Mapping lineage-resolved scRNA-seq data with spatial transcriptomics using TemSOMap

Xinhai Pan[1], Alejandro Danies-Lopez[1], and Xiuwei Zhang[1,*]

1 Georgia Institute of Technology, Atlanta GA 30332, USA

Abstract. Spatial transcriptomics (ST) has become a powerful technique that advances the study of cell spatial organization and cell-cell interactions. While ST can preserve location information of cells or spots, limitations of such technologies include lower number of genes, and lower resolution compared to scRNA-seq datasets. These limitations can be alleviated by integrating scRNA-seq data with the ST data. By mapping the single cells onto the spatial data, we can infer the spatial coordinates of the cells from the scRNA-seq dataset. We consider leveraging temporal information in this challenging task of spatial location inference. During tissue formation, cells divided from the same ancestor are likely to be located close to each other in the tissue, thus the cell clonal or lineage information can improve cell location inference. CRISPR/Cas9-based lineage tracing technologies have enabled paired sequencing of cells' gene expression and lineage barcodes. The lineage barcodes can be used to reconstruct the cell lineage tree, which represents cells' clonal relationships. In order to incorporate this information, we developed TemSOMap (**Tem**poral dynamics guided **S**patial **O**mics **Map**ping), which infers the spatial coordinates of cells by mapping a paired gene expression and lineage barcode dataset onto a spatial transcriptomics dataset. TemSOMap utilizes a machine learning framework to infer a cell-to-spot mapping matrix by minimizing a loss function based on expression and lineage. We show that TemSOMap more accurately infers the spatial location of single cells compared to state-of-the-art baseline methods under various scenarios, using both simulated and real datasets. The resulting lineage-resolved ST data can help us better understand the spatio-temporal dynamics of cells in a tissue. TemSOMap is publicly available at `https://github.com/ZhangLabGT/TemSOMap`.

Keywords: Spatial transcriptomics · single cell lineage tracing · machine learning · spatio-temporal dynamics.

1 Introduction

Spatial Transcriptomics (ST) has become one of the most popular technologies in single-cell multi-omics. This technology has enabled the study of the spatial distribution of gene expression patterns within complex biological structures, providing valuable insights into the organization and function of cells in their native environment [1,7,28]. The state-of-the-art ST technologies can be summarized into two categories: sequencing-based and imaging-based. The 10x Visium technology [9] is a widely adopted sequencing-based method with reasonable cost. This technology can sequence the whole transcriptome, but the measured gene expression is not at the single-cell level; instead, it measures gene expression profiles at each *spot*, which often covers multiple cells. Stereo-seq [18], on the other hand, can measure sub-cellular resolution spots but still does not provide cell-level gene expression data. For imaging-based methods (such as CosMx [11], 10x Xenium [12,16], MERFISH [6], STARmap [27]), since gene expression is captured *in situ*, the cell location information can be obtained, but these technologies can only capture up to hundreds of genes at the same time.

To overcome the limitations of the ST technologies, researchers have explored the possibilities of integrating or mapping ST data and other single-cell omics data, and in particular, scRNA-seq data. Methods have been proposed to integrate scRNA-seq and ST data, which serves various objectives including predicting locations of cells in the scRNA-seq data, imputing missing gene expression in the ST data, and estimating cell type proportions of the spots in low-resolution ST data [19,17]. In this paper, we focus on the task of predicting cell locations from the scRNA-seq data. Although our method can work with ST data from various technologies, we focus on the "low-resolution" spot-based ST data (like those from 10x Visium) when describing our methods, as these data are the most abundant while also being the most challenging to use. Tangram [2] infers a cell-by-spot mapping matrix utilizing gene expression dissimilarity loss between mapped and the reference ST data. SpaOTsc [3] is a method developed based on optimal transport (OT) that infers spatial coordinates for scRNA-seq data. CeLEry [29] uses a variational autoencoder to learn the mapping between scRNA-seq and ST data. Though these methods vary in their methodologies, they share the fundamental ideas of using the expression levels of common genes in both datasets as a bridge to map cells in scRNA-seq data to locations in ST data. However, due to large batch effects between ST and scRNA-seq data, the large search space of potential spatial locations for cells in the scRNA-seq data, and the reduced number of shared features between ST and scRNA-seq data, inferring the spatial locations of cells based solely on gene expression data is still a very challenging problem, and the accuracy of state-of-the-art methods is unknown due to the lack of ground truth data.

We consider that the temporal dynamics of cells are closely coupled with spatial information of cells, therefore, the temporal information of cells can inform spatial locations of cells. In particular, tissues are formed through generations of cell divisions, which is a temporal process. The fundamental concept of this work is that the information on the temporal process can inform the spatial organization of cells in the tissue. In this paper, we assume that cells divided from the same ancestral cells tend to be located closely in space unless the cells randomly migrate to distant locations after division. To obtain the *cell lineage* or *cell clonal* information, we take advantage of the CRISPR/Cas9-based *lineage tracing* datasets [26,4,24], which consist of paired gene expressions and lineage barcodes in single cells. That is, a lineage tracing dataset consists of a scRNA-seq count matrix, and a set of lineage barcodes, each for a single cell. For any two cells, the difference between their lineage barcodes should reflect their distance on the cell division tree. Therefore, the lineage barcodes can be used to reconstruct the cell lineage tree, representing the clonal relationships between cells. To incorporate the lineage and clonal information of cells to improve the prediction of cell spatial locations, we developed TemSOMap (**Tem**poral dynamics guided **S**patial **O**mics **Map**ping), where we assume that closer cells on the lineage tree are likely to have closer spatial coordinates. On both simulated and real datasets, we show that TemSOMap outperforms other methods in inferring the spatial coordinates of cells, while accurately inferring the spatial distribution of gene expressions. TemSOMap is the first method that integrates lineage information with ST data. In addition to improved accuracy of spatial location prediction, applying TemSOMap to real data can output a spatiotemporal map of single cells, which provides both lineage, spatial, and gene expression information of every single cell, and allows for the analysis of the spatiotemporal dynamics of cells. For example, following the lineage from a progenitor, the spatial migration

pattern for its descendant cells can be analyzed. Potentially, one can predict the spatial distribution of cells at earlier development time using the spatial coordinates of leaf cells and the lineage tree.

2 Methods

2.1 Related background

Cell lineage, cell clone, and spatial coordinates A *cell lineage* refers to the tree representing the cell division history from a root to present-day cells. A *cell clone* refers to a group of cells that are derived from the same ancestor in the cell lineage tree, thus a cell clone corresponds to leaf cells of a subtree in a cell division tree. The spatial coordinates of cells represent the cells' relative locations in tissue. In TemSOMap, we assume that cells that are closer on the cell lineage tree are more likely to have closer spatial coordinates. If this is true in a tissue, we say that this tissue has *clonal pattern*.

Biological processes related to cells' spatio-temporal dynamics The formation and growth of tissue and organs is a crucial question for developmental biology. We consider two key biological models contributing to cell spatial coordinates in a tissue: *cell division* and *cell migration*.

Cell divisions refer to duplicating a parent cell into two daughter cells. Cell migration refers to the process of cells' movement in tissue, guided by complex biological factors including cell state, cell microenvironment, and overall tissue environment. Cell migration activities can change the clonal pattern in the tissue. Therefore, when testing TemSOMap, we use a variety of migration rates to test the robustness of TemSOMap against cell migrations.

2.2 Overview of TemSOMap

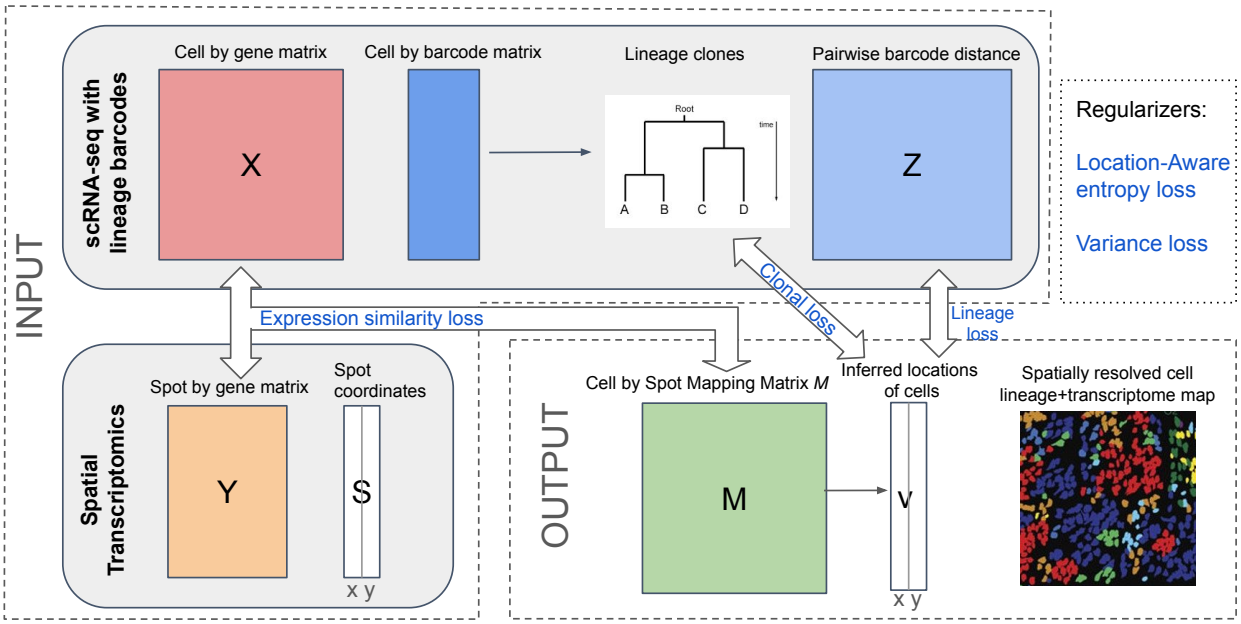


Fig. 1. Overview of TemSOMap. The input to TemSOMap includes a scRNA-seq data matrix with single-cell lineage barcodes and a spatial transcriptomic (ST) dataset. TemSOMap outputs a mapping matrix M, which can be used to obtain the inferred location of cells. Blue fonts highlight the loss terms (See Methods).

As shown in Fig. 1, TemSOMap takes in a lineage tracing dataset with both scRNA-seq count matrix (denoted by X) and lineage barcodes, and a ST dataset with both a spot-level gene expression matrix (denoted by Y) and spatial coordinates of the spots (denoted by S). The lineage barcodes can be used to

obtain cell clones, where each cell has a clone label, and pairwise barcode distance between cells (denoted by Z). The objective is to infer M, a cell-by-spot mapping matrix that represents the probability for each cell in X being mapped to each spot in Y. This definition of the mapping matrix is used in previous work [2]. Then $M^T X$ becomes a ST data matrix converted from the scRNA-seq count matrix X. Predicted cell coordinates of cells in X can be obtained from MS. We design an objective function to infer M.

The objective function of TemSOMap consists of three major loss terms and two regularization terms (Fig. 1). The three major terms are: expression similarity loss, which calculates the dissimilarity between $M^T X$ (the ST data converted from scRNA-seq) and the given ST data Y; lineage loss, which aims to maintain the similarity of the pairwise distances between cells on the inferred spatial coordinates and on the lineage tree; and clonal loss, which enforces the spatial clustering of cell clones, *i.e.*, cells in the same clone should locate close to each other.

We have added regularization terms each aiming at maintaining a desired property of M. These are: location-aware entropy loss, which forces the probability distribution of every cell to be sparse and spatially concentrated; and variance loss, which penalizes large variance for the spatial distribution of a cell's possible coordinates. Then, the total loss is calculated as the weighted sum of the losses described above. Total loss is minimized via stochastic gradient descent to find the optimal mapping matrix M. The predicted locations of cells in scRNA-seq data are then obtained from M. These steps are described in detail in the following sections.

2.3 Notation definitions

The input data to TemSOMap are denoted as follows: $X \in \mathcal{R}^{n_{cells} \times n_{genes}}$ is a cell-by-gene matrix (scRNA-seq count matrix); $Y \in \mathcal{R}^{n_{spots} \times n_{genes}}$, a spot-by-gene matrix representing the ST data; and $Z \in \mathcal{R}^{n_{cells} \times n_{cells}}$, a cell-cell pairwise distance matrix, indicating the distance between single cells in the cell lineage tree. $c \in \mathcal{N}^{n_{cells}}$ is a vector representing discrete clonal classes for every cell. The spatial coordinates of the spots is denoted as $S \in \mathcal{R}^{n_{spots} \times 2}$ (a spot-by-coordinate matrix).

$M \in \mathcal{R}^{n_{cells} \times n_{spots}}$ is a cell-by-spot mapping matrix that encodes the soft assignment for each cell to each spot, where $\sum_j^{n_{spots}} M_{ij} = 1$ and $M_{ij} >= 0$ for every cell i and every spot j.

2.4 Calculating cell clones and cell lineages from lineage barcode data

Calculating cell clones and cell lineages from lineage barcode data can be viewed as a pre-processing step to prepare Z and c for TemSOMap. The lineage barcodes in lineage tracing data are strings of fixed length (at the scale of tens or hundreds). At the root of the cell lineage tree, the barcode has only unmutated characters. During cell divisions, daughter cells inherit the parent cell's barcode while some new mutations are introduced. Dropouts, which are missing data in the barcodes, can exist in the final barcode readouts in cells.

Given the barcodes for all the cells, to obtain Z, we calculate the pairwise weighted Hamming distances between all pairs of barcodes, following practice in existing work [22,10]. In Z, Z_{ij} represents the weighted hamming distance of the lineage barcodes between cell i and cell j. To obtain clone labels c for the cells, we use the Neighbor-Joining method [25] to infer a lineage tree T, which is a binary tree graph where the leaf nodes are the cells present in the lineage barcode data. In TemSOMap, the clonal IDs are learned automatically based on the inferred lineage tree. To obtain balanced clones from the inferred lineage tree, we first re-balanced the lineage tree by rooting the tree at an internal node that has the closest number of cells under the left and right subtrees. Then, the clonal IDs can be obtained by cutting the tree at the specific generation $\log_2 n_{clone}$ on the re-balanced bifurcating tree, where n_{clone} is the number of clones set by users.

2.5 TemSOMap mapping algorithm

To constrain values in matrix M to be probabilities, we utilize the softmax function to transform a given matrix M:

$$M_{i,j} = softmax(M)_{i,j} = \frac{e^{M_{i,j}}}{\sum_j^{n_{spots}} e^{M_{i,j}}} \tag{1}$$

The total loss function of the TemSOMap mapper consists of the following components:

Expression similarity loss Following Biancalani *et al* [2], we utilize the cosine similarity function to enforce similarity between $M^T X$ and Y. We apply the cosine similarity function to both rows of Y, which correspond to the spot-wise expression patterns (Eq. 2), and columns of Y, which correspond to gene-wise expression patterns (Eq. 3). The loss terms can be formulated as follows:

$$\Phi_s(M|X,Y) = -\sum_{j=1}^{N} cos_{sim}[(M^T X)_{j,*}, Y_{j,*}] \tag{2}$$

$$\Phi_g(M|X,Y) = -\sum_{k=1}^{K} cos_{sim}[(M^T X)_{*,k}, Y_{*,k}] \tag{3}$$

where $cos_{sim}(A,B) = \frac{A \cdot B}{|A| \times |B|}$.

Lineage loss Using the mapping matrix M, we can calculate the pairwise distance matrix between the inferred spatial location of cells. The inferred spatial location of cells can be calculated as $\hat{v} = MS$, where for each cell, its inferred location is the weighted mean of all locations, with weights being probabilities in M. Therefore, the pairwise Euclidean distances between inferred locations of cells can be represented as follows:

$$D = \{D_{i,j} | D_{i,j} = ||\hat{v}_{i,*} - \hat{v}_{j,*}||\} \tag{4}$$

and the lineage loss is the MSE loss between the inferred pairwise distances of cells and the pairwise Hamming distances of cells' lineage barcodes (with normalization):

$$\Phi_l(M|Z) = MSE(D,Z) = \frac{1}{n_{cells}} \sum_{i}^{n_{cells}} (\frac{D}{||D||^2} - \frac{Z}{||Z||^2})^2 \tag{5}$$

Clonal loss The clonal loss considers cell clones, where cells are assigned clonal labels (described in Sec 2.4) that are similar to cluster labels. In contrast to lineage loss, where all pairwise distances between cells are considered, in the clonal loss, there are two types of distances: intra-clone distance and inter-clone distance. A vector $c \in \mathcal{N}^{n_{cells}}$ represents the clonal ID for the cells. Considering that cells in the same clone should be located close to each other in space, the clonal loss can be calculated as follows:

$$\Phi_c(M|c) = \frac{\sum_{c_i = c_j} D_{i,j} - \sum_{c_i \neq c_j} D_{i,j}}{\sum_{i,j}^{n_{cells}} D_{i,j}} \tag{6}$$

where we try to minimize intra-clone distances and maximize inter-clone distances.

Location-aware (LA) entropy loss While the mapping matrix M indicates soft assignment, we enforce an entropy loss on the inferred probability distribution of each cell to promote sparsity. Furthermore, we want the mapping matrix to better reflect realistic cell-spot mapping relationships, that is, a cell's probability among spots should be spatially concentrated. Therefore, we first smooth the mapping matrix using a Gaussian kernel based on the pairwise distances between spots,

$$G^S = \left\{ G^S_{i,j} | {G^S}_{i,j} = exp(\frac{||S_{i,*} - S_{j,*}||^2}{2\sigma^2}) \right\} \tag{7}$$

and then calculate the entropy on the smoothed mapping matrix M_s, which can be formulated as follows:

$$\Phi_e(M) = \frac{1}{n_{cells}} \sum_{i}^{n_{cells}} \sum_{j}^{n_{spots}} M^s_{i,j} \log(M^s_{i,j}) \tag{8}$$

where $\hat{M}_s = M \cdot G^S$, indicating the smoothed mapping matrix.

Variance loss Each row of the M matrix represents a probability distribution among spots for each cell. For each cell, it is desired for spots with high probabilities of including it to be located closely in 2-D space. This can be achieved by minimizing the following weighted location variance that we designed. Given the mapping matrix M and the spot coordinates $S = (\mathbf{x}, \mathbf{y})$, indicating the two axes for the coordinates, we calculate the variance loss as the average weighted variance of the spatial distribution of all cells:

$$\Phi_v(M|S) = \frac{1}{n_{cells}} \sum_{i}^{n_{cells}} (M_{i,*}(\mathbf{x} - \overline{x_i})^2 + M_{i,*}(\mathbf{y} - \overline{y_i})^2)/2 \tag{9}$$

where $\overline{\mathbf{x}} = M\mathbf{x}$ and $\overline{\mathbf{y}} = M\mathbf{y}$, representing the weighted mean locations of cells on x and y axes, respectively.

Total loss and optimization The total loss is therefore calculated as:

$$\begin{aligned}\Phi(M|X, Y, Z, c, S) = \lambda_1[\Phi_s(M|X, Y) + \Phi_g(M|X, Y)] + \lambda_2\Phi_l(M|Z) + \lambda_3\Phi_c(M|c) \\ + \lambda_4\Phi_e(M) + \lambda_5\Phi_v(M|S)\end{aligned} \tag{10}$$

where each loss is weighted using a hyperparameter (λ_{1-5}). For the tests we run in this work, we use a default set of hyperparameters without tuning. The default values are chosen so that the loss terms are of similar magnitude while prioritizing fitting the gene expression and lineage patterns. Details about the hyperparameter settings are in Supp. Info. Sec. 3.2. The minimization of the total loss is achieved via the Adam optimizer using the Pytorch library.

Once the final M is calculated, the inferred coordinates of a cell i is calculated as

$$v_i = (M[i, *]\mathbf{x}, M[i, *]\mathbf{y}) \tag{11}$$

which is the weighted mean of all locations, with probabilities in M as weights.

2.6 Simulating lineage-resolved single-cell spatial transcriptomics

To benchmark TemSOMap's performances and compare with other state-of-the-art methods, ground truth information about cells' lineage identity, spatial coordinates, and gene expression is needed. Due to the lack of real datasets that contain all three modalities, we developed SpaTedSim, a computational simulation framework that generates cells' gene expressions, spatial coordinates, and lineage barcodes simultaneously. The simulation process is based on a ground-truth cell division tree, and along the generations of cell divisions, we simulate the lineage barcodes and gene expressions based on our previously published method, TedSim [23] (Supp. Fig. 1a). In SpaTedSim, we simulate the cells' movement in space from cell division and cell migration. For cell division, starting with the root cell, SpaTedSim iteratively generates the daughter cells' (cell u and cell v) initial coordinates given the parent's coordinates (x_p, y_p):

$$(x_u = x_p + r_g \cos\theta + \mathcal{N}(0, \sigma^2\delta),\ y_u = y_p + r_g \cos\theta + \mathcal{N}(0, \sigma^2\delta))$$

$$(x_v = x_p - r_g \cos\theta + \mathcal{N}(0, \sigma^2\delta),\ y_v = y_p - r_g \cos\theta + \mathcal{N}(0, \sigma^2\delta))$$

where r_g is the division radius, which depends on the current generation of divisions. The more the cells divide, the smaller the division radius becomes, indicating less freedom of movement when the tissue becomes more crowded and differentiated. θ determines the angle of the dividing cells, which leads to opposite moving directions for the two daughter cells. Lastly, $\mathcal{N}(0, \sigma^2\delta)$ represents the 2-D Brownian motion term (Gaussian random walk), where δ is the step size and σ is the standard deviation.

After cell division, a cell's spatial coordinates can further change due to cell migration. In real tissues, cell migration is a complex biological process, controlled by various intracellular and extracellular factors. In SpaTedSim, the migration behavior of cells is dependent on their cell types. Given the spatial coordinates of cells of the same cell type, $v = (v_1, v_2, ..., v_n) \in R^{n\times 2}$, we first estimate the spatial density of the cell type:

$$\hat{f}_h(\mathbf{v}) = \frac{1}{nh\sigma}\frac{1}{\sqrt{2\pi}}\sum_i^n exp(\frac{-(\mathbf{v}-v_i)^2}{2h^2\sigma^2})$$

To simplify, we set $\sigma = 1$, and h is the bandwidth hyperparameter defined in density estimation. After each generation of cell divisions, the Gaussian density estimation is performed on every cell type, and cell migration is the process of migrating a cell's coordinates to a nearby location based on the cell type density map. For a cell's coordinates after cell division, v, and a migration radius r_m, the migrated coordinates, v', are sampled from the cell type densities:

$$v' \sim \frac{\hat{f}_h(\mathbf{v})}{\sum_{\mathbf{v}} \hat{f}_h(\mathbf{v})}\ , \forall\ \mathbf{v} \text{ that satisfies } (\mathbf{v}-v)^2 \leq r_m^2$$

Lastly, after generating all spatial coordinates of the leaf cells, we perform an additional postprocessing step to make the overall coordinates more realistic; that is, cells should have a relatively uniform spatial distribution in the simulated tissue region. More details about the simulation algorithms and parameter settings are in Supp. Info. Sec. 2.3.

3 Results

3.1 Evaluating TemSOMap on synthetic datasets using SpaTedSim

To quantify the performances of TemSOMap and baseline methods, we use datasets simulated by SpaTedSim. SpaTedSim generates scRNA-seq data and ST data with ground truth cell locations in the scRNA-seq data, and it can vary the amount of batch effects between scRNA-seq and ST data (Supp. Fig. 1b, Supp. Info. Sec. 2.2). By varying migration rate, SpaTedSim can generate datasets with different spatial clustering patterns in terms of cell types or clonal IDs (Supp. Fig. 1c). With a lower migration rate, the cell migration occurs less, and the cells will show better clonal clustering pattern spatially. On the contrary, a higher migration rate results in a better cell-type clustering pattern. To test the performances of TemSOMap, we mask the spatial coordinates of single cells and try to reconstruct them using spot-level spatial transcriptomics and paired gene expressions and barcodes. An example of visualization of TemSOMap's reconstructed spatial map of single cells is shown in Fig. 2a, in comparison with the ground truth.

We compare TemSOMap with existing methods that estimate the spatial location of cells: Tangram, CeLEry, and SpaOTsc. Several methods that integrate scRNA-seq with ST data, including Cell2location [13], do not output inferred locations of cells. To compare inferred cell locations with ground truth location information, we used four metrics: Mean Squared Errors (MSE), Mean Absolute Errors (MAE), Average Jaccard Index, and Pearson's Correlation. Detailed procedures and settings of data generation, methods, and metrics can be found in Supp. Info. Sec. 3.

Fig. 2b shows the comparison between TemSOMap and baseline methods using datasets simulated with a wide range of migration rates. For each migration rate, 20 simulated datasets were generated. We can observe that TemSOMap consistently outperforms other methods on all metrics and migration rates, indicating more accurate inferences of the cells' spatial locations. From the accuracy changes of TemSOMap with migration rate, we observe that the performances tend to be worse for the migration rate at 0.5 and better when the migration rate is either small or large. This is because when the migration rate is low or high, the spatial distribution of cells will be more clustered based on either clones or cell types. Due to the partial consistency between clones and cell types [23], either pattern will benefit TemSOMap's performance. When the migration rate is around 0.5, both cell type and clonal patterns are worse, making it more difficult for TemSOMap. However, TemSOMap is still able to outperform other methods under such circumstances. We provide some sample visualizations of the inferred spatial maps of cells under different migration rates (Supp. Fig. 2), which further shows the advantage of TemSOMap in inferring cells' coordinates with similar clonal and cell-type patterns.

We also show the performances of TemSOMap compared with other methods on larger datasets (4096 cells, Supp. Fig. 3a) and with worse lineage barcode quality (25% dropouts, Supp. Fig. 3b). The comparisons are consistent overall with those on 1024 cells datasets, even though the gap between TemSOMap and other

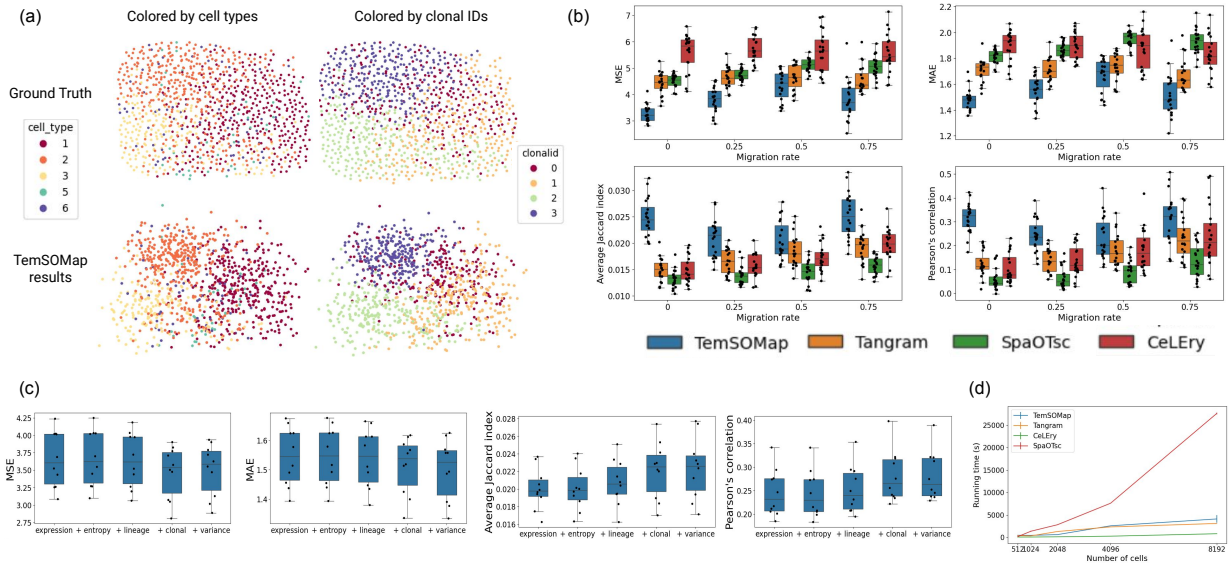


Fig. 2. TemSOMap results on SpaTedSim-simulated data. (a) Visualization of the spatial coordinates of single cells, comparing ground truth and TemSOMap inferred results. Left column: cells colored by cell types; right column: cells colored by clonal IDs. The migration rate is 1 for this plot. (b) Comparisons of TemSOMap and baseline methods on SpaTedSim-simulated datasets (1024 cells), with varying migration rates. For MSE and MAE, lower values mean better accuracy. For the Average Jaccard Index and Pearson's Correlation, higher values mean better accuracy. No dropouts are added to the lineage barcode data in these results and each migration rate has 20 datasets. (c) Ablation test results of TemSOMap. (d) Running time comparisons for the four methods. With a fixed spot size, we increase the number of cells which accordingly increases the number of spots (Approximately $n_spots = \log_2(n_cells)$). For each dataset size, we record the average running time across five datasets. More details are in Supp. Info. Sec. 3.1.

methods decreases. This is because in both cases, the quality of the lineage barcodes is worse with dropouts, or with the same amount of mutations (the number of characters in the lineage barcode matrix) but with increased sample size (1024 cells to 4096 cells). We also compared the methods under different levels of batch effects in Supp. Fig. 3c, on selected migration rates. With the same migration rate, methods generally perform better with small batch effects compared to large batch effects, which is expected. We investigated the change in the values of loss terms during optimization and found that under all different scenarios, training on the total loss using Adam optimizer, different loss terms in TemSOMap converge simultaneously and reach a cell-spot mapping matrix that fits both the gene expressions and spatial patterns (Supp. Fig. 3d).

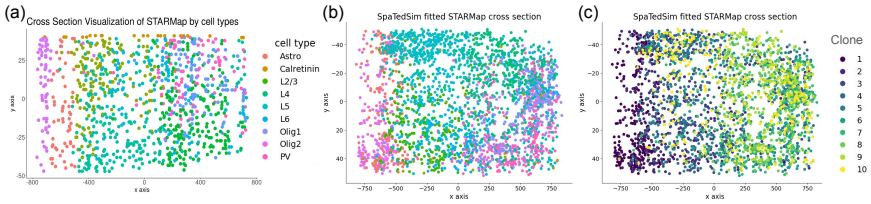
Using simulated datasets also allows us to perform an ablation test to quantitatively assess the contribution of different loss terms of TemSOMap. We started using only the expression similarity loss, then sequentially added the LA entropy loss, lineage loss, clonal loss, and variance loss. We can see that with the inclusion of each loss term, the performance of TemSOMap generally increases considering all four metrics (Fig. 2c).

Additionally, we compare the running time of TemSOMap with baseline methods (Fig. 2d). Compared with SpaOTsc, the other three methods scale better to large datasets, with CeLEry performing the fastest and TemSOMap being similar to Tangram. Detailed descriptions of the computational resources and other parameter settings for running all methods are in the Supp. Info. Sec. 3.4.

3.2 Testing TemSOMap on fitted mouse cortex data

In the previous section, we generated purely synthetic data to perform a comprehensive test of the methods with different clonal patterns (migration rates), dataset size, and lineage barcode quality. Here we use a real ST dataset as a reference, to generate synthetic ST data with lineage barcodes that mimic the real ST

dataset. The mammalian cerebral cortex is a suitable system to generate such datasets, as although there is no real dataset with both spatial coordinates and lineage barcode information, its layered structure is well studied, and previous studies have shown the early development of the mammalian prefrontal cortex, with the birthtime of cells labeled at different embryonic times [14]. With this knowledge, we can simulate lineage barcode data for a real ST dataset on the cortex under the model of cell growth from inner to outer layers. We use the STARmap mouse brain cortex data from the Giotto toolbox [27,8] with labeled cell types as the ST reference dataset (Fig. 3a). We use SpaTedSim to generate lineage barcodes and to fit the real STARmap dataset, where the spatial distribution of cell types is preserved.



(d)

Method	benchmarking metrics			
	MSE	MAE	J-index	Pearson
TemSOMap	$\mathbf{5.13 \times 10^4}$	**138**	$\mathbf{9.17 \times 10^{-3}}$	**0.48**
Tangram	6.37×10^4	157	7.96×10^{-3}	0.37
CeLEry	7.22×10^4	155	8.82×10^{-3}	0.40
SpaOTsc	8.64×10^4	191	7.71×10^{-3}	0.28

Fig. 3. TemSOMap results on SpaTedSim-fitted STARmap mouse cortex data and comparison with other methods. (a) 2-D spatial visualization of a slice in the STARmap dataset. (b) 2-D visualization of SpaTedSim-fitted STARmap data, colored by cell types. (c) 2-D visualization of SpaTedSim-fitted STARMap data, colored by clones. (d) Accuracy metrics of inferred single-cell coordinates by TemSOMap and the other state-of-the-art methods. Detailed descriptions of the methods and metrics are in Supp. Info. Sec. 3.

From the 3-D mouse cortex dataset, we first obtain a slice of 2-D distribution of cells by taking a cross section along the layers of the tissue (Supp. Fig. 4a). On the 2-D slice, we first separate spatial locations into clones based on the layered structure, and obtained ten clones [14] (Supp. Fig. 4b). For the area corresponding to each clone, we use a moving window to perform cell type density estimation to simulate realistic spatial distribution. Then, we merge the clones to get the final fitted data. The simulated dataset also aims to preserve other features of the reference STARmap data, such as cell type percentages (Supp. Fig. 4c) and average counts per cell type (Supp. Fig. 4d). Fig. 3b-c shows the visualization of generated data in 2-D space. Finally, we generated spot-level ST data with a fixed size (procedure in Supp. Info. Sec. 3.1) from the simulated STARmap data which is at single-cell resolution (Supp. Fig. 4e).

We used TemSOMap and other state-of-the-art methods to recover the hidden spatial coordinates of single cells by mapping the single-cell gene expressions onto the spot-level spatial data. We compared the performances of the methods using the same four metrics: MSE, MAE, Average Jaccard index, and Pearson's correlation between the inferred cell coordinates and the ground truth. TemSOMap consistently outperformed the other methods and more accurately inferred the spatial coordinates of single cells (Fig. 3d).

3.3 Evaluating TemSOMap on E9.5 mouse embryo datasets

Understanding the spatiotemporal dynamics of early mammalian embryogenesis — that is, how a single omnipotent fertilized egg divides and differentiates into an embryo — is a fundamental problem in developmental biology. Integrating lineage tracing data and spatial transcriptomic data has the potential to uncover how cell types emerge and migrate in time and space. In this study, we attempt to integrate a lineage tracing dataset and a ST dataset of the mouse E9.5 embryo using TemSOMap. The lineage tracing dataset consists of paired scRNA-seq and CRISPR/Cas9-induced lineage barcodes of E9.5 mouse embryo cells [4]. The lineage barcodes are used to calculate the pairwise lineage distance between cells and clonal labels of cells (Supp. Fig. 5a).

For the ST data, we used a Stereo-seq dataset [5] where spot-level ST data of an E9.5 mouse embryo with labeled cell types was provided. To map the single cells from the lineage tracing data onto the ST data, We first find the highly variable genes that are present in both datasets (based on the dispersion of genes, Supp. Fig. 5b). We then applied TemSOMap and obtained the mapping matrix M, which includes location information of cells in the lineage tracing dataset.

The Stereo-seq paper [5] provided cell type annotations of the high-resolution spots which clearly correspond to different organs in the embryo (Fig. 4a, left). To visualize the lineage tracing data with cells in

their inferred locations and compare with the ST data visualization in Fig. 4a, we need to annotate cells in the lineage tracing data with the same set of cell type labels. Therefore, we performed label transfer from the ST data to the lineage tracing data, by projecting the PCAs of lineage tracing data onto the PCA of the ST data, such that the two datasets are aligned in their gene expression space (Supp. Fig. 5c-d).

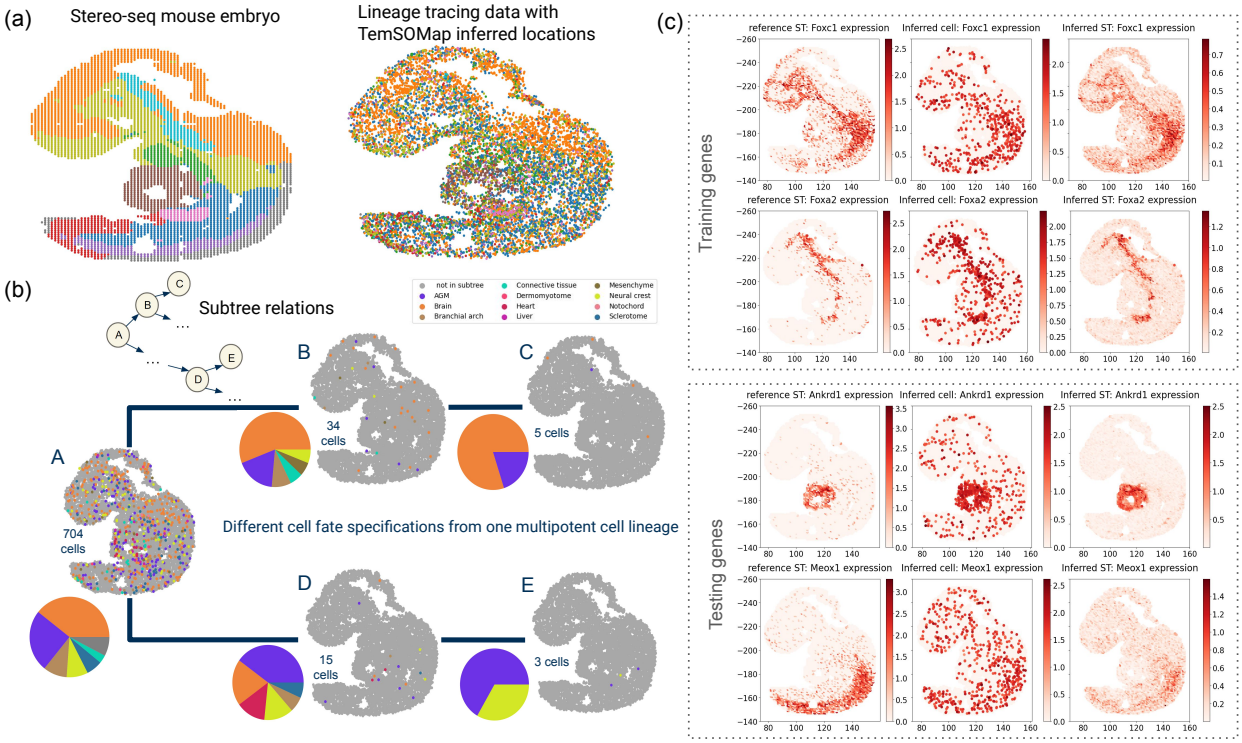


Fig. 4. TemSOMap results on E9.5 mouse embryo data. (a) 2-D visualization of the input Stereo-seq mouse embryo data (spot level) and the TemSOMap-inferred mouse embryo data (single-cell level). Colors represent cell types from the Stereo-seq annotation. (b) Spatiotemporal analysis on the cell fate specifications on the cell lineage. Each embryo plot represents the spatial distribution of cells for each subtree, with pie plots showing the cell type percentages of the leaves. (c) Comparing spatial maps of gene expressions of the reference ST data (left), inferred cell-level data (middle), and inferred spot-level data (right). Foxc1 and Foxa2 are used in the training of TemSOMap; Ankrd1 and Meox1 are masked from the training of TemSOMap and their spatial distribution is predicted using TemSOMap.

When obtaining inferred locations of cells from the M matrix, we have been using Eq. 11, which means a weighted mean location is used as the inferred location. An alternative way of determining inferred locations from M is to take the location with the highest probability for each cell, *i.e.*, taking the Maximum A Posteriori (MAP) for each cell from the mapping matrix. To best explore this complex mouse embryo dataset, we used both modes, mean and MAP, as described above, to obtain inferred cell locations.

To compare with baseline methods, we also ran CeLEry and Tangram on this dataset. We did not run SpaOTsc, because this method does not scale to this large dataset with 9707 cells in the lineage tracing dataset and 5031 spots in the ST dataset. For a fair comparison, we also used the two modes, mean and MAP to determine inferred cell locations for Tangram, as Tangram also outputs the M mapping matrix. Visualizations of the lineage tracing data with inferred locations using all methods and modes are shown in Supp. Fig. 6. We can see that results from CeLEry and Tangram-mean, though showing certain clustering patterns, preserve little shape of the embryo. TemSOMap-MAP and Tangram-MAP preserve the embryo and organ shapes best, and show overall comparable quality, while TemSOMap-MAP shows slightly better clustering patterns than Tangram-MAP on some cell types, like liver and heart.

While being better than other results, the organ patterns in TemSOMap-mean results are rather noisy. In addition to inference errors, there are other factors that pose difficulties for the integration. First, only a

slice of the embryo is used in the reference ST data, and the lineage tracing dataset has all the cells in the embryo. Second, there can be a large batch effect between the two datasets, including individual differences from the two mice embryos.

Nevertheless, the lineage tracing dataset with inferred cell locations allows us to obtain more insights into the spatiotemporal changes of cells' locations along with their cell division histories (Fig. 4b). With results from Tangram-MAP (Fig. 4a, right), we analyze a subtree of the whole lineage tree, which corresponds to a large multipotent progenitor (node A in Fig. 4b). This progenitor is the common ancestor of 704 observed present-day cells, whose spatial locations and cell types can be visualized (Fig. 4b). Going down the lineage from this node on the lineage tree, we can observe that different daughter cells take on different cell fate specifications; while one lineage ($A \rightarrow B \rightarrow C$) differentiates into mostly brain cells, another lineage ($A \rightarrow D \rightarrow E$) differentiates into a majority of AGM cells.

Using inferred cell locations from TemSOMap, we can also show spatial gene expression distributions of individual genes in the lineage tracing dataset. Furthermore, we can predict the spatial distribution of unseen genes or test genes. In Fig. 4c, the top panel shows examples of genes in the training set, while the bottom panels show examples of genes in the test set. For each gene, the left plot shows its expression level in the ST dataset, the middle plot shows its expression in the lineage tracing dataset, with inferred cell locations and the right plot shows the gene expression converted to spot resolution by calculating $M^T X[*, j]$, which results in a vector of length n_{spot} that represents the expression of gene j across spots. Details on performing this test and splitting the training and test sets are in Supp. Info. Sec. 3.3.

More spatial gene comparisons are shown in Supp. Fig. 7 (for genes in the training set) and Supp. Fig. 8 (for genes in the test set). From the visualizations, we can see that for genes in both the training set and the test set, the TemSOMap predicted spatial distribution of gene expression in the lineage tracing dataset highly resembles that in the ST dataset, especially the inferred ST level data (the rightmost plots for each gene). These results not only support the inferred cell locations, but also indicate the potential of TemSOMap in compensating for the lack of gene throughput or resolution of ST technologies.

4 Discussion

We presented TemSOMap, a method that can incorporate cell lineage (temporal) information when mapping scRNA-seq data to ST data. This not only allows for more accurate inference of cell spatial locations in scRNA-seq data, but also generates lineage-resolved spatial data; that is, a dataset that has three modalities: gene expression, lineage and clonal information, and spatial locations of cells. We have shown that such spatiotemporal data can be used to study cell fate and cell location specifications.

In this paper, we primarily used lineage tracing data to obtain cell lineage and clonal information. Other types of data can potentially be used to obtain this information. For example, literature has suggested that scATAC-seq data or scRNA-seq data can include information on mitochondria somatic mutations, which can be used to trace cell lineages [20,15,21]. Moreover, for cancer tissues, copy number variation (CNV) that can be detected from scRNA-seq or scATAC-seq data can also be used to trace cell lineages [20].

Given the assumption that cells from the same clone should be located closely in space, TemSOMap is most applicable to solid tissues where cells do not typically migrate frequently to distant locations, like the brain, skin, and embryos. However, our simulation tool SpaTedSim allows us to generate data with different migration rates, which is the probability for a cell to migrate from its sister cell. We can see that TemSOMap still outperforms baseline methods even when the migration rate is high. This means that TemSOMap exploits clonal and lineage patterns even when such signals are very noisy. Overall, although the method is developed with this assumption, it can work when clonal patterns are weak.

Finally, TemSOMap can work with datasets from a wide range of ST technologies. In this paper, we have used STARmap data and Steoro-seq data, and the simulated datasets were designed to simulate larger spatial spots like those in 10x Visium (around 50 μm [9]). The Steoro-seq data has subcellular resolution, though we have used the pre-processed data in the Steoro-seq paper where high-resolution spots are binned into larger spots. But TemSOMap can also work with subcellular resolution data such as Xenium($0.5\mu m$ [12]) since the mapping matrix in TemSOMap can be implied as a subcellular-level mapping of cells' transcripts from cells to space.

References

1. Atta, L., Fan, J.: Computational challenges and opportunities in spatially resolved transcriptomic data analysis. Nat. Commun. **12**(1), 5283 (Sep 2021)
2. Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C.R., Segerstolpe, A., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N.B., Fanelli, D., Zhuang, X., Macosko, E.Z., Regev, A.: Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. Nature Methods **18**(11), 1352–1362 (Oct 2021). https://doi.org/10.1038/s41592-021-01264-7, http://dx.doi.org/10.1038/s41592-021-01264-7
3. Cang, Z., Nie, Q.: Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat. Commun. **11**(1), 2084 (Apr 2020)
4. Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T.M., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G., Khodaverdian, A., Yosef, N., Meissner, A., Weissman, J.S.: Molecular recording of mammalian embryogenesis. Nature **570**(7759), 77–82 (Jun 2019). https://doi.org/10.1038/s41586-019-1184-5, https://doi.org/10.1038/s41586-019-1184-5
5. Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., Wang, X., Lu, H., Chen, X., Liu, X., Huang, X., Li, Z., Hong, Y., Jiang, Y., Peng, J., Liu, S., Shen, M., Liu, C., Li, Q., Yuan, Y., Wei, X., Zheng, H., Feng, W., Wang, Z., Liu, Y., Wang, Z., Yang, Y., Xiang, H., Han, L., Qin, B., Guo, P., Lai, G., Muñoz-Cánoves, P., Maxwell, P.H., Thiery, J.P., Wu, Q.F., Zhao, F., Chen, B., Li, M., Dai, X., Wang, S., Kuang, H., Hui, J., Wang, L., Fei, J.F., Wang, O., Wei, X., Lu, H., Wang, B., Liu, S., Gu, Y., Ni, M., Zhang, W., Mu, F., Yin, Y., Yang, H., Lisby, M., Cornall, R.J., Mulder, J., Uhlén, M., Esteban, M.A., Li, Y., Liu, L., Xu, X., Wang, J.: Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. Cell **185**(10), 1777–1792.e21 (May 2022). https://doi.org/10.1016/j.cell.2022.04.003, http://dx.doi.org/10.1016/j.cell.2022.04.003
6. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., Zhuang, X.: RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells. Science **348**(6233), aaa6090 (Apr 2015)
7. Dries, R., Chen, J., Del Rossi, N., Khan, M.M., Sistig, A., Yuan, G.C.: Advances in spatial transcriptomic data analysis. Genome Res. **31**(10), 1706–1718 (Oct 2021)
8. Dries, R., Zhu, Q., Dong, R., Eng, C.H.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., George, R.E., Pierson, N., Cai, L., Yuan, G.C.: Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biology **22**(1) (Mar 2021). https://doi.org/10.1186/s13059-021-02286-2, http://dx.doi.org/10.1186/s13059-021-02286-2
9. 10x Genomics: Spatial gene expression. https://www.10xgenomics.com/products/spatial-gene-expression, accessed: 2024-10-1
10. Gong, W., Kim, H.J., Garry, D.J., Kwak, I.Y.: Single cell lineage reconstruction using distance-based algorithms and the R package, DCLEAR. BMC Bioinformatics **23**(1), 103 (Mar 2022)
11. He, S., Bhatt, R., Brown, C., Brown, E.A., Buhr, D.L., Chantranuvatana, K., Danaher, P., Dunaway, D., Garrison, R.G., Geiss, G., Gregory, M.T., Hoang, M.L., Khafizov, R., Killingbeck, E.E., Kim, D., Kim, T.K., Kim, Y., Klock, A., Korukonda, M., Kutchma, A., Lewis, Z.R., Liang, Y., Nelson, J.S., Ong, G.T., Perillo, E.P., Phan, J.C., Phan-Everson, T., Piazza, E., Rane, T., Reitz, Z., Rhodes, M., Rosenbloom, A., Ross, D., Sato, H., Wardhani, A.W., Williams-Wietzikoski, C.A., Wu, L., Beechem, J.M.: High-plex multiomic analysis in FFPE at subcellular level by spatial molecular imaging. bioRxiv p. 2021.11.03.467020 (Nov 2021)
12. Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., Nilsson, M.: In situ sequencing for RNA analysis in preserved tissue and cells. Nat. Methods **10**(9), 857–860 (Sep 2013)
13. Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M.S., Park, J.S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., Bayraktar, O.A.: Cell2location maps fine-grained cell types in spatial transcriptomics. Nat. Biotechnol. (Jan 2022)
14. Kolk, S.M., Rakic, P.: Development of prefrontal cortex. Neuropsychopharmacology **47**(1), 41–57 (Oct 2021). https://doi.org/10.1038/s41386-021-01137-9, http://dx.doi.org/10.1038/s41386-021-01137-9
15. Lareau, C.A., Liu, V., Muus, C., Praktiknjo, S.D., Nitsch, L., Kautz, P., Sandor, K., Yin, Y., Gutierrez, J.C., Pelka, K., Satpathy, A.T., Regev, A., Sankaran, V.G., Ludwig, L.S.: Mitochondrial single-cell atac-seq for high-throughput multi-omic detection of mitochondrial genotypes and chromatin accessibility. Nature Protocols **18**(5), 1416–1440 (Feb 2023). https://doi.org/10.1038/s41596-022-00795-3, http://dx.doi.org/10.1038/s41596-022-00795-3

16. Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M., Yang, J.L., Lee, H.S., Aach, J., Zhang, K., Church, G.M.: Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat. Protoc. **10**(3), 442–458 (Mar 2015)
17. Li, B., Zhang, W., Guo, C., Xu, H., Li, L., Fang, M., Hu, Y., Zhang, X., Yao, X., Tang, M., Liu, K., Zhao, X., Lin, J., Cheng, L., Chen, F., Xue, T., Qu, K.: Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat. Methods (May 2022)
18. Liu, C., Li, R., Li, Y., Lin, X., Zhao, K., Liu, Q., Wang, S., Yang, X., Shi, X., Ma, Y., Pei, C., Wang, H., Bao, W., Hui, J., Yang, T., Xu, Z., Lai, T., Berberoglu, M.A., Sahu, S.K., Esteban, M.A., Ma, K., Fan, G., Li, Y., Liu, S., Chen, A., Xu, X., Dong, Z., Liu, L.: Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. Dev. Cell **57**(10), 1284–1298.e5 (May 2022)
19. Longo, S.K., Guo, M.G., Ji, A.L., Khavari, P.A.: Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat. Rev. Genet. **22**(10), 627–644 (Oct 2021)
20. Ludwig, L.S., Lareau, C.A., Ulirsch, J.C., Christian, E., Muus, C., Li, L.H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J.H., Boland, G.M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M.J., Buenrostro, J.D., Regev, A., Sankaran, V.G.: Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. Cell **176**(6), 1325–1339.e22 (Mar 2019)
21. Nitsch, L., Lareau, C.A., Ludwig, L.S.: Mitochondrial genetics through the lens of single-cell multi-omics. Nat. Genet. **56**(7), 1355–1365 (Jul 2024)
22. Pan, X., Li, H., Putta, P., Zhang, X.: LinRace: cell division history reconstruction of single cells using paired lineage barcode and gene expression data. Nat. Commun. **14**(1), 8388 (Dec 2023)
23. Pan, X., Li, H., Zhang, X.: Tedsim: temporal dynamics simulation of single-cell rna sequencing data and cell division history. Nucleic Acids Research **50**(8), 4272–4288 (Apr 2022). https://doi.org/10.1093/nar/gkac235, `http://dx.doi.org/10.1093/nar/gkac235`
24. Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., Schier, A.F.: Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nature Biotechnology **36**(5), 442–450 (May 2018). https://doi.org/10.1038/nbt.4103, `https://doi.org/10.1038/nbt.4103`
25. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4), 406–425 (Jul 1987)
26. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., Junker, J.P.: Simultaneous lineage tracing and cell-type identification using crispr–cas9-induced genetic scars. Nature Biotechnology **36**(5), 469–473 (May 2018). https://doi.org/10.1038/nbt.4124, `https://doi.org/10.1038/nbt.4124`
27. Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G.P., Bava, F.A., Deisseroth, K.: Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science **361**(6400) (Jul 2018)
28. Yue, L., Liu, F., Hu, J., Yang, P., Wang, Y., Dong, J., Shu, W., Huang, X., Wang, S.: A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. Comput. Struct. Biotechnol. J. **21**, 940–955 (Jan 2023)
29. Zhang, Q., Jiang, S., Schroeder, A., Hu, J., Li, K., Zhang, B., Dai, D., Lee, E.B., Xiao, R., Li, M.: Leveraging spatial transcriptomics data to recover cell locations in single-cell RNA-seq with CeLEry. Nat. Commun. **14**(1), 4050 (Jul 2023)