SC2Spa: a deep learning based approach to map transcriptome to spatial origins at cellular resolution

Linbu Liao[1,$], Esha Madan[2,3,4,5,$], António M. Palma[2,3,4,5,6], Hyobin Kim[7], Amit Kumar[3,4,5], Praveen Bhoopathi[3,4,5], Robert Winn[2,3], Jose Trevino[2,3], Paul Fisher[3,4,5], Cord Herbert Brakebusch[1], Rajan Gogna[2,3,4,5,#], Kyoung Jae Won[1,7,#]

1- Biotech Research and Innovation Centre (BRIC), University of Copenhagen, 2200 Copenhagen N, Denmark, Faculty of Health and Medical Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark,

2- Department of Surgery, Virginia Commonwealth University, School of Medicine, 1200 E Broad St. P.O. Box 980011, Richmond VA, 23298, USA

3-Massey Cancer Center, Virginia Commonwealth University, Richmond, VA, 23298, USA

4-Institute of Molecular Medicine, Department of Human and Molecular Genetics, Virginia Commonwealth University, School of Medicine, Richmond, VA, 23298, USA

5-Department of Human and Molecular Genetics, Virginia Commonwealth University, School of Medicine, Richmond, VA, USA

6-Instituto Superior Tecnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal

7- Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA

$-These authors contributed equally to this work

#-To whom correspondence should be addressed.

RG (Tel: +1 628-252-6501; E-mail: rajangogna@gmail.com; rajangogna@vcuhealth.org)

and

KJW (Tel: +1 310-423-3967; Email: kyoungjae.won@cshs.org).

Abstract

Integrating single cell RNAseq (scRNAseq) and spatial transcriptomics (ST) data is still challenging especially when the spatial resolution is poor. For cellular resolution spatial mapping, we have developed deep learning-based SC2Spa to learn the intricate spatial mapping rules from the transcriptome to its location from ST data. Benchmarking tests show that SC2Spa uniquely recapitulates tissue architecture from scRNAseq. SC2Spa successfully mapped scRNAseq even to various low resolution Visium data. SC2Spa identified spatially variable genes and suggested negative regulatory relationships between genes. SC2Spa armored with deep learning provides a new way to map the transcriptome to its spatial location and perform subsequent analyses.

Keyword: Spatial Transcriptomics, spatial inference, negative regulatory relationships, deep learning

Background

Cells function within the context of tissue architecture and continuously interact with their local environment. The transcriptome of a cell varies in accordance with the location and the environment. For instance, cells express genes related to cell communication (such as ligands and receptors) and genes involved in downstream pathways [1]. Interaction with the neighboring cells influences the transcriptome of a cell [2-5]. Also, there are genes whose spatial expression patterns are associated with tissue architecture [6] supporting the fact that cell transcription is influenced also by the global tissue structure. On a larger scale, several genes showed gradual pattern along the developing drosophila embryo body [7].

Single-cell RNA sequencing (scRNAseq) portrays a snapshot of the gene expression of a cell. This expression profile has been used to identify cell types and dynamic transcriptomic changes [8-11]. However, loss of spatial location makes it hard to interpret scRNAseq with respect to the influence of cellular neighborhood and spatial location within a tissue. There have been attempts to study cell-cell communication from scRNAseq. For instance, expression of ligand-receptor pairs from donor and recipient cells has been used to predict the cell types communicating frequently [12]. We previously showed that the neighboring cell type can be predicted from scRNAseq [3].

Considering the rich information of cells contained in the transcriptome of cells, can the transcriptome of cells be used to predict their location within tissue and infer tissue architecture? Several systemic approaches have tried to predict spatial origin from the transcriptome [13-20]. Many algorithms have relied on transcriptomic similarities to map a cell to its potential spatial origins [13, 15, 21]. Seurat used mixture models and fitted their parameters by estimating the posterior probability to maximize transcriptomic similarity [14]. SpaOTsc applied the optimal transportation algorithm for spatial reconstruction, followed by cell-cell communication inference using ligand-receptor co-expression [17]. NovoSpaRc reconstructed spatial map assuming continuity in gene expression [16, 20]. Despite these efforts, it is still arduous to map single cells to their spatial origins at a single cell resolution. Previous spatial inference tools lack the ability to generalize over large datasets. In addition, the design of previous

studies neglected the relative position between spatial voxels, resulting in the loss of information. The algorithms based on transcriptomic similarity and continuity cannot promise accurate detection of cell location at cellular resolution. Ligand-receptor co-expression can only explain frequently interacting cell types and cannot provide the precise location of a cell.

To overcome these challenges, we suggest using deep learning as a method to learn the rules to map cells to their respective location within a tissue. Deep learning has the power to represent complex high-dimensional relationships by combining multiple layers of nonlinear functions [22]. There were attempts to use deep learning in predicting location from transcriptome. DEEPsc was developed to map single cells to spatial voxels based on the transcriptome using deep learning. DEEPsc defined mapping as a binary classification task [19], resulting in modest performance. Tangram applied deep learning to map scRNAseq to reference ST data [18]. To achieve this, Tangram generates an initial random spatial placement of scRNAseq then it re-aligns cells in the spatial dimension to maximize the total spatial correlation.

We noted that previous deep learning based approaches did not train their network to predict the cell locations directly. To capture the rules of location mapping from the transcriptome to the location in the weight space of the deep neural networks, we developed SC2Spa. SC2Spa consists of a fully connected neural network (FCNN) designed to predict the location of a cell from its transcriptome. SC2Spa uses ST data such as Slide-Seq V2 [23] or Visium [24] to learn the absolute spatial coordinate (output) from transcriptome (input). The trained model can be used to predict the original location of transcriptome (e.g. scRNAseq).

Benchmarking tests show that previous approaches fail to predict the location of single cells correctly and the original tissue structure is completely lost in the predicted results. In contrast, SC2Spa predicted the approximate location of cells and recapitulated the original tissue structures from independent ST data and scRNAseq data. Even SC2Spa trained on low-resolution Visium [24] data successfully recovered the spatial location of key marker genes from scRNAseq.

SC2Spa can suggest spatially variable genes (SVGs) using the trained weight. Among the SVGs identified by SC2Spa, we found gene pairs showing exclusive patterns when visualized. We devised a score to measure the spatial exclusiveness between two genes from ST data and predicted potential negative regulatory relationships. Experimental work supported the negative relationships that we predicted.

RESULTS

Overview of SC2Spa

SC2Spa is designed to infer spatial location of cells using deep learning (**Figure. 1a**). The core component of SC2Spa is a FCNN. To learn the mapping rules from transcriptome to the location, SC2Spa uses ST data such as Visium [24], Slide-seqV2 [23] or image-based ST data [15]. For training, SC2Spa uses the transcriptome of ST data as the input and the spatial location of the ST data as the output. Once trained, SC2Spa can predict the spatial location from the transcriptome of scRNAseq or ST (**Figure. 1a**).

The FCNN of SC2Spa has a flexible structure. By default, SC2Spa uses 6 hidden layers and the number of nodes in these layers was set to decrease by 4 folds from 4096 (i.e. 4096, 1024, 256, 64, 16, and 4, respectively). The input layer is designed to accept the entire transcriptome of a cell or a bead and the output layer is composed of 2 nodes for spatial coordinates. The rectified linear unit (ReLU) function is used for the first five hidden layers. To avoid overfitting, the input layer and the first 5 hidden layers were randomly deactivated during the training with a dropout rate of 0.05 and were normalized. The output layer uses the sigmoid function. The Adam algorithm [25], an algorithm for first-order gradient-based optimization of stochastic objective functions, was used to train the weights (Method).

The trained SC2Spa model can predict the spatial origin by using the transcriptome of scRNAseq or ST data. The predicted outputs are the coordinates on the reference ST assay which was used to train the model (**Figure. 1a**). Multiple cells or ST spatial indices can be mapped to a spatial index on the reference slide. For visualization, we merged the weighted predicted outputs when they are within a certain range

from the center of a spatial index. The weights are inversely proportional to the distance from the center of the spatial index in the reference assay (Method).

To provide a level of interpretability to the trained FCNN, SC2Spa trains the weights while penalizing the sum of weights by using L1 regularization (Method). The weights of the network that do not contribute to the location prediction tend to shrink to zero during training. SC2Spa determines which genes contribute to location prediction by backtracking the most informative weights (**Figure. 1b**), providing an alternative approach to detecting SVGs. Based on the spatial expression patterns, SC2Spa further measures the degree of exclusiveness between two genes and identifies potential negative gene regulatory relationships (**Figure. 1c**).

SC2Spa outperforms other spatial inference algorithms in the cross-validation tests

We assessed the performance of SC2Spa by benchmarking against other location predictors including Tangram [18], Novosparc [16, 20], and SpaOTsc [17]. Additionally, we included a simple logistic regression (LR) [26]. For all approaches, 5-fold cross-validation was implemented using various ST datasets.

First, we used the Slide-seqV2 datasets in the mouse brain [23] focusing on the beads for 8 cell types including CA1 principal cells, CA2 principal cells, dentate principal cells, interneurons, neurons, entorhinal cortex, oligodendrocytes, and astrocytes. After training, the coordinates for the transcriptome of the test set were visualized for all predictors (**Figure. 2a**). The prediction by SC2Spa showed striking similarity with the reference test set. For instance, SC2Spa roughly recapitulated the Cornus Ammonis (CA) composed of the CA1 and CA3 regions as well as the Dentate gyrus (**Figure. 2a**). Compared with it, other algorithms we tested failed in predicting the location of cells and could not retrieve the structure of the reference image (**Figure. 2a**). Additional cross-validation tests using Slide-seqV2 from mouse embryo [23] further confirmed that SC2Spa accurately predicted the distribution of epithelial cells, myocytes and stromal cells while other approaches failed in predicting their distribution **(Supplementary Figure 1)**. To

interpret these results further, we calculated Euclidean distances between the predicted and the original location. While SC2Spa successfully identified the location for all regions in general, Tangram, NovoSpaRc and SpaOTsc showed worse performance for the transcriptome located at the periphery of the ST assay (**Supplementary Figure 2,3)**.

We also performed cross-validation using the Stereo-seq data from mouse embryo [27]. Stereo-seq is a sequencing based ST data with a resolution (<1μm) higher than Slide-seq. SC2Spa again recapitulated the reference location of the transcriptomes while other approaches failed in predicting them (**Supplementary Figure 4**). Again, Thangram, NovoSpaRc and SpaOTsc showed worse prediction performance for the transcriptome located in the periphery of the assay when we investigated Euclidian distances between the reference and the predicted locations (**Supplementary Figure 5**).

For comprehensive tests, we also tested using the three-dimensional (3D) fluorescence in situ hybridization (FISH) data from drosophila embryos [15]. These FISH data contains the transcriptomic profile of 84 genes from the drosophila embryo [15]. All approaches except for LR predicted the rough layout of fruit fly embryos (**Supplementary Figure 6, 7**).

To provide quantitative assessment, we used various measures including Pearson's correlation coefficient (*r*), root-mean-square error (RMSE) and the coefficient of determination (R^2) between the reference ST data and the prediction on it (Methods). R^2 indicates how well a statistical model predicts an outcome (Methods). Besides, we added pairwise Pearson's correlation coefficient (*r*) and pairwise RMSE of the predicted and original distances of all possible cell/bead pairs. All approach including RL showed Pearson's *r* larger than 0.75 for 3D FISH dataset (**Figure 2b**). In all these criteria, SC2Spa showed the best Pearson's *r*, R^2, and pairwise Pearson's *r* with the smallest RMSE and the pairwise RMSE values (**Figure 2b**).

SC2Spa outperforms in predicting the spatial origin from independent ST data

We further tested if SC2Spa can be generalized to predict the spatial origin of the transcriptome from untrained independent ST data. For this test, we trained SC2Spa and other approaches using a Slide-SeqV2 puck (ID: puck_200115_08) and tested using another puck (Query set, ID: puck_191204_01) from the mouse brain [23].

As an example, we show the beads for CA1 and CA3 pyramidal cells and the dentate gyrus for both the training and the query set (**Figure. 3**). SC2Spa, trained on one puck (**Figure. 3a**), processed the transcriptome from the Query set (**Figure. 3b**), and predicted the location for the query data (**Figure. 3c**). In this test, SC2Spa predicted the structure of CA1, CA3, and dentate gyrus regions. It is notable that the orientation of the predicted results followed the training set rather than the test set. Considering that the orientation from the query set is lost and only transcriptome is used as the input for SC2Spa, it is natural that the orientation of the predicted image follows that of the training set (**Figure. 3c**).

To assess the spatial similarity of the gene expression we used bivariate Moran's I (BVI) [28, 29] (Method). BVI measures the spatial autocorrelation between two genes. If BVI=0, there is no spatial autocorrelation. Visualization of individual genes showed that the spatial distribution of genes such as *Hpca* (BVI = 0.35) and *Pcp4* (BVI=0.09) match well with the spatial patterns observed in the training as well as the Query set (**Figure. 3d and 3e**). A deeper analysis revealed mild expression levels of *Hpca* in the Dentate Gyrus, which becomes stronger in CA regions with very strong expression levels in the CA2 region. These tendencies were observed in both the training set and the prediction images (**Figure. 3d; marked by an arrow**). The gradual changes of the *Pcp4* expression were also observed in the training set and the predicted image (**Figure. 3e**). This suggests that SC2Spa can capture even subtle changes in gene expression related to cell location (**Figure. 3d and 3e**). Compared with the results of SC2Spa, other competitors could not retrieve the tissue structures in their prediction. The evaluation using the pairwise Pearson's *r* and the pairwise RMSE further confirmed the outperforming performance of SC2Spa (**Supplementary Figure. 8).**

We further investigated the mapping for each cell type. For this, we visualized the cell types from the reference ST (**Supplementary Figure. 9).** We performed clustering and the results are visualized in the query ST assay **(Supplementary Figure. 10)**. The predicted results of the clusters identified in the query ST data show that the prediction matches well with the cell types in the reference data **(Supplementary Figure. 11)**. For instance, cluster 14, 16 and 11 of the predicted results matched well with the dentate, CA3 and CA1 principal cells of the reference data.

As an additional test, we investigate the performance when testing with 10X Genomics Visium [24] data. Visium [24] is a commercially available ST approach but with lower resolution (around 55 μm in diameter per spot) compared with Slide-Seq (10 μm). For training we used brain data from a mouse (ID: ST8059051 and ST8059052) and tested it using the Visium data from another mouse (ID: ST8059048). Benchmarking shows that SC2Spa achieves the best pairwise RMSE and Pearson's r **(Supplementary Figure 12a)**. SC2Spa reconstructed the rough structure of clusters of the query data while other algorithms failed **(Supplementary Figure 12b)**. Genes such as *Nrgn* (BVI = 0.68), *Prkcd* (BVI = 0.72) and *Slc17a6* (BVI = 0.6) show spatial distribution similar betwwn the training, the query and the prediction **(Supplementary Figure. 13)**.

SC2Spa recovered the spatial locations from scRNAseq

scRNAseq loses the location information during the dissociation into single cells. Annotating cell location can help our understanding about cell heterogeneity. Besides, integrating scRNAseq with low-resolution Visium data can enhance the usage of Visium data.

We used SC2Spa trained with the cellular-resolution Slide-Seq V2 data from the mouse brain [23] and predicted the location of scRNAseq using the publicly available data from the mouse brain [30]. In total, we predicted the location for 24 annotated cell types from scRNAseq [30] **(Figure. 4** and **Supplementary Figure. 14)**. The prediction results matched well with the known location for each cell type. For instance, the prediction for CA1, CA2, CA3 pyramidal cells and dentate gyrus granule cells

matched well with the corresponding regions (**Figure. 4a**). Spatial expression of *Nrgn* (BVI=0.28), *Fibcd1* (BVI=0.27), *Tshz2* (BVI=0.29), and *Hs3st4* (BVI=0.26) confirmed the similarity between the reference ST data and the predicted spatial image (**Figure. 4b**).

We further evaluated if SC2Spa can still be used to predict the location of single cells even when trained on low-resolution Visium data. We used public Visium 10X for training and single nucleus RNA sequencing (snRNAseq) datasets obtained from 2 mouse brains for testing (one from the same mouse and another from the other mouse) [31].

We assessed the location mapping of oligodendrocytes (localized mainly in the white matter (WM)), inhibitory neurons (mainly in hypothalamus (HYP)), and excitatory neurons (mainly in thalamus (THA)) that express marker genes such as *Mbp*, *Gad2*, and *Slc17a6*, respectively [31]. The visualization of the Visium image data confirmed spatial expression for these genes in the associated regions (**Figure. 5a**). We were intrigued to find that the location mapping of the scRNAseq from the same mouse brain exhibited the cell localization patterns matching with our prior knowledge **(Figure. 5b)**. The test using the scRNAseq from an independent mouse brain also showed the same cell localization patterns **(Figure. 5c)**. Genes such as *Rgs9* (BVI=0.2), *Eps8l2* (BVI=0.16), *Stard8* (BVI=0.05), and *Otof* (BVI=0.14) further confirmed the similarity between the predicted results and the Visium data. We further showed the predicted locations for the cell types identified using snRNAseq [31] **(Supplementary Figure. 15, 16)**.

We also applied SC2Spa to the human kidney cancer Visium data and mapped the location using the accompanying scRNAseq data [32]. SC2Spa successfully recapture the spatial gene expression patterns in the kidney cancer tissue. For instance, *DMKN* (BVI = 0.32), *EGR1* (BVI = 0.18), *KCNJ1* (BVI = 0.35), *MT1H* (BVI = 0.24), *RASD1* (BVI = 0.23) and *RPL37* (BVI = 0.20) show high spatial expression similarity between the original Visium data and prediction **(Supplementary Figure 17)**.

We further trained SC2Spa using human breast cancer Visum data [33] and tested the mapping performance using the 2 scRNAseq datasets from formalin-fixed paraffin-embedded (scFFPE-seq) obtained using 3’ and 5’ protocol by 10X Genomics, respectively. We found robust performance of SC2Spa

regardless of the scRNAseq protocol. For example, *CYB561*, *FASN* and *PVALB* showed high spatial similarity among the reference and two predictions **(Supplementary Figure. 18)**. We also noted that the prediction results using 3' scFFPE-seq and 5' ssFFPE-seq are consistent for these genes with the BVI>0.4. When checking BVI values for all genes, we found that more than 10,000 genes showed BVIs higher than the ones for the randomly shuffled gene expression values **(Supplementary Figure. 19)**.

SC2Spa predicts SVGs

A key feature of SC2Spa is the ability to score the SVGs from their weight space. We compared the SVGs obtained with SC2Spa to those obtained by SpatialDE [34]. SpatialDE uses statistical tests to identify genes with spatial variable expression patterns and calculates the log-likelihood ratio (LLR) score against a random distribution. We plotted the SpatialDE's LLR values against the scores indicating contribution to location prediction from SC2Spa and found positive relationships **(Supplementary Figure. 20a)**. Genes such as *Ttr*, *Pcp4*, *Calb2* and *Tshz2* showed a high score for both of the methods **(Supplementary Figure. 20b)**. We also found genes such as *Gm13872*, *Gm26724*, and *2010005H15Rik* were only observed by SpatialDE **(Supplementary Figure. 20c)**.

Moran's I has been used to assess the distribution of spatial gene expression [35, 36]. We found that the scores for contribution to location prediction by SC2Spa better correlate with Moran's I than SpatialDE scores **(Supplementary Figure. 20c)**. Visual inspection showed that SVGs specific to SpatialDE showed a near random distribution with a very low number of beads with their expression. Their absolute Moran's I score was less than 0.001 **(Supplementary Figure. 20a)**. Compared with them, the SVGs commonly observed by the two methods showed Moran's I value higher than 0.17 **(Supplementary Figure. 20b)**.

Similar outcomes were observed in the second Slide-seqV2 mouse hippocampus dataset and a Slide-seqV2 mouse embryo dataset **(Supplementary Figure. 21)** and **(Supplementary Figure. 22)**. Some genes (*Tert*, *Gm19684*, and *Gm10640* for mouse hippocampus and *Krt75*, *Tmem171*, and *Has1* for mouse

embryo) with a high SpatialDE score do not exhibit any spatial pattern and have an absolute Moran's I score of less than 0.0001 **(Supplementary Figure. 21a and 21b)** and **(Supplementary Figure. 22a and 22b)**. In particular, the SVGs highly observed only in SC2Spa such as *Cxcl14* showed informative spatial expression in the mouse embryo data **(Supplementary Figure. 22c)**. Again, the score for SC2Spa better correlated with Moran's I in these datasets **(Supplementary Figure. 20c)**. Our results show that SC2Spa can suggest an alternative way to obtain SVGs.

SC2Spa identified potential negative gene regulation from spatial gene expression patterns.

We noticed that some genes share similar spatial expression patterns. Clustering using SpatialDE [34], we found 7 major clusters for the SVGs using the mouse hippocampus datasets **(Supplementary Figure. 23)**. The spatial clustering could mean co-regulated genes. Similarly, we found spatially exclusive gene expression expression patterns (e.g. cluster 2 and 6).

To search for mutually exclusive gene expression, we used a balanced mutual exclusivity (BME) score and the statistical significance of each gene (Method) **(Supplementary Figure. 24)**. BME calculates the exclusiveness of the expression for a gene pair while emphasizing the balance between the number of spots expressing them. We successfully identified 414 gene-pairs whose expression patterns were mutually exclusive (BME>0.45) **(Supplementary Table 1)** including the pair of *Plp1* and *Camk2a* (**Fig. 6a**). The superimposed expression for *Plp1*(blue) and *Camk2a* (red) showed that beads with high expressions of both genes are rare. We further investigated their gene expression using scRNAseq from adult mouse brain [30] and neuron-associated cells in Tabula Muris Senis [8]. These experiments confirmed the exclusive relationships between these two genes **(Figure. 6b)** and **(Supplementary Figure. 25)**.

Camk2a (calcium/calmodulin-dependent protein kinase 2 alpha) is an important enzyme for synaptic plasticity [37-39]. Previous studies have suggested *Camk2a* is involved in various disorders including bipolar disorder [40], Alzheimer disease [41], schizophrenia, and depression [42]. *Plp1* (Proteolipid protein 1) is the predominant myelin protein present in the central nervous system (CNS) [43]. Mutations in *Plp1* are associated with Pelizaeus–Merzbacher disease [43]. Previous study had observed a

negative correlation between *Plp1* and *Camk2a* in the PFC region of the brain from depressed suicides [44], suggesting a negative regulatory relationship between these two genes.

We experimentally investigated the regulatory relationships between Plp1 and Camk2a by performing Lentivirus-mediated overexpression as well as shRNA-mediated knockdown of both genes in freshly isolated murine hippocampal astrocytes (Methods). As predicted by SC2Spa, overexpression of Plp1 lead to a 2-fold reduction of *Camk2a* compared to its expression in wild-type cells, whereas Plp1 knockdown led to a 6-fold increase of *Camk2a* expression (Figure 6c). Similarly, genetic manipulation of Camk2a expression via overexpression resulted in approximately 2-fold decrease in the expression of Plp1, while Camk2a knockdown increased the expression of Camk2a by 7-fold (**Figure 6d**).

Discussion

By providing both transcriptomic profiles and spatial location, ST is becoming more popular for many areas in biology [45-49]. However, there are still some limitations in using ST data. Imaging-based ST techniques such as seqFISH+ [50], MERFISH [51] and EEL FISH [52] can obtain cellular resolution ST data. However, they are limited by the number of genes they capture simulanously. Barcode-based ST sequencing techniques target RNA species unbiasedly. However, a spatial barcode can be mapped to multiple cells or a portion of a cell depending on the resolution [23, 24, 27, 53-57]. Even without the spatial component, scRNAseq is still useful as it provides an unbiased transcriptome measure of cell. Plenty of resources have been accumulated in the form of scRNAseq. Location prediction for scRNAseq will be extremely useful to explain cell heterogeneity, communication and function in a given tissue architecture.

A number of approaches have been developed to predict the spatial origin of scRNAseq. Majority of them tried to map the transcriptome to the location by minimizing the distances between spatial reference and the scRNAseq data [13-17, 20]. A deep learning based Tangram also tried to minimize the distance between the meta-ST data it generates and the reference [18]. These approaches worked for 3D FISH data where transcriptomic changes are distinct for each spatial position **(Figure. 2** and **Supplementary Figure**

8, 9). However, for ST with complicated tissue structures, they failed in recapitulating accurate cell mapping **(Figure. 2** and **Supplementary Figure. 1,2,3,4,5)**.

Using a deep neural network framework, SC2Spa was developed to learn intricate mapping rules from transcriptome to spatial coordinates directly from ST data. For all cross-validation tests, SC2Spa showed far outstanding performance compared to the other approaches we tested. SC2Spa successfully retrieved the spatial pattern of cells and the associated genes when applied to high-resolution SlideSeq V2 data **(Figure. 2)**. On the other hand, the predicted patterns by other methods were not even close to the reference location information **(Figure. 2)**.

The ultimate goal is to predict the spatial origin of scRNAseq. SC2Spa successfully predicted the location for unseen SlideSeq V2 data and scRNAseq data from mouse brain **(Figure. 3 and 4)**. It is also notable that SC2Spa trained on even low-resolution Visium data still accurately predicts the location of scRNAseq data **(Figure. 5, Supplementary Figure 15,16,17,18)**. Our predictions indicate that scRNAseq retains information about the location of the cells in the tissue before cells are isolated. The gene expression changes due to location may be subtle but SC2Spa successfully modeled them using deep learning. These results suggest that cell heterogeneity is not always random and indicates a specific role given the structure of a tissue. SC2Spa was trained using various ST data and tested on scRNAseq data **(Supplementary Table 2)**.

The mapping of scRNAseq into Visium can suggest potential use of SC2Spa in enhancing the spatial resolution of Visium data. scRNAseq has been used to study cell-cell communication [12, 58]. However, it is of note that additional spatial information can boost the performance of predicting cell-cell communication [59]. We applied SpaTalk [59], a software to detect cell-cell communication from ST data, using the spatial information provided by SC2Spa. We found the interaction of interneuron and astrocyte through ErbB4 and Hbegf in the spatial domain **(Supplementary Figure. 26)** and we further identified downstream genes **(Supplementary Figure. 27a)** and related biological processes **(Supplementary Figure. 27b).**

Deep learning often suffers from overfitting. We examined the overfitting using the 5-fold cross-validation. We calculated the RMSE values for the training data and the test data while we train SC2Spa. The RMSE values for the training data gradually decreased (**Supplementary Figure 28,29,30,31**). Even though the RMSE values for the test data fluctuated over the training epoch, we observed general decrease of the RMSE values for the tests we performed (**Supplementary Figure 28,29,30,31**). This indicates that overfitting does not seriously influence the performance of SC2Spa.

There are a number of methods to decompose spatial beads into the cell type proportions composing the transcriptomic profile [60-66]. Our approach is different in that we try to find the best position that a cell can be mapped based on transcriptome information and do not aim to decompose spatial transcriptome.

The trained weights also provide additional knowledge about the genes. SC2Spa scores the genes according to the contribution towards location prediction. We regarded them as SVGs even though the weights were not originally designed to. The SVGs from SC2Spa better correlated with Moran's I than those recovered by SpatialDE [34]. The clustering of genes further found spatially exclusive patterns in gene expression. We found potential negative regulatory relationships between genes including *Plp1* and *Camk2a* **(Figure. 6)**. It is very intriguing that pattern analysis leads to identification of negative gene regulation. Our study suggests that ST can be used to infer gene regulatory rules.

SC2Spa is designed to accept one ST data. It is possible to use multiple ST datasets if multiple ST data are merged when the ST data have similar spatial expression patterns and share a similar coordinate system. In our tests, SC2Spa and logistic regression were run with a 24GB RTX Titan GPU. The other CPU-based softwares were run with a 18-core Intel® Xeon(R) W-2195 CPU of 2.30GHz. We compared SC2Spa with existing algorithms for the applications where the locations of the query data are available (**Supplementary Table 3)**. As SC2Spa maps gene expression directly to coordinates the computational complexity of SC2Spa increases linearly. The complexity of optimal transport-based methods increases linearly with the magnitude of the reference and query data. When both the quantity of the reference data and the query data increases, the complexity of the other methods increases quadratically. As the size of

the input data increases, SC2Spa will exhibits a distinct advantage over other algorithms in terms of execution time.

Conclusions

We developed a new framework SC2Spa to map scRNAseq to spatial locations. Both high-resolution and low-resolution spatial transcriptomics (ST) references can be used for the spatial mapping with SC2Spa.The location prediction by SC2Spa demonstrates the emergence of expression patterns contributing tissue architecture contained within scRNAseq. SC2Spa also provides functions for the identification of spatially variable genes and potential negative gene regulation. SC2Spa is a tool that can extract new knowledge from scRNAseq data.

Methods

Data processing for SC2Spa

To normalize the ST data, the unique molecular identifier (UMI) count per cell is normalized to 10,000 and then logarithmically transformed. SC2Spa can choose either polar or Cartesian coordinates. Polar coordinates are useful for circular ST data. The spatial coordinates of the beads were normalized between 0 and 1 using min-max normalization. The FCNN is trained to predict the normalized coordinates of beads.

Training neural network

SC2Spa is composed of 8 layers (input, output, and 6 hidden layers). The number of nodes for the input layer is the number of the transcriptome and the number of the output node is 2 for the spatial coordination. By default, the number of nodes in these layers was set to 4096, 1024, 256, 64, 16, and 4, respectively. The ReLU function is used for the first five hidden layers.

$$ReLU(z) = \max(0, z) \tag{1}$$

The output layer uses the sigmoid function:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

For training, SC2Spa uses the Adam algorithm [25]. The RMSE is the loss function used to train the FCNN. L1 regularization was used to penalize dense layers.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}{N}} \tag{3}$$

$$loss = RMSE + \lambda \sum_{i=0}^{I} \sum_{j=0}^{J_i} \sum_{k=0}^{K_i} |W_{ijk}| \tag{4}$$

, where

N: Number of voxels/cells from the reference ST data.

y_t: Spatial coordination in the reference ST t.

$\hat{y}_t$: Predicted spatial location on the reference ST t

λ: coefficient used to adjust the L1 penalty

W_{ijk}: The k-th value in the j-th row of the i-th weight matrix

I: Number of weight matrices

J_i: Number of rows in the i-th matrix

K_i: Number of columns of the i-th matrix.

To handle the batch effect, we used Wasserstein distance [67] was used to obtain genes with comparable distributions between them:

$$l_1(u, v) = \int_{-\infty}^{+\infty} |U - V| \quad (5)$$

, where

u: The expression values of a gene in the reference data

v: The expression values of a gene in the query data

U: The cumulative distribution function (CDF) of u

V: The CDF of v.

For the spatial inference of Visium data [31-33], genes with Wasserstein distances greater than 0.1 were selected for training and subsequent analysis.

Glorot normalization [68] is used to initialize the parameters of the neural network. Default learning rate 0.001 of Keras [69] is used. Dropout rate is set to 0.05 for the input and the first 5 hidden layers. A dropout layer is used to arbitrarily eliminate a certain number of input nodes during prediction. The batch size is set to 16 for the mouse olfactory bulb dataset, 128 for the other Visium data and 4096 for the other data. Number of epochs is set to 500. The L1 regularization scale factor is set to 0.00001. Learning rate reduction is configured with a minimal learning rate of 0.00001, reduction factor of 0.5 and patient of 20. Early stopping is configured with a patience parameter of 50. The patience parameter refers to the max number of epochs where the model doesn't achieve any improvement.

Benchmarking

We compared SC2Spa with LR, Novosparc [16, 20], SpaOTsc [17], and Tangram [18] in predicting the coordinates of voxels using 5-fold cross validation. Predicted and original coordinates of the ST spots in the leave-out set were used for calculating metrics. LR was performed using Scikit-learn 0.23.1. All parameters of LR were set with default values. Pearson's *r*, RMSE, R^2, pairwise Pearson's r and pairwise RMSE were used to assess the performance. Pairwise refers to all possible cell pairs of the leave-out spots/cells of the 5-fold cross validation tests. Pairwise Pearson's r and RMSE are the Pearson's r and

RMSE of the predicted and the reference distances of the cell pairs. The RMSE values were calculated using min-max normalized coordinates. R^2 is calculated as follow:

$$\bar{y} = \frac{1}{n}\sum_{t=0}^{N} y_t \tag{6}$$

$$SS_{res} = \sum_{t=0}^{N} (y_t - \hat{y}_t)^2 \tag{7}$$

$$SS_{tot} = \sum_{t=0}^{N} (y_i - \bar{y})^2 \tag{8}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{9}$$

, where SS_{res} is the residual sum of squares. SS_{tot} is the total sum of squares. R^2 was calculated using the "r2_score" function from the Scikit-learn [70] package.

NovoSpaRc [16, 20], SpaOTsc [17] and Tangram [18] were initially designed to find the optimal mapping probability matrix between single cells to spatial voxels. For the 5-fold cross-validation. The leave-out voxels were used as pseudo scRNAseq data. The remaining data served as the ST reference. The question was reformulated as searching for the optimal transport between the leave-out data and the training data. The spatial coordinates predictions of the three tools were obtained by summing up the probability weighted coordinates of training voxels.

When we apply SpaOTsc using all genes, SpaOTsc generated results similar to random mapping. Therefore, we selected the top 1,800 highly variable genes identified using SCANPY [71] for SpaOTsc for all tests except for the drosophila embryo FISH data [15]. For the drosophila embryo FISH data, we used all 84 genes. For the Stereo-seq mouse embryo data, due to the limitation of Random Access Memory (RAM) and the heavy computation cost of SpaOTsc, 55,000 cells out of 74342 cells were randomly selected and mapped to the 9300 cells in the leave-out set.

ST data can be represented with multiple vectors across multiple dimensions. Pearson's *r* is calculated for all vectors across the dimension and averaged. RMSE were calculated for all vectors across the dimension and summed. "multioutput" parameter of the "r2_score" function was set to "variance_weighted" to sum up the variance-weighted R^2 values of multiple dimensions [70]. The variance-weighted R^2 sum was used for benchmarking.

Fine mapping and ST bead reconstruction for spatial origin prediction for scRNAseq

SC2Spa predicts the spatial coordinates from the transcriptome. The predicted position may not correspond to the center of a bead and multiple single cells can be mapped to a bead. For visualization, the reconstructed expression levels of a bead are the weighted average of the scRNAseq mapped to the bead. The weight is determined based on the distance from the center of a bead to the predicted cell location. Only the cells within a certain distance to the center of a bead are selected for the ST bead expression reconstruction. We set the distance cutoff as 15μm for Slide-seqV2 and 110μm for Visium. The weight of a cell *j* for a bead is defined as:

$$\frac{1/d_j}{\sum_{i=1}^{k} 1/d_i} \tag{10}$$

, where

k: Number of single cells within a range. The range is the radius of the bead for the ST technology

d_i: Distance between cell *i* and the bead

d_j: Distance between cell *j* and the bead $(j = 1,2,\ldots k)$.

Quantify the consistency of the prediction with the reference ST data

To evaluate the quality of the prediction, we computed BVI [28, 29] for the gene expression levels between the reconstructed and the reference ST data. The BVI is defined as below:

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{i,j}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad (14)$$

, where

N: Number of cells

x_i: The expression value of the gene x in cell i of the original ST data

$\bar{x}$: Mean of gene x expression value in all cells of the reference ST data

y_i: The expression value of the gene x in cell j of the predicted ST data

$\bar{y}$: Mean of gene x expression value in all cells of the predicted ST data

$w_{i,j}$: Weight of a neighboring cell pair cell i and cell j

W: Sum of all w_{ij}.

The weight of cell pairs with a distance less than a specified threshold is set to $\frac{100}{\text{disij}}$, where "disij" is the distance between a cell i and a cell j. 100 is a scale factor. The weight of the remaining pairs of cells is assigned to zero.

Prioritizing location predictive genes

SC2Spa predicts genes that contribute to the location predictions from the trained weights. SC2Spa selects only the nodes with high weights (top 50% as the default) from the back of the trained FCNN until it reaches the input layer (**Figure. 1b**). Each node in the input layer corresponds to a gene. SC2Spa ranks the genes based on their contribution to the location prediction by summing the weights towards the nodes

in the input layer. Consequently, the expressions of the genes contributing location prediction are spatially variable.

An importance score is assigned to each gene by aggregating weight matrices of the neural network backward.

$$Mask(W_{i,:,j}) = \begin{cases} 1, if\ S_{i+1,j} > Percentile(S_{i+1}, r), \\ 0, else, \end{cases} \tag{11}$$

$$S_{i,j} = \begin{cases} \sum_{k=1}^{K_i} |W_{i,j,k}|, if\ i = 1, \\ \sum_{k=1}^{K_i} Mask(W_{i,j,k}) |W_{i,j,k}|, else, \end{cases} \tag{12}$$

, where

W_{ijk}: The *k*-th value in the *j*-th row of the *i*-th weight matrix

$W_{i,:,j}$: The *j*-th column of the *i*-th weight matrix

$S_{i,j}$:The sum of the selected values of the *j*-th row of the *i*-th weight matrix

$Percentile(S_{i+1}, r)$: The r-th percentile of Si+1. r was set as 50 in our study

I: Number of weight matrices, which is 1 more than the number of hidden layers

K_i: Number of columns of the *i*-th weight matrix.

The calculation was performed backwards from the last weight matrix to the first weight matrix. The output from the first weight matrix is S_1. S_1 assesses the significance of genes in determining the spatial location of beads. The importance score of the *j*-th gene S_{1j} will be normalized subsequently:

$$S_{1,jnorm} = \frac{S_{1,j}}{\sum_{i=1}^{n} S_{1,i}} . \bar{r}. Scaler \tag{13}$$

, where

$S_{1,jnorm}$: The normalized importance score of the j-th gene

$S_{1,i}$:The importance score of the *i*-th gene

$\bar{r}$: The average Pearson correlation coefficient between the prediction and reference spatial coordination

Scaler: A factor for scaling the importance score. Default 1000.

Prioritizing spatially variable genes.

Moran's I of each gene was calculated using the following equation:

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (14)$$

, where

N: Number of cells

x_i: The expression value of the gene x in cell i

$\bar{x}$: Mean of gene x expression value in all cells

$w_{i,j}$: Weight of a neighboring cell pair cell i and cell j

W: Sum of all w_{ij}.

Squidpy [36] was used to calculate Moran's I with the default settings.

Cell type annotation

The cell types of ST beads in the Slide-seq V2 datasets were annotated by RCTD [63]. The scRNAseq datasets were acquired from publicly available repositories. For the embryo, the single cells at E12.5 were extracted, which is the same as the developmental stage of the mouse embryo of Slide-seq data. The scRNAseq dataset has 37 cell types and consists of 26,183 genes and 270,197 cells. For the hippocampus, it has 17 cell types and consists of 27,953 genes and 113,507 cells. Based on the scRNAseq datasets, the cell types of the ST beads in the Slide-seq V2 datasets for mouse embryo and hippocampus were inferred, respectively. Furthermore, the cell type annotation of ST beads was validated by the marker genes of cell types accompanied with the scRNAseq datasets [30, 72].

Balanced mutual exclusivity

We defined balanced mutual exclusivity (BME) to identify mutually exclusive expression patterns between two genes. BME score is calculated as follow:

$$f(x1, x2) = \begin{cases} 1, if\ x1 < cutoff\ and\ x2 \geq cutoff, \\ 0, else, \end{cases} \tag{15}$$

$$e_1 = \sum_{i=1}^{n} f(G_1^i, G_2^i) \tag{16}$$

$$e_2 = \sum_{i=1}^{n} f(G_2^i, G_1^i) \tag{17}$$

$$P = \frac{(e_1 + e_2)^2}{2({e_1}^2 + {e_2}^2)} \tag{18}$$

$$BME = \frac{e_1 + e_2}{n} \times P^C \tag{19}$$

, where

G_1^i: Expression level of gene 1 in cell *i*

n: Total number of cells

e_1: Number of cells where gene 1 is expressed but gene 2 is not expressed

e_2: Number of cells where gene 2 is expressed but gene 1 is not expressed

P: A penalty term used to prevent a gene from contributing too much to the score.

C: A scaling factor for *P*'s influence. By default, 1.

The BME score ranges from 0 to 1. The confidence interval and standard error of BME were calculated using the bootstrapping algorithm [73] implemented by SciPy [74]. The number of resamples for bootstrapping was set to 200.

Isolation of hippocampal astrocytes and maintenance.

Briefly, 300 µm thick hippocampal slices (cut in the vibratome) were prepared from the brains of postnatal day 5-30 C57Bl/6J mice and maintained in a Ca^{2+}-free solution containing (in mm): 124 NaCl, 5 KCl, 2 $MgSO_4$, 1 Na-pyruvate, 26 $NaHCO_3$, and 10 glucose (bubbled with 95% O_2–5%CO_2 for 1 hr) as described previously [75]. The slices were then (4-5 from each mouse) were transferred into standard artificial CSF solution containing 24 U/ml papain and 0.24 mg/ml cysteine and bubbled with 95% O_2–5%CO_2 at 22°C for 30 min. The enzymatic digestion was halted by thorough washing and incubation of the slices in the Ca^{2+}-free solution for at least 1 hour before trituration [75]. After washing, the CA1 region of the hippocampus was dissected under the microscope and triturated in HEPES-buffer solution using wide-bore pipettes, yielding freshly isolated astrocytes (FIAs). The cell suspension containing the astrocytes were then identified by their characteristic "bushy" morphology, based on the fluorescence in the case of dissociation from hGFAP-EGFP or staining of astrocytes using SR101 as described previously [76]. FIAs were maintained in AstroMACS medium and incubated at 37 °C in a humidified atmosphere of 95% air and 5% CO2 and analyzed for *Camk2a* and *Plp2* expression post knockdown and overexpression studies.

Transfections and preparation of lentiviral particles for cDNA overexpression and shRNA knockdown

Knockdown and overexpression of *Plp1* and *Camk2a* was performed using lentiviral constructs (Origene). The lentiviral packaging was performed in 293FT cells using ViraSafe Lentivirus Packaging System (Pantropic Cell Biolabs), as per the manufacturer's instructions and as previously described [5]. The viral particles were concentrated by centrifuging the supernatant at 22,000 rpm. for 2 h at 20 °C. The pellet was re-suspended in 200 µl 1× HBSS, purified by layering the 200 µl of viral preparation on 1.5 ml of 20% sucrose (in HBSS) and centrifugation at 22,000 rpm for 2 h at 20 °C, and then the pellet was re-suspended in 100 µl 1× HBSS. The titre of lentiviral particles was determined using a p24 ELISA kit (Cell

Biolabs); 10^7 TU ml^{-1} was used for experiments. For infection of target cells, lentiviral particles were diluted in medium containing 5 μg ml^{-1} polybrene (Thermo Fisher Scientific).

RNA isolation and quantitative PCR (qPCR)

Total RNA from the FIAs was isolated using RNeasy micro kit (Qiagen). The RNA quantity was assessed using the NanoDrop 1000 and RNA integrity was assessed using an Agilent Bioanalyzer 2100. 10 ng of total RNA was reverse transcribed to cDNA using Superscript Vilo cDNA synthesis kit. Real-time PCR (qPCR) was performed with PowerUp SYBR Green master mix (Thermo Fisher Scientific) using Quant Studio 5 (ABI). The following primers were used for the study: *Camk2a* forward, 5′-AGCCATCCTCACCACTATGCTG-3′ and reverse, 5′- GTGTCTTCGTCCTCAATGGTGG-3′; *Plp1* forward, 5′- CCTAGCAAGACCTCTGCCAGTA-3′ and reverse, 5′-GGACAGAAGGTTGGAGCCACAA-3′; *GAPDH* forward, 5′- CATCACTGCCACCCAGAAGACTG-3′ and reverse, 5′- ATGCCAGTGAGCTTCCCGTTCAG-3′. All reactions were set up in triplicates. Relative mRNA quantification was obtained using the comparative Ct method (ΔΔCt), and *GAPDH* served as an internal control, as described previously [4, 5].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

SC2Spa is available at https://github.com/linbuliao/SC2Spa. The analysis for this manuscript is available at https://github.com/linbuliao/SC2Spa_Notebooks. The data and trained model for running SC2Spa is deposited at Figshare: https://doi.org/10.6084/m9.figshare.21829905.v7. All data and analysis involved in this manuscript is deposited at Zenodo: 10.5281/zenodo.8252714.

Competing interests

The authors declare no competing interests.

Funding

KJW is supported by the Lundbeck Foundation Ascending Investigator grant R313-2019-421 and the startup fund by Cedars-Sinai Medical Center.

Author contributions

K.J.W, R.G. and L.L. conceived the project. L.L. developed the software. L.L. performed data analysis. H.K., K.J.W. and C.H.B. helped with data analysis. E.M., AMP, AK, PB, RW, JT, PF, RG carried out experiments. R.G. and E.M. helped with the experimental details. All wrote the manuscript.

Acknowledgements

Not applicable

References

1. Browaeys R, Saelens W, Saeys Y: **NicheNet: modeling intercellular communication by linking ligands to target genes.** *Nat Methods* 2020, **17:**159-162.
2. Fischer DS, Schaar AC, Theis FJ: **Modeling intercellular communication in tissues using spatial graphs of cells.** *Nat Biotechnol* 2022.
3. Kim J, Rothová MM, Madan E, Rhee S, Weng G, Palma AM, Liao L, David E, Amit I, Hajkarim MC, et al: **Neighbor-specific gene expression revealed from physically interacting cells during mouse embryonic development.** *bioRxiv* 2022**:**2021.2012.2002.470916.
4. Madan E, Parker TM, Pelham CJ, Palma AM, Peixoto ML, Nagane M, Chandaria A, Tomas AR, Canas-Marques R, Henriques V, et al: **HIF-transcribed p53 chaperones HIF-1alpha.** *Nucleic Acids Res* 2019, **47:**10212-10234.
5. Madan E, Pelham CJ, Nagane M, Parker TM, Canas-Marques R, Fazio K, Shaik K, Yuan Y, Henriques V, Galzerano A, et al: **Flower isoforms promote competitive growth in cancer.** *Nature* 2019, **572:**260-264.
6. Martin PCN, Kim H, Lovkvist C, Hong BW, Won KJ: **Vesalius: high-resolution in silico anatomization of spatial transcriptomic data using image analysis.** *Mol Syst Biol* 2022, **18:**e11080.
7. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM: **Systematic determination of patterns of gene expression during Drosophila embryogenesis.** *Genome Biol* 2002, **3:**RESEARCH0088.
8. Almanzar N, Antony J, Baghel AS, Bakerman I, Bansal I, Barres BA, Beachy PA, Berdnik D, Bilen B, Brownfield D, et al: **A single-cell transcriptomic atlas characterizes ageing tissues in the mouse.** *Nature* 2020, **583:**590-+.
9. Han XP, Zhou ZM, Fei LJ, Sun HY, Wang RY, Chen Y, Chen HD, Wang JJ, Tang HN, Ge WH, et al: **Construction of a human cell landscape at single-cell level.** *Nature* 2020, **581:**303-+.
10. Ren X, Wen W, Fan X, Hou W, Su B, Cai P, Li J, Liu Y, Tang F, Zhang F, et al: **COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas.** *Cell* 2021, **184:**5838.
11. Winkler EA, Kim CN, Ross JM, Garcia JH, Gil E, Oh I, Chen LQ, Wu D, Catapano JS, Raygor K, et al: **A single-cell atlas of the normal and malformed human brain vasculature.** *Science* 2022, **375:**eabi7377.
12. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R: **CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.** *Nat Protoc* 2020, **15:**1484-1506.
13. Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC: **High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.** *Nat Biotechnol* 2015, **33:**503-509.
14. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of single-cell gene expression data.** *Nature Biotechnology* 2015, **33:**495-U206.

15. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, Zinzen RP: **The Drosophila embryo at single-cell transcriptome resolution.** *Science* 2017, **358:**194-199.
16. Nitzan M, Karaiskos N, Friedman N, Rajewsky N: **Gene expression cartography.** *Nature* 2019, **576:**132-137.
17. Cang Z, Nie Q: **Inferring spatial and signaling relationships between cells from single cell transcriptomic data.** *Nat Commun* 2020, **11:**2084.
18. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe A, Zhang M, et al: **Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram.** *Nat Methods* 2021, **18:**1352-1362.
19. Maseda F, Cang Z, Nie Q: **DEEPsc: A Deep Learning-Based Map Connecting Single-Cell Transcriptomics and Spatial Imaging Data.** *Front Genet* 2021, **12:**636743.
20. Moriel N, Senel E, Friedman N, Rajewsky N, Karaiskos N, Nitzan M: **NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport.** *Nat Protoc* 2021, **16:**4177-4200.
21. Bageritz J, Willnow P, Valentini E, Leible S, Boutros M, Teleman AA: **Gene expression atlas of a developing tissue by single cell expression correlation analysis.** *Nature Methods* 2019, **16:**750-+.
22. Sarker IH: **Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions.** *SN Comput Sci* 2021, **2:**420.
23. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, Chen F: **Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2.** *Nature Biotechnology* 2020.
24. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al: **Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.** *Science* 2016, **353:**78-82.
25. Kingma DP, Ba J: **Adam: A Method for Stochastic Optimization.** 2015.
26. DeMaris A: **A Tutorial in Logistic Regression.** *Journal of Marriage and Family* 1995, **57:**956-968.
27. Chen A, Liao S, Cheng MN, Ma KL, Wu L, Lai YW, Qiu XJ, Yang J, Xu JS, Hao SJ, et al: **Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays.** *Cell* 2022, **185:**1777-+.
28. Li Z, Wang T, Liu P, Huang Y: **SpatialDM: Rapid identification of spatially co-expressed ligand-receptor reveals cell-cell communication patterns.** *bioRxiv* 2023**:**2022.2008.2019.504616.
29. Wartenberg D: **Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis.** *Geographical Analysis* 1985, **17:**263-283.
30. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang SY, et al: **Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain.** *Cell* 2018, **174:**1015-+.
31. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Elmentaite R, Lomakin A, Kedlian V, Gayoso A, et al: **Cell2location maps fine-grained cell types in spatial transcriptomics.** *Nat Biotechnol* 2022, **40:**661-671.

32. Li R, Ferdinand JR, Loudon KW, Bowyer GS, Laidlaw S, Muyas F, Mamanova L, Neves JB, Bolt L, Fasouli ES, et al: **Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer.** *Cancer Cell* 2022, **40:**1583-1599 e1510.
33. Janesick A, Shelansky R, Gottscho AD, Wagner F, Rouault M, Beliakoff G, Oliveira MFd, Kohlway A, Abousoud J, Morrison CA, et al: **High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue.** *bioRxiv* 2022**:**2022.2010.2006.510405.
34. Svensson V, Teichmann SA, Stegle O: **SpatialDE: identification of spatially variable genes.** *Nat Methods* 2018, **15:**343-346.
35. Moran PA: **Notes on continuous stochastic phenomena.** *Biometrika* 1950, **37:**17-23.
36. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, et al: **Squidpy: a scalable framework for spatial omics analysis.** *Nat Methods* 2022, **19:**171-178.
37. Lisman J, Schulman H, Cline H: **The molecular basis of CaMKII function in synaptic and behavioural memory.** *Nat Rev Neurosci* 2002, **3:**175-190.
38. Zhang H, Webb DJ, Asmussen H, Niu S, Horwitz AF: **A GIT1/PIX/Rac/PAK signaling module regulates spine morphogenesis and synapse formation through MLC.** *J Neurosci* 2005, **25:**3379-3388.
39. Lucchesi W, Mizuno K, Giese KP: **Novel insights into CaMKII function and regulation during memory formation.** *Brain Res Bull* 2011, **85:**2-8.
40. Xing G, Russell S, Hough C, O'Grady J, Zhang L, Yang S, Zhang LX, Post R: **Decreased prefrontal CaMKII alpha mRNA in bipolar illness.** *Neuroreport* 2002, **13:**501-505.
41. Amada N, Aihara K, Ravid R, Horie M: **Reduction of NR1 and phosphorylated Ca2+/calmodulin-dependent protein kinase II levels in Alzheimer's disease.** *Neuroreport* 2005, **16:**1809-1813.
42. Novak G, Seeman P, Tallerico T: **Increased expression of calcium/calmodulin-dependent protein kinase IIbeta in frontal cortex in schizophrenia and depression.** *Synapse* 2006, **59:**61-68.
43. Inoue K: **PLP1-related inherited dysmyelinating disorders: Pelizaeus-Merzbacher disease and spastic paraplegia type 2.** *Neurogenetics* 2005, **6:**1-16.
44. Fuchsova B, Alvarez Julia A, Rizavi HS, Frasch AC, Pandey GN: **Altered expression of neuroplasticity-related genes in the brain of depressed suicides.** *Neuroscience* 2015, **299:**1-17.
45. Rao A, Barkley D, Franca GS, Yanai I: **Exploring tissue architecture using spatial transcriptomics.** *Nature* 2021, **596:**211-220.
46. Anderson AC, Yanai I, Yates LR, Wang L, Swarbrick A, Sorger P, Santagata S, Fridman WH, Gao Q, Jerby L, et al: **Spatial transcriptomics.** *Cancer Cell* 2022, **40:**895-900.
47. Moses L, Pachter L: **Museum of spatial transcriptomics.** *Nat Methods* 2022, **19:**534-546.
48. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A: **An introduction to spatial transcriptomics for biomedical research.** *Genome Med* 2022, **14:**68.
49. Zhang L, Chen D, Song D, Liu X, Zhang Y, Xu X, Wang X: **Clinical and translational values of spatial transcriptomics.** *Signal Transduct Target Ther* 2022, **7:**111.

50. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan G-C, Cai L: **Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+.** *Nature* 2019, **568:**235-239.
51. Moffitt JR, Zhuang X: **RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH).** *Methods Enzymol* 2016, **572:**1-49.
52. Borm LE, Mossi Albiach A, Mannens CCA, Janusauskas J, Ozgun C, Fernandez-Garcia D, Hodge R, Castillo F, Hedin CRH, Villablanca EJ, et al: **Scalable in situ single-cell profiling by electrophoretic capture of mRNA using EEL FISH.** *Nat Biotechnol* 2022.
53. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ: **Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution.** *Science* 2019, **363:**1463-1467.
54. Vickovic S, Eraslan G, Salmen F, Klughammer J, Stenbeck L, Schapiro D, Aijo T, Bonneau R, Bergenstrahle L, Navarro JF, et al: **High-definition spatial transcriptomics for in situ tissue profiling.** *Nat Methods* 2019, **16:**987-990.
55. Cho CS, Xi J, Si Y, Park SR, Hsu JE, Kim M, Jun G, Kang HM, Lee JH: **Microscopic examination of spatial transcriptome using Seq-Scope.** *Cell* 2021, **184:**3559-3572 e3522.
56. Fu X, Sun L, Chen JY, Dong R, Lin Y, Palmiter RD, Lin S, Gu L: **Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture Efficiency.** *bioRxiv* 2021**:**2021.2003.2017.435795.
57. Xia K, Sun HX, Li J, Li J, Zhao Y, Chen L, Qin C, Chen R, Chen Z, Liu G, et al: **The single-cell stereo-seq reveals region-specific cell subtypes and transcriptome profiling in Arabidopsis leaves.** *Dev Cell* 2022, **57:**1299-1310 e1294.
58. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q: **Inference and analysis of cell-cell communication using CellChat.** *Nat Commun* 2021, **12:**1088.
59. Shao X, Li C, Yang H, Lu X, Liao J, Qian J, Wang K, Cheng J, Yang P, Chen H, et al: **Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk.** *Nat Commun* 2022, **13:**4429.
60. Andersson A, Bergenstrahle J, Asp M, Bergenstrahle L, Jurek A, Fernandez Navarro J, Lundeberg J: **Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography.** *Commun Biol* 2020, **3:**565.
61. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Lomakin A, Kedlian V, Jain MS, Park JS, et al: **Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics.** *bioRxiv* 2020**:**2020.2011.2015.378125.
62. Brbić M, Cao K, Hickey JW, Tan Y, Snyder MP, Nolan GP, Leskovec J: **Annotation of Spatially Resolved Single-cell Data with STELLAR.** *bioRxiv* 2021**:**2021.2011.2024.469947.
63. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, Irizarry RA: **Robust decomposition of cell type mixtures in spatial transcriptomics.** *Nat Biotechnol* 2021.
64. Dong R, Yuan GC: **SpatialDWLS: accurate deconvolution of spatial transcriptomic data.** *Genome Biol* 2021, **22:**145.

65. Song Q, Su J: **DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence.** *Brief Bioinform* 2021, **22**.

66. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Elmentaite R, Lomakin A, Kedlian V, Gayoso A, et al: **Cell2location maps fine-grained cell types in spatial transcriptomics.** *Nat Biotechnol* 2022.

67. Vallender SS: **Calculation of the Wasserstein Distance Between Probability Distributions on the Line.** *Theory of Probability & Its Applications* 1974, **18:**784-786.

68. Glorot X, Bengio Y: **Understanding the difficulty of training deep feedforward neural networks.** In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Yee Whye T, Mike T eds.), vol. 9. pp. 249--256. Proceedings of Machine Learning Research: PMLR; 2010:249--256.

69. Chollet F, others: **Keras.** GitHub; 2015.

70. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al: **Scikit-learn: Machine Learning in Python.** *Journal of Machine Learning Research* 2011, **12:**2825-2830.

71. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis.** *Genome Biol* 2018, **19:**15.

72. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, et al: **The single-cell transcriptional landscape of mammalian organogenesis.** *Nature* 2019, **566:**496-502.

73. Efron B, Tibshirani R, Taylor, Francis: *An introduction to the bootstrap.* Boca Raton, Florida: Chapman & Hall/CRC; 1994.

74. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al: **SciPy 1.0: fundamental algorithms for scientific computing in Python.** *Nat Methods* 2020, **17:**261-272.

75. Zhou M, Kimelberg HK: **Freshly isolated hippocampal CA1 astrocytes comprise two populations differing in glutamate transporter and AMPA receptor expression.** *J Neurosci* 2001, **21:**7901-7908.

76. Kafitz KW, Meier SD, Stephan J, Rose CR: **Developmental profile and properties of sulforhodamine 101--Labeled glial cells in acute brain slices of rat hippocampus.** *J Neurosci Methods* 2008, **169:**84-92.

Figure Legends

Figure Legend-1. Overview of SC2Spa.

a. SC2Spa is composed of a FCNN to map transcriptome to the spatial location. For training, SC2Spa uses the transcriptome of ST beads as the input and the spatial location of them on the ST assay as the output. The trained FCNN is used to predict spatial location for the transcriptome from either ST beads or scRNAseq data. The predicted location is further visualized on the ST assay.

b. The weights of the FCNN are used to identify SVGs

c. SC2Spa scores the spatial exclusiveness between two genes and predicts potential negative gene regulatory rules.

Figure Legend-2. SC2Spa outperformed other predictors in predicting spatial origin from transcriptomics data for cross-validation tests.

a. The reference Slide-seqV2 data and the predicted results were visualized for eight cell types in mouse brain. One of the five-fold cross-validation results is shown.

b. Benchmarking results for the four ST datasets. Performance was compared for a mouse hippocampus Slide-seqV2 dataset, a mouse embryo Slide-seqV2 dataset, a drosophila embryo 3D FISH dataset and a mouse embryo Stereo-seq dataset. Pearson's *r*, RMSE, R^2, Pairwise RMSE and Pairwise Pearson's *r* were calculated for algorithms we tested.

Figure Legend-3. The trained SC2Spa recovers the locations for unseen ST data.

a. The locations of dentate gyrus, CA1, CA2 and CA3 pyramidal cells in the training ST data (puck_200115_08) . **b.** in the independent query set (puck_191204_01) and **c.** prediction **d.** The spatial expression of *Hpca* and *Pcp4* in the training, and the testing sets as well as the predicted results. BVI is

calculated for *Hpca* (0.35) and *Pcp4* (0.09) between the training set and the prediction. The color bar is to show the expression levels for the genes.

Figure Legend-4. SC2Spa predicted the locations for scRNAseq data.

a. The predicted locations for CA1, CA2, CA3 pyramidal cells and dentate principal cells from scRNAseq. The histology image is from [66] and is manually labeled.

b. Spatial expression patterns of *Nrgn*, *Fibcd1*, *Tshz2* and *Hs3st4* show the similarity between the reference and the predicted results. BVIs were calculated for these genes between the reference ST and the prediction by SC2Spa

Figure Legend-5. SC2Spa trained on Visium data predicted the locations for scRNAseq.

a. The spatial expression profiles of *Mbp*, *Gad2*, and *Slc17a6* on the Visium data.

b. The predicted localization of oligodendrocytes, inhibitory neurons and excitatory neurons in thalamus from the scRNAseq from the same mouse brain.

c. Spatial reconstruction using the scRNAseq data from an independent mouse brain. **d.** The expression of *Rgs9* (BVI=0.2), *Eps8l2* (BVI=0.16), *Stard8* (BVI=0.25), and *Otof* (BVI=0.14) in the reference data and our predicted results.

Figure Legend-6. BME identified mutually exclusive gene expression using spatial transcriptomic data from mouse hippocampus.

a. *Plp1* and *Camk2a* showed exclusive spatial distribution (BME=0.495). The expression of *Plp1* and *Camk2a* is shown with red and blue color, respectively.

b. The scatter density plots of *Plp1* and *Camk2a* in scRNAseq for neuroepithelial cells, neurons and astrocytes, respectively. The color indicates the Gaussian density estimation of cells.

c. The analysis of *Camk2a* expression through RT-qPCR shows that *Plp1* knockdown increases the expression of Camk2a by 6-fold change (p=1.07 e^{-3}) compared to wild-type murine hippocampal astrocytes. Plp1 overexpression decreases by 2-fold the expression of this gene (p = 3.44 e^{-5}) when compared to wild-type astrocytes and a 12-fold decrease when compared to murine hippocampal astrocytes with *Plp1* knockdown (p 8.52 e^{-4}) (N=3). We performed two-tailed Student t-test to determined significant differences in Camk2a expression.

d. The analysis of *Plp1* expression through RT-qPCR shows that *Camk2a* knockdown increases the expression of *Plp1* by 7-fold change (p=2.01 e^{-3}) compared to wild-type murine hippocampal astrocytes. *Camk2a* overexpression decreases nearly by 2-fold the expression of this gene (p =1.45 e^{-5}) when compared to wild-type astrocytes and a 13-fold decrease when compared to murine hippocampal astrocytes with *Camk2a* knockdown (p=8.52 e^{-4}) (N=3). We performed two-tailed Student t-test to determined significant differences in *Camk2a* expression.

Supplementary Figure and Table Legends

Supplementary Figure Legend-1. The reference image data and the predicted location for SC2Spa and other algorithms using a SlideSeq-V2 mouse embryo.

The reference image and the predicted location for SC2Spa, Tangram, Novosparc, SpaOTsc and LR were visualized for the leave-out beads for 8 cell types including myocytes, stromal cells, isthmic organizer cells, epithelial cells, limb mesenchymal call, jaw and tooth progenitors, melanocytes and primitive erythroid lineage cells.

Supplementary Figure Legend-2. Euclidean distance between the predicted and the original location for Slide-seqV2 mouse hippocampus data.

The color bar represents Euclidian distance. The color represents Euclidean distance.

Supplementary Figure Legend-3. Euclidean distance between the predicted and the original location for Slide-seqV2 mouse embryo dat.

The color represents Euclidian distance.

Supplementary Figure Legend-4. The reference image data and the predicted location of SC2Spa and other algorithms for a Stereo-seq mouse embryo dataset.

The reference image and the predicted location for SC2Spa, Tangram, Novosparc, SpaOTsc and LR were visualized for the leave-out single cells of 8 cell types including cavity, connective tissue, muscle, liver, smooth muscle, cartilage primordium, meninges and brain.

Supplementary Figure Legend-5. Euclidean distance between the predicted and the original location for Stereo-seq data from mouse embryo.

The color represents Euclidian distance.

Supplementary Figure Legend-6. The original and predicted locations of the leave-out beads from one round of cross-validation for the 3D FISH drosophila embryo data.

The original and predicted locations of ST beads in eight clusters were represented graphically. The drosophila embryo beads retained three-dimensional space information. The beads were viewed from three angles: lateral, anterior and posterior, as well as ventral and dorsal.

Supplementary Figure Legend-7. Euclidean distance between the predicted and the original location for the 3D FISH drosophila embryo data.

The color represents Euclidian distance.

Supplementary Figure Legend-8. Performance comparison when testing with an independent Slide-seqV2 mouse brain dataset

a, The pairwise RMSE and the pairwise Pearson's *r* for SC2Spa, Tangram, NovoSpaRc, SpaOTsc and LR between reference Slide-seqV2 and an independent Query data

b, The Query SlideSeqV2 data and the predicted results.

Supplementary Figure Legend-9. The locations of various cell types in the reference Slide-seqV2 mouse hippocampus dataset (puck_200115_08).

Supplementary Figure Legend-10. The locations of clusters of the query Slide-seqV2 mouse hippocampus dataset (puck_191204_01). We obtained 26 clusters. The beads for each cluster are visualized in the query Slide-SeqV2 data.

Supplementary Figure Legend-11. Predicted locations of the spots from the query Slide-seqV2 data by SC2Spa(puck_191204_01). The prediction is shown for the beads for each cluster. The location shown in the clustering results (Supplementary Figure 10) largely match each other. For instance, cluster 14, 16 and 11 of the prediction matched well with the dentate, CA3 and CA1 principal cells of the reference data.

Supplementary Figure Legend-12. Performance comparison when testing with an independent Visium mouse brain dataset.

a, The pairwise RMSE and the pairwise Pearson's *r* between the reference Visium data and the predicted results for SC2Spa, Tangram, NovoSpaRc, SpaOTsc and LR.

b, The Visium data used for Query and the predicted results. From the Qurery we obtained the clustering information. The beads for 8 clusters were visualized.

Supplementary Figure Legend-13. The spatial expression profiles of *Nrgn*, *Prkcd* and *Slc17a6* in the reference, query and predicted Visium data.

Supplementary Figure Legend-14. SC2Spa recovered the spatial locations of single cells

SC2Spa was trained using the SlideSeqV2 mouse brain data. The locations of 24 cell types were predicted: CA1, CA2, CA3 principal cell, dentate principal cell, lateral CA3 principal cell, dentate hilum, astrocyte, subiculum, oligodendrocyte, neuron, interneuron, endothelial tip, anterior subiculum, neurogenesis, polydendrocyte, ependymal, endothelial stalk, myelin, deep layer subiculum, resident macrophage, mural, microglia, medial entorhinal cortex and post subiculum.

Supplementary Figure Legend-15. Predicted locations of the single cells of 36 cell types by SC2Spa for mapping a snRNAseq dataset to a Visium mouse brain dataset.

Supplementary Figure Legend-16. Predicted locations of the single cells of the remaining 35 cell types by SC2Spa for mapping a snRNAseq dataset to a Visium mouse brain dataset.

Supplementary Figure Legend-17. SC2Spa reconstructs the spatial locations of scRNAseq.

SC2Spa is trained using the Visium dataset for human kidney cancer. The spatial expression for *DMKN, EGR1, KCNJ1, MT1H, RASD1, RPL37* is visualized in the reference ST data and the prediction.

Supplementary Figure Legend-18. SC2Spa reconstructs spatial location of scFFPE-seq.

SC2Spa is trained using the Visium dataset for human breast cancer. The prediction results using 3’ and 5’ scFFPE-seq are shown for *CYB561, FASN* and *PVALB*. BVI is calculated between the reference ST and the prediction. The BVI between the two predictions (for 3’ and 5’ scFFPE-seq) is also shown to show the consistency.

Supplementary Figure Legend-19. The rank plot of the BVI values between the predictions from 3’ and 5’ scFFPE-seq human breast cancer data.

The BVI values between the predictions based on the 3' and 5' scFFPEseq displays a high difference from the BVI values of randomized predictions. More than 10,000 genes showed high BVI values than randomly shuffled genes.

Supplementary Figure Legend-20. Assessing SVGs obtained by SC2Spa and SpatialDE.

a, c, The scatter plot for SC2Spa scores (contribution to location prediction) versus SpatialDE LLR values (PCC=0.562). The color gradient represents the percentage of cells where a gene is expressed. **b.** The SVGs commonly observed by SpatialDE and SC2Spa. **c,** The three genes with high SpatialDE LLR values but low SC2Spa scores. **d.** The PCC for SC2Spa and SpatialDE with Moran's I for three Slide-seqV2 datasets.

Supplementary Figure Legend-21. SC2Spa identified LPGs that SpatialDE and Moran's I failed to emphasize on the second mouse hippocampus Slide-seqV2 dataset (Puck_191204_01).

a, A scatter plot of SC2Spa score (contribution to location prediction) versus SpatialDE LLR value for all genes is shown. The color represents the proportion of cells that express a particular gene. *Tert* (I=-5.89e-5), *Gm19684* (I=-1.16e-4) and *Gm10640* (I=-3.15e-4) have high SpatialDE LLR scores but low SC2Spa scores. **b,** The top four genes identified by SC2Spa.

Supplementary Figure Legend-22. SC2Spa identified LPGs that SpatialDE and Moran's I failed to emphasize on a mouse embryo Slide-seqV2 dataset.

a, A scatter plot of SC2Spa score (contribution to location prediction) versus SpatialDE LLR value for all genes is presented. The hue indicates the proportion of cells that express a gene. *Krt75* (I=-6.90e-5), *Tmem171* (I=-6.94e-5) and *Has1* (I=-1.01e-4) have high SpatialDE LLR scores but low SC2Spa scores. **b,** The top four genes identified by SC2Spa.

Supplementary Figure Legend-23. SpatialDE identified seven clusters based on spatial gene expression patterns in a mouse hippocampus dataset.

Supplementary Figure Legend-24. Workflow to calculate BME score.

Supplementary Figure Legend-25. Scatter density plots of Plp1 and Camk2a from six cell types from adult mouse hippocampus.

The six cell types are neurogenesis cell, ependymal cell, mural cell, endothelial tip cell, microglia and resident macrophage. The color indicates the Gaussian density estimation of cells.

Supplementary Figure Legend-26. The spatial expression of Hbegf in interneurons and ErbB4 in Astrocyte. The locations of interneurons and astrocytes were predicted using SC2Spa. The cell-cell communication was predicted using SpaTalk [59] based on the results by SC2Spa.

Supplementary Figure Legend-27. Downstream genes and related pathways of Hbegf-ErbB4 interaction.

Supplementary Figure Legend-28. RMSE histories of training and test data of the 5-fold cross-validation of Stereo-seq mouse embryo.

Supplementary Figure Legend-29. RMSE histories of training and test data of the 5-fold cross-validation of Slide-seqV2 mouse hippocampus.

Supplementary Figure Legend-30. RMSE histories of training and test data of the 5-fold cross-validation of Slide-seqV2 mouse embryo.

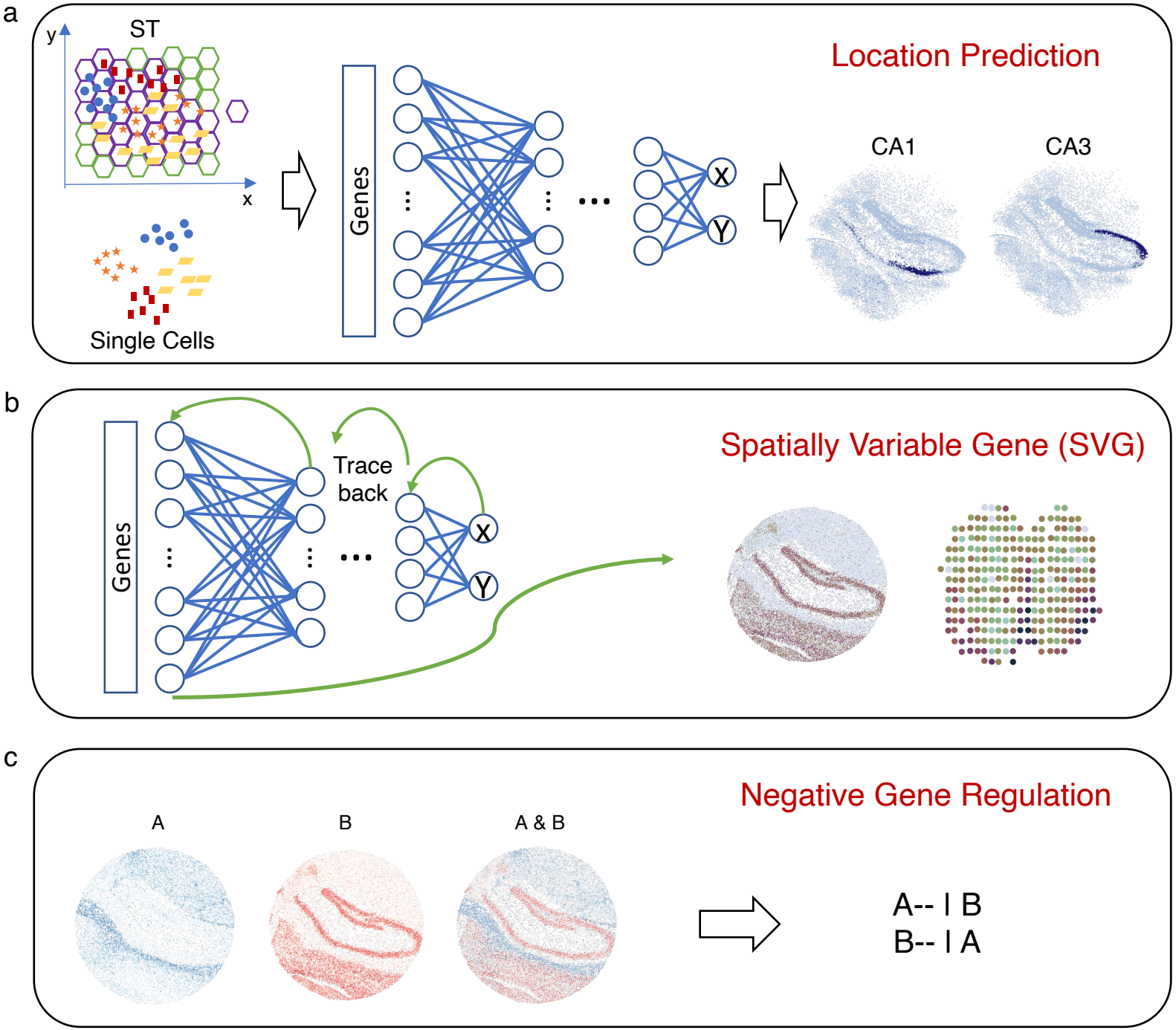
Supplementary Figure Legend-31. RMSE histories of training and test data of the 5-fold cross-validation of 3D FISH drosophila embryo.

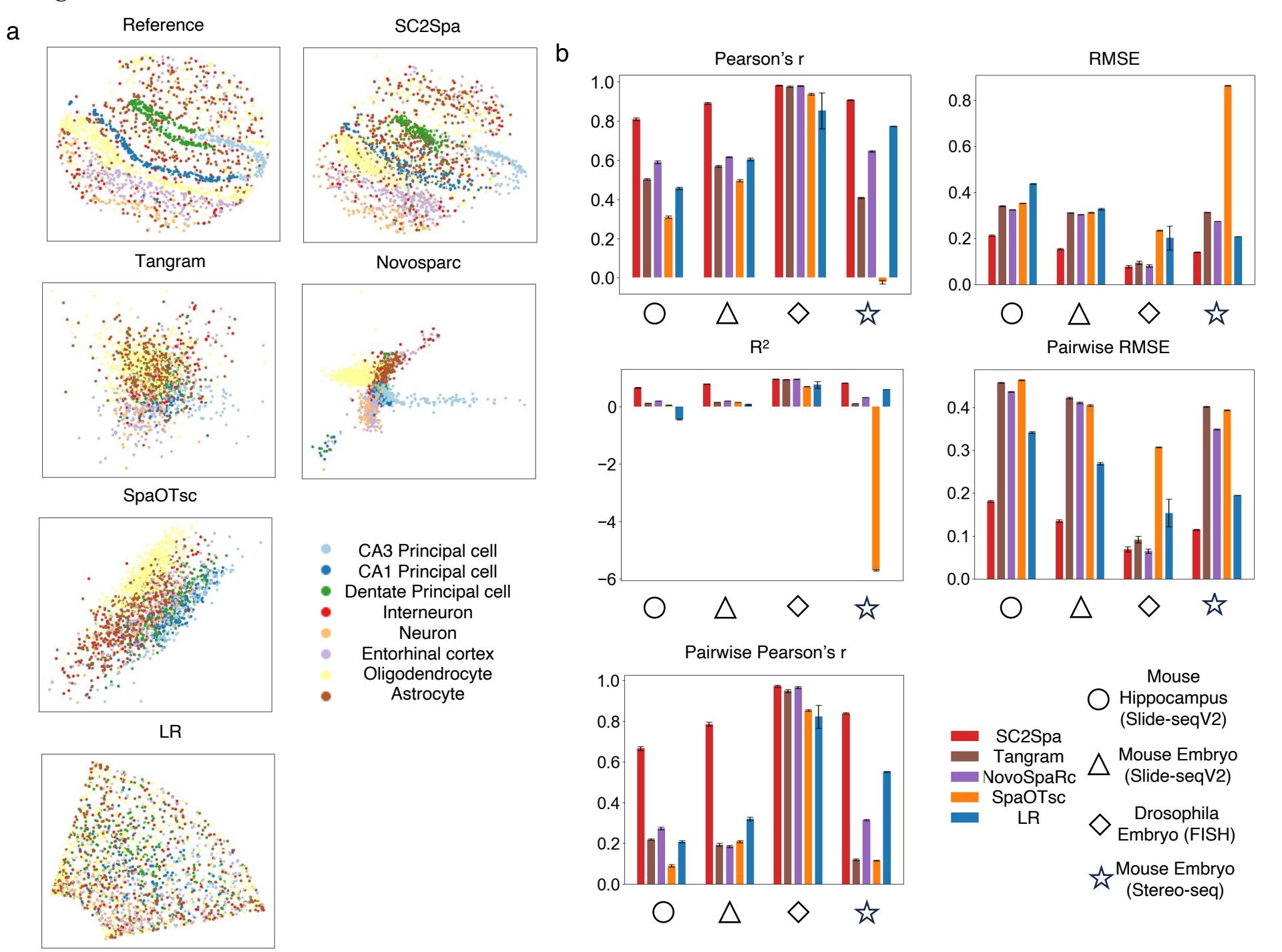
Supplementary Table Legend-1. Mutually exclusive gene pairs (BME>0.45) identified in a Slide-seqV2 dataset of mouse hippocampus.

Supplementary Table Legend-2. Single-cell and ST data used in this study.

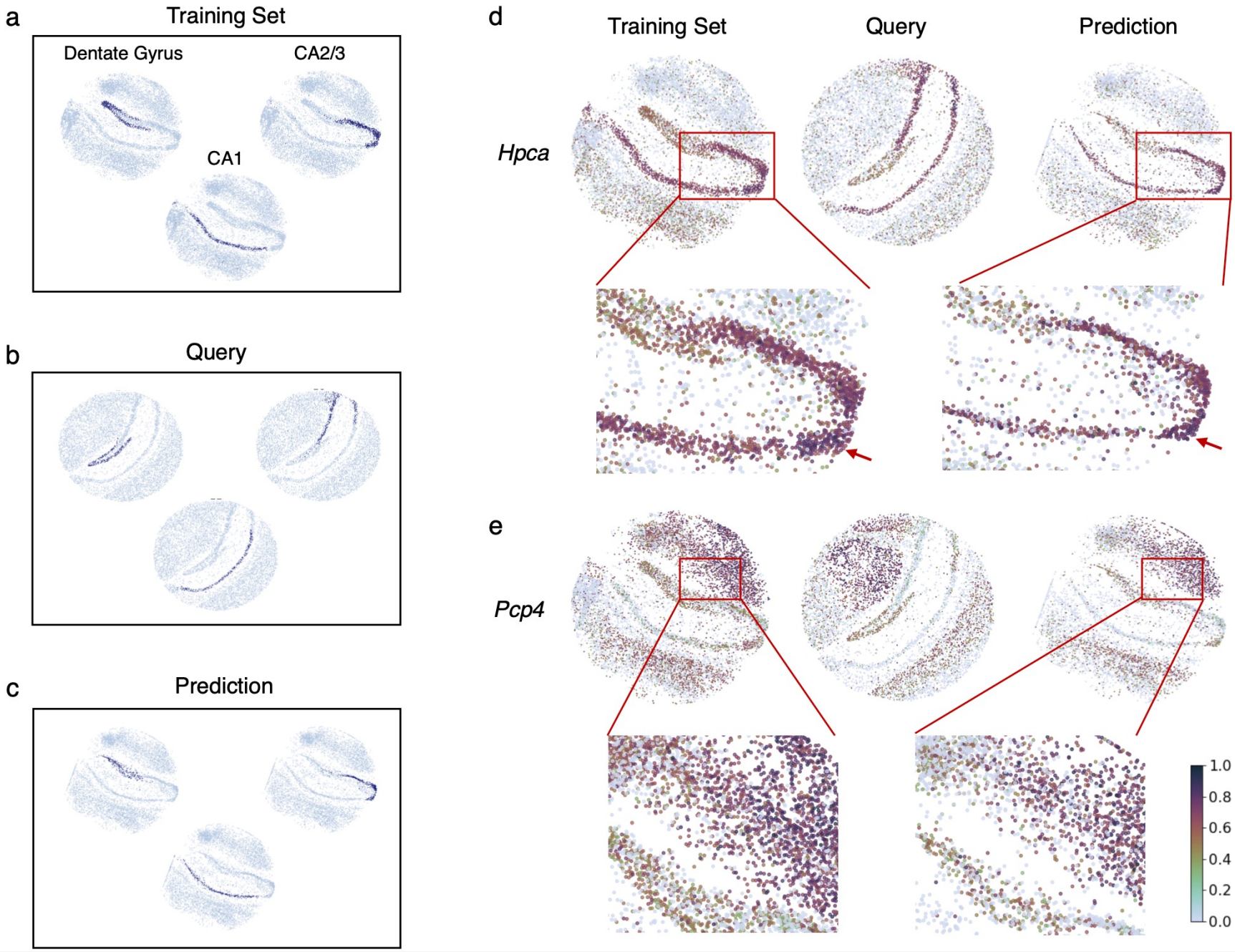
Supplementary Table Legend-3. Comparison of running time

Liao et. al.
Figure 1

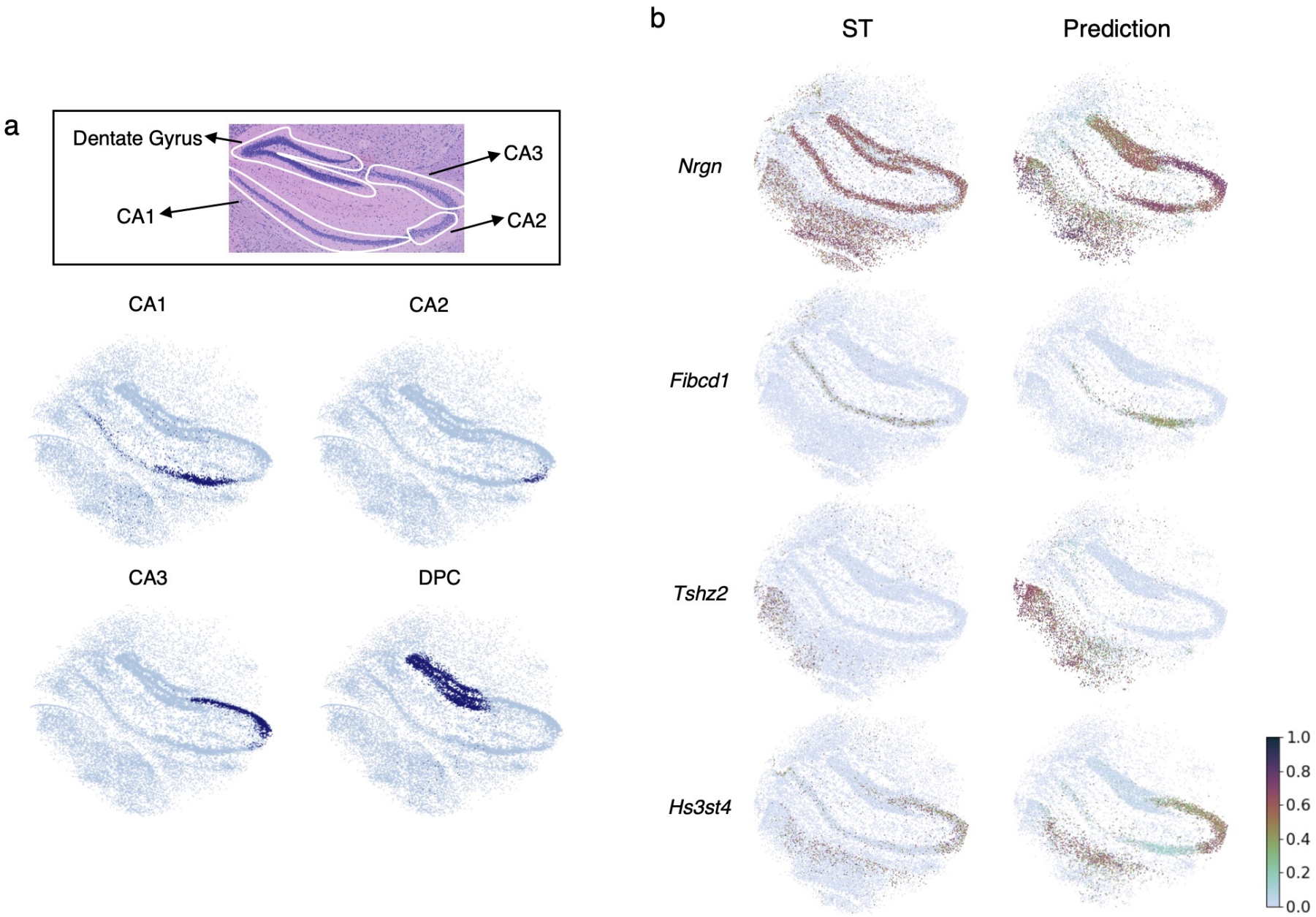



a
Reference
SC2Spa
Tangram
Novosparc
SpaOTsc
LR
CA3 Principal cell
CA1 Principal cell
Dentate Principal cell
Interneuron
Neuron
Entorhinal cortex
Oligodendrocyte
Astrocyte
b
Pearson's r
RMSE
R²
Pairwise RMSE
Pairwise Pearson's r
SC2Spa
Tangram
NovoSpaRc
SpaOTsc
LR
Mouse Hippocampus (Slide-seqV2)
Mouse Embryo (Slide-seqV2)
Drosophila Embryo (FISH)
Mouse Embryo (Stereo-seq)

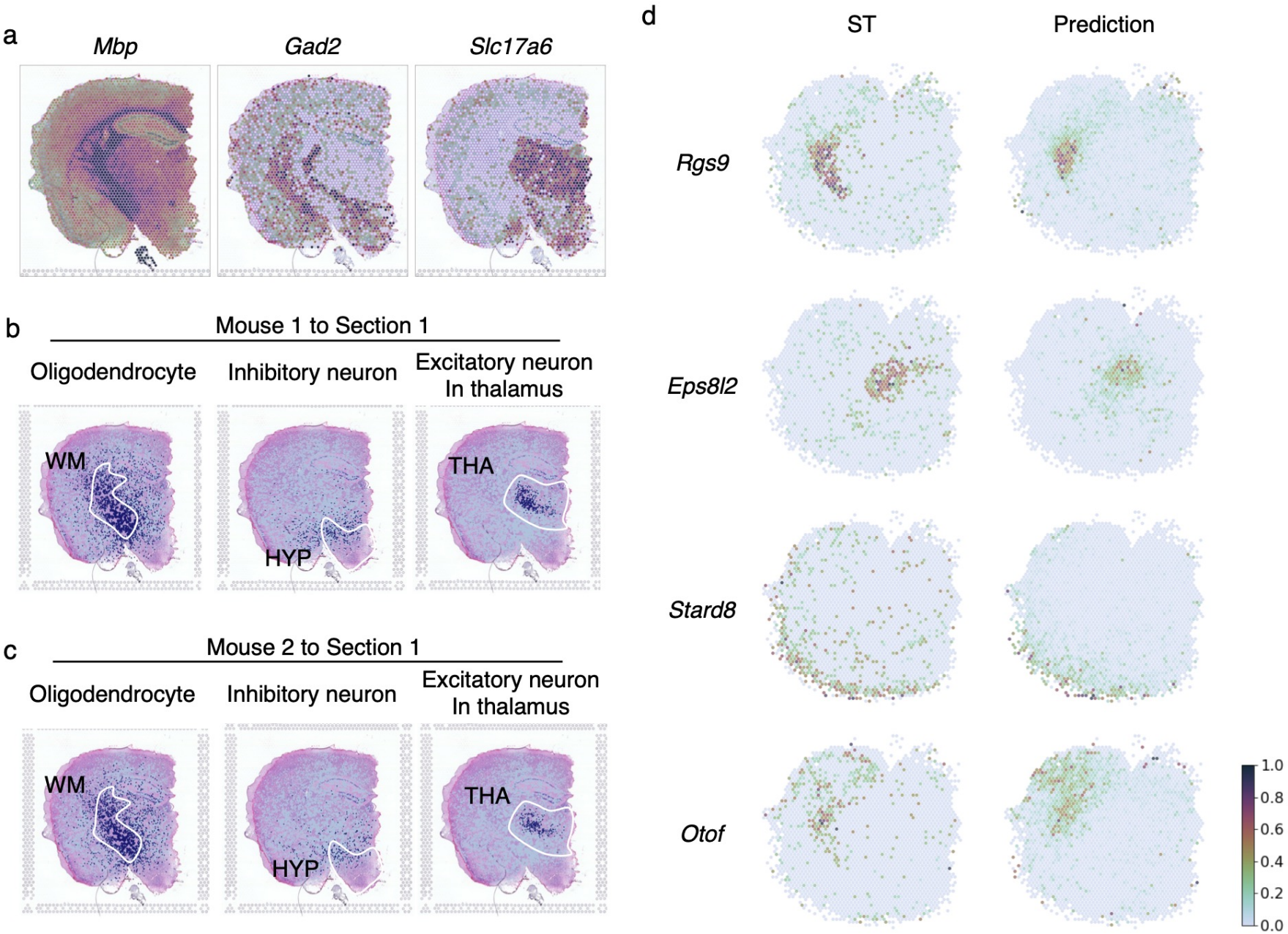
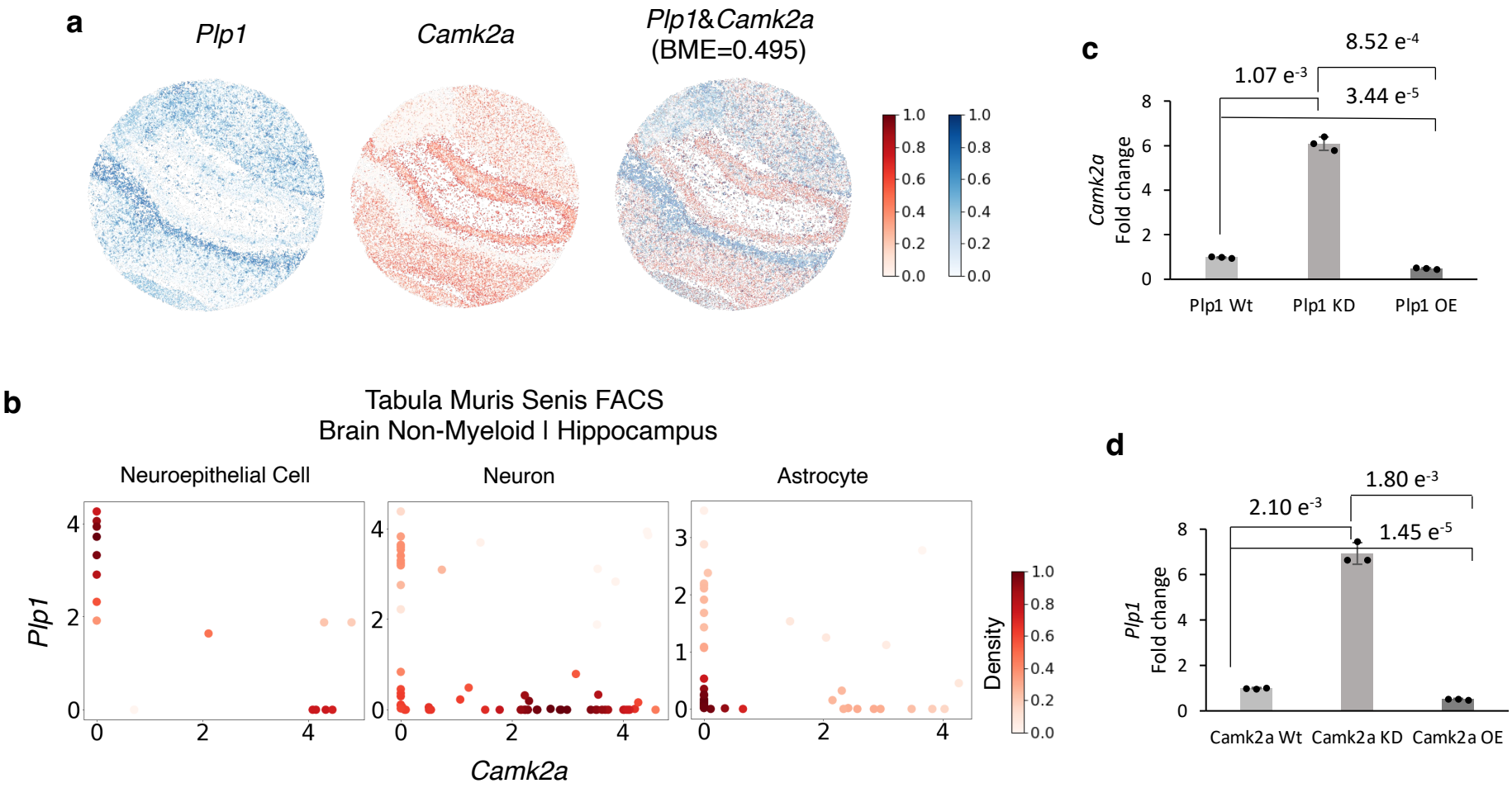
Liao et. al. Figure 3



Liao et. al. Figure 4



Liao et. al. Figure 5



Liao et. al. Figure 6