Artificial Intelligence in the Life Sciences 3 (2023) 100068

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Elucidating dynamic cell lineages and gene networks in time-course single cell differentiation

Mengrui Zhang [c,1], Yongkai Chen [a,1], Dingyi Yu [b], Wenxuan Zhong [a], Jingyi Zhang [b,*], Ping Ma [a,*]

[a] Department of Statistics, University of Georgia, Athens, GA 30602, United Stated
[b] Department of Industrial Engineering, Center for Statistical Science, Tsinghua University, Beijing, China
[c] Surrozen, Inc., South San Francisco, CA, USA

ARTICLE INFO

Keywords:
scRNA-seq
Optimal Transport
Smoothing Spline
Dynamic Gene Networks

ABSTRACT

Single cell RNA sequencing (scRNA-seq) technologies provide researchers with an unprecedented opportunity to exploit cell heterogeneity. For example, the sequenced cells belong to various cell lineages, which may have different cell fates in stem and progenitor cells. Those cells may differentiate into various mature cell types in a cell differentiation process. To trace the behavior of cell differentiation, researchers reconstruct cell lineages and predict cell fates by ordering cells chronologically into a trajectory with a pseudo-time. However, in scRNA-seq experiments, there are no cell-to-cell correspondences along with the time to reconstruct the cell lineages, which creates a significant challenge for cell lineage tracing and cell fate prediction. Therefore, methods that can accurately reconstruct the dynamic cell lineages and predict cell fates are highly desirable.

In this article, we develop an innovative machine-learning framework called Cell Smoothing Transformation (CellST) to elucidate the dynamic cell fate paths and construct gene networks in cell differentiation processes. Unlike the existing methods that construct one single bulk cell trajectory, CellST builds cell trajectories and tracks behaviors for each individual cell. Additionally, CellST can predict cell fates even for less frequent cell types. Based on the individual cell fate trajectories, CellST can further construct dynamic gene networks to model gene-gene relationships along the cell differentiation process and discover critical genes that potentially regulate cells into various mature cell types.

1. Introduction

A comprehensive understanding of complex biological processes such as tissue development and regeneration requires the investigation of cell differentiation across a wide range of samples and experimental time points [1]. The cell differentiation process includes the differentiation of stem cells into different mature cell types [2,3]. Such a process is dynamic and continuous, including rapid changes in gene expressions and cell types over time. To profile such cell differentiation behaviors, single cell RNA-seq sequencing (scRNA-seq) technology has been developed rapidly [4–7]. In particular, scRNA-seq enables researchers to observe the gene expressions of all cells simultaneously (Fig. 1a) in both static or time-course experiments (Fig. 1b). The static scRNA-seq experiment takes a snapshot of all cells and their gene expressions at one time [8,9], whereas the time-course scRNA-seq experiments take snapshots at multiple time points. Using scRNA-seq, researchers can observe the behavior of individual cells in cell differentiation processes over time. Cell lineage tracing has been widely used to predict dynamic cell fates by indicating the ancestor and posterity cells in cell differentiation processes. For example, during a stem cell differentiation process, the multipotent stem cells can develop into multiple cell lineage endpoints (Fig. 1c). Despite the effectiveness, quantifying the dynamic cellular changes of cell development is still challenging due to the following technical limitations [10]. In time-course scRNA-seq experiments, cells are sacrificed and sequenced at each time point. Thus there is no cell-to-cell correspondence information for cells between two time points, which creates a significant challenge in constructing cell lineages and elucidating the dynamic cell behaviors in the differentiation process. Moreover, it is very challenging to align different cells sequenced in two adjacent time points since expressions of cells are high-dimensional and noisy, and the number of cells in each time point is large. Such a large sample and high-dimensional and noisy data problem render many classical methods, such as Euclidean distance or Pearson correlation, invalid [11,12].

* Corresponding authors.
E-mail addresses: mengrui@surrozen.com (M. Zhang), yongkaichen@uga.edu (Y. Chen), yudy21@mails.tsinghua.edu.cn (D. Yu), wenxuan@uga.edu (W. Zhong), jingyizhang@tsinghua.edu.cn (J. Zhang), pingma@uga.edu (P. Ma).
[1] Mengrui Zhang and Yongkai Chen contributed equally.

https://doi.org/10.1016/j.ailsci.2023.100068
Received 21 December 2022; Received in revised form 17 March 2023; Accepted 22 March 2023
Available online 25 March 2023

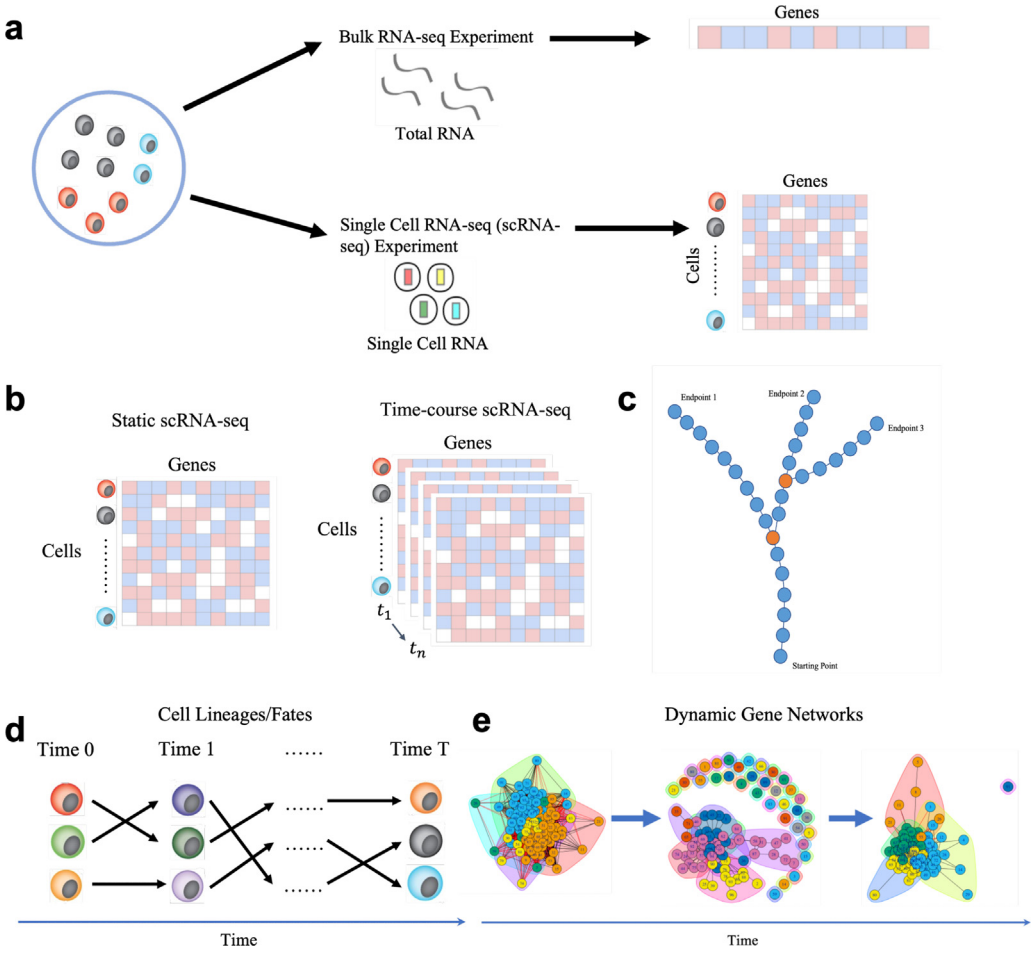

Fig. 1. Single cell analysis and cell smoothing transformation (CellST) overview: **a**: The advantage of scRNA-seq analysis over bulk RNA-seq analysis. **b**: Data structures for static scRNA-seq experiments and time-course scRNA-seq experiments. Cells in time-course experiments have been marked with experimental time points. **c**: The multipotent stem cells can develop into multiple cell lineage endpoints. **d**: Cell lineages are constructed by connecting individual cells over time. Cell fate trajectories are constructed by smoothing the connected cell lineages. **e**: Dynamic gene networks are constructed based on the calculated dynamic relationship between genes.

One natural approach to surmount the challenges is to order cells into a continuous cell trajectory. Many methods have been proposed to achieve this goal in static scRNA-seq experiments. In these methods, researchers construct a pseudotime to order cells chronologically [13–19]. Despite their effectiveness, such methods may fail in the following circumstances [20]. First of all, most existing trajectory inference methods construct a bulk cell trajectory, i.e., the mean trajectory of the population cells across time rather than that of individual cells.

However, some individual cells' behaviors may oscillate up and down around their mean expressions or severely deviate from them. Cell differentiation behaviors are dominated by cells with major cell types, and patterns with less frequent might be hidden in the dataset. Second, individual cell developing trajectories may follow different complex topologies, including loops or alternative paths during the development. For example, analysis approaches in [21], and [22] used dimension reduction methods to identify a low-dimensional space of the gene expression space before constructing cell trajectories [23,24]. Those methods may introduce a significant bias and are hard to validate, as cells are ordered based only on the selected reduced dimensions. Finally, the cells may not be synchronized at the same developing time points. Cells within the same time point can be expressed at different developing stages. In this situation, the bulk cell trajectory that takes the average pattern of cells at different stages might result in unreliable scientific discovery.

In this article, we propose a novel analysis framework named Cell Smoothing Transformation (CellST) to overcome the aforementioned limitations. The CellST framework elucidates dynamic cell fates and constructs gene networks in the cell differentiation process. In the CellST framework, we propose a cell lineage tracing method, which aligns two individual cells between any adjacent two time points via the optimal transport technique. The optimal transport technique is a mathematical tool initially developed to identify the transport map between two probability measures with the minimum transport cost [25]. Recently, this technique has gained increasing importance in various applications such as subsampling [61], sufficient dimension reduction [59], feature screening [57], and graph comparison [55]. Consequently, it has become a powerful tool for modeling cell dynamics [27–29]. See Zhang et al. [29,60] for recent reviews. Those aligned cells can potentially represent individual cell lineages, tracing cell differentiation behaviors by constructing cell-to-cell trajectories (Fig. 1d). We then use a smoothing spline model to predict cell fate trajectories and reduce both cell-cell variations. The smoothing spline method models the gene expression patterns in the aligned cell lineages from the previous step and builds the estimated individual cell fate trajectories. Lastly, we narrow down our focus to utilize the gene expression patterns from those cell fate trajectories to construct dynamic gene networks (Fig. 1e). The dynamic gene networks are constructed by estimating the dynamic relationship of pairwise gene expression patterns using the functional concurrent mod-

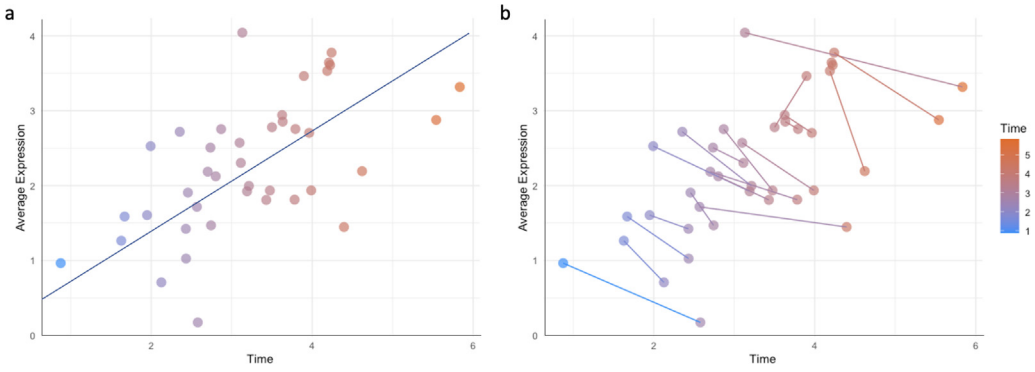

Fig. 2. Example of cell-to-cell linking: **a**: Cell differentiation over time (x-axis) reflects an increasing trend in average cell expressions (y-axis). **b**: The individual cell correspondences at different time points reflect a decreasing trend in average cell expression (y-axis).

els [30] and smoothing spline models [31]. The dynamic gene networks can be used to find critical genes by profiling genes with significantly different patterns from other genes.

Our major contribution is developing the first analysis framework (CellST) to construct cell lineages and predict dynamic cell fates at the individual cell level, which can help researchers better observe cell behaviors in the differentiation process. In contrast, the existing methods only estimate the bulk trajectory in scRNA-seq experiments. Those analysis methods may overlook the hidden patterns in the cell differentiation process to create a spurious cell differentiation trajectory. We illustrate this problem by using a simulated time-course cell dataset indicating the disadvantage of bulk cell trajectories (Fig. 2). The cell-to-cell trajectories are able to overcome the disadvantage and identify the real gene expression patterns in cell development. Under some cell development and differentiation circumstances, cells' average expressions show an increasing pattern if we only construct one average cell trajectory to order cells (Fig. 2a). However, when individual cells are aligned at different time points, the individual cell lineages' average expressions reflect unique decreasing patterns, which are in contrast to the bulk trajectory (Fig. 2b). This means some cells start at a lower expression level, and the expression keeps going down over time. Those cell development patterns can be easily misled by the average cell trajectory and thus reflect spurious cell differentiation behaviors. Furthermore, we propose the dynamic gene networks based on the individual cell fate trajectories to estimate the dynamic gene-gene relationship and critical genes in the differentiation process. The empirical performance of the proposed framework is evaluated by several simulated and real experiment studies.

2. Method

In this section, we introduce the Cell Smooth Transformation (CellST) method, which constructs the cell fate trajectories and dynamic gene networks for time-course scRNA-seq data.

2.1. Cell lineage & Individual cell fate trajectories

To construct the cell fate trajectories, we first align the cells at different time points to construct the cells' lineages information between time points. We then smooth the gene expression pattern for each gene over time and extract the "mean curve" of all individual gene expression patterns in single cell fate trajectories to obtain the general gene expression pattern.

2.1.1. Cell-to-cell lineages by optimal transport

Regarding cells at different time points as cells with genes of different domain spaces, we transform the problem of aligning cells at different time points into a problem of domain adaptation. Specifically, we denote the normalized gene expression for cell i at time t by a d-dimensional vector $\mathbf{x}_i^t$; each dimension of $\mathbf{x}_i^t$ represents a gene expression[1]. We write $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, where n_t indicates the number of cells at time t in single cell RNA-seq dataset. Our goal is to learn the transformation between the domain spaces by aligning the distribution of $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$.

As a powerful tool to learn the transformation from one probability measure to another, optimal transport has been applied to solve the domain adaptation problem [32]. We thus apply optimal transport to obtain the domain adaptive coupling between $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$. In other words, we transform the cell alignment problem into an optimal transport problem. In particular, we formulate the problem as a Monge optimal transport by minimizing the cost for transporting a gene expression distribution μ_t and μ_{t+1} using a map $\mathbf{T}_t$:

$$\min_{\mathbf{T}_t} \int_t c(x, \mathbf{T}_t(x)) \mathrm{d}\mu_t(x), \tag{1}$$

where $\mathbf{T}_t \# \mu_t = \mu_{t+1}$, # represents the push-forward operator, such that for any measurable x, $\mathbf{T}_t \# \mu_t(x) = \mu_t(\mathbf{T}_t^{-1}(x))$, μ_t and μ_{t+1} are probability distribution of $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$ in $\mathbb{R}^d$, where d is the dimension. We define the optimal transport map $\mathbf{T}_t : \mathbb{R}^d \to \mathbb{R}^d$, where $\mathbb{R}^d$ can be interpreted as the domain space for $\mathbf{x}_i^t$ or $\mathbf{x}_i^{t+1}$. In this optimal transport problem, one constraint for the transportation map $\mathbf{T}_t$ from a measure μ_t to a measure μ_{t+1} is the so-called measurement-preserving, i.e., $\mathbf{T}_t \# \mu_t = \mu_{t+1}$. Among all the measurement-preserving maps, the optimal $\mathbf{T}_t$ is the one that minimizes the transportation cost.

Since we can only observe gene expressions for sample cells at each time point, we focus on the case where the probability distributions are discrete. The distributions μ_t and μ_{t+1} for gene features at time points t and $t+1$ are defined as:

$$\mu_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta_{t_i} \quad \text{and} \quad \mu_{t+1} = \frac{1}{n_{t+1}} \sum_{j=1}^{n_{t+1}} \delta_{t+1_j}, \tag{2}$$

where δ_{t_i} and δ_{t+1_j} are the Dirac measures at location $\mathbf{x}_{t_i}$ and $\mathbf{x}_{t+1_j}$ respectively. Denote the positions of the supporting points $\mathbf{X}_t = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \dots, \mathbf{x}_{t_{n_t}})^T$. In discrete cases, the transport $\mathbf{T}_t$ from μ_t to μ_{t+1} can be denoted as $\mathbf{T}_t(\mathbf{X}_t) = \Sigma \mathbf{X}_t$, where Σ is an $n_{t+1} \times n_t$ matrix. For simplicity, we first consider the equal-size mapping, i.e., $n_t = n_{t+1} = n$. Notice that in this case, the transport between $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$ is a one-to-one assignment with permutation, Σ then can be regarded as a permutation matrix with the (i, j)th element:

$$\Sigma_{i,j} = \begin{cases} 1 & \text{if } \mathbf{T}_t(\mathbf{x}_{t_j}) = \mathbf{x}_{t+1_i}, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

[1] For $\mathbf{x}_i^t$, we (1) use all available genes, (2) select highly expressed genes, or (3) apply dimension reduction methods such as the principal component analysis (PCA). In the first two cases, each gene expression represents an individual gene; while in the last case, each gene feature represents a combination of all genes.

Furthermore, the transportation cost $C(\mathbf{T}_t)$ defined in (1) can be calculated as:

$$C(\mathbf{T}_t) = \sum_{i=1}^{n}\sum_{j=1}^{n} c\left(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_i}\right)\Sigma_{i,j}, \tag{4}$$

where $c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_j})$ can be interpreted as the energy required to transform an individual cell from the stage as $\mathbf{x}_j^t$ to the stage as $\mathbf{x}_i^{t+1}$. The optimal transport map $\mathbf{T}_t$ then can be calculated through:

$$\min_{\Sigma} \sum_{i=1}^{n}\sum_{j=1}^{n} c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_i})\Sigma_{i,j}. \tag{5}$$

where $c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_i}) = \left\|\mathbf{x}_{t_j} - \mathbf{x}_{t+1_i}\right\|^{\alpha}$ and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. We set $\alpha = 2$ in this paper. The minimum of the optimization problem (5) is called the L^{α}-Wasserstein distance (to the power α) and is denoted by $W_{\alpha}(\mu_t, \mu_{t+1})^{\alpha}$. The W_{α} defines a distance on the set of distributions (cells) that have moments of order α. In general, the cell lineage construction by optimal transport can be summarized as three steps: Estimating empirical gene feature distributions μ_t and μ_{t+1} as in (2). Finding an optimal transport map $\mathbf{T}_t$ from μ_t to μ_{t+1} through (5). Applying $\mathbf{T}_t$ to obtain the cell-to-cell coupling from $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$.

When dealing with large-scale studies involving numerous sequenced cells or measured genes, several more efficient algorithms can be used to calculate the optimal transport map, including the Sinkhorn algorithm (Cutur., 2013), sliced-Wasserstein methods [26,54], Hilbert curve-based method [56], fast Sinkhorn method [58].

It's important to note that the optimal transport map discussed above could be unsuitable in some cases when the one-to-one cell differentiation assumption does not hold. However, our optimal transport framework can be modified to account for more general cell-to-cell relationships. Specifically, we consider the following two general scenarios. First, the number of cells may vary at different time points, and as a result, some cells may need to be reused when constructing cell-to-cell lineages. This can lead to multiple lineages passing through a single cell at specific time points. Second, cells may exhibit different proliferation rates at the same time points, i.e., the numbers of new cells produced by two cells at the same time point could be significantly different. This could result in the varying proportions of cell groups across time. To address these general assumptions, we present a comprehensive discussion of our generalized methods and experimental results in the supplementary materials.

2.1.2. Individual cell fate trajectories by smoothing spline model

After we align the cells from different time points, we can obtain the individual cell lineages at time points t and $t+1$. We then align cells for all time points based on the cell couplings to construct each cells coarse cell fate trajectories across the timeline. Those cell fate trajectories are smoothed to reduce the estimation variance in CellST by utilizing the smoothing spline models. The smoothing spline model is a versatile family of smoothing methods that are suitable for both univariate and multivariate problems [33]. To construct the proposed smoothed cell trajectories, we use equation 6 to model the behavior patterns of the gene expression along the cell fate trajectories. Let t represent the time points in the time-course dataset, and g_i represent the gene expression for each gene within an aligned cell fate trajectory. For co-expressed genes, we model the gene expression patterns using a smoothing spline mix-effect model with $\{g_i, t_i\}_{i=1}^n$ as the observations [31]:

$$g_i = \eta(t_i) + \mathbf{z}_i^T\mathbf{b} + \varepsilon_i \tag{6}$$

$i = 1, \ldots, n$, where the regression function $\eta(t_i)$ is assumed to be a smooth function on the genes domain space in a cell. $\eta(t_i)$ are the fixed effects and $\mathbf{z}_i^T\mathbf{b}$ are the random effects with $\mathbf{b} \sim N(\mathbf{0}, B)$ and $\varepsilon_i \sim N(0, \sigma^2)$. The random effects are used to account for the co-expressed genes in one individual cell trajectory. The model terms $\eta(t)$ or $\eta(t) + \mathbf{z}^T\mathbf{b}$ will be estimated using the penalized (unweighted) least squares method through the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\left(g_i - \eta(t_i) - \mathbf{z}_i^T\mathbf{b}\right)^2 + \frac{1}{n}\mathbf{b}^T\Sigma\mathbf{b} + \lambda J(\eta), \tag{7}$$

where the first term measures the goodness-of-fit, $J(\eta) = \int(\eta''(t))^2 dt$ quantifies the smoothness of η, and λ is the smoothing parameter controlling the trade-off between the goodness-of-fit and the smoothness of η [33,34]. Consider the minimization of the least squares estimation (equation 7) in a space with basis $\{\xi_1, \ldots, \xi_q\}$, function η can be expressed as

$$\eta(t) = \sum_{j=1}^{d} c_j\xi_j(t) = \boldsymbol{\xi}^T(t)\mathbf{c}. \tag{8}$$

Plugging equation 8 into equation 7, thus η can be estimated by minimizing:

$$(\mathbf{g} - R\mathbf{c} - Z\mathbf{b})^T(\mathbf{g} - R\mathbf{c} - Z\mathbf{b}) + \mathbf{b}^T\Sigma\mathbf{b} + n\lambda\mathbf{c}^T Q\mathbf{c}. \tag{9}$$

With the standard formulation of penalized least squares regression, the minimization of equation 7 is performed in a so-called reproducing kernel Hilbert space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a square seminorm, and the solution resides in the space $\mathcal{N}_J \oplus \text{span}\left\{R_J(t_i, \cdot), i = 1, \ldots, n\right\}$, where $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ is the null space of $J(\eta)$ and $R_J(\cdot,\cdot)$ is the so-called reproducing kernel in $\mathcal{H} \ominus \mathcal{N}_J$. The solution has an expression:

$$\eta(t) = \sum_{i=1}^{m} d_v\phi_v(t) + \sum_{i=1}^{n} \tilde{c}_i R_J(t_i, t) \tag{10}$$

where $\{\phi_v\}_{v=1}^m$ is a basis of $\mathcal{N}_J$. It follows that $R = (S, \tilde{Q})$, where S is $n\times m$ with the (i,v)th entry $\phi_v(t_i)$ and $\tilde{Q}$ is $n\times n$ with the (i,j) th entry $R_J(t_i, t_j)$. In the smoothing spline model, the estimation of η is highly related to the choosing of the smoothing parameter λ. We choose the smoothing parameter λ and estimate random effect $\mathbf{b}$ by Generalized Cross-Validation (GCV) [31,34]. Since there are d gene expression patterns over t time points for the cell fate trajectories, the smoothing spline model estimates one expression pattern for individual cells and smooths the expression patterns.

2.1.3. Dynamic gene networks

We consider the connection of two genes to be dynamic and the relationship may smoothly change. Suppose we want to study the dynamic relationship of the lth gene and sth gene, where $1 \le l, s \le p, l \ne s$. Denote $X_i^{\langle l\rangle}(t)$ and $X_i^{\langle s\rangle}(t)$ as the lth gene and sth gene's expression values of cell fate trajectories i, and $i = 1, \cdots, n$. By taking lth gene as the response and sth gene as the covariate, we consider the functional concurrent linear model,

$$X_i^{\langle l\rangle}(t) = \beta^{\langle l,s\rangle}(t)X_i^{\langle s\rangle}(t) + \varepsilon_{i,t}^{\langle l,s\rangle}, \tag{11}$$

where $\beta^{\langle l,s\rangle}(t)$ models the dynamic linear relationship between two genes, $\varepsilon_{i,t}^{\langle l,s\rangle}$s are i.i.d. random errors with mean zero and constant variance. We estimate $\beta^{\langle l,s\rangle}(t)$ by minimizing the following penalized least squares function,

$$\frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(X_i^{\langle l\rangle}(t_{ik}) - \beta^{\langle l,s\rangle}(t_{ik})X_i^{\langle s\rangle}(t_{ik})\right)^2 + \lambda J(\beta^{\langle l,s\rangle}). \tag{12}$$

With the representer theorem [34], the optimizer of 12 can be written as

$$\hat{\beta}_{\lambda}^{\langle l,s\rangle}(t) = \sum_{v=1}^{m} d_v\psi_v(t) + \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}R_1(t_{ik}, t) \tag{13}$$

where $\{\psi_v\}_{v=1}^m$ is the basis function of the m-dimensional null space $\mathcal{H}_0$, and $R_J(\cdot,\cdot)$ is the reproducing kernel of $\mathcal{H}_1$. Moreover, d_v and c_{ik} are the coefficients to be estimated. By Plugging equation 13 to equation 12, we can yield the estimations of $\boldsymbol{c} = (c_1, \cdots, c_{1K}, \cdots, c_{n1}, \cdots, c_{nK})^T$ and $\boldsymbol{d} = (d_1, \cdots, d_{1K}, \cdots, d_{n1}, \cdots, d_{nk})^T$, which follow

$$\begin{aligned}\boldsymbol{c} &= \left(\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\right)\mathbf{X}^{\langle s\rangle}\boldsymbol{X}^{\langle l\rangle}\\ \boldsymbol{d} &= \left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\boldsymbol{X}^{\langle l\rangle}\end{aligned} \tag{14}$$

where $\mathbf{X}^{\langle s\rangle} = diag((\boldsymbol{X}_1^{\langle s\rangle T}, \cdots, \boldsymbol{X}_n^{\langle s\rangle T}))$ with the vector $\boldsymbol{X}_i^{\langle s\rangle T} = (X_i^{\langle s\rangle}(t_{i1}), \cdots, X_i^{\langle s\rangle}(t_{iK}))^T$, $\boldsymbol{X}^{\langle l\rangle} = (\boldsymbol{X}_1^{\langle l\rangle T}, \cdots, \boldsymbol{X}_n^{\langle l\rangle T})^T$ with the vector $\boldsymbol{X}_i^{\langle l\rangle} = (X_i^{\langle l\rangle}(t_{i1}), \cdots, X_i^{\langle l\rangle}(t_{iK}))^T$, $\mathbf{S} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)^T$ with the (k, v)th entry of the $K \times m$ matrix $\mathbf{S}_i$ equals to $\psi_v(t_{ik})X_i^{\langle s\rangle}(t_{ik})$, $\mathbf{M} = \mathbf{X}^{\langle s\rangle}\mathbf{Q}\mathbf{X}^{\langle s\rangle} + n\lambda\mathbf{I}$ and $\mathbf{Q}$ is the $nK \times nK$ block matrix with the (i, j)th block is the $K \times K$ matrix with the (k, u)th entry equals to $R_1(t_{ik}, t_{ju})$. Thus, the estimation of $\beta^{\langle l,s\rangle}(t)$ can be written as

$$\hat{\beta}^{\langle l,s\rangle}(t) = \boldsymbol{\psi}^T\boldsymbol{d} + \boldsymbol{\xi}^T\boldsymbol{c} \tag{15}$$

where $\boldsymbol{\psi} = (\psi_1(t), \cdots, \psi_m(t))^T$ and $\boldsymbol{\xi} = (R_1(t_{11}, t), \cdots, R_1(t_{1K}, t), \cdots, R_1(t_{n1}, t), \cdots, R_1(t_{nK}, t))^T$. Note that $\beta^{\langle l,s\rangle}(t)$ models the dynamic linear relationship between lth gene and sth gene, and $\beta^{\langle l,s\rangle}(t_0) = 0$ means the correlation between gene l and gene s to be 0 at the time point t_0. We then derive the $100(1-\alpha)\%$ confidence band of $\beta^{\langle l,s\rangle}(t)$. We adopt the Bayes model in [33] and get the posterior variance of $\beta^{\langle l,s\rangle}(t)$ satisfies

$$\mathrm{Var}\left[\beta^{\langle l,s\rangle}(t) \mid \mathbf{X}, \mathbf{X}^{\langle l\rangle}\right] = \frac{\sigma^2}{nK\lambda}\left(R_1(t,t) + \boldsymbol{\psi}^T\left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\boldsymbol{\psi} - 2\boldsymbol{\psi}^T\boldsymbol{d}_\xi - \boldsymbol{\xi}^T\boldsymbol{c}_\xi\right) \tag{16}$$

where

$$\begin{aligned}\boldsymbol{c}_\xi &= \left(\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\right)\mathbf{X}^{\langle s\rangle}\boldsymbol{\xi} \\ \boldsymbol{d}_\xi &= \left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\mathbf{X}^{\langle s\rangle}\boldsymbol{\xi}\end{aligned} \tag{17}$$

Using equation (16), we can estimate the posterior variance of $\beta^{\langle l,s\rangle}(t_0)$ and write as $\gamma^{\langle l,s\rangle}(t_0)$. We then construct the $100(1-\alpha\%)$ Bayesian confidence interval (BCI) of $\beta^{\langle l,s\rangle}(t)$: $BCI_{\langle l,s\rangle}(t) := \hat{\beta}^{\langle l,s\rangle}(t) \pm z_{\alpha/2}\sqrt{\gamma^{\langle l,s\rangle}(t)}$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile for standard normal distribution.

We use Bayesian confidence intervals to construct the dynamic graph, where a node represents a gene, and an edge between two nodes exists if the two corresponding genes follow the model (11) with non-zero coefficient $\beta^{\langle l,s\rangle}(t)$.

2.2. Test differentially expressed genes

We integrate a functional ANOVA test method [35] in our framework to estimate deferentially expressed genes based on the constructed cell fate trajectories. For each gene in those cell fate trajectories, we consider independent vectors of the random function $\mathbf{X}_{ki}(t) = \left(X_{ki1}(t), \ldots, X_{kid}(t)\right)^\top$, where k indicates the number of trajectory groups, i indicates cells and d indicates the number of genes in one individual cell trajectory, defined over the interval I. In the multivariate analysis of variance problem for functional data (FMANOVA), we test the following hypothesis

$$\begin{aligned}&H_0 : \boldsymbol{\mu}_1(t) = \cdots = \boldsymbol{\mu}_k(t), t \in I, \\ &H_A : \boldsymbol{\mu}_1(t) \neq \cdots \neq \boldsymbol{\mu}_k(t), t \in I.\end{aligned} \tag{18}$$

Wilks lambda test statistics for testing significantly different genes are approximated using the fdANOVA method [35]. The null distributions of test statistics are approximated by $F_{(l-1)\kappa,(n-l)\kappa}$-distribution, κ are estimated by the naive and biased-reduced methods [36]. The p-value is given by $P\left(F_{(l-1)\kappa,(n-l)\kappa} > F_n\right)$, where F_n denotes the test statistic. P-values for all genes tested are corrected by Benjamini & Yekutieli method [37].

3. Results

We evaluated the performance of the CellST framework on cell lineage tracing and cell fate prediction in both simulated and real scRNA-seq experiments. The simulation analysis was conducted in two scenarios: Firstly, we simulated scRNA-seq datasets with cells at two time points to only investigate the accuracy of constructed cell-to-cell correspondence between time points. Secondly, we simulated a time-course scRNA-seq dataset with multiple time points to examine individual cell differentiation patterns in the cell fate trajectories. For real scRNA-seq experiments, we conducted cell lineage tracing in a single-cell mouse hematopoietic system experiment [38]. Moreover, we evaluated the entire proposed framework on a scRNA-seq experiment for zebrafish cell embryogenesis [39].

3.1. Simulation

3.1.1. Reconstruct cell-to-cell correspondence along time points

To investigate the accuracy of cell aligning, we simulated scRNA-seq experiments with only two different time points. The simulated datasets, which contain the same number of cells and cell types, were generated independently for each time point. These simulation datasets contain five same cell types in both time points. In the simulation setting, the number of cells ranges from 200 to 600, and the number of genes in one cell ranges from 100 to 500. The cell alignment and cell-to-cell correspondences were constructed using the CellST based only on the gene expression information of cells at each time point and no information on the benchmark labels of cell types. Specifically, we estimated an empirical transportation cost for the individual cell alignment between two time points using the gene expressions in cells. We aligned cells by selecting pairs with the smallest transportation cost. Since cell dynamics is a continuous development process and cells within the same cell type tend to have similar gene expression profiles, the cell aligning accuracy can be validated by counting the number of aligned cell pairs with the same cell type (Fig. 3b).

We noticed that the accuracy of the cell aligning method has an increasing trend as we added more genes in cells for the simulated data. This observation is due to the fact that the CellST gets more information to learn the patterns of genes when more genes are simulated in each cell. Similarly, increasing cell numbers will also increase the aligning accuracy since cells can be treated as information replicates to enhance the accuracy. We also compared the accuracy of coupled cells with the Euclidean distance and Pearson's correlation. Those two methods are the most commonly used distances or similarity measures for gene expression analysis. [40–42]. The accuracy comparison results (Fig. 3b) show the CellST method achieves the best cell aligning accuracy in the simulation settings. In summary, the cell aligning method achieves high accuracy and captures the significant gene expressions when aligning cells and constructing individual cell correspondences at two different time points. Accurate cell alignment is crucial for the down-streaming individual cell fate prediction when aligned cells are transformed into a trajectory over time.

3.1.2. CellST estimate cell fates in two simulated pathways

To investigate the effectiveness in predicting the individual cell fates, we simulated a time-course scRNA-seq data with 13 experimental time stamps, and 160 cells were simulated at each time stamp. This simulation dataset has two pathways with distinct development expression patterns, and each pathway contains 100 genes. The first pathway was created using the contact inhibition genes that keep cells growing into only a layer one cell thick (mono-layer) [43,44]. The growth of cells' average expression in this simulated pathway diminishes and approaches an equilibrium expression over time. We simulated the second pathway according to the cellular division process, which is more active in cells under mitosis and less active in cells in interphase [45]. Eighty cells contain only the contact inhibition pathway at each time point, and eighty cells contain only the cellular division pathway.

To observe and predict the dynamic cell fates, we utilized cells' experimental time information and built individual cell fate trajectories using CellST. The cell lineages between adjacent time points were constructed using cell lineage tracing in CellST (Fig. 4a). Those connected cells were then smoothed using the smoothing spline technique in CellST to estimate the cell fate trajectories. Fig. 4b illustrates the estimated individual cell fate trajectories (red curves). Based on the two distinct pathways, the two types of cells are automatically well separated by CellST. The expressions of cell fate trajectories were compared

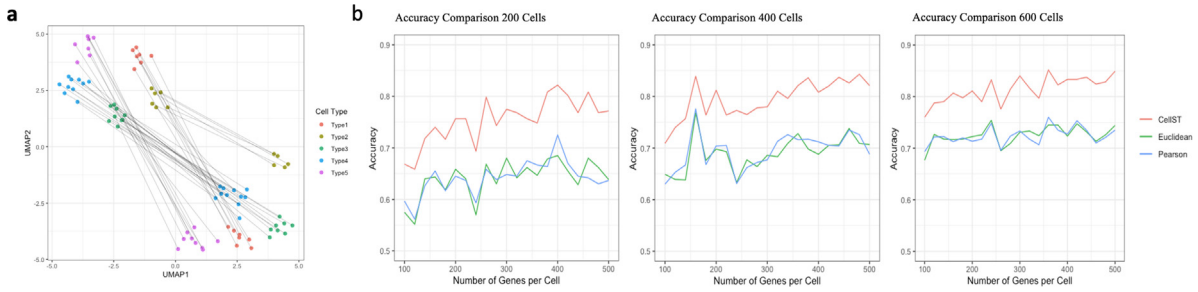

Fig. 3. A simulation example of cell aligning process at two time points. **a**: Cell-to-cell alignment with five (right) cell types in both time points. **b**: Accuracy comparison of the cell aligning process (red) with other gene similarity measurements (Pearson correlation (blue) and Euclidean distance (green)).

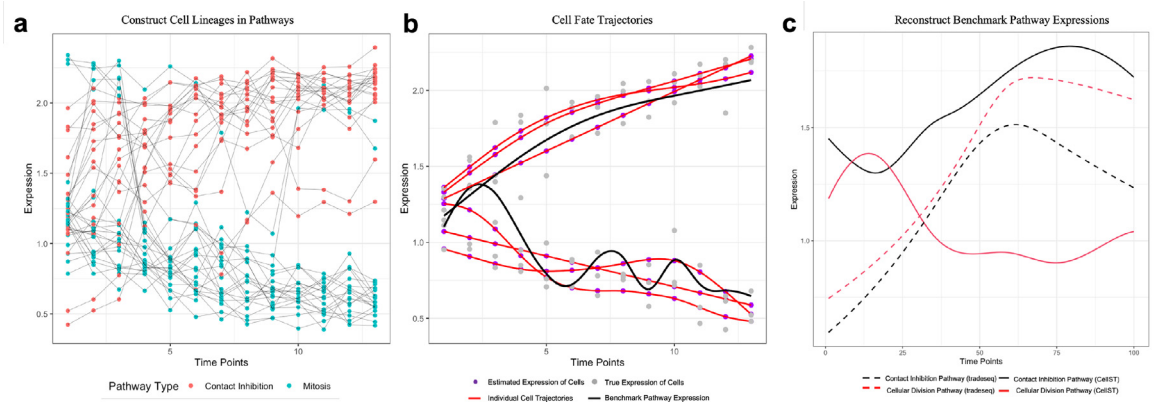

Fig. 4. **a**: Cell couplings through all time points constructed by the CellST method. The cells are classified by the pathway they contain. **b**: The cell fate trajectories (red curves) built by CellST and Benchmark average expression cell trajectory(black curve). **c**: Development expression patterns for a simulated gene (m_67). The red and black curves estimated by the CellST method indicate the gene expression in two different pathways. The dotted two curves are constructed by the tradeseq method.

with the benchmark pathway expression patterns (black curves). The expression of cell fate trajectories illustrates consistent patterns with the benchmark expression of the two simulated pathways over time. In addition to the consistency, we observed that cells have unique behaviors over time from the cell fate trajectories. Some cells grow slower and have lower expression values, while others grow faster and have higher expression values than the simulated average development patterns. The cell fate trajectories predict the unique cell development behaviors by smoothing the constructed cell lineages to reduce cell-cell variance.

Next, we performed a comparative analysis of CellST with the existing trajectories analysis method “tradeseq” [46]. The “tradeseq” is a trajectory-based method to estimate the dynamic expressions of differentially expressed genes. By comparing the gene expression patterns constructed by CellST and tradeseq (Fig. 4c), we notice that the tradeseq method constructed two similar expression patterns for a simulated gene expression (dotted curves), while the CellST method built two distinct expression patterns (black and red curves). Those constructed dynamic gene expression curves by CellST are also consistent with the simulated benchmark expressions by showing distinct expression patterns. When constructing the cell fate trajectories, the CellST method can automatically classify cells that contain different pathway expressions and construct cell correspondences within the same pathway.

3.2. Real scRNA-Seq Experiments

3.2.1. CellST construct accurate cell lineages

We applied the proposed method on a single cell mouse hematopoietic system experiment to evaluate the effectiveness of constructing cell lineages [47]. The dataset includes three experimental time points and the cells defined a continuous state map spanning from multipotent progenitors (MPPs) to nine mature cell types, including erythrocytes (Er), megakaryocytes (Mk), basophils (Ba), mast cells (Ma), eosinophils (Eos), neutrophils (Neu), monocytes (Mo), plasmacytoid dendritic cells (pDCs), Ccr7+ migratory DCs (migDCs), and lymphoid precursors (Ly) (Fig. 5a). We constructed the cell developing lineages for neutrophils (Neu) cell type and compared the results with the benchmark cell pseudo-time from the original experiment (Fig. 5a). CellST connected cells through the three experimental time points to represent the developing process’s penitential cell lineages. Specifically, we estimated an empirical transportation cost for the individual cell correspondence between two time points using the gene expressions in cells and then aligned cells by selecting pairs with the smallest transportation cost.

Since cell development is a gradual process and cells within the adjacent time points tend to have similar gene expression profiles, we measured the differences from cell pseudo-time in all constructed cell lineages. We compared the distribution of cell pseudo-time with Pearson’s correlation method, which has been widely used to measure similarity between two cells [40–42]. Comparing with Pearson’s correlation method, we observed that CellST constructs cell lineages with higher cell similarities, in which the changes in pseudo-time of connected cells are gathering near zero (Fig. 5b and c). The results indicate that CellST can observe the gradually step-wise developing behaviors of cells at each time point. CellST constructed cell lineages based on the cell-cell correspondence connection to represent the cell-developing behaviors throughout the time points.

3.3. Discover critical genes in zebrafish cell embryogenesis

To further investigate individual cell differentiation behaviors and gene-gene relationships, we performed CellST on a zebrafish embryogenesis scRNA-seq dataset. This dataset contains 38,731 cells and 11,588

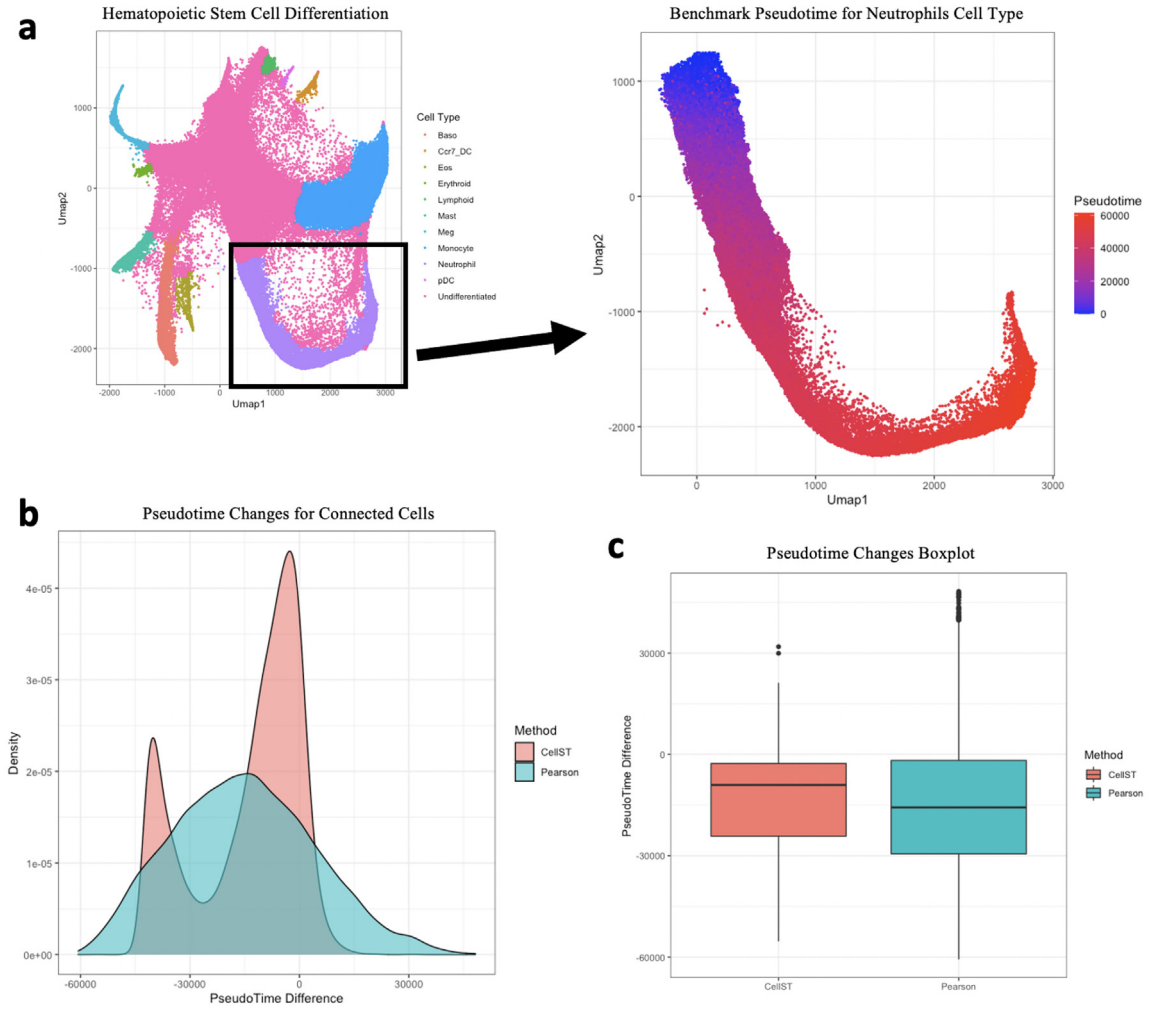

Fig. 5. Constructing individual cell lineages with mouse hematopoietic system. **a**: scRNA-seq dataset for hematopoietic stem cell differentiation. We specifically focus on the cell differentiation of neutrophils (Neu) mature cell types. **b**: The distribution pseudo time difference between aligned individual cells between two time points. The distribution has been compared with the Pearson Correlation method, which measures the similarities between cells. **c**: Boxplot comparison for the pseudo time difference between CellST and Pearson correlation method.

genes of early zebrafish development using Drop-seq [39]. Samples in the dataset are from the high blastula stage (3.3 hours post-fertilization) when most cells are pluripotent, to the six-somite stage (12 hours post-fertilization), when many cells have differentiated into different cell types. We observed that cells were clustered together at the beginning high blastula stage and differentiated into different cell types in later development stages. Since we are constructing cell trajectories for multiple cell types simultaneously, the proliferation rates potentially vary across cells. To address this, we utilize our generalized CellST method to construct cell fate trajectories (Fig. 6a), which capture the unique individual cell development behaviors under more general scenarios. This method also enables us to estimate the proliferation rate for each individual cell. The histograms of the normalized proliferation rates at different time points are presented in Fig. 6b. Note that the mean of the normalized proliferation rates at each time point equals 1, while the variance indicates the heterogeneity level of the proliferation rates among cells at the corresponding time point. We notice that this heterogeneity is significantly high in the early stages of the cell differentiation process and decreases gradually over time.

Unlike the bulk cell trajectory, the CellST cell fate trajectories achieved full cell development coverage for all cells. The full coverage indicates that the cell fate trajectories can reveal less frequent cell development patterns overlooked by the bulk cell trajectory. The CellST constructed cell fate trajectories throughout the stages and illustrated the unique individual cell development behaviors. The cell fate trajectories return each cell's potential cell fate paths into different cell types throughout the 12 developmental stages.

Furthermore, as cells developed into multiple cell types at the 12.0-6-somite stage (last developmental stage), we built trajectory groups to those cell fate paths by CellST according to the cell types in the last developmental stage. We constructed the dynamic gene networks (Fig. 6c) for each group of cell fate trajectories. In those dynamic networks, we observed some genes that behave significantly differently from other genes (Fig. 6c) as cells developed into different cell types. For instance, MYF5, MYOD1, and KRT18 genes appeared to behave differently in two of the trajectory groups in later developmental stages. The MYF5 is a protein with a key role in regulating muscle differentiation or myogenesis, specifically the development of skeletal muscle [48,49], and MyoD1 is a key regulator that orchestrates skeletal muscle differentiation through the regulation of gene expression [50,51]. Moreover, KRT18 regulates the epithelial cell differentiation process [52,53].

We then visualized and validated the expressions of those critical genes (Fig. 6d and Fig. 6e). The expressions of MYF5 and MYOD1 genes are significantly higher in trajectory group 5 versus in other trajectory groups, which is consistent with the discovery in the CellST dynamic

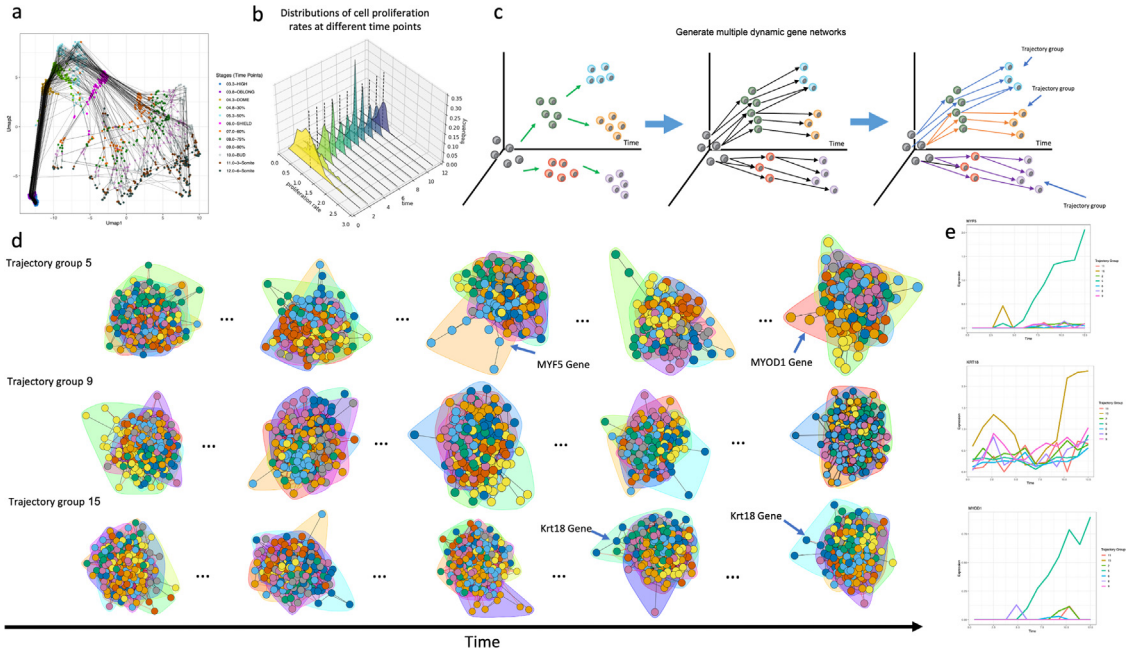

Fig. 6. **a**: We constructed the individual cell fate trajectories by connecting cells through the 12 developing stages. **b**: Histograms of the distribution of the cell proliferation rates at different time points. **c**: Illustration of identifying multiple cell trajectory groups based on the cell types at the 12.0-6-somite stage (last developmental stage). **d**: Example of critical genes identified using CellST dynamic gene network. **e**: Dynamic gene networks constructed by CellST for the cell trajectory groups.

Table 1
Top five gene functional annotation groups.

Gene Ontology (GO) annotations	Count	P-value
Multicellular organism development	133	2.100393e-45
Cell Differentiation	86	4.8e-25
Differentiation	36	8.6e-10
Cell fate specification	13	1.3e-5
Cell fate commitment	12	6.1e-5

networks. KRT18 is highly expressed in trajectory group 15, which is also consistent with the CellST dynamic network results. Additionally, we performed functional deferentially expressed gene tests based on the CellST cell fate trajectories. We discovered a total of 268 differentially expressed genes in this zebrafish cell development process dataset. We performed gene ontology annotations to those genes (Table 1), and the function of those genes is highly related to regulating the cell development/differentiation process.

Those results proved that the cell fate trajectories and dynamic gene networks in the CellST method can be used to discover critical genes in a cell differentiation process. We also demonstrated the CellST cell fate trajectories have full coverage on different cell lineages even in some rare cell types since the trajectories track individual cell behaviors. Lastly, those individual cell fate trajectories reflect unique gene expression patterns when cells develop into different mature cell types.

4. Discussion

Understanding the dynamic of cell differentiation throughout a period is crucial for future research in scRNA-Seq analysis. We developed a novel machine learning analysis framework, CellST, to build cell fate trajectories and dynamic gene networks for time-course scRNA-seq datasets. The cell fate trajectories enabled researchers to observe the individual cell development behaviors and better use the benefit of the single cell sequencing technology. Compared to the existing bulk single-cell trajectory, we brought the cell development analysis into a more precise and unprecedented resolution. The dynamic gene networks estimated the dynamic relationship of genes and discovered potential critical genes during cell differentiation processes.

There are three major advantages of the CellST analysis framework. Firstly, the cell lineages were constructed with high accuracy and provides unique individual cell differentiation behaviors between time points. Secondly, since the trajectory tracks individual cells, the cell fate trajectories will have full coverage on different cell lineages even in some rare cell types (Fig. 6a). Thirdly, the dynamic gene networks analysis in the CellST framework can accurately estimate gene-gene relationships and discover critical genes in the cell differentiation process. Through the simulation and real dataset analysis, We constructed cell fate trajectories in single cell RNA-seq experiment and various dynamic cell differentiation behaviors were observed.

Code availability

The CellST R package and example for constructing the cell fate trajectories and dynamic gene networks analysis can be found on GitHub (https://github.com/zhanzmr/CellST).

Author information

Corresponding authors: Correspondence to Jingyi Zhang and Ping Ma.
Co-first author: Mengrui Zhang, Yongkai Chen

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data used in this paper are publicly available datasets. The mouse hematopoietic system experiment dataset can be found on

GitHub (https://github.com/AllonKleinLab/paper-data). The zebrafish embryogenesis dataset from the original paper [24] can be found in the NCBI database with accession number GSE106587.

Acknowledgements

This work was supported by National Science Foundation grants DMS-1903226, DMS-1925066, DMS-2124493 and NIH grants R01GM1222080. The authors acknowledge the contribution of Dr. Jingyi Zhang for the grant "National Key R&D Program of China (No. 2021YFA1001300)".

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100068.

References

[1] Spiller DG, Wood CD, Rand DA, White MR. Measurement of single-cell dynamics. Nature 2010;465(7299):736–45.
[2] Guo J, Grow EJ, Yi C, Mlcochova H, Maher GJ, Lindskog C, Murphy PJ, Wike CL, Carrell DT, Goriely A, et al. Chromatin and single-cell rna-seq profiling reveal dynamic signaling and metabolic transitions during human spermatogonial stem cell development. Cell Stem Cell 2017;21(4):533–46.
[3] Burrows N, Bashford-Rogers RJ, Bhute VJ, Peñalver A, Ferdinand JR, Stewart BJ, Smith JE, Deobagkar-Lele M, Giudice G, Connor TM, et al. Dynamic regulation of hypoxia-inducible factor-1α activity is essential for normal b cell development. Nature Immunol 2020;21(11):1408–20.
[4] Nawy T. Single-cell sequencing. Nature Methods 2013;11(1):18.
[5] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Rev Genet 2013;14(9):618–30.
[6] Grün D, Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell 2015;163(4):799–810.
[7] Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature 2017;541(7637):331–8.
[8] Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, Zhou A, Eyob H, Balakrishnan S, Wang C-Y, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature 2015;526(7571):131–5.
[9] Hrvatin S, Hochbaum DR, Nagy MA, Cicconet M, Robertson K, Cheadle L, Zilionis R, Ratner A, Borges-Monroy R, Klein AM, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nature Neurosci 2018;21(1):120–9.
[10] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature Rev Genet 2015;16(3):133–45.
[11] Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 2020;182(1):145–61.
[12] Ren G, Jin W, Cui K, Rodrigez J, Hu G, Zhang Z, Larson DR, Zhao K. Ctcf-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. Mol Cell 2017;67(6):1049–58.
[13] Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mrna quantification and differential analysis with census. Nature Method 2017;14(3):309.
[14] Cannoodt R, Saelens W, Saeys Y. Computational methods for trajectory inference from single-cell transcriptomics. Eur J Immunol 2016;46(11):2496–506.
[15] Trapnell C. Defining cell types and states with single-cell genomics. Genome Res 2015;25(10):1491–8.
[16] Ji Z, Ji H. Tscan: pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. Nucl Acid Res 2016;44(13). e117–e117
[17] Chen H, Albergante L, Hsu JY, Lareau CA, Bosco GL, Guan J, Zhou S, Gorban AN, Bauer DE, Aryee MJ, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. Nature Commun 2019;10(1):1–14.
[18] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnol 2014;32(4):381.
[19] Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R, Chen T. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nature Commun 2017;8(1):1–9.
[20] Tritschler S, Büttner M, Fischer DS, Lange M, Bergen V, Lickert H, Theis FJ. Concepts and limitations for learning developmental trajectories from single cell genomics. Development 2019;146(12).
[21] Moon KR, Stanley III JS, Burkhardt D, van Dijk D, Wolf G, Krishnaswamy S. Manifold learning-based methods for analyzing single-cell rna-sequencing data. Curr Opin Syst Biol 2018;7:36–46.
[22] Dai K, Damodaran K, Venkatachalapathy S, Soylemezoglu AC, Shivashankar G, Uhler C. Predicting cell lineages using autoencoders and optimal transport. PLoS Comput Biol 2020;16(4):e1007828.
[23] Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nature Biotechnol 2019;37(5):547–54.
[24] Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science 2018;360(6392):981–7.
[25] Villani C. Topics in optimal transportation. American Mathematical Soc; 2003.
[26] Meng C, Ke Y, Zhang J, Zhang M, Zhong W, Ma P. Large-scale optimal transport map estimation using projection pursuit. In: Advances in Neural Information Processing Systems; 2019. p. 8116–27.
[27] Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, Gould J, Liu S, Lin S, Berube P, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell 2019;176(4):928–43.
[28] Tong A, Huang J, Wolf G, van Dijk D, Krishnaswamy S. Trajectorynet: a dynamic optimal transport network for modeling cellular dynamics. arXiv Preprint arXiv:200204461 2020.
[29] Zhang J, Zhong W, Ma P. A review on modern computational optimal transport methods with applications in biomedical research. arXiv preprint arXiv:200802995 2020.
[30] Wang J-L, Chiou J-M, Müller H-G. Functional data analysis. Annual Rev Stat Appl 2016;3:257–95.
[31] Gu C, Ma P. Optimal smoothing in nonparametric mixed-effect models. Annal Stat 2005;33(3):1357–79.
[32] Courty N, Flamary R, Tuia D. Domain adaptation with regularized optimal transport. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2014. p. 274–89.
[33] Gu C. Smoothing spline ANOVA models, vol 297. Springer Science & Business Media; 2013.
[34] Wahba G. Spline models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. Philadelphia: SIAM; 1990.
[35] Górecki T, Smaga Ł. fdanova: an r software package for analysis of variance for univariate and multivariate functional data. Comput Stat 2019;34(2):571–97.
[36] Zhang J. Analysis of variance for functional data. Monograph Stat Appl Probab 2014;127:127.
[37] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annal Stat 2001:1165–88.
[38] Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. Fundamental limits on dynamic inference from single-cell snapshots. Proc Natl Acad Sci 2018;115(10):E2467–76.
[39] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161(5):1202–14.
[40] Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature Method 2016;13(3):229–32.
[41] Klimovskaia A, Lopez-Paz D, Bottou L, Nickel M. Poincaré maps for analyzing complex hierarchies in single-cell data. Nature Commun 2020;11(1):1–9.
[42] Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single–cell transcriptomics. Nature Method 2019;16(5):381–6.
[43] Pavel M, Renna M, Park SJ, Menzies FM, Ricketts T, Füllgrabe J, Ashkenazi A, Frake RA, Lombarte AC, Bento CF, et al. Contact inhibition controls cell survival and proliferation via yap/taz-autophagy axis. Nature Commun 2018;9(1):1–18.
[44] Mendonsa AM, Na T-Y, Gumbiner BM. E-cadherin in contact inhibition and cancer. Oncogene 2018;37(35):4769–80.
[45] Tomasetti C, Durrett R, Kimmel M, Lambert A, Parmigiani G, Zauber A, Vogelstein B. Role of stem-cell divisions in cancer risk. Nature 2017;548(7666):E13–14.
[46] Van den Berge K, De Bezieux HR, Street K, Saelens W, Cannoodt R, Saeys Y, Dudoit S, Clement L. Trajectory-based differential expression analysis for single-cell sequencing data. Nature Commun 2020;11(1):1–13.
[47] Weinreb C, Rodriguez-Fraticelli A, Camargo FD, Klein AM. Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science 2020;367(6479):eaaw3381.
[48] Esteves de Lima J, Relaix F. Master regulators of skeletal muscle lineage development and pluripotent stem cells differentiation. Cell Regenerat 2021;10:1–13.
[49] Agarwal M, Bharadwaj A, Mathew SJ. Tle4 regulates muscle stem cell quiescence and skeletal muscle differentiation. J Cell Sci 2022;135(4):jcs256008.
[50] Blum R, Vethantham V, Bowman C, Rudnicki M, Dynlacht BD. Genome-wide identification of enhancers in skeletal muscle: the role of myod1. Genes Dev. 2012;26(24):2763–79.
[51] Agaram NP, LaQuaglia MP, Alaggio R, Zhang L, Fujisawa Y, Ladanyi M, Wexler LH, Antonescu CR. Myod1-mutant spindle cell and sclerosing rhabdomyosarcoma: an aggressive subtype irrespective of age. a reappraisal for molecular classification and risk stratification. Modern Pathol 2019;32(1):27–36.
[52] Jiang P, Gil de Rubio R, Hrycaj SM, Gurczynski SJ, Riemondy KA, Moore BB, Omary MB, Ridge KM, Zemans RL. Ineffectual type 2–to–type 1 alveolar epithelial cell differentiation in idiopathic pulmonary fibrosis: persistence of the krt8hi transitional state. Am J Respirat Crit Care Med 2020;201(11):1443–7.
[53] Liu Y, Li J, Yao B, Wang Y, Wang R, Yang S, Li Z, Zhang Y, Huang S, Fu X. The stiffness of hydrogel-based bioink impacts mesenchymal stem cells differentiation toward sweat glands in 3d-bioprinted matrix. Mater Sci Eng: C 2021;118:111387.
[54] Bonneel Nicolas, Rabin Julien, Peyré Gabriel, Pfister Hanspeter. Sliced and radon wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision 2015;51:22–45.
[55] Li Mengyu, Yu Jun, Xu Hongteng, Meng Cheng. Efficient approximation of gromov-wasserstein distance using importance sparsification. Journal of Computational and Graphical Statistics 2023a:1–25 just-accepted.
[56] Tao Li, Cheng Meng, Jun Yu, and Hongteng Xu. Hilbert curve projection distance for distribution comparison. arXiv preprint arXiv:2205.15059, 2022.

[57] Li Tao, Yu Jun, Meng Cheng. Scalable model-free feature screening via sliced-wasserstein dependency. Journal of Computational and Graphical Statistics 2023b:1–24 just-accepted.
[58] Liao Qichen, Chen Jing, Wang Zihao, Bai Bo, Shi Jin, Wu Hao. arXiv preprint; 2022.
[59] Meng Cheng, Yu Jun, Zhang Jingyi, Ma Ping, Zhong Wenxuan. Sufficient dimension reduction for classification using principal optimal transport direction. Advances in Neural Information Processing Systems 2020;33:4015–28.
[60] Zhang Jingyi, Ma Ping, Zhong Wenxuan, Meng Cheng. Projection-based techniques for highdimensional optimal transport problems. Wiley Interdisciplinary Reviews: Computational Statistics 2022:e1587 page.
[61] Zhang Jingyi, Meng Cheng, Yu Jun, Zhang Mengrui, Zhong Wenxuan, Ma Ping. An optimal transport approach for selecting a representative subsample with application in efficient kernel density estimation. Journal of Computational and Graphical Statistics 2023;32(1):329–39.