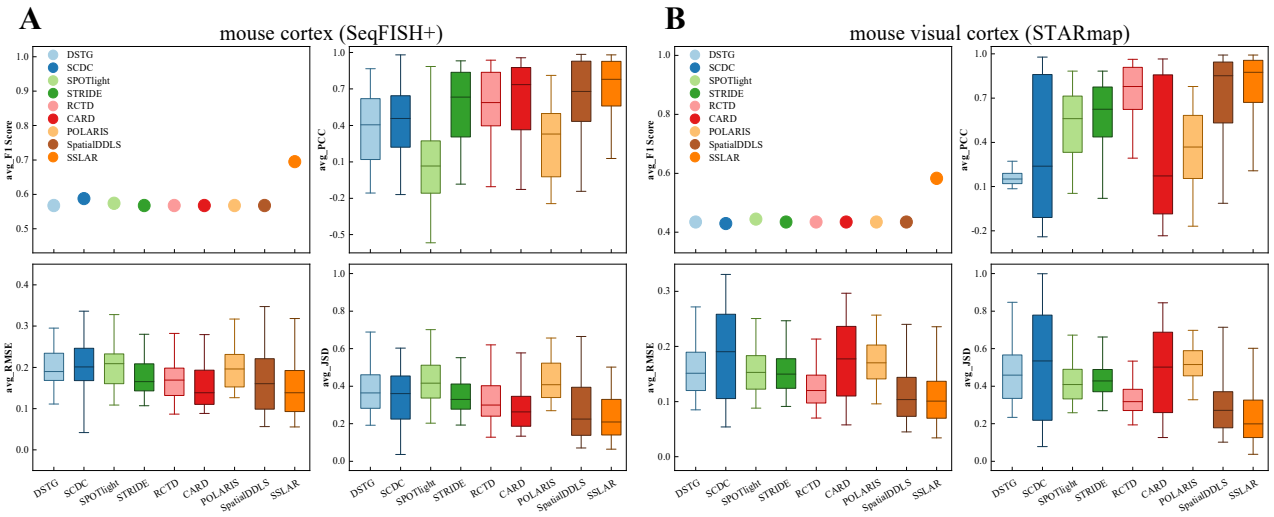


Figure S2: Performance of 9 deconvolution methods on all spots. (A) mouse cortex (SeqFISH+). (B) mouse visual cortex (STARmap)

RMSE and JSD in most cases. Although the SSLAR's deconvolution performance declined with the gradually decreasing sequencing depths, it was still the most robust to the sequencing depths among all eight methods. Collectively, SSLAR more accurately deciphered the cell type proportions in a spot, and was more insensitive to spatially co-localized cell type distribution and low sequencing depth.

scRNA-seq data produced by different protocols have vastly variable quality, which severely affects downstream applications including deconvolution. Thus, we used 13 PBMC datasets from different sequencing platforms to evaluate if different scRNA-seq technologies used to generate synthetic ST data affect the cell type identification accuracy. Figure S12B illustrates F1 Score, PCC, RMSE, and JSD of SSLAR and other 7 methods on PBMC (Smart-seq2). Figure S13 and S14 demonstrate their performance on other 12 PBMC datasets. Table S3 and S4 elucidate F1 Score, JSD, PCC, and RMSE computed by the 8 methods on 13 synthetic datasets from different sequencing platforms. Using F1 core, PCC, RMSE, and JSD as evaluation metrics, SSLAR obtained the optimal cell type annotation performance on 13 different sequencing platforms under majority of conditions, fully demonstrating its deconvolution performance.

In addition, we down-sampled ST data (20,000 reads per cell) and trained the SSLAR model on synthetic mixtures from each sequencing platform. Figure S12C elucidates F1 Score, JSD, PCC, RMSE, and running time of SSLAR on 13 PBMC datasets. Red upper triangle, purple circle, and blue lower triangle denote the third quartile (75th percentile), median (50th percentile), and the first quartile (25th percentile) of the performance list, respectively. SSLAR obtained the best performance on Quartz-Seq2, Smart-Seq2, and Chromium protocols. Notably, its performance on single-nucleus (sn) sequencing platform (Chromium sn) was comparable to scRNA-seq platform although we implemented down-sampling. In general, SSLAR obtained the optimal cell type annotation performance on scRNA-seq data with known labels and corresponding marker genes. But it also carried out accurate predictions on other commonly used sc/snRNA-seq platforms.

2.3.2 The robustness analysis on different tissues

To examine the robustness of SSLAR and other 7 ST analysis methods (except CARD) on different tissues, we used 4 scRNA-seq datasets from different tissues to generate the benchmarking synthetic data. The 4 tissues were from head and neck squamous cell carcinomas (HNSCC) (SMART-Seq2), melanoma (SMART-Seq2), human pancreas (Cel-Seq2), and human kidney tumors (10x Genomics), respectively. Similar to SPOTlight [12], each of the 4 scRNA-seq datasets was used to generate 10 ST datasets. Figure S15 shows the computed F1 Score, JSD, PCC, and RMSE on the 4 different tissues. The results showed that SSLAR and RCTD computed the higher F1 Score and PCC and the lower JSD and RMSE on the 4 tissues in most cases, demonstrating their powerful cell type annotation ability.

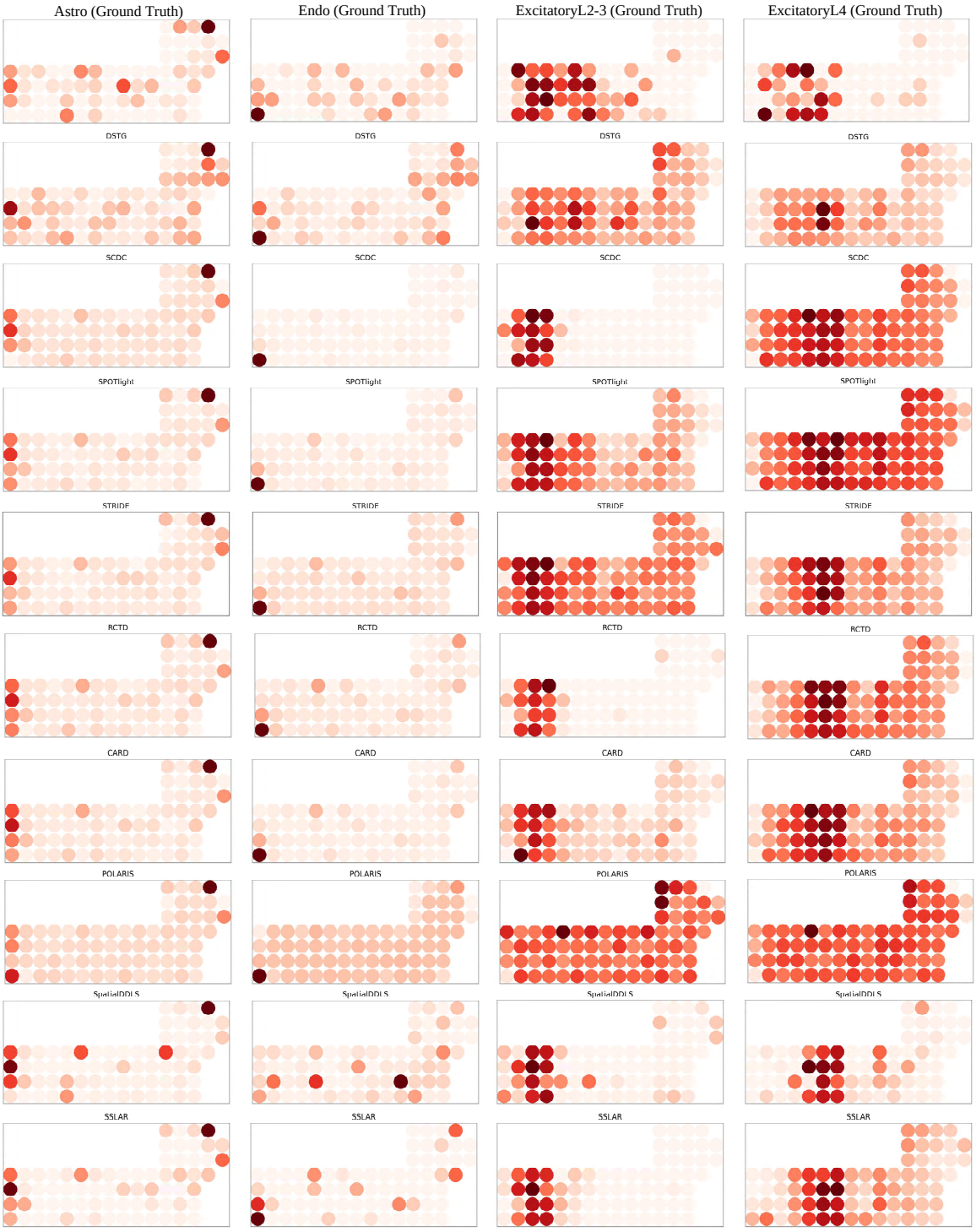


Figure S3: The proportion of four cell types (Astro, Endo, ExcitatoryL2and3, and ExcitatoryL4) in the spots simulated from mouse cortex, containing the ground truth and predictions.

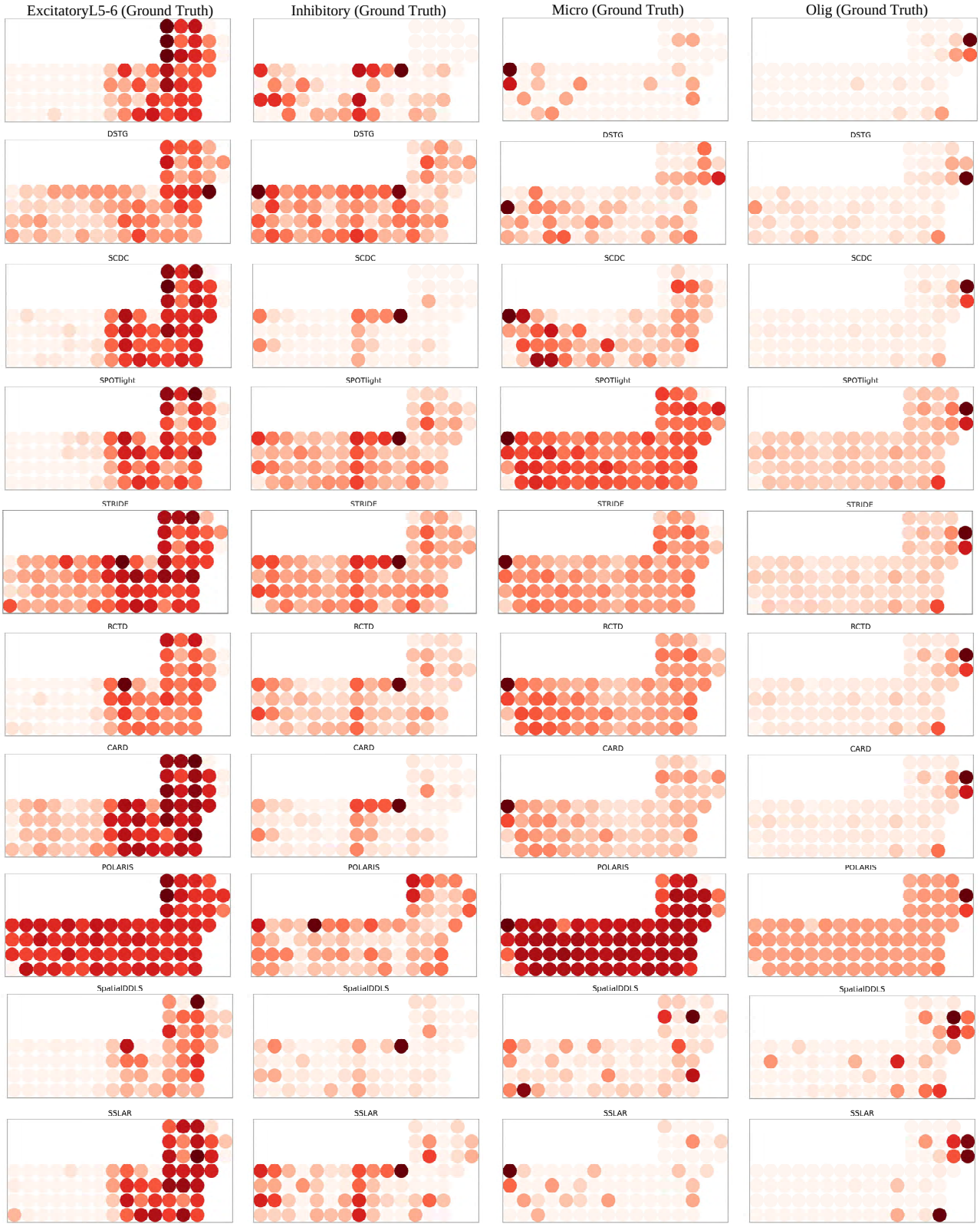


Figure S4: The proportion of four cell types (ExcitatoryL5and6, Inhibitory, Micro, and Olig) in the spots simulated from mouse cortex (SeqFISH+), containing the ground truth and predictions.

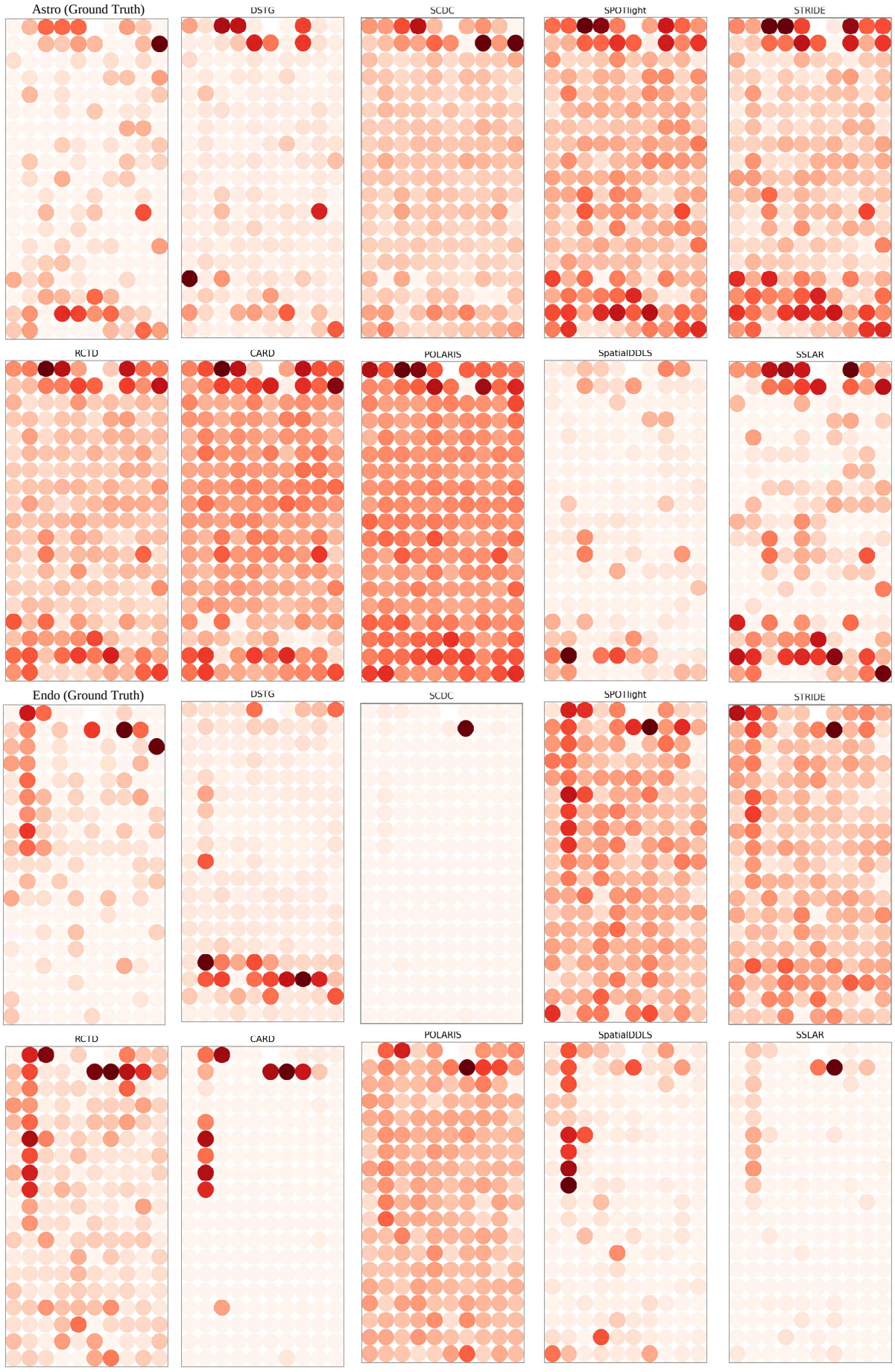


Figure S5: The proportion of two cell types (Astro and Endo) in the spots simulated from mouse visual cortex (STARmap), containing the ground truth and predictions.

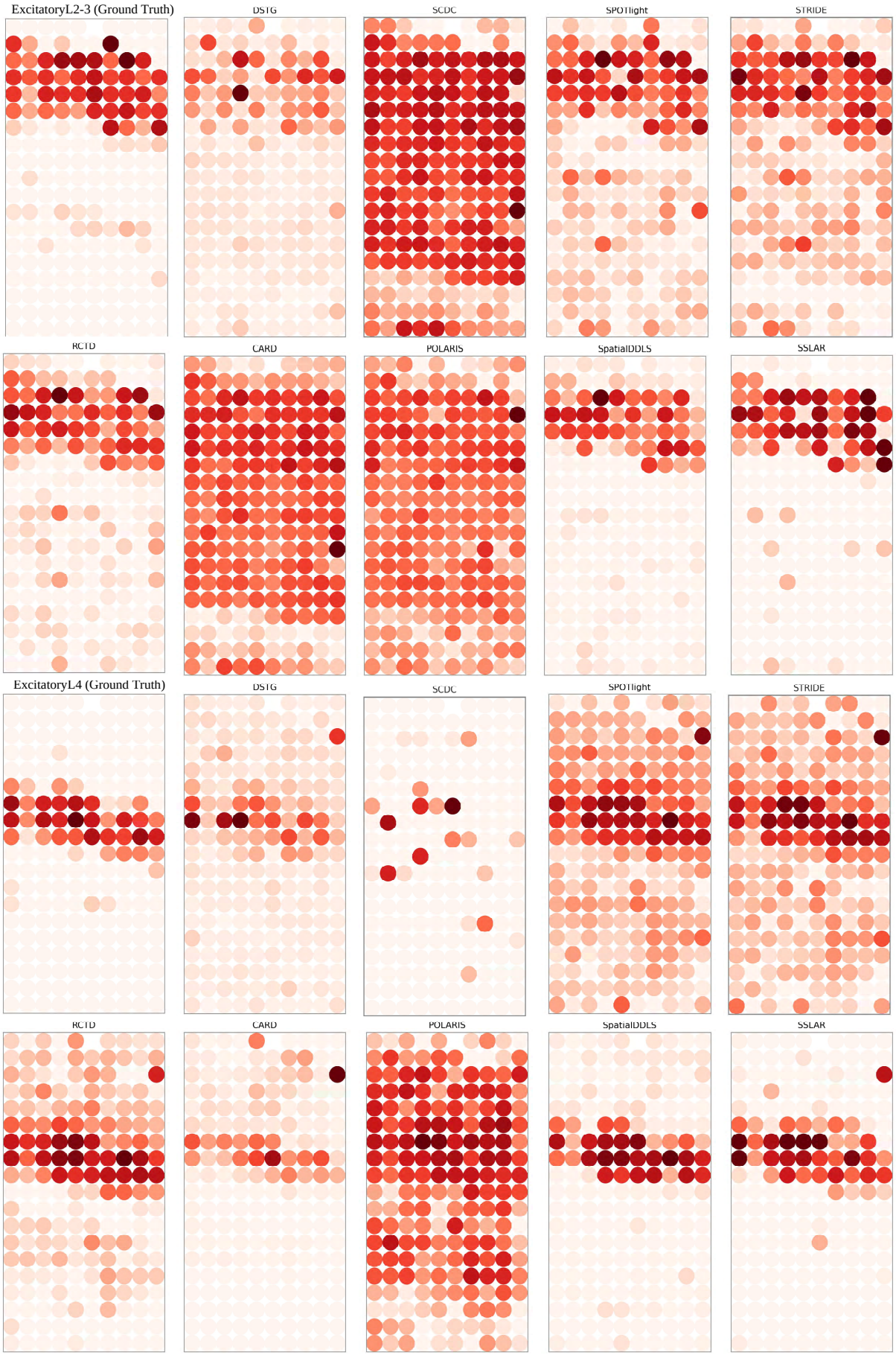


Figure S6: The proportion of two cell types (ExcitatoryL2-3 and ExcitatoryL4) in the spots simulated from mouse visual cortex (STARmap), containing the ground truth and predictions.

Figure S7: The proportion of two cell types (ExcitatoryL5 and ExcitatoryL6) in the spots simulated from mouse visual cortex (STARmap), containing the ground truth and predictions.

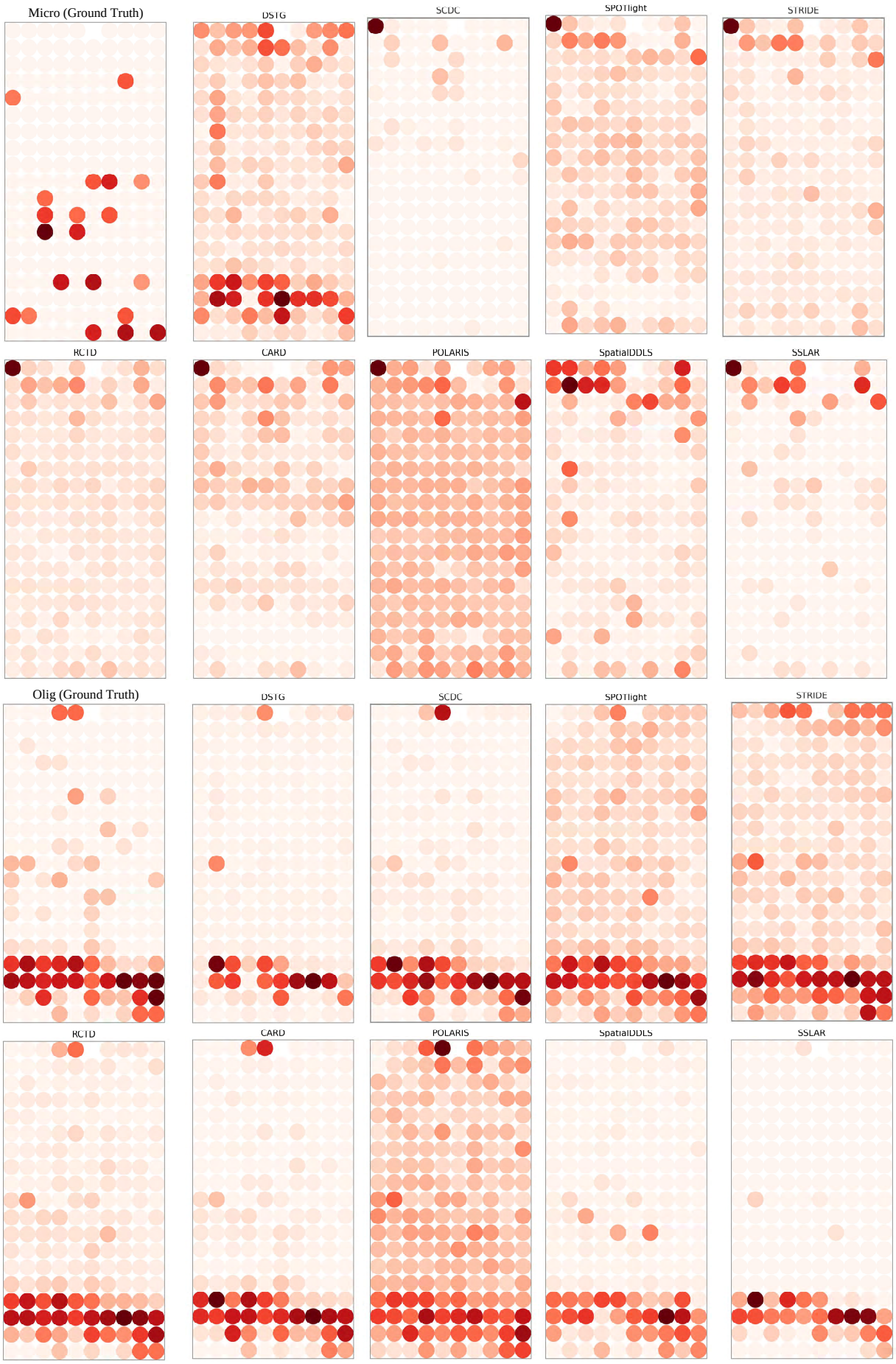


Figure S8: The proportion of two cell types (Micro and Olig) in the spots simulated from mouse visual cortex (STARmap), containing the ground truth and predictions.

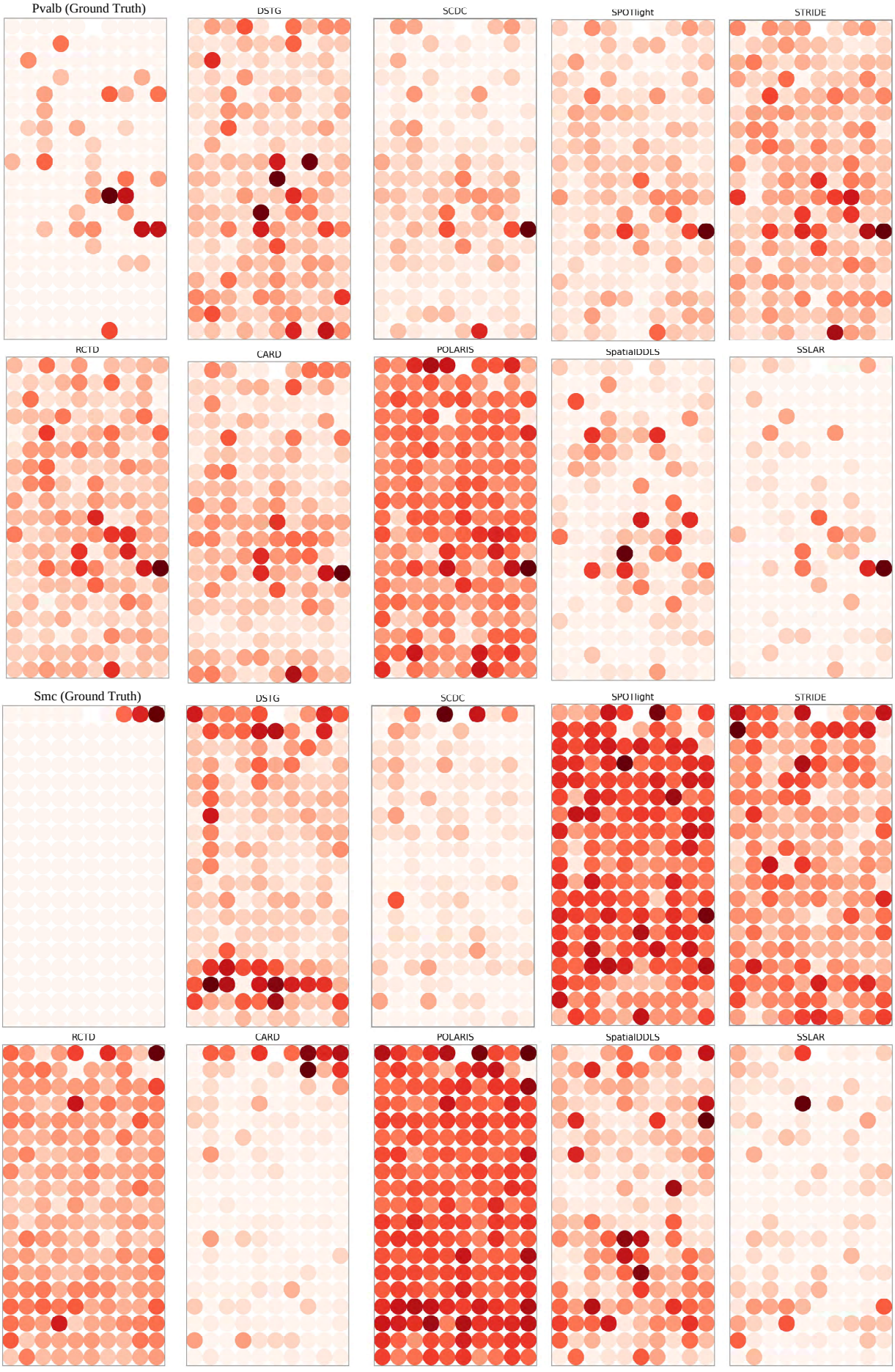


Figure S9: The proportion of two cell types (Pvalb and Smc) in the spots simulated from mouse visual cortex (STARmap), containing the ground truth and predictions.

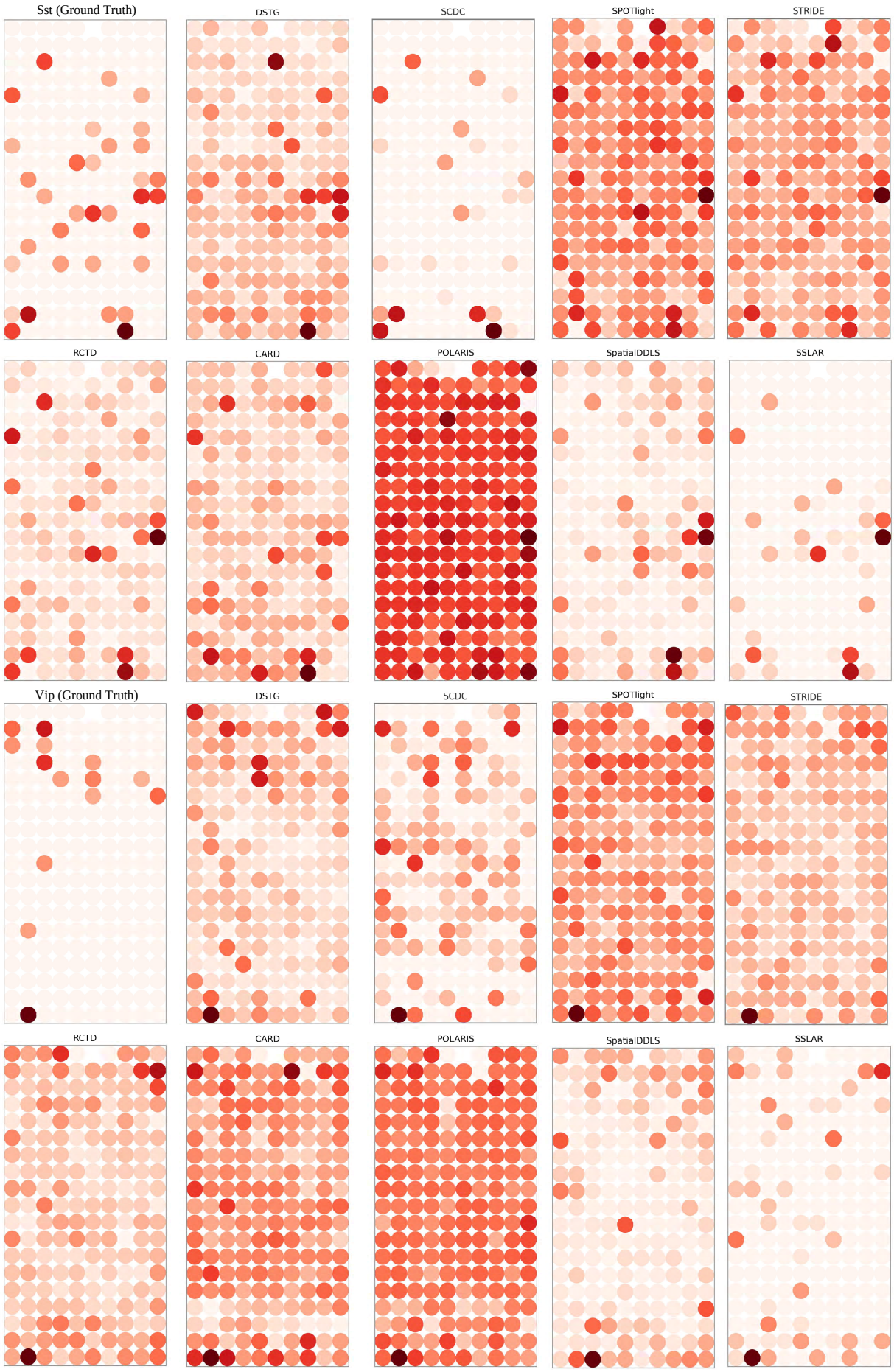


Figure S10: The proportion of two cell types (Sst and Vip) in the spots simulated from mouse visual cortex (STARmap), containing the ground truth and predictions.

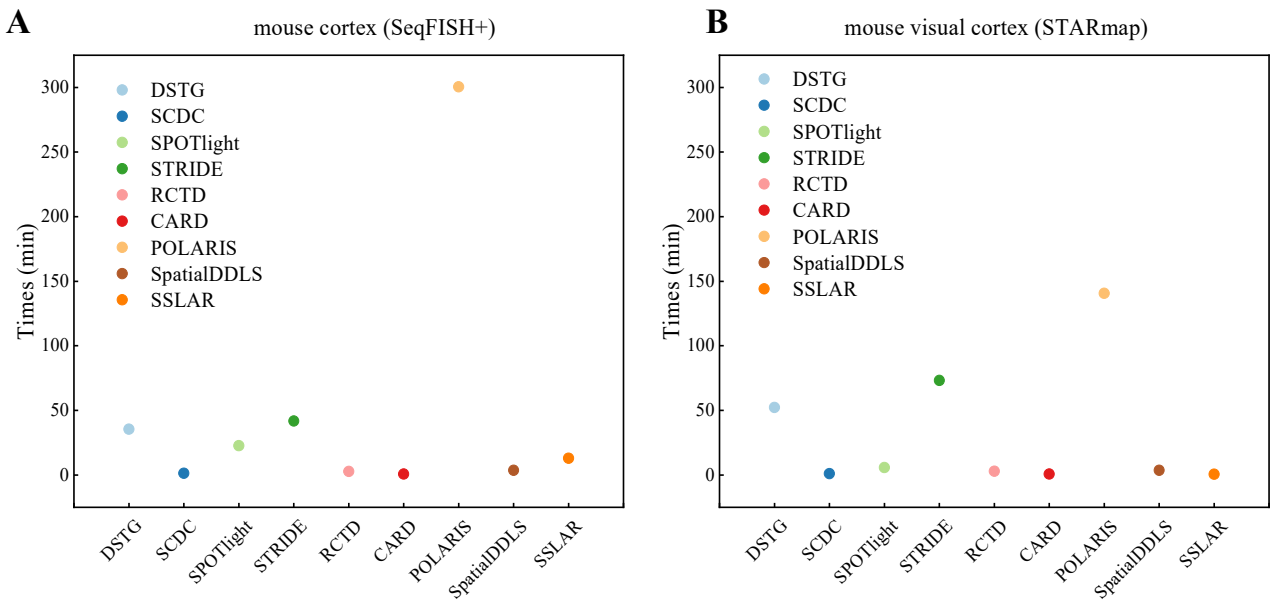


Figure S11: (A) Running time per method in mouse corte. (B) Running time per method in mouse visual corte.

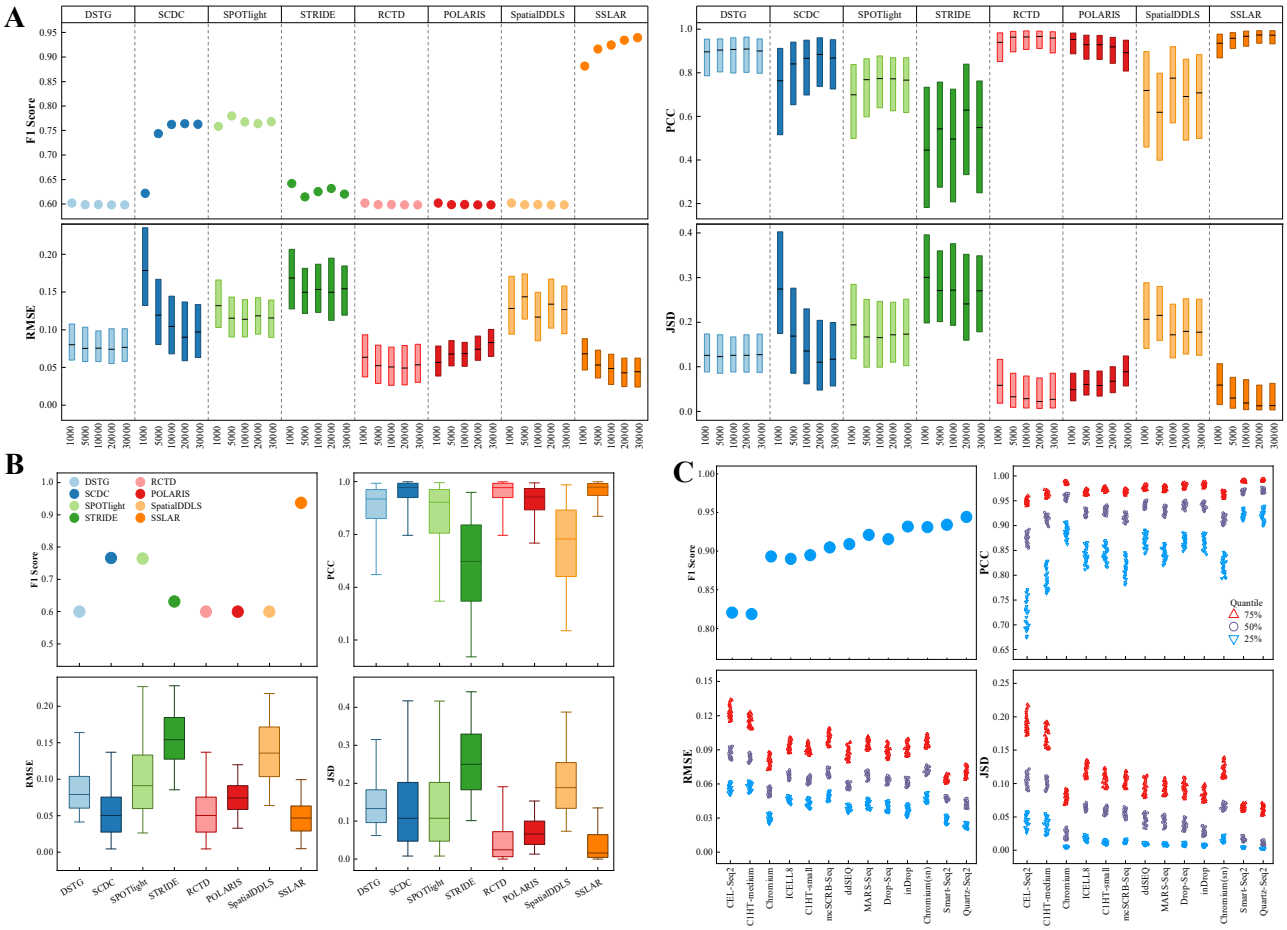


Figure S12: Benchmarking SSLAR. (A) The performance of SSLAR and other 7 methods on different sequencing depths. (B) The performance of SSLAR and other 7 methods on PBMC (Smart-seq2) dataset. (C) The SSLAR performance on 13 PBMC datasets.

Table S2: F1 score, PCC, RMSE, and JSD of DSTG, SCDC, SPOTlight, STRIDE, RCTD, POLARIS, SpatialDDLS, and SSLAR on synthetic PBMC (Smart-seq2) dataset with 5 different sequencing depths

	depth	DSTG	SCDC	SPOTlight	STRIDE	RCTD	POLARIS	SpatialDDLS	SSLAR
F1 score	1,000	0.602+0.0054	0.6219±0.0068	0.7585±0.0175	0.6419±0.0153	0.602±0.0054	0.602±0.0054	0.602±0.0054	**0.8815±0.0028**
	5,000	0.5987+0.0048	0.7436±0.0045	0.7797±0.0089	0.6145±0.0159	0.5987±0.0048	0.5987±0.0048	0.5987±0.0048	**0.9163±0.002**
	10,000	0.599+0.0026	0.7624±0.0042	0.7675±0.0046	0.6253±0.0133	0.599±0.0026	0.599±0.0026	0.599±0.0026	**0.9244±0.0033**
	20,000	0.5982+0.005	0.764±0.0049	0.7641±0.0026	0.6316±0.0133	0.5982±0.005	0.5982±0.005	0.5982±0.005	**0.9343±0.005**
	30,000	0.5983+0.0037	0.7628±0.0041	0.768±0.0026	0.6205±0.023	0.5983±0.0037	0.5983±0.0037	0.5983±0.0037	**0.9392±0.0049**
	Ave.	0.5992	0.7309	0.7676	0.6268	0.5992	0.5992	0.5992	**0.9191**
PCC	1,000	0.8899+0.0083	0.7639±0.0097	0.5986±0.1041	0.4749±0.0604	0.9395±0.0044	**0.9451±0.0043**	0.6191±0.0536	0.9352±0.0039
	5,000	0.8945+0.0323	0.8446±0.0099	0.7736±0.0151	0.2297±0.3313	0.964±0.0029	0.9329±0.0046	0.704±0.0442	**0.9602±0.0026**
	10,000	0.9034+0.0068	0.8767±0.007	0.7736±0.0064	0.4069±0.3033	0.9668±0.002	0.9306±0.0049	0.7086±0.0456	**0.9667±0.0024**
	20,000	0.8952+0.0346	0.8834±0.006	0.7719±0.005	0.5354±0.0616	0.9661±0.0012	0.9215±0.0055	0.7192±0.0356	**0.9714±0.0025**
	30,000	0.9036+0.0171	0.8845±0.0108	0.7736±0.0065	0.285±0.3449	0.9635±0.0025	0.9042±0.0055	0.7248±0.0281	**0.9743±0.0022**
	Ave.	0.8973	0.8506	0.7383	0.3864	0.96	0.9269	0.6951	**0.9616**
RMSE	1,000	0.0818+0.0025	0.1802±0.0041	0.1486±0.0198	0.1657±0.0087	0.0642±0.0014	**0.0597±0.0016**	0.1467±0.0097	0.0677±0.001
	5,000	0.0809+0.0134	0.1188±0.0022	0.1147±0.0027	0.1864±0.0363	**0.0508±0.0015**	0.0665±0.0016	0.1295±0.009	0.0525±0.0012
	10,000	0.0769+0.0015	0.0997±0.0023	0.116±0.0021	0.1654±0.0325	0.0491±0.0013	0.0675±0.0015	0.1273±0.0098	**0.0482±0.0014**
	20,000	0.0801+0.0147	0.092±0.0025	0.116±0.0017	0.1557±0.0059	0.0506±0.0011	0.0721±0.003	0.1278±0.0063	**0.045±0.0017**
	30,000	0.0774+0.0061	0.0916±0.0038	0.1157±0.0011	0.1819±0.0316	0.0521±0.0014	0.0805±0.0016	0.125±0.0068	**0.0424±0.0017**
	Ave.	0.0794	0.1165	0.1222	0.171	0.0534	0.0693	0.1313	**0.0512**
JSD	1,000	0.1326+0.0059	0.2823±0.0126	0.231±0.0421	0.2941±0.0207	0.0587±0.0027	**0.051±0.0022**	0.2354±0.0173	0.0619±0.0024
	5,000	0.134+0.0234	0.1667±0.0046	0.162±0.0094	0.3832±0.128	0.0321±0.002	0.0569±0.0028	0.1879±0.0171	**0.0285±0.0024**
	10,000	0.1282+0.0034	0.1282±0.0046	0.17±0.0043	0.3071±0.1147	0.0268±0.0015	0.0569±0.0031	0.1857±0.0139	**0.0193±0.002**
	20,000	0.1359+0.0286	0.1143±0.0054	0.172±0.0043	0.2671±0.0196	0.0261±0.0021	0.0634±0.005	0.1736±0.0125	**0.0145±0.0024**
	30,000	0.129+0.0151	0.1087±0.0056	0.1722±0.0035	0.3571±0.1247	0.0259±0.0017	0.0792±0.0054	0.1673±0.0124	**0.0118±0.0017**
	Ave.	0.1319	0.16	0.1814	0.3217	0.0339	0.0615	0.19	**0.0272**

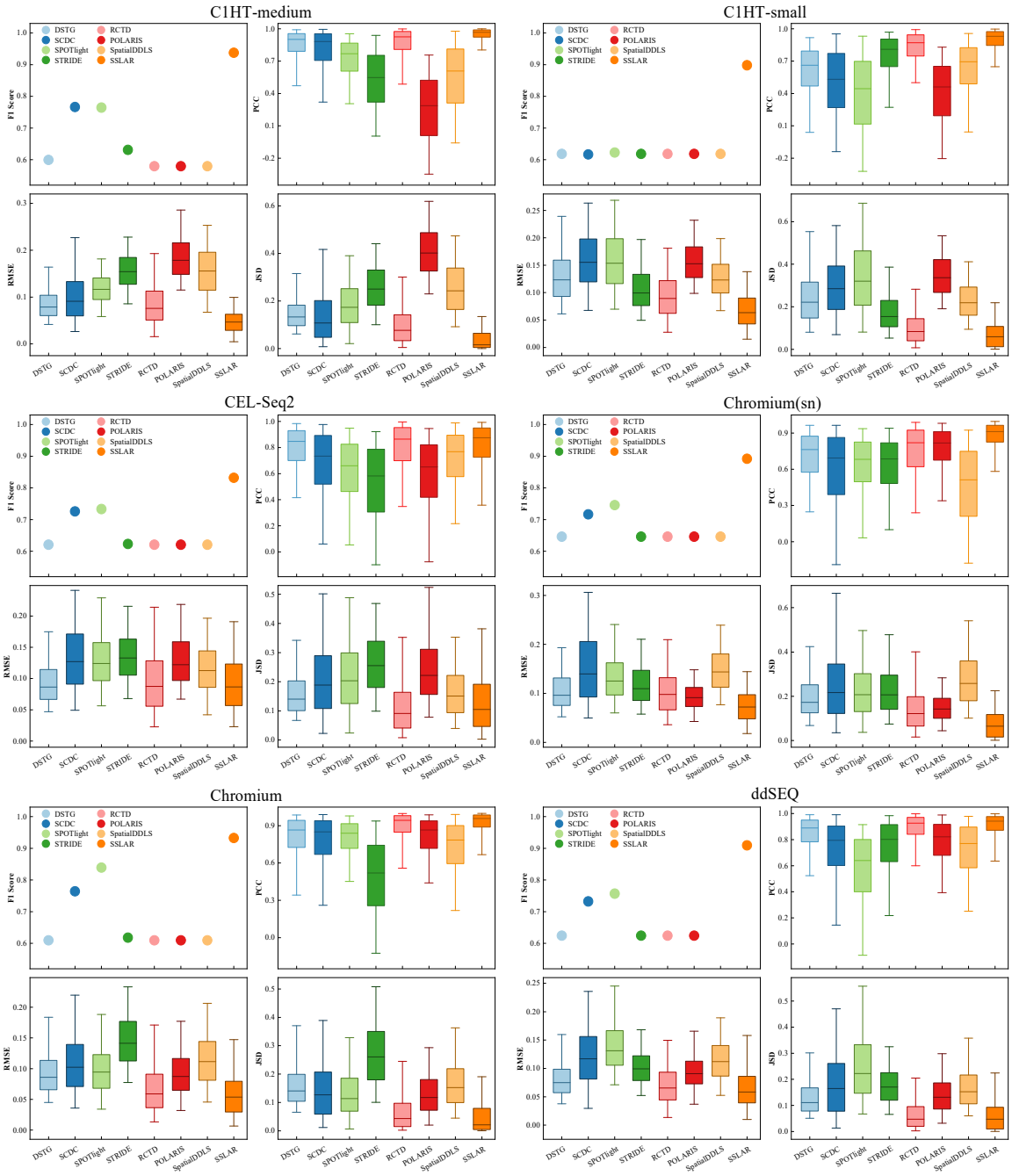


Figure S13: The SSLAR performance on six PBMC datasets from different sequencing platforms (Cel-Seq2, Chromium, Chromium(sn), C1HT-medium, C1HT-Small, ddSeq) on PBMC dataset.

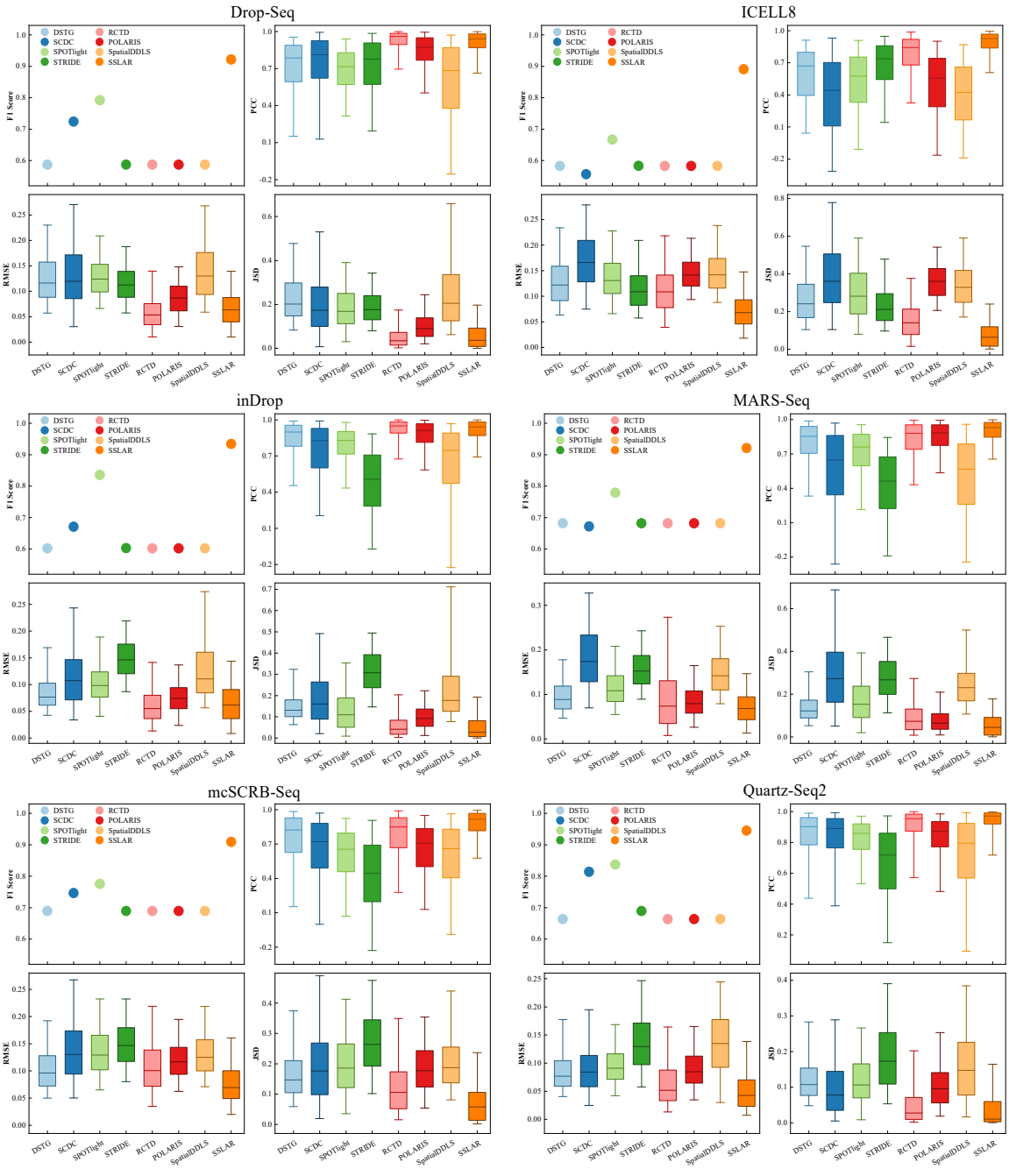


Figure S14: The SSLAR performance on other six PBMC datasets from different sequence platforms (Drop-Seq, ICELL8, inDrop, MARS-Seq, mcSCRB-Seq, Qqartz-Seq2) on PBMC dataset.

Table S3: F1 score and PCC of DSTG, SCDC, SPOTlight, STRIDE, RCTD, POLARIS, SpatialDDLS, and SSLAR on 13 synthetic datasets. PBMC 1-13 denote Cel-Seq2, Chromium, Chromium(sn), C1HT-medium, C1HT-Small, ddSeq, Drop-Seq, ICELL8, inDrop, MARS-Seq, mcSCRB-Seq, and Qqartz-Seq2.

	Dataset	DSTG	SCDC	SPOTlight	STRIDE	RCTD	POLARIS	SpatialDDLS	SSLAR
F1	PBMC1	0.5797+0.0034	0.6088±0.0048	0.6227±0.0038	0.5798±0.0033	0.5797±0.0034	0.5797±0.0034	0.5797±0.0034	**0.8128±0.0072**
	PBMC2	0.6184+0.0042	0.6167±0.005	0.6229±0.0071	0.6184±0.0042	0.6184±0.0042	0.6184±0.0042	0.6184±0.0042	**0.8981±0.0042**
	PBMC3	0.6208+0.0028	0.7259±0.0034	0.7333±0.0087	0.6231±0.0048	0.6208±0.0028	0.6208±0.0028	0.6208±0.0028	**0.8319±0.0053**
	PBMC4	0.6098+0.0043	0.7642±0.004	0.8391±0.0045	0.618±0.0078	0.6098±0.0043	0.6098±0.0043	0.6098±0.0043	**0.9325±0.0044**
	PBMC5	0.6462+0.004	0.7165±0.0035	0.7458±0.0038	0.6462±0.004	0.6462±0.004	0.6462±0.004	0.6462±0.004	**0.8917±0.0033**
	PBMC6	0.6244+0.0034	0.7324±0.0055	0.7569±0.011	0.6244±0.0034	0.6244±0.0034	0.6244±0.0034	0.6244±0.0034	**0.9095±0.0035**
	PBMC7	0.5872+0.0051	0.7241±0.0044	0.7918±0.0056	0.5873±0.0051	0.5872±0.0051	0.5872±0.0051	0.5872±0.0051	**0.9217±0.0043**
	PBMC8	0.5832+0.0058	0.5569±0.0042	0.667±0.0066	0.5835±0.0058	0.5832±0.0058	0.5832±0.0058	0.5832±0.0058	**0.8908±0.0046**
	PBMC9	0.6021+0.005	0.6711±0.0046	0.8355±0.0041	0.603±0.0043	0.6021±0.005	0.6021±0.005	0.6021±0.005	**0.9342±0.0021**
	PBMC10	0.6819+0.0029	0.6719±0.0041	0.7792±0.0045	0.6819±0.0029	0.6819±0.0029	0.6819±0.0029	0.6819±0.0029	**0.921±0.001**
	PBMC11	0.6893+0.0042	0.7463±0.0041	0.7757±0.0044	0.6893±0.0042	0.6893±0.0042	0.6893±0.0042	0.6893±0.0042	**0.9093±0.0034**
	PBMC12	0.6637+0.0049	0.8142±0.0039	0.837±0.0052	0.6895±0.0051	0.6637±0.0049	0.6637±0.0049	0.6637±0.0049	**0.945±0.0022**
	PBMC13	0.5994+0.0046	0.7662±0.0058	0.7644±0.0062	0.631±0.0108	0.5994±0.0046	0.5994±0.0046	0.5994±0.0046	**0.9373±0.0024**
	Ave.	0.6235	0.7012	0.7516	0.6289	0.6235	0.6235	0.6235	**0.9028**
PCC	PBMC1	0.9082+0.0239	0.5135±0.0216	0.4779±0.0105	0.9072±0.0105	**0.9236±0.0061**	0.2817±0.0077	0.5813±0.025	0.9102±0.0067
	PBMC2	0.8276+0.0789	0.5596±0.0131	0.4364±0.0172	0.786±0.0159	0.8761±0.0043	0.4552±0.0124	0.7027±0.0262	**0.9337±0.0033**
	PBMC3	0.8281+0.0177	0.7519±0.0109	0.643±0.0126	0.6266±0.0305	0.8721±0.0063	0.6625±0.0148	0.7582±0.0181	**0.8844±0.0083**
	PBMC4	0.8715+0.0093	0.8522±0.0068	0.8476±0.0046	0.5373±0.0493	0.9437±0.0026	0.863±0.0052	0.7808±0.0086	**0.9619±0.0019**
	PBMC5	0.7213+0.1059	0.6635±0.0132	0.6896±0.0101	0.625±0.0675	0.8036±0.0097	0.8106±0.0084	0.5173±0.0181	**0.9113±0.0043**
	PBMC6	0.8887+0.0082	0.7962±0.0067	0.7095±0.0255	0.8657±0.0242	0.9248±0.0033	0.8198±0.008	0.7701±0.0136	**0.9396±0.0024**
	PBMC7	0.8891+0.0382	0.814±0.006	0.7262±0.0074	0.807±0.0247	**0.9609±0.0017**	0.8799±0.0073	0.7013±0.0214	0.9478±0.0033
	PBMC8	0.7744+0.08	0.427±0.0124	0.5606±0.0111	0.7062±0.0355	0.8234±0.0099	0.5576±0.0101	0.3992±0.0583	**0.9172±0.0057**
	PBMC9	0.8757+0.0523	0.8166±0.0128	0.8256±0.0064	0.4599±0.0714	**0.9497±0.0036**	0.9074±0.0078	0.7659±0.0265	0.9408±0.0022
	PBMC10	0.8238+0.067	0.6387±0.009	0.7559±0.0101	0.693±0.0737	0.8804±0.0031	0.8816±0.0036	0.548±0.0448	**0.9302±0.0032**
	PBMC11	0.7185+0.1438	0.7239±0.0146	0.6607±0.0075	0.4919±0.038	0.8494±0.0046	0.7152±0.0144	0.6783±0.0313	**0.9231±0.0037**
	PBMC12	0.8799+0.0801	0.8855±0.0055	0.8609±0.005	0.743±0.0324	0.9539±0.0029	0.8739±0.0064	0.7777±0.0114	**0.9695±0.002**
	PBMC13	0.8819+0.0523	0.8873±0.0054	0.7734±0.0043	0.5592±0.0314	0.966±0.0017	0.9161±0.0037	0.688±0.0335	**0.9724±0.002**
	Ave.	0.8376	0.7177	0.6898	0.6775	0.9021	0.7403	0.6668	**0.9340**

Table S4: RMSE and JSD of DSTG, SCDC, SPOTlight, STRIDE, RCTD, POLARIS, SpatialDDLS, and SSLAR on 13 synthetic datasets. PBMC 1-13 denote Cel-Seq2, Chromium, Chromium(sn), C1HT-medium, C1HT-Small, ddSeq, Drop-Seq, ICELL8, inDrop, MARS-Seq, mcSCRB-Seq, and Qqartz-Seq2.

	Dataset	DSTG	SCDC	SPOTlight	STRIDE	RCTD	POLARIS	SpatialDDLS	SSLAR
RMSE	PBMC1	0.0874+0.0105	0.1715±0.0017	0.1678±0.0016	0.0856±0.0034	**0.0789±0.0022**	0.1807±0.0017	0.1555±0.0046	0.0842±0.0024
	PBMC2	0.0919+0.0162	0.1526±0.0018	0.1549±0.0021	0.103±0.0027	0.0884±0.0014	0.1534±0.002	0.1201±0.0042	**0.0634±0.0014**
	PBMC3	0.092+0.0043	0.1245±0.0017	0.1376±0.0053	0.1274±0.0045	0.0852±0.0017	0.1215±0.0021	0.1141±0.0034	**0.0822±0.0024**
	PBMC4	0.0839+0.0024	0.101±0.0017	0.0914±0.0016	0.1406±0.0041	0.0588±0.0012	0.0878±0.001	0.1107±0.0025	**0.05±0.0017**
	PBMC5	0.1057+0.0163	0.1416±0.0028	0.1248±0.0016	0.1233±0.0101	0.0996±0.0014	0.0931±0.0013	0.1454±0.0024	**0.072±0.001**
	PBMC6	0.0753+0.0021	0.112±0.0024	0.1212±0.0035	0.0816±0.0071	0.0654±0.0015	0.0916±9e-04	0.1078±0.0034	**0.0603±9e-04**
	PBMC7	0.0843+0.0122	0.1236±0.0023	0.1238±0.0016	0.1112±0.0043	**0.0523±0.0012**	0.0858±0.0026	0.1284±0.0037	0.0602±0.0018
	PBMC8	0.0986+0.0139	0.1666±0.0019	0.1312±0.0017	0.1114±0.0056	0.1088±0.0015	0.1404±0.0017	0.1438±0.0043	**0.0684±0.0017**
	PBMC9	0.0849+0.0116	0.107±0.0029	0.0994±0.0016	0.1506±0.0056	**0.0563±0.0012**	0.0746±0.0021	0.1122±0.0052	0.0628±0.0013
	PBMC10	0.0949+0.0132	0.1766±0.0021	0.1114±0.0019	0.1221±0.0095	0.0858±0.0017	0.0802±0.0015	0.1438±0.0044	**0.0666±9e-04**
	PBMC11	0.1105+0.0221	0.1296±0.002	0.1288±0.0018	0.144±0.0048	0.1002±0.0014	0.1149±0.0022	0.1239±0.0058	**0.0683±0.0011**
	PBMC12	0.0812+0.0196	0.0852±9e-04	0.0904±0.002	0.1253±0.0049	0.0527±0.0015	0.0847±0.0025	0.1366±0.004	**0.0444±0.0012**
	PBMC13	0.0852+0.019	0.091±0.0021	0.1159±0.0018	0.1496±0.0044	0.0505±0.0013	0.0739±0.0012	0.1326±0.0067	**0.0442±0.0013**
	Ave.	0.0904	0.1294	0.1230	0.1212	0.0756	0.1064	0.1288	**0.0636**
JSD	PBMC1	0.1501+0.0251	0.3051±0.0069	0.329±0.0045	0.1266±0.0058	**0.078±0.0027**	0.4082±0.004	0.2468±0.0088	0.0996±0.0057
	PBMC2	0.1526+0.0373	0.2769±0.0047	0.3267±0.0085	0.1699±0.008	0.0822±0.0024	0.3382±0.0052	0.214±0.0102	**0.0567±0.0025**
	PBMC3	0.1517+0.0091	0.1795±0.007	0.2199±0.0099	0.2419±0.0156	**0.0862±0.0046**	0.2149±0.0057	0.157±0.0109	0.0944±0.005
	PBMC4	0.1345+0.0064	0.1226±0.004	0.1079±0.004	0.2603±0.0132	0.0423±0.0016	0.1135±0.0031	0.1466±0.0046	**0.0193±0.0025**
	PBMC5	0.1959+0.0419	0.2239±0.0046	0.2038±0.0043	0.2506±0.0335	0.1291±0.006	0.1445±0.0032	0.266±0.0083	**0.0669±0.0024**
	PBMC6	0.1151+0.006	0.1518±0.0061	0.1955±0.0099	0.1129±0.0206	**0.0478±0.0025**	0.1334±0.0042	0.1472±0.0077	0.0487±0.0032
	PBMC7	0.1353+0.0261	0.1769±0.0031	0.1733±0.0044	0.185±0.0139	0.0329±0.0013	0.0951±0.0054	0.2019±0.0074	**0.0309±0.0044**
	PBMC8	0.1847+0.0368	0.3664±0.0059	0.2834±0.0047	0.2163±0.0116	0.1402±0.0039	0.3547±0.0048	0.3342±0.0137	**0.0656±0.0032**
	PBMC9	0.1444+0.0206	0.1624±0.0068	0.1117±0.004	0.3206±0.0224	0.0424±0.0023	0.0945±0.0044	0.1826±0.0103	**0.0282±0.0026**
	PBMC10	0.1393+0.0286	0.2759±0.0082	0.1566±0.0043	0.1933±0.0218	0.0726±0.0026	0.0653±0.003	0.2341±0.0136	**0.0413±0.0028**
	PBMC11	0.1749+0.0453	0.1758±0.0046	0.1855±0.0032	0.2583±0.0115	0.1024±0.0025	0.174±0.0048	0.183±0.0089	**0.0528±0.0025**
	PBMC12	0.112+0.0373	0.0794±0.0034	0.1038±0.0048	0.1683±0.0144	0.0276±0.0015	0.0932±0.0034	0.1512±0.0045	**0.0113±0.0011**
	PBMC13	0.1422+0.0351	0.1112±0.0052	0.1731±0.0054	0.2557±0.0114	0.0258±0.0015	0.0664±0.0022	0.1823±0.0131	**0.0136±0.0012**
	Ave.	0.1487	0.2006	0.1977	0.2123	0.0700	0.1766	0.2036	**0.0484**

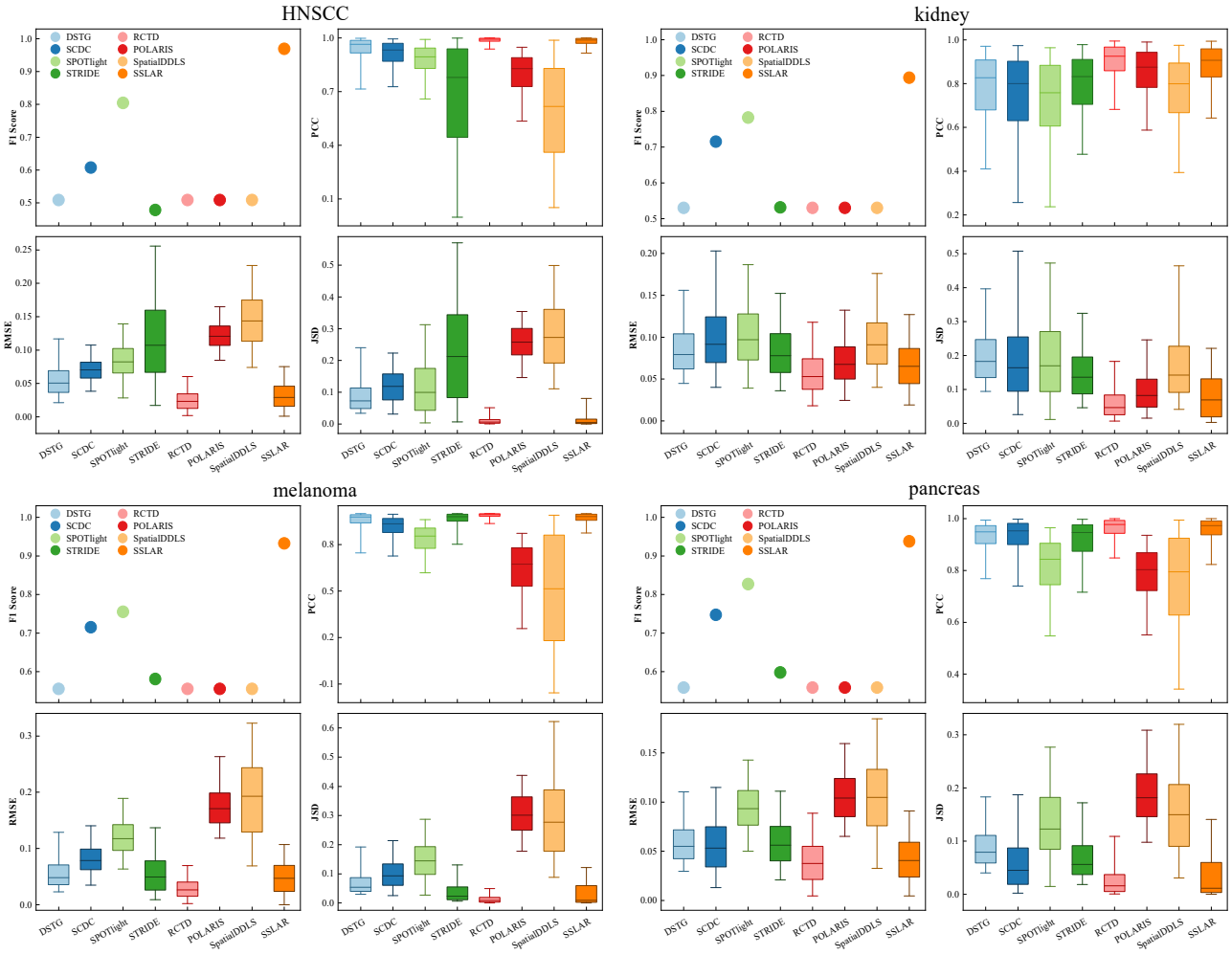


Figure S15: The robustness analysis on different tissues. (A) HNSCC (Smart-seq2). (B) human kidney tumors (10x Genomics). (C) human melanoma (Smart-seq2). (D) human pancreas (Cel-Seq2).

Table S5: The F1 score, PCC, RMSE, and JSD of nsNMF, NNLS, and SSLAR on 17 synthetic datasets and 2 grided datasets.

	Dataset	nsNMF	NNLS	unrefined	SSLAR
F1	17 synthetic	0.7698	0.9063	\	**0.9100**
	2 grided	0.5964	0.5277	0.5747	**0.6157**
PCC	17 synthetic	0.7302	0.9325	\	**0.9408**
	2 grided	0.8126	0.50405	0.828	**0.8294**
JSD	17 synthetic	0.1806	0.0476	\	**0.0427**
	2 grided	0.23635	0.3608	0.2249	**0.2172**
RMSE	17 synthetic	0.1178	0.0631	\	**0.0593**
	2 grided	0.1184	0.16265	0.1221	**0.113**

2.3.3 Ablation study

To evaluate the contributions of ssNMF, LARS, and label refinement to our model, we conduct ablation studies across 19 datasets. As the ablated variants, SPOTlight utilized non-smooth NMF (nsNMF) [28] to for matrix factorization and NNLS for linear regression during deconvolution. To investigate the affects of different NMF method and linear regression model on the cell type annotation performance, we replaced ssNMF using nsNMF during topic identification and replaced LARS using NNLS during topic distribution analysis, respectively. Finally, we compared SSLAR with nsNMF and NNLS in addition to the above five classical models on 17 synthetic datasets and 2 grided datasets.

Table S5 delineates the results of nsNMF, NNLS, and SSLAR. Notably, SSLAR, which aggregates SSNMF and NNLS, cooperated well and greatly improved the cell type annotation performance. Moreover, SSLAR effective learned supervised information by fully leveraging maker genes. LARS, as a novel step-wise regression model, facilitates to prioritize the selection of cell type topics with the most relevance to the current spot, rather than minimizing residuals as in classical linear regression. Additionally, refining labels based on spatial information can also improve the prediction performance to some extent.

2.4 SSLAR on mouse brain ST data

The combination of unpaired scRNA-seq and ST data greatly advances single-cell atlas automated interpretation. Mouse brain demonstrates well-defined structures with position-specific types although it has high complexity. Here, we used a mouse brain reference scRNA-seq dataset provided by the Allen Institute to implement its cell type annotation. The dataset contains 73,363 cells sampled from the hippocampus and multiple cortical areas as well as 42 annotated cell types by sequencing with SMART-Seq (GSE185862) [7, 6]. Figure S16 depicts T-SNE projection of the 73,363 cells. Cells were labeled and colored based on known 42 annotations.

ST data on adult mouse brain from anterior and posterior sagittal slices were downloaded from 10x Genomics. To investigate the predicted spatial distribution of cell types within mouse brain tissues, we adopted marker genes with known cell types at spot-level resolution based on *in situ hybridization* images from the Allen Mouse Brain Atlas [29]. Similar to Ref. [30], we used the following Dentate gyrus (DG) cells and corresponding marker genes *Prox1* in the hippocampus. Its *in situ hybridization* images was downloaded at `mouse.brain-map.org/experiment/siv?id=69289763&imageId=69177644&initImage=ish&coordSystem=pixel&x=5416.5&y=3720.5&z=1`.

As shown in Figure S17, SSLAR and 8 comparison methods used canonical marker genes to characterize the hippocampus architecture DG (*Prox1*) [30]. Gene *prox1* has been validated to be able to clearly annotate DG by *in situ hybridization* images (S17A). In particular, Figure S17B shows that the expression of *Prox1* is noisy on the spatial level. Compared to other 8 methods, SSLAR still accurately identified DG in the presence of noise canonical marker gene on the ST data, elucidating the SSLAR powerful cell type annotation ability. Figure S18 illustrates cell type distribution identified by 9 deconvolution methods within each spot on mouse brain, respectively.

Figure S19A delineates the predicted cell types on adult mouse brain ST data. Scatter pie plots depict the cell type proportions within each spot. SSLAR identified multiple substructures of pre-defined specific cell types in the mouse brain. The results demonstrated that SSLAR accurately annotated spatial cell types of mouse

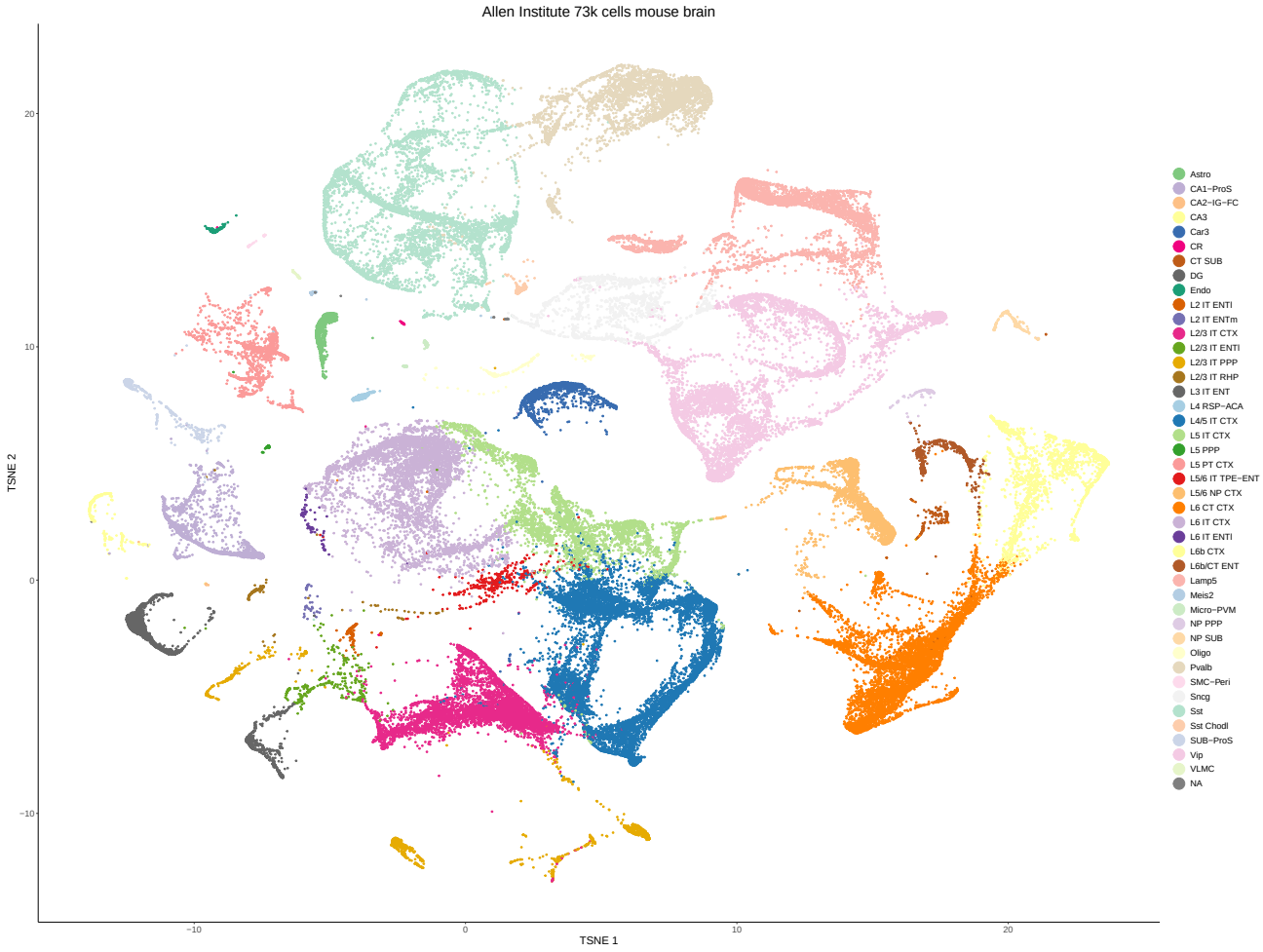


Figure S16: T-SNE projections of 73,363 cells from mouse brain scRNA-seq reference atlas provided by the Allen Institute. Cells are labeled and colored based on known 42 cell type annotations.

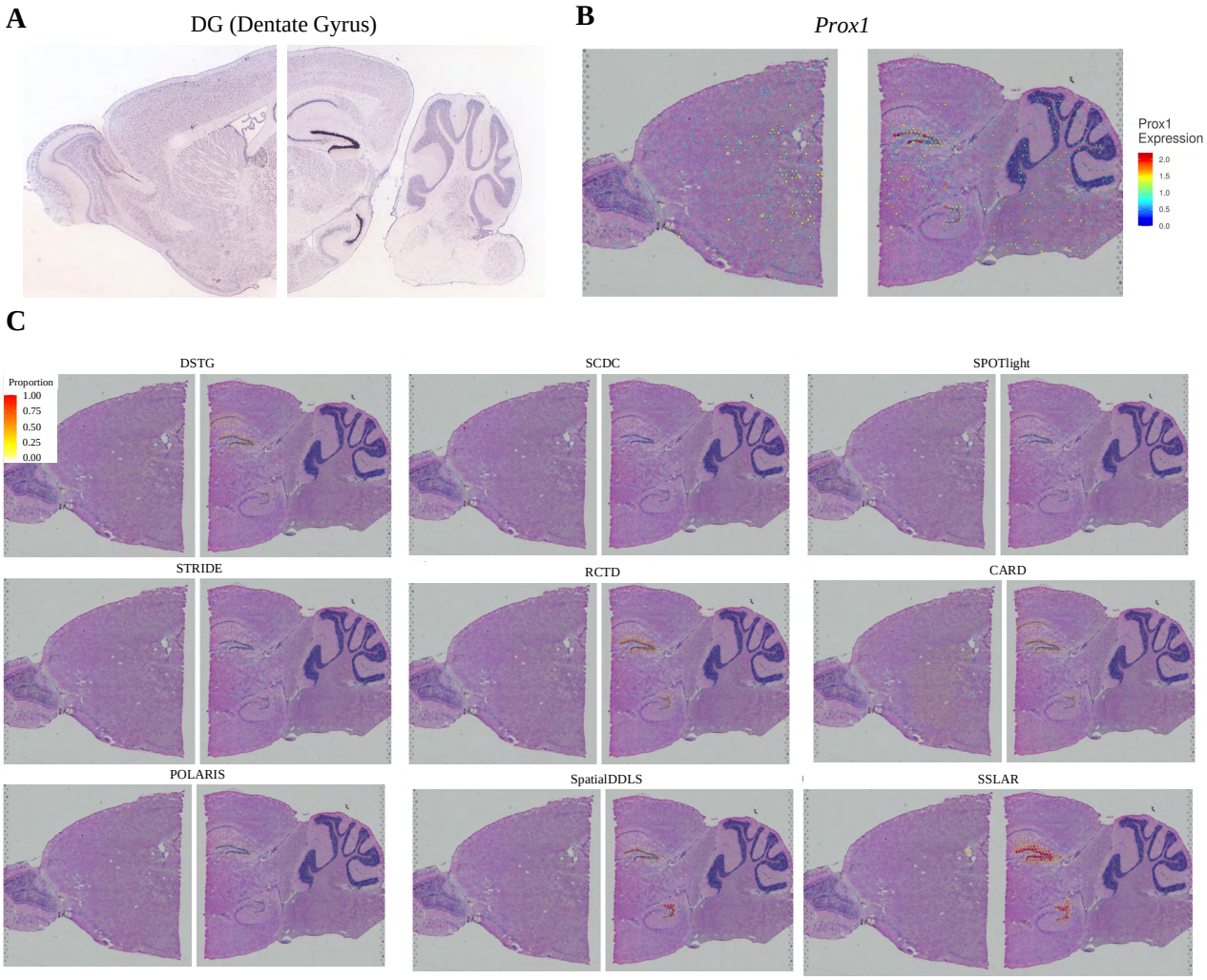


Figure S17: Visualization of hippocampal structure identified by SSLAR. (A) Reference in situ Hybridization labelling of the canonical marker genes from the Allen Institute to validate the cell types are predicted on their correspondent structure. (B) Canonical marker gene *Prox1* expression in each capture location (C) The proportion of each of the DG inferred by 9 methods on each spatial location.

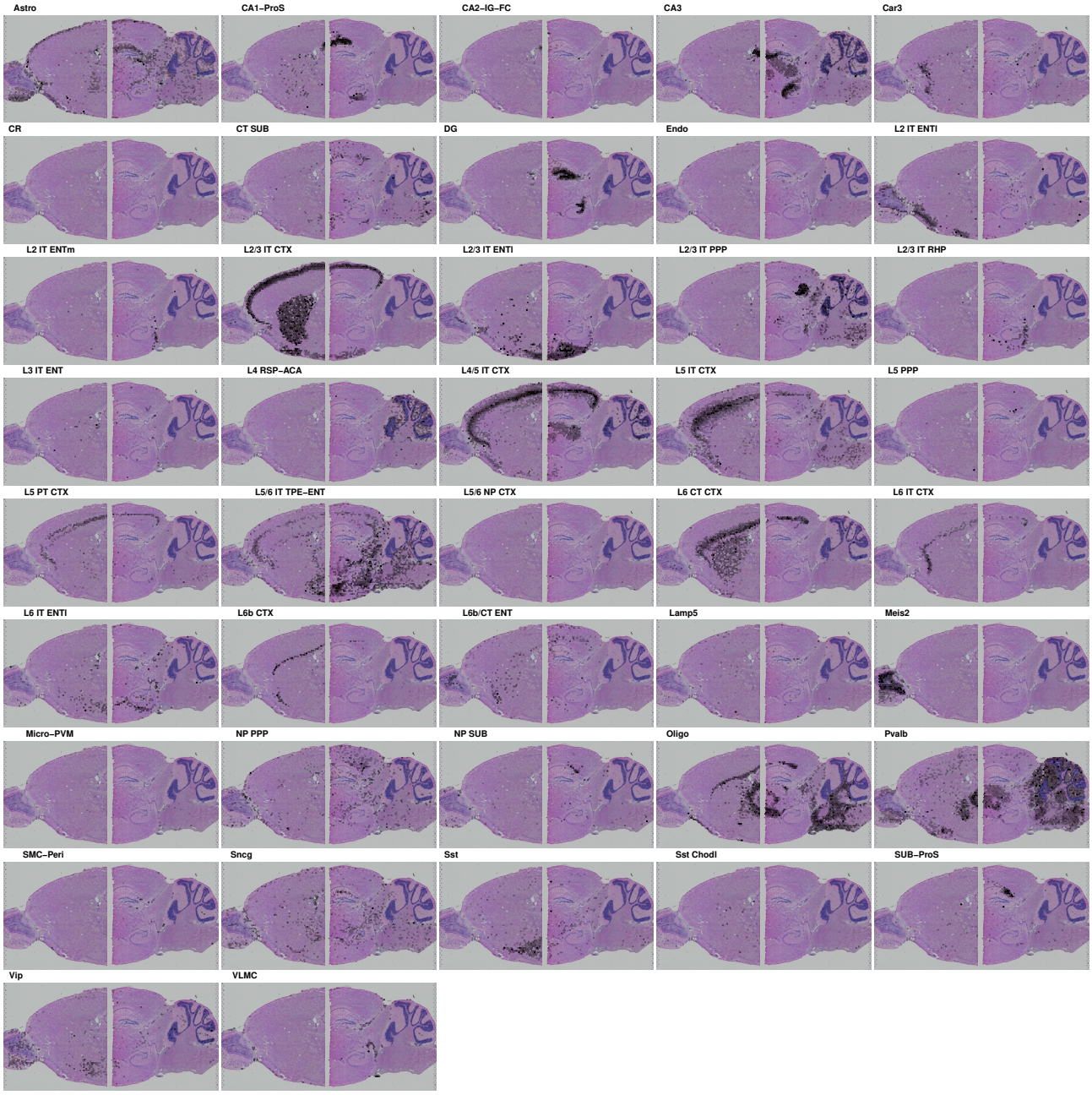


Figure S18: The SSLAR predicted proportions of cell types within each spot on mouse brain.

brain.

Figure S19B-I illustrates the spatial organization with different cortical layers (L2 to L6) including layer-specific neuronal subtypes. The proportion of each specific cortical neuron types (L2-L6) in each spot was characterized. Consistent with the real layered structure in mouse cortex, the predicted subtypes from L2 to L6 were successively aligned. Particularly, Figure S18 showed that L6 captured multiple neuronal subpopulations that were precisely classified to the layer substructures. The ability to distinguish between cortical neuronal subpopulations demonstrated that SSLAR had better sensitivity in complex tissues with similar cell types and states.

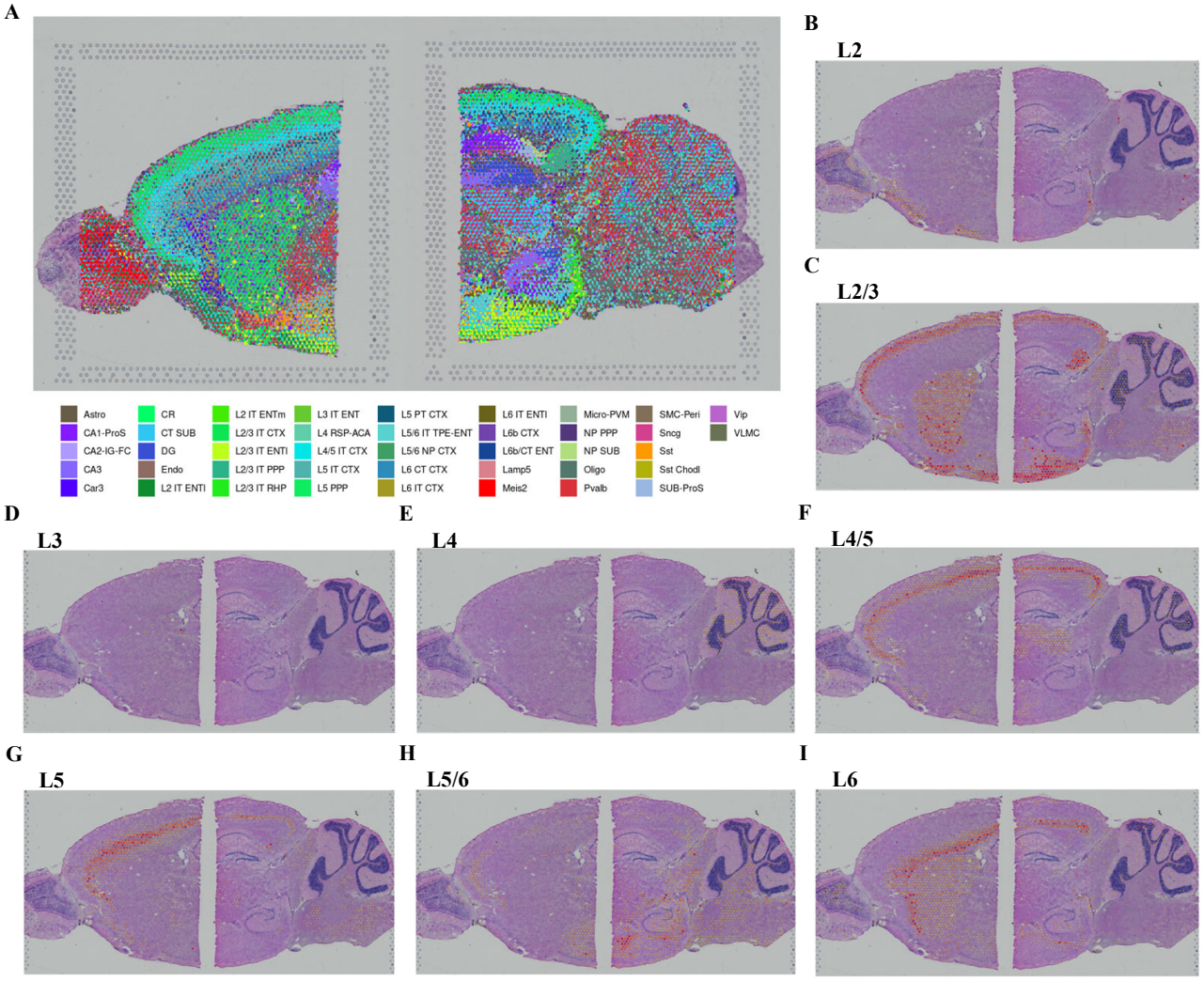


Figure S19: Cell type annotation on adult mouse brain ST data. (A) Substructures of anatomical regions in mouse brain. (B-I) Proportion of each specific cortical neuron type (L2-L6) within each spot.

In summary, the identified cell types manifested better enrichment in distinct layers (e.g. cortical areas) or specific regions (e.g. hippocampus) on the mouse brain structure. Furthermore, specific cell types demonstrated the regional enrichment on mouse brain, suggesting high accuracy and sensitivity of the SSLAR predictions.

2.5 SSLAR on pancreatic ductal adenocarcinoma ST data

The automated interpretation of ST data in tissue sections from patients contributes to the digitization of pathology and the improvement of patient stratification. To further verify a wider application of SSLAR in complex human tissues, we applied it on PDAC ST data which were generated by an ST protocol version differed from the platform. PDAC ST data are publicly available at the Gene Expression Omnibus (GSE111672) [8]. The data contain 10 spatial slides from 6 PDAC samples, two of which (i.e., PDAC-A and PDAC-B) provide paired scRNA-seq data.

First, SSLAR used PDAC-A and PDAC-B to select sections that contain both normal and tumor areas. For PDAC-A, SSLAR used GSM3036911 (ST1) and GSM3036909, GSM3036910, GSM3405527, GSM3405528, GSM3405529, and GSM3405530 (inDrops). Similar to Ref. [8], scRNA-seq data were filtered and preprocessed to keep cells with ≥ 1000 UMIs, $\leq 20\%$ mitochondrial transcripts, and $\leq 30\%$ ribosomal transcripts. For PDAC-B, SSLAR used GSM4100723 (ST2) and GSM3405531, GSM3405532, and GSM3405533 (inDrops). SSLAR removed ductal cells with low UMIs and high mitochondrial content. The labels of cell types on PDAC scRNA-seq datasets were obtained from Ref. [8].

To produce a comprehensive immune cellular reference atlas for PDAC, SSLAR re-analyzed scRNA-seq data from Ref. [9] (Genome Sequence Archive ID: PRJCA001063). Cells with < 100 UMIs and $> 20\%$ of mitochondrial content were removed. Next, SSLAR extracted, normalized, and scaled highly variable genes and implemented PCA analysis for the rest of cells before deconvolution. Cell types were annotated based on marker genes provided by Peng et al. [9]. All tumor and non-immune cells were removed through marker gene analysis and only immune cells were remained.

To stratify the PDAC tissue into tumoral and non-tumoral regions, tumoral spots were set to $> 40\%$ cancer cell proportion. Cell type proportions within tumoral spots and non-tumoral spots were compared and the biological significance was evaluated using the Mann-Whitney test. The proportion of spots containing each cell type was used to measure cell type enrichment between tumoral and non-tumoral regions. The significance between cell type proportions and spot proportion was analyzed with a permutations test by randomly building cell type-specific statistic distribution for 10,000 times for each cell type. Bonferroni adjusted P-values were used to assess if immune cells associate with tumor cells on PDAC. In addition, SSLAR constructed four cell type-cell type interaction networks, where the weight of each edge represents the proportion of cell type co-localization within spots.

As illustrated in Figure S20, SSLAR mapped cell subpopulations within the PDAC-A tissue, and delineated tumoral and immune cell distribution to detect the difference within the immune microenvironments in tumoral versus non-tumoral areas. Figure S20A charts the PDAC cellular composition. Figure S20B illustrates cell type distribution within normal and tumor tissues by integrating scRNA-seq and ST data (PDAC-A). The results shed light on a discrete regional enrichment between the two tissues. Notably, cell types within the normal pancreas tissues were mainly excluded from the tumor section and further partitioned to acinar and ductal regions. As shown in Figure S21, we further analyzed the PDAC tumor tissues. Centroacinar ductal populations were found in duct epithelium, while terminal ductal populations emerged in duct epithelium and co-localized in the cancerous part within the PDAC tissues. Figure S20C elucidates the inferred proportion of cancer clone S100A4, cancer clone TM4SF1, acinar cells, and terminal ductal cells within each spot. SSLAR captured an intersection between two different cancer clones and an enrichment of a ductal population in the PDAC cancer region. Figure S20D analyzed the distribution of pancreatic immune cell types within the PDAC immune microenvironment. Briefly, the used scRNA-seq data contain 41,986 cells from 24 PDAC patients. Notably, SSLAR captured 10,623 immune cells. As shown in Figure S20E and F, SSLAR was trained on PDAC immune cell dataset and then applied on the PDAC-A ST sections. Finally, it obtained a prominent local enrichment of tumor specific cells. Figure S21 illustrates spatial distributions of the remaining 6 cell types. Figure S20G depicts a remarkable segmentation of cancer and normal cells, which was consistent with the spatial distribution of immune cell types in the PDAC slides. Figure S20H and I demonstrated that acinar cells were enriched in the normal pancreas tissue while ductal cell type 2 was obviously increased in PDAC-A tumor ($P < 0.01$).

For PDAC-B, as illustrated in Figure S22, ductal cell type 2 and fibroblast cells again co-localized within the PDAC-B tumor regions, while acinar and endothelial cells were vanished from the tumor area. In addition, the enrichment of fibroblast cells could not be detected solely in one region. Fibroblast cells were limited to the PDAC-A tumor region while were highly abundant in all PDAC-B regions. This finding considerably boosted the need to accurately deconvolute spot composition for precise pathology evaluation. The difference between tumor and non-tumor regions and local enrichment of immune cell types advanced us to construct cell-cell communication networks using the cells' co-localization in the PDAC slides. Figure S23 visualizes four intercellular communication networks on PDAC. The four networks helped us to understand the concerted interactions between immune cells in the PDAC region and further provided insight into the PDAC-relevant tumor microenvironments.

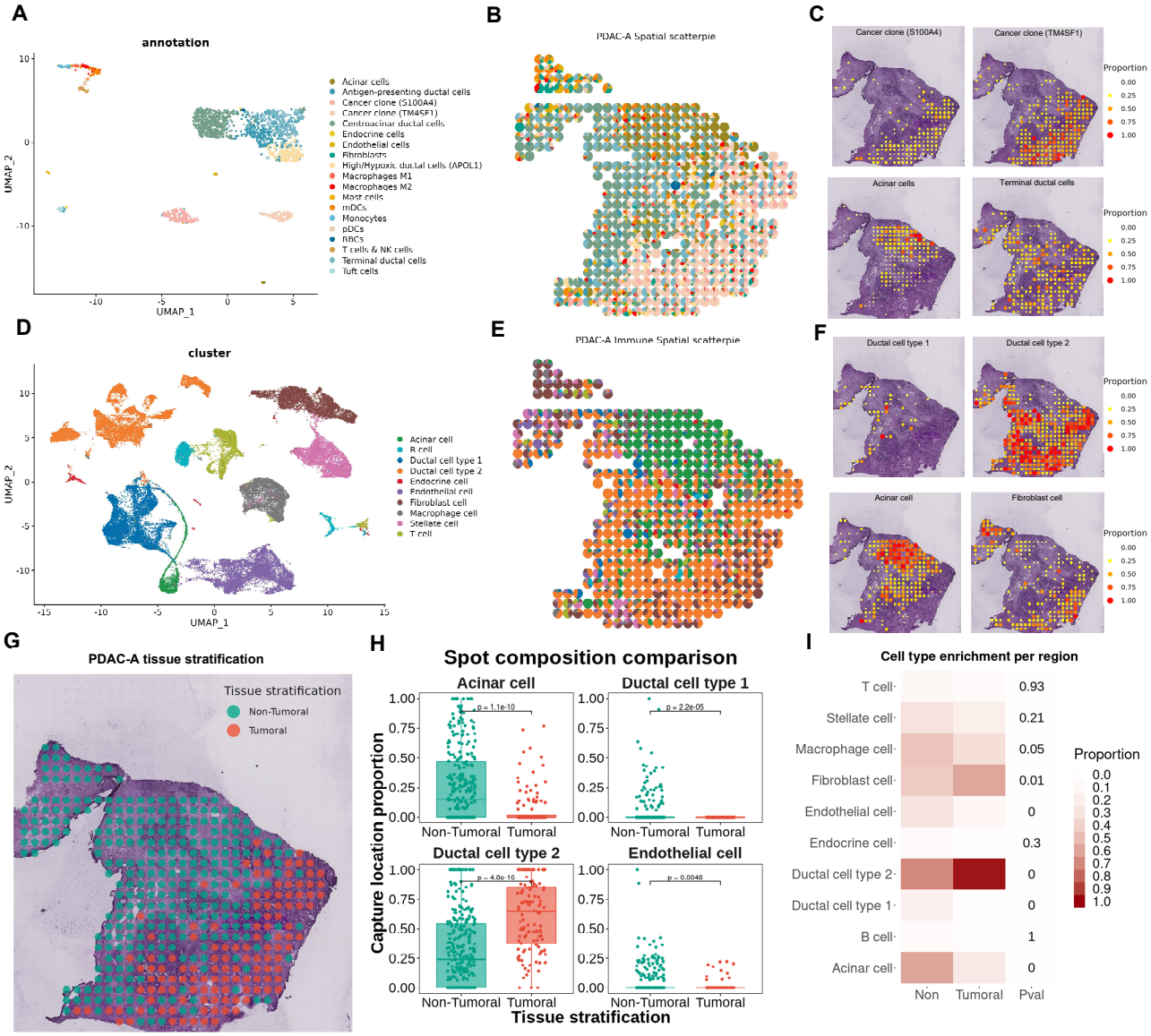
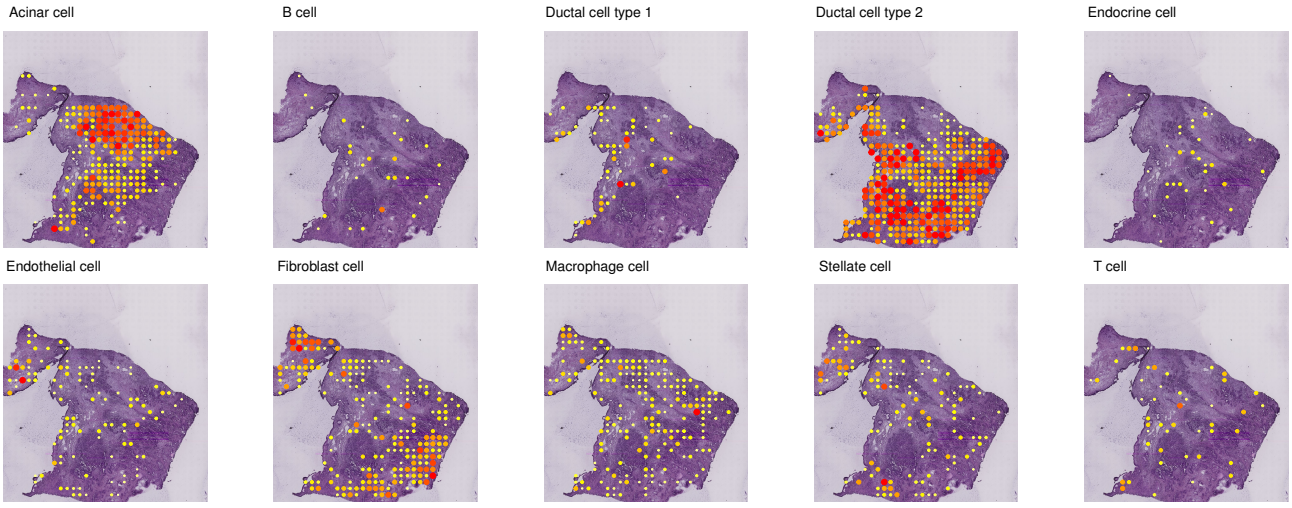


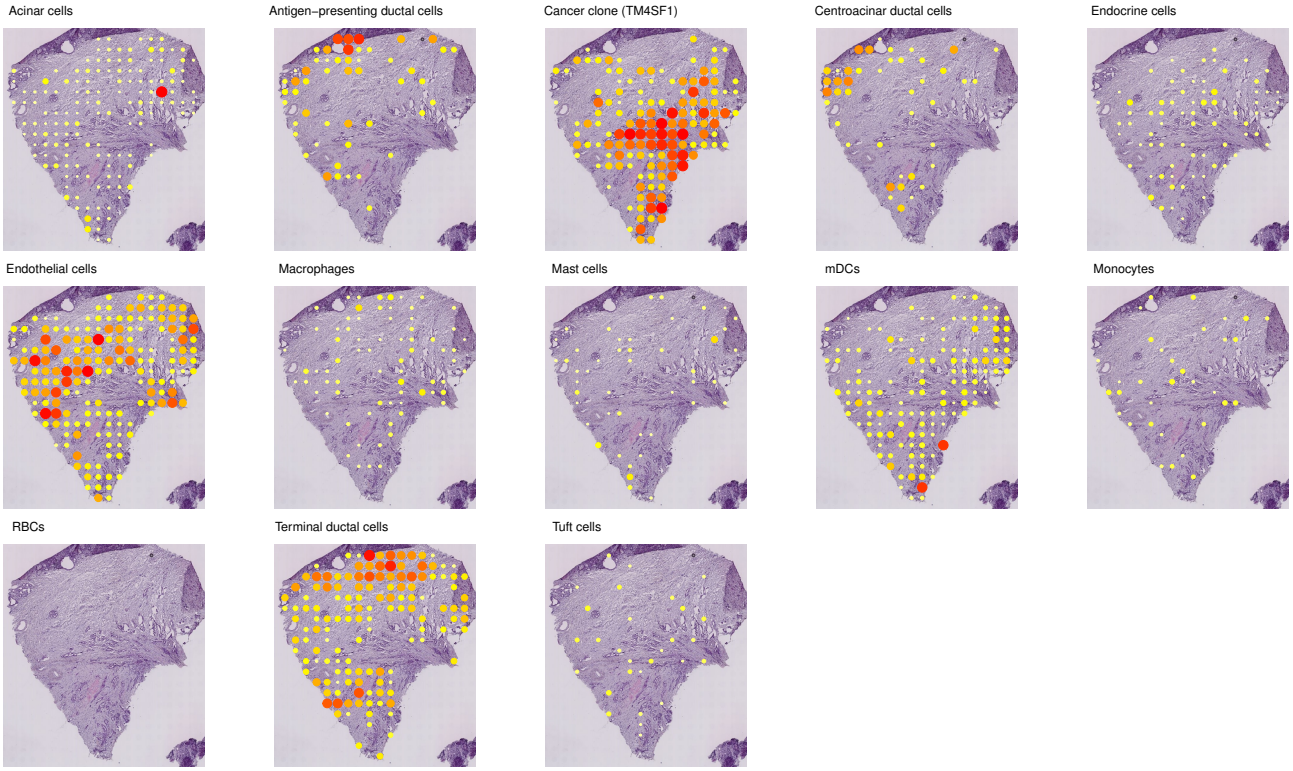
Figure S20: Visualization of cell subpopulations within the PDAC-A tissue. (A) UMAP atlas of 1,926 PDAC-A cells from the same tissue slices. The cells were labeled and colored based on the corresponding cell type annotations. (B) Spatial scatter pie plot denoting the proportions of all cell types within each spot. (C) The identified weights of four cell types within each spot (cancer clone S100A4, cancer clone TM4SF1, acinar cells, and terminal ductal cells). (D) UMAP atlas of pancreatic immune cells. (E) Spatial scatter pie plot denoting the weights of pancreatic immune cell types within each spot. (F) The identified weights of four immune cell types within each spot (ductal cell type 1, ductal cell type 2, acinar cells, and fibroblast cells). (G) Tissue stratification of tumoral-non-tumoral regions. (H) Proportion comparison of the four immune cell type within each spot between tumoral and non-tumoral regions. (I) Proportion of spots comprising all cell types within the tumoral and non-tumoral regions.

(a)

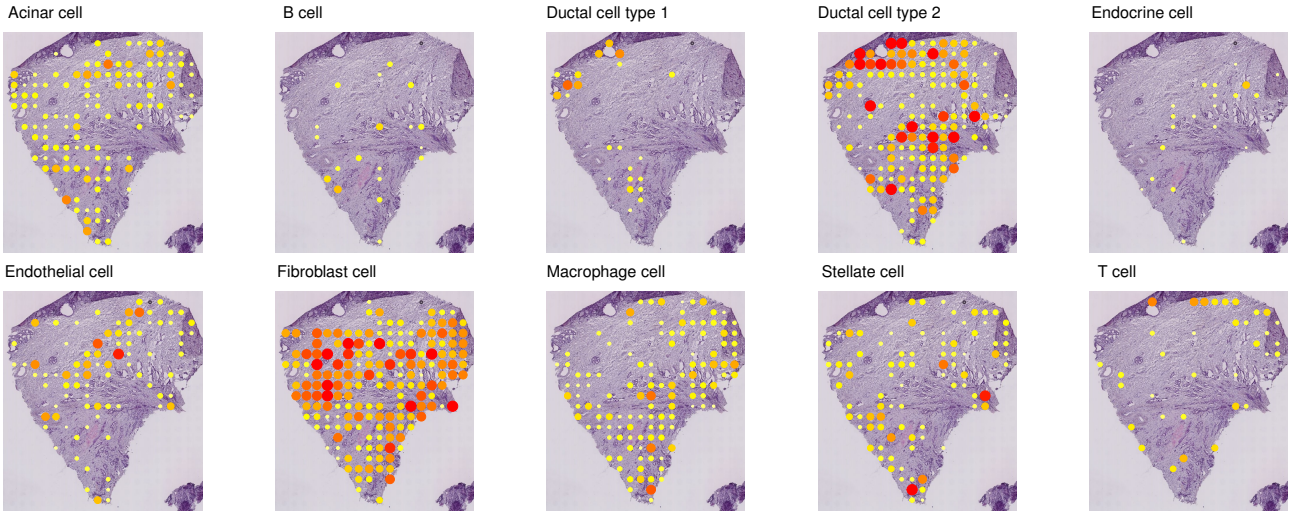


(b)

Figure S21: (a) The predicted proportions of cell types within each spot on the paired PDAC-A inDrops dataset. (b) The predicted proportions of cell types within each spot on the PDAC-A immune reference dataset.



(a)



(b)

Figure S22: (a) The predicted proportions of cell types within each spot on the paired PDAC-B inDrops dataset. (b) The predicted proportions of cell types within each spot on the PDAC-B immune reference dataset.

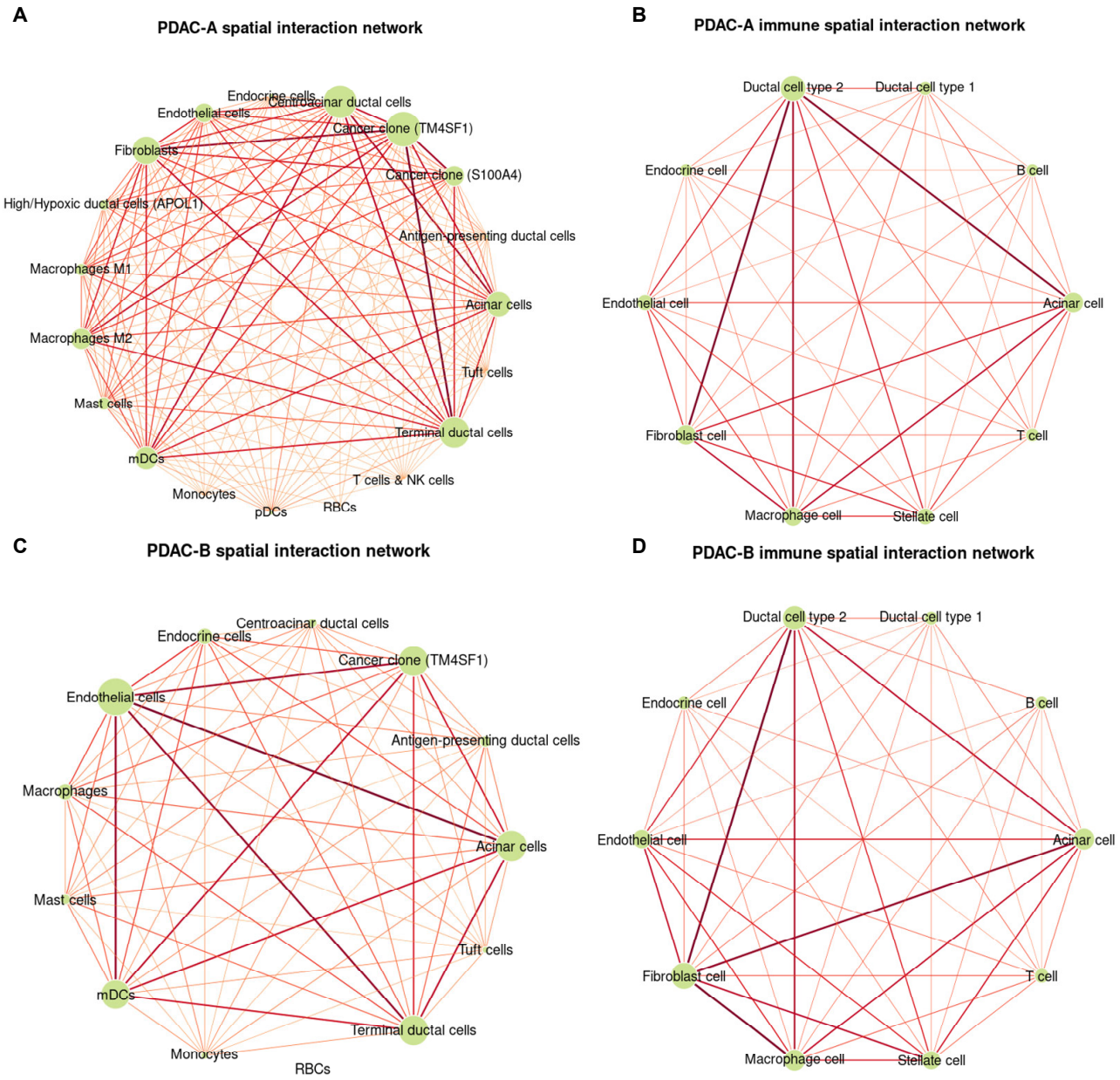


Figure S23: Spatial interaction networks characterizing the degree with which cell types co-localized together. (A) spatial interaction network on PDAC-A cell types. (B) spatial interaction work in PDAC-A immune cell types. (C) spatial interaction network on PDAC-B cell types. (D) spatial interaction work in PDAC-B. An bolder edge denotes more higher frequent of cell type co-localization.

Table S6: Key Resource Table

RESOURCE	SOURCE	IDENTIFIER OR LINK
Deposited data (sc/sn-RNAseq)		
PBMC	Gene Expression Omnibus	GSE133549
HNSCC	Gene Expression Omnibus	GSE103322
melanoma	Gene Expression Omnibus	GSE72056
pancreas	Gene Expression Omnibus	GSE85241
kidney tumor	Github	https://github.com/xuebaliang/scziDesk/tree/master/dataset/Young
mouse primary visual cortex	Allen Institute for Brain Science	https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq.
mouse brain	Gene Expression Omnibus	GSE185862
PDAC-A	Gene Expression Omnibus	GSM3036909, GSM3036910, GSM3405527, GSM3405528, GSM3405529, GSM3405530
PDAC-B	Gene Expression Omnibus	GSM3405531, GSM3405532, GSM3405533
pancreatic immune	Genome Sequence Archive	PRJCA001063
Deposited data (spatial transcriptomics)		
mouse cortex	Github	https://github.com/CaiGroup/seqFISH-PLUS
mouse visual cortex	STARmap	https://starmapresources.net/data
mouse brain posterior	10X Genomics	https://www.10xgenomics.com/resources/datasets
mouse brain anterior	10X Genomics	https://www.10xgenomics.com/resources/datasets
PDAC-A	Gene Expression Omnibus	GSM3036911
PDAC-B	Gene Expression Omnibus	GSM3405534
Software and algorithms		
Seurat	R Satija, et al.	https://github.com/satijalab/seurat
STRIDE	D Sun, et al.	https://github.com/wanglabtongji/STRIDE
DSTG	Q Song, et al.	https://github.com/Su-informatics-lab/DSTG
RCTD	D M. Cable, et al.	https://github.com/dmcable/RCTD
SpatialDDLS	D Mananes, et al.	https://github.com/diegommcc/SpatialDDLS
SCDC	M Dong, et al.	https://github.com/meichendong/SCDC/
SPOTlight	M Elosua, et al.	https://github.com/MarcElosua/SPOTlight
POLARIS	J Chen, et al.	https://zenodo.org/records/7302022
CARD	Y Ma, et al.	https://github.com/YMa-lab/CARD
SSLAR	This paper	https://github.com/plhhnu/SSLAR

References

[1] Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J McCarthy, Adrián Álvarez-Varela, Eduard Batlle, Dominic Grün, Julia K Lau, Stéphane C Boutet, et al. Benchmarking single-cell rna-sequencing protocols for cell atlas projects. *Nature biotechnology*, 38(6):747–755, 2020.

[2] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.

[3] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

[4] Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon Van Gurp, Marten A Engelse, Francoise Carlotti, Eelco Jp De Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3(4):385–394, 2016.

[5] Matthew D Young, Thomas J Mitchell, Felipe A Vieira Braga, Maxine GB Tran, Benjamin J Stewart, John R Ferdinand, Grace Collord, Rachel A Botting, Dorin-Mirel Popescu, Kevin W Loudon, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *science*, 361(6402): 594–599, 2018.

[6] Zizhen Yao, Cindy TJ van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

[7] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

[8] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C Devlin, Maayan Baron, Cristina H Hajdu, Diane M Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology*, 38(3):333–342, 2020.

[9] Junya Peng, Bao-Fa Sun, Chuan-Yuan Chen, Jia-Yi Zhou, Yu-Sheng Chen, Hao Chen, Lulu Liu, Dan Huang, Jialin Jiang, Guan-Shen Cui, et al. Single-cell rna-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research*, 29(9):725–738, 2019.

[10] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.

[11] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018.

[12] Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research*, 49(9):e50–e50, 2021.

[13] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[14] Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8, 2019.

[15] D Seung and L Lee. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

[16] Paolo Favaro and Stefano Soatto. *3-d shape estimation and image restoration: Exploiting defocus and motion-blur*. Springer Science & Business Media, 2007.

[17] Jamie Haddock, Lara Kassab, Sixian Li, Alona Kryshchenko, Rachel Grotheer, Elena Sizikova, Chuntian Wang, Thomas Merkh, RWMA Madushani, Miju Ahn, et al. Semi-supervised nmf models for topic modeling in learning tasks. *arXiv preprint arXiv:2010.07956*, 2020.

[18] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[19] Brendan F Miller, Dhananjay Bambah-Mukku, Catherine Dulac, Xiaowei Zhuang, and Jean Fan. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome research*, 31(10):1843–1855, 2021.

[20] Dongqing Sun, Zhaoyang Liu, Taiwen Li, Qiu Wu, and Chenfei Wang. Stride: accurately decomposing and integrating spatial transcriptomics using single-cell rna sequencing. *Nucleic acids research*, 50(7):e42–e42, 2022.

[21] Qianqian Song and Jing Su. Dstg: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics*, 22(5):bbaa414, 2021.

[22] Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40 (4):517–526, 2022.

[23] Diego Mañanes, Inés Rivero-García, Carlos Relaño, Miguel Torres, David Sancho, Daniel Jimenez-Carretero, Carlos Torroja, and Fátima Sánchez-Cabo. Spatialddls: an r package to deconvolute spatial transcriptomics data using neural networks. *Bioinformatics*, 40(2):btae072, 2024.

[24] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in bioinformatics*, 22(1):416–427, 2021.

[25] Jiawen Chen, Tianyou Luo, Minzhi Jiang, Jiandong Liu, Gaorav P Gupta, and Yun Li. Cell composition inference and identification of layer-specific spatial transcriptional profiles with polaris. *Science Advances*, 9(9):eadd9818, 2023.

[26] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.

[27] Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature methods*, 19(6):662–670, 2022.

[28] Alberto Pascual-Montano, Jose Maria Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE transactions on pattern analysis and machine intelligence*, 28(3):403–415, 2006.

[29] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[30] Mark S Cembrowski, Lihua Wang, Ken Sugino, Brenda C Shields, and Nelson Spruston. Hipposeq: a comprehensive rna-seq database of gene expression in hippocampal principal neurons. *elife*, 5, 2016.