scEGOT: Single-cell trajectory inference framework based on entropic Gaussian mixture optimal transport

Toshiaki Yachimura [a,*] Hanbo Wang [b] Yusuke Imoto [c] Momoko Yoshida [d]
Sohei Tasaki [e] Yoji Kojima [f] Yukihiro Yabuta [b,c] Mitinori Saitou [b,c,f]
Yasuaki Hiraoka [c,*]

[a] Mathematical Science Center for Co-creative Society, Tohoku University, Sendai 980-0845, Japan
[b] Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
[c] Institute for the Advanced Study of Human Biology, Kyoto University Institute for Advanced Study, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
[d] Faculty of Science, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan
[e] Department of Mathematics, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan
[f] Center for iPS Cell Research and Application, Kyoto University, 53 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan

[*] To whom correspondence should be addressed. E-mail: toshiaki.yachimura.a4@tohoku.ac.jp or hiraoka.yasuaki.6z@kyoto-u.ac.jp

Abstract

Time-series single-cell RNA sequencing (scRNA-seq) data have opened a door to elucidate cell differentiation processes. In this context, the optimal transport (OT) theory has attracted attention to interpolate scRNA-seq data and infer the trajectories of cell differentiation. However, there remain critical issues in interpretability and computational cost. This paper presents scEGOT, a novel comprehensive trajectory inference framework for single-cell data based on entropic Gaussian mixture optimal transport (EGOT). By constructing a theory of EGOT via an explicit construction of the entropic transport plan and its connection to a continuous OT with its error estimates, EGOT is realized as a generative model with high interpretability and low computational cost, dramatically facilitating the inference of cell trajectories and dynamics from time-series data. The scEGOT framework provides comprehensive outputs from multiple perspectives, including cell state graphs, velocity fields of cell differentiation, time interpolations of single-cell data, space-time continuous videos of cell differentiation with gene expressions, gene regulatory networks, and reconstructions of Waddington's epigenetic landscape. To demonstrate that scEGOT is a powerful and versatile tool for single-cell biology, we applied it to time-series scRNA-seq data of the human primordial germ cell-like cell (human PGCLC) induction system. Using scEGOT, we precisely identified the PGCLC progenitor population and the bifurcation time of the segregation. Our analysis suggests that a known marker gene *TFAP2A* alone is not sufficient to identify the PGCLC progenitor cell population, but that *NKX1-2* is also required. In addition, we found that *MESP1* and *GATA6* may also be crucial for PGCLC/somatic cell segregation.

Keywords: optimal transport, Gaussian mixture model, single-cell biology, trajectory inference, epigenetic landscape

1 Introduction

The "epigenetic landscape" proposed by the renowned biologist C. H. Waddington is a well-known metaphor for describing cell differentiation and is a key concept in developmental biology [47]. In this conceptual model, cells begin as stem cells at the top of this landscape and differentiate into more specialized cell types as they move down the valleys during the development, with the ridges representing potential barriers that prevent transitions between cell types.

Although a useful concept, the actual shapes of the landscapes during differentiation processes have remained unclear in many biological systems. However, recent advances in genome-scale high-dimensional single-cell technologies, such as single-cell RNA sequencing (scRNA-seq) [16, 23], have opened an avenue for inferring the dynamics of cell differentiation in a data-driven manner, as well as for reconstructing Waddington's landscape. This has made trajectory inference for cell differentiation a central topic in current single-cell and systems biology [21, 41, 48].

Many methods [4, 34] for trajectory inference have been developed using a single snapshot of the scRNA-seq data. In spite of the snapshot data, due to the heterogeneity of the cell population, these methods can identify changes in gene expression levels along the pseudo-time [12, 32, 39, 43, 53]. However, the dynamics of overall cell differentiation are very complex, causing the static trajectory inference described above to have obvious limitations [51]. Recently, time-series scRNA-seq data have been used to overcome this difficulty and to gain more insight into the dynamics of cell differentiation. Nevertheless, the destruction of cells at each measurement severely impedes the identification of cell populations between adjacent time points. To deal with this issue, trajectory inference methods based on the optimal transport theory have attracted attention in recent years.

The optimal transport (OT) is a mathematical theory that provides distances and optimal matchings between probability distributions [36, 45, 46]. It has recently been applied in biology [55], especially in single-cell analysis [17, 27, 28], as well as in several other research fields [31]. Among them, Waddington-OT [38] is a well-known method for inferring the cell lineages by applying a static unbalanced optimal transport to the time-series scRNA-seq data. While it can predict cell lineages from the optimal matching of cell populations, since it does not learn the continuous distributions of cell populations (i.e., a non-generative model), we cannot gain much insight into the intermediate states in the differentiation process. It is also known that such optimal transports between cells do not sufficiently reflect those between the cell distributions [50].

On the other hand, optimal transport methods with generative models based on neural networks have also been reported, such as TrajectoryNet [42], JKONET [2], and PRESCIENT [56]. They can generate data in the intermediate states of the differentiation process. However, the neural networks used there introduce black boxes into these methods, making biological interpretation difficult.

In general, the computational cost of solving optimal transport problems, including the above methods, is very high and can be a bottleneck for trajectory inference. GraphFP [15] is a method that addresses this problem by combining dynamic optimal transports on cell state graphs with the nonlinear Fokker–Planck equation. The key to reducing the computational cost is the formulation using inter-cluster optimal transport. Owing to its construction, this method achieves high biological interpretability and low computational cost. However, since it deals with only the cell lineages of the cell clusters, it cannot infer the continuous state of the differentiation process (e.g., transitions during the merging/separation of cell clusters).

In this paper, we present scEGOT, a novel trajectory inference framework based on entropic Gaussian mixture optimal transport (EGOT). It aims to provide a comprehensive trajectory inference framework to infer the dynamics of cell differentiation from time-series single-cell data. The methodology is based on an inter-cluster optimal transport, where clustering and learning of the distributions are performed on cell populations in the gene expression space using the Gaussian mixture model (GMM), and each Gaussian distribution corresponds to a cell type. The main feature of this method is that it has a clear and rigorous correspondence to a continuous optimal transport. Moreover, its computational cost is significantly low owing to the inter-cluster

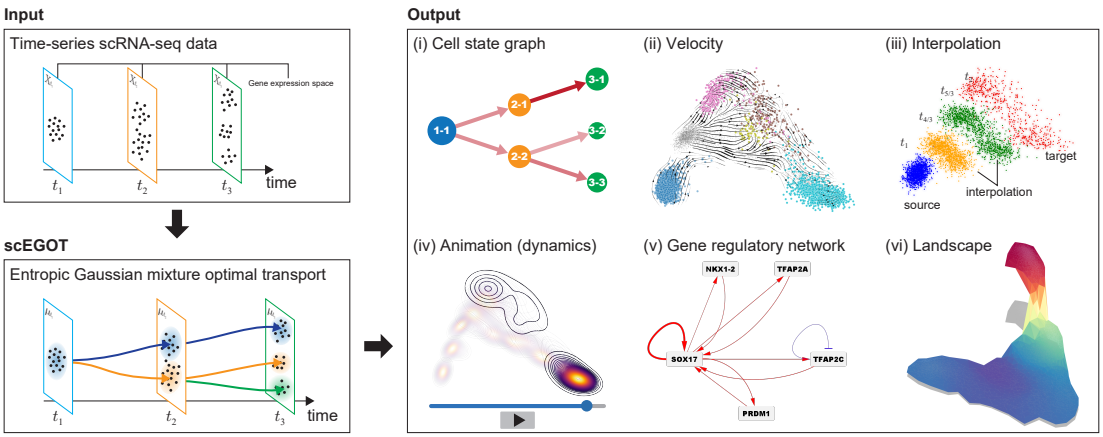


Figure 1: Sketch of the framework of scEGOT. ScEGOT provides six different functions based on entropic Gaussian mixture optimal transport (EGOT).

optimal transport. Accordingly, we can continuously infer the intermediate states of the cell differentiation process at low computational cost. As a comprehensive framework, scEGOT can construct (i) cell state graphs, (ii) velocity fields of cell differentiation, (iii) time interpolations of single-cell data, (iv) space-time continuous videos of cell differentiation with gene expressions, (v) gene regulatory networks (GRNs), and (vi) Waddington's landscape (Fig. 1).

As a biological application, we apply scEGOT to the time-series scRNA-seq data of the human primordial germ cell-like cell (human PGCLC) induction system. We demonstrate that scEGOT provides insights into the molecular mechanism of PGCLC/somatic cell segregation. In particular, using the functions of scEGOT, we elucidate the dynamics of the PGCLC differentiation and identify the PGCLC progenitor cell population. Furthermore, we find key genes such as *NKX1-2*, *MESP1*, and *GATA6* that may play crucial roles during human PGCLC differentiation.

2 Theory

We present here the mathematical foundation of EGOT and its application to single-cell biology, called scEGOT. By generalizing [7], the EGOT is formulated by an entropic regularization of the discrete optimal transport, which is a coarse-grained model derived by taking each Gaussian distribution as a single point.

We first summarize the properties of the solution of the EGOT and then discuss the entropic transport plan constructed from the solution of EGOT. Furthermore, we show a correspondence between EGOT and the continuous optimal transport presented by [8]. This theoretical compatibility enables us to present a novel trajectory inference framework in scEGOT. Specifically, this framework allows us to construct a cell state graph and infer the intermediate states and velocity of the cell differentiation process, from which we can further infer the dynamic GRNs between adjacent time points and reconstruct Waddington's landscape in the gene expression space, with high biological interpretability and low computational cost.

All proofs of the mathematical statements made here are provided in the supporting information.

2.1 Gaussian mixture model

Let $\{X_{t_i}\}_{1\leq i\leq I}$ denote the time-series data of the point clouds in $\mathbb{R}^n (n \geq 1)$ with I time stages (possibly with different sample sizes). For each point cloud X_{t_i}, we apply a Gaussian mixture

model (GMM) [26]. Then, we obtain

$$\mu_{t_i} = \sum_{k=1}^{K_i} \pi_i^k \mu_i^k, \tag{1}$$

where K_i denotes the number of clusters in the distribution at t_i, $\mu_i^k = \mathcal{N}(m_i^k, \Sigma_i^k)$ is a Gaussian distribution with a mean vector m_i^k and a covariance matrix Σ_i^k, and π_i^k is a non-negative weight of each Gaussian distribution satisfying $\sum_{k=1}^{K_i} \pi_i^k = 1$. In the scRNA-seq data, X_{t_i} denotes a cell population in the gene expression space $\mathbb{R}^n$ at the time t_i, where each cell $x = (x_1, \ldots, x_n) \in X_{t_i}$ is characterized by the n values of the gene expressions, or those in a principal component analysis (PCA) space after the dimension reduction. Then, each Gaussian distribution and its weight are regarded as a certain cell type and its existence probability, respectively.

2.2 Entropic Gaussian mixture Optimal Transport (EGOT)

EGOT connects these Gaussian mixture distributions between adjacent times by solving the following optimization problem for the weights $\pi_i = (\pi_i^k)_{1 \le k \le K_i}$ and $\pi_{i+1} = (\pi_{i+1}^l)_{1 \le l \le K_{i+1}}$:

$$d_G^\varepsilon(\mu_{t_i}, \mu_{t_{i+1}}) = \min_{w \in \Pi(\pi_i, \pi_{i+1})} \sum_{k,l} w_{k,l} W_2^2(\mu_i^k, \mu_{i+1}^l) - \varepsilon H(w), \tag{2}$$

where

$$\Pi(\pi_i, \pi_{i+1}) = \{w \in \mathbb{R}_+^{K_i \times K_{i+1}} : w\mathbf{1}_{K_{i+1}} = \pi_i, w^T \mathbf{1}_{K_i} = \pi_{i+1}\}$$

is a subset of $K_i \times K_{i+1}$ matrices of non-negative real numbers; $\mathbf{1}_{K_i}$ expresses the K_i-dimensional vector whose entries are all 1; and $(\bullet)^T$ denotes the transpose. Here, W_2 is the L^2-Bures–Wassersterin distance [9, 11, 30]

$$W_2^2(\mu_i^k, \mu_{i+1}^l) = \|m_i^k - m_{i+1}^l\|_2^2 + \mathcal{B}^2(\Sigma_i^k, \Sigma_{i+1}^l)$$

using the squared Bures metric [3]

$$\mathcal{B}^2(\Sigma_i^k, \Sigma_{i+1}^l) = \mathrm{tr}\left(\Sigma_i^k + \Sigma_{i+1}^l - 2\left((\Sigma_i^k)^{\frac{1}{2}} \Sigma_{i+1}^l (\Sigma_i^k)^{\frac{1}{2}}\right)^{\frac{1}{2}}\right),$$

and H is an entropy function defined as

$$H(w) = -\sum_{k,l} w_{k,l}(\log w_{k,l} - 1).$$

The solution w of the EGOT represents how close the Gaussian distributions of the adjacent time points are to each other. Thus, it indicates the similarity between the cell populations when applied to the scRNA-seq data. Since the discrete entropy H is a strongly concave function, the objective function of (2) is a strongly convex function. Therefore, the optimization problem (2) has a unique optimal solution. In addition, by considering the Lagrangian associated with the problem (2) and the first-order optimality condition, the following proposition holds:

Proposition 1. *The optimization problem* (2) *has a unique solution* w^ε *given by*

$$w_{k,l}^\varepsilon = u_k \, e^{-\frac{W_2^2(\mu_i^k, \mu_{i+1}^l)}{\varepsilon}} \, v_l, \tag{3}$$

where $(u, v) \in \mathbb{R}_+^{K_i} \times \mathbb{R}_+^{K_{i+1}}$ *are derived from the dual variables of the Lagrangian associated with the problem* (2).

We emphasize that by coarse-graining point clouds with Gaussian mixture distributions, the EGOT is significantly less computationally expensive than directly analyzing the optimal transports with full point clouds (e.g., Waddington-OT [38]). Furthermore, as we will see subsequently, EGOT can recover the solution and distance of a continuous optimal transport between Gaussian mixtures, which will provide deeper insights into the stochastic dynamics of cells in the gene expression space.

2.3 Convergence of EGOT and its connection to continuous OT

In this section, we clarify the connection between EGOT and the following continuous optimal transport proposed by [8]:

$$MW_2^2(\mu_{t_i}, \mu_{t_{i+1}}) = \inf_{\gamma \in \hat{\Pi}(\mu_{t_i}, \mu_{t_{i+1}}) \cap G_{2n}(\infty)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 \, d\gamma, \tag{4}$$

where $\hat{\Pi}(\mu_{t_i}, \mu_{t_{i+1}})$ is a set of probability measures on $\mathbb{R}^n \times \mathbb{R}^n$ with the given marginals μ_{t_i} and $\mu_{t_{i+1}}$; $G_{2n}(K)$ is a set of probability measures on $\mathbb{R}^n \times \mathbb{R}^n$, which can be written as Gaussian mixtures with K or less components; and $G_{2n}(\infty) = \bigcup_{K \geq 1} G_{2n}(K)$.

Let us first consider the convergence of the EGOT (2) and the solution (3) as $\varepsilon \to 0$. By using an estimate similar to that for discrete entropic regularized optimal transports with Shannon entropy by Weed [49], we can prove the following proposition.

Proposition 2. *The solution w^ε (3) exponentially converges to an optimal transport plan of $d_G := d_G^0$ with the maximal entropy as $\varepsilon \to 0$. That is, there exists $M > 0$ independent of $\varepsilon > 0$ such that*

$$\|w^\varepsilon - w^*\|_1 \leq e^{-\frac{M}{\varepsilon}}, \tag{5}$$

where w^* is the solution of

$$w^* = \operatorname*{arg\,max}_{w \in \Pi(\pi_i, \pi_{i+1})} \{H(w) : \textit{solutions of } (2) \textit{ with } \varepsilon = 0\} \tag{6}$$

and $\|\cdot\|_1$ denotes the ℓ_1 norm of vectors (viewing w as a vector). Moreover, $d_G^\varepsilon \to d_G$ as $\varepsilon \to 0$ with the following estimate:

$$|d_G^\varepsilon - d_G| \leq \max_{k,l} W_2^2(\mu_i^k, \mu_{i+1}^l) e^{-\frac{M}{\varepsilon}} + (1 + \log K_i K_{i+1})\varepsilon. \tag{7}$$

Next, we define an entropic transport plan

$$\gamma^\varepsilon(x, y) = \sum_{k,l} w_{k,l}^\varepsilon g_{m_i^k, \Sigma_i^k}(x) \, \delta_{y = T_{k,l}(x)}, \tag{8}$$

where $g_{m_i^k, \Sigma_i^k}$ is the multivariate Gaussian distribution

$$g_{m_i^k, \Sigma_i^k}(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_i^k)}} e^{-\frac{1}{2}(x - m_i^k)^T (\Sigma_i^k)^{-1} (x - m_i^k)}$$

and $T_{k,l}$ is the optimal transport map between μ_i^k and μ_{i+1}^l given by

$$T_{k,l}(x) = m_{i+1}^l + A_{k,l}(x - m_i^k)$$

with

$$A_{k,l} = (\Sigma_i^k)^{-\frac{1}{2}} \left((\Sigma_i^k)^{\frac{1}{2}} \Sigma_{i+1}^l (\Sigma_i^k)^{\frac{1}{2}} \right)^{\frac{1}{2}} (\Sigma_i^k)^{-\frac{1}{2}}.$$

By applying the convergence result (5) and the estimate (7), we obtain the main theorem in this paper, which provides a correspondence between EGOT and the continuous optimal transport (4) (Fig. 2).

Theorem 3. *The continuous optimal transport (4) has a minimizer*

$$\gamma^*(x, y) = \sum_{k,l} w_{k,l}^* g_{m_i^k, \Sigma_i^k}(x) \, \delta_{y = T_{k,l}(x)}$$

and the entropic transport plan γ^ε converges to γ^* in the sense of the narrow convergence, i.e., for any bounded continuous functions $\varphi \in C_b(\mathbb{R}^n \times \mathbb{R}^n)$,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(x, y) \, d\gamma^\varepsilon(x, y) \to \int_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(x, y) \, d\gamma^*(x, y).$$

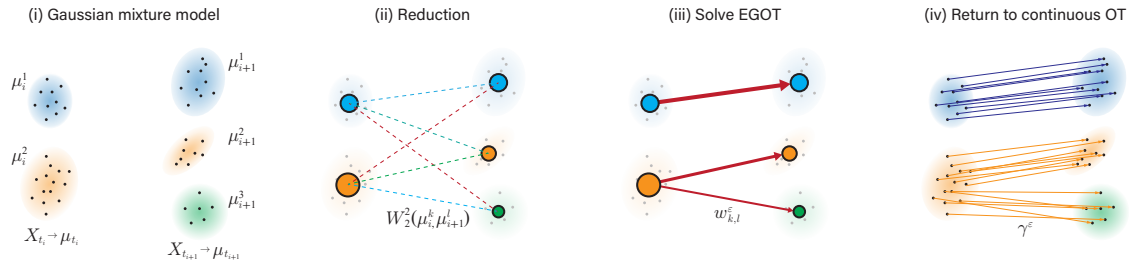


Figure 2: Illustration of the connection between EGOT (2) and continuous optimal transport (4).

Moreover, $d_G^\varepsilon \to MW_2^2$ *as* $\varepsilon \to 0$ *with the following estimate:*

$$|d_G^\varepsilon - MW_2^2| \leq \max_{k,l} W_2^2(\mu_i^k, \mu_{i+1}^l) e^{-\frac{M}{\varepsilon}} + (1 + \log K_i K_{i+1})\varepsilon.$$

Theorem 3 enables us to go back and forth between the discrete and continuous optimal transports through the entropic transport plan (8). Thus it provides a versatile trajectory inference framework (from continuous OT) with the low computational complexity (from discrete OT) (Fig. 3).

2.4 Cell state graph by EGOT

In the case of point clouds obtained from the time-series scRNA-seq data, each Gaussian distribution and its weight can be regarded as a certain cell type and its existence probability, respectively. Then, the solution of EGOT, which represents how much weight is transported between the Gaussian distributions at adjacent time points, can be interpreted as the degree of relationship between these cell types.

Based on this interpretation, we define a cell state graph of cell differentiation as follows. For each adjacent times t_i, t_{i+1}, a complete weighted bipartite graph (V_i, W_i) in $\mathbb{R}^n$ is constructed as

$$\begin{aligned} V_i &= \{m_i^k\}_{1\leq k\leq K_i} \cup \{m_{i+1}^l\}_{1\leq l\leq K_{i+1}}, \\ W_i &= \left\{ w_{k,l}^\varepsilon / \pi_i^k \right\}_{1\leq k\leq K_i, 1\leq l\leq K_{i+1}}, \end{aligned}$$

where the set V_i of nodes consists of the mean vectors at times t_i and t_{i+1} representing the locations of the cell types, and the set W_i of the weighted edges are given by the normalized solution of the EGOT (called transport rate) corresponding to the degree of relationship of the cell types. Then, the cell state graph (V, W) is constructed by combining (V_i, W_i) for all the time stages. From this cell state graph, we can study the state transitions of the cell population in the temporal evolution.

2.5 Entropic barycentric projection map and cell velocity

The entropic transport plan $\gamma^\varepsilon(x, y)$ can be regarded as the weights to be transported from the point x to y. By averaging over y, we can define an entropic barycentric projection map of the EGOT, which transports the least cost from x, as follows:

$$T^\varepsilon(x) = \int_{\mathbb{R}^n} y \, d\gamma_x^\varepsilon(y), \tag{9}$$

where $\gamma_x^\varepsilon(y)$ is the disintegration of the entropic transport plan γ^ε with respect to the first marginal μ_{t_i}, i.e., $d\gamma^\varepsilon(x, y) = d\gamma_x^\varepsilon(y) d\mu_{t_i}(x)$. Then, by (8), we can compute the entropic barycentric projection map T^ε explicitly.

Proposition 4. *The entropic barycentric projection map with respect to γ^ε is expressed as*

$$T^\varepsilon(x) = \sum_{k,l} P^\varepsilon_{k,l}(x) T_{k,l}(x), \tag{10}$$

where $P^\varepsilon_{k,l}$ denotes

$$P^\varepsilon_{k,l}(x) = \frac{w^\varepsilon_{k,l} g_{m^k_i, \Sigma^k_i}(x)}{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x)}.$$

From the definition, the entropic barycentric projection map T^ε represents where a cell $x \in X_{t_i}$ at time t_i moves at time t_{i+1} in the gene expression space. Accordingly, we define the rate of change of the gene expression for a cell x from the time t_i to t_{i+1}, called the cell velocity, as

$$v(x) = T^\varepsilon(x) - x. \tag{11}$$

The cell velocity $v(x)$ expresses which direction and how much the gene expression of the cell x changes. A high speed $|v|$ implies a significant change of the cell in cell differentiation, whereas a low speed $|v|$ indicates that the gene expression hardly changes and is close to the steady state.

2.6 Entropic displacement interpolation and gene expression animation

One of the great advantages of the optimal transport theory is that we can obtain the displacement interpolations between the probability distributions, allowing us to study the dynamics of cell differentiation. We define the following entropic displacement interpolation between μ_{t_i} and $\mu_{t_{i+1}}$

$$\mu^\varepsilon_{t_{i+s}} = \sum_{k,l} w^\varepsilon_{k,l} \mu^{k,l}_{t_{i+s}} \tag{12}$$

at $t_{i+s} = (1-s)t_i + s t_{i+1}$, $s \in [0,1]$. Here $w^\varepsilon_{k,l}$ is the solution (3), and $\mu^{k,l}_{t_{i+s}} = \mathcal{N}(m^{k,l}_{i+s}, \Sigma^{k,l}_{i+s})$ with $m^{k,l}_{i+s} = (1-s)m^k_i + s m^l_{i+1}$ and

$$\Sigma^{k,l}_{i+s} = \left((1-s)I_d + sB_{k,l}\right) \Sigma^k_i \left((1-s)I_d + sB_{k,l}\right),$$

where I_d is the $d \times d$ identity matrix and

$$B_{k,l} = (\Sigma^l_{i+1})^{\frac{1}{2}} \left((\Sigma^l_{i+1})^{\frac{1}{2}} \Sigma^k_i (\Sigma^l_{i+1})^{\frac{1}{2}} \right)^{-\frac{1}{2}} (\Sigma^l_{i+1})^{\frac{1}{2}}.$$

Theorem 5. *The probability distribution*

$$\mu^*_{t_{i+s}} = \sum_{k,l} w^*_{k,l} \mu^{k,l}_{t_{i+s}} \tag{13}$$

is the geodesic with the maximum entropy between μ_{t_i} and $\mu_{t_{i+1}}$ in the space $G_{2n}(\infty)$ equipped with the distance MW_2, where w^* is the solution (6). Moreover, for any $s \in [0,1]$, the entropic displacement interpolation $\mu^\varepsilon_{t_{i+s}}$ converges to $\mu^*_{t_{i+s}}$ narrowly.

From Theorem 5, we see that the entropic displacement interpolation (12) is an approximation of the geodesic (13) in the space $G_{2n}(\infty)$ equipped with the distance MW_2. Through this entropic displacement interpolation (12), we can generate an interpolation distribution between μ_{t_i} and $\mu_{t_{i+1}}$, which means that the interpolation of the adjacent scRNA-seq data X_{t_i} and $X_{t_{i+1}}$ can be generated. Therefore, by successively constructing the interpolations for all the time stages, we can create an animation of the gene expressions to track the time-varying dynamics of cell differentiation.

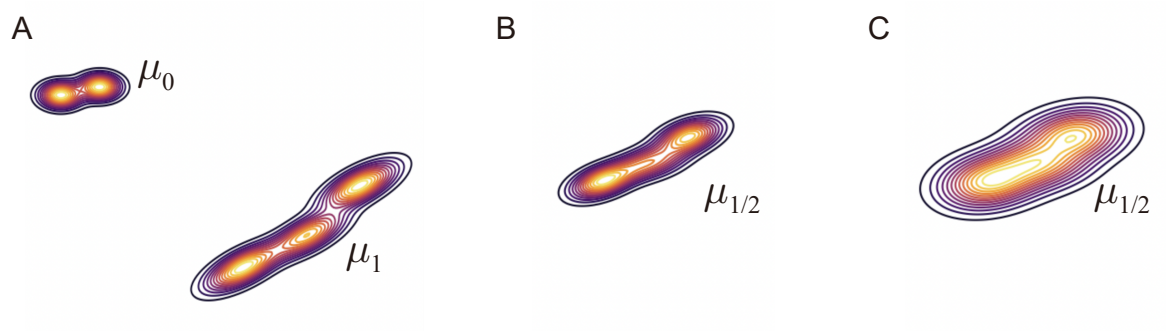


Figure 3: Comparison of computation time between interpolated distributions generated by EGOT and continuous OT. The experiments were performed on a MacBook Pro (14-inch, 2021) equipped with Apple M1 Max and 64GB RAM. (A) Source and target Gaussian mixture distributions μ_0 and μ_1 in 2D with two and three clusters, respectively. The space was divided into 40,000 segments. (B) Interpolation distribution $\mu_{1/2}$ generated by EGOT. The computation time was 41ms. (C) Interpolation distribution $\mu_{1/2}$ generated by using the POT library [10], specifically the ot.bregman.barycenter module. The computation time was 1h 44min 49s.

2.7 Estimation of GRNs

Once the cell velocity (11) representing the changes in the gene expressions is obtained, it can be used to estimate the gene regulatory network (GRN) that drives the cell differentiation dynamics. Let $\hat{X} = \{x^j\}_{1\le j\le D} \in \mathbb{R}^{n\times D}$ be a cell population consisting of D cells for which the GRN is to be obtained. We assume that the gene expressions in the cell population $\hat{X}$ are driven by linear dynamics

$$\frac{d\hat{X}}{dt} = A\hat{X},$$

where $A \in \mathbb{R}^{n\times n}$ is the matrix characterizing the GRN and represents the effect of the expression level of each gene on the expression dynamics of the other genes. By the cell velocity (11), we may assume that $v(\hat{X}) = \frac{d\hat{X}}{dt}$. Thus, we obtain

$$v(\hat{X}) = A\hat{X}. \tag{14}$$

Based on (14), the GRN matrix A^* is estimated by solving the following regression problem:

$$A^* = \underset{A\in\mathbb{R}^{n\times n}}{\arg\min} \|v(\hat{X}) - A\hat{X}\|_2^2 + \lambda\|A\|_2^2, \tag{15}$$

where $\lambda > 0$ is a regularization parameter. In terms of each cell x^j, this can be expressed as

$$A^* = \underset{A\in\mathbb{R}^{n\times n}}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{D} \left| v_i(x^j) - \sum_{s=1}^{n} A_{is}x_s^j \right|^2 + \lambda\|A\|_2^2, \tag{16}$$

where $v_i(x^j)$ is the i-th component of $v(x^j)$. By solving the problem (15) or (16) at each adjacent time point, we can estimate the dynamic GRNs driving cell differentiation.

2.8 Construction of Waddington's landscape

From the cell velocity (11), we can construct Waddington's landscape in the gene expression space. The Helmholtz–Hodge decomposition implies that the cell velocity can be written as

$$v = -\nabla\varphi + q, \quad \operatorname{div} q = 0, \tag{17}$$

where φ is the gradient potential and q is the divergence-free part. Since Waddington's landscape is a gradient system, the potential φ can be regarded as the realization of Waddington's landscape

in the gene expression space. To construct φ, we take the divergence operator in the equation (17) and obtain the following differential equation

$$-\Delta\varphi = \operatorname{div} v. \tag{18}$$

The partial differential equation (18) is not uniquely solvable in general since the boundary conditions are undefined. In the following, instead of solving (18) directly, we consider it as an equation on the k-nearest neighbor graph of cells and look for its least-squares and least-norm solution. In other words, we consider the potential φ as the solution to the following minimization problem:

$$\varphi^* = \arg\min_{\varphi \in S} \|\varphi\|_2^2, \tag{19}$$

where S is the set of solutions to the least-squares minimization problem

$$\min_{\varphi} \|(-\hat{\Delta})\varphi - \operatorname{div} v\|_2^2.$$

Here, $\hat{\Delta}$ denotes the graph Laplacian on the k-nearest neighbor graph of the cells. The solution of the minimization problem (19) can be written as

$$\varphi^* = (-\hat{\Delta})^{\text{p-inv}} \operatorname{div} v, \tag{20}$$

where $(-\hat{\Delta})^{\text{p-inv}}$ denotes the Moore–Penrose pseudoinverse of $(-\hat{\Delta})$. Note that the divergence of the cell velocity v on the right-hand side in (20) can also be computed concretely, as the cell velocity (11) is obtained explicitly. Thus, the following theorem holds.

Theorem 6. *The potential φ^* is represented by*

$$\varphi^* = (-\hat{\Delta})^{\text{p-inv}} F,$$

with

$$F := \sum_{k,l} \{(C_k, T_{k,l}) + \operatorname{tr}(A_{k,l})\} P^{\varepsilon}_{k,l} - n,$$

where $(\cdot,\cdot)$ denotes the standard Euclidean inner product of vectors and C_k denotes

$$C_k = -(\Sigma_i^k)^{-1}(x - m_i^k) + \frac{\sum_j \pi_i^j g_{m_i^j, \Sigma_i^j}(x)(\Sigma_i^j)^{-1}(x - m_i^j)}{\sum_j \pi_i^j g_{m_i^j, \Sigma_i^j}(x)}.$$

Through this procedure, we can construct the gradient potential (20) of the cells as in Waddington's landscape.

3 Biological application of scEGOT to human PGCLC induction system dataset

This section presents a biological application of scEGOT to time-series single-cell gene expression data and its validation. In particular, we consider the human primordial germ cell-like cell (PGCLC) induction system in vitro. In this in vitro culture system, previous studies have shown that genes such as *EOMES*, *GATA3*, *SOX17*, *TFAP2C* are critical genes for PGCLC differentiation [14, 18, 19, 37, 40]. However, the induction rate of PGCLC is only approximately 10 to 40 %, and the precise molecular mechanisms underlying the PGCLC differentiation and PGCLC/somatic cell segregation remain poorly understood. Using the functions of scEGOT, we characterize the progenitor cell population and elucidate the molecular mechanism of PGCLC/somatic cell segregation.

3.1 Dataset

The induction starts with the induction of human induced pluripotent stem cells (hiPSC) into the incipient mesoderm-like cells (iMeLC) in a 2D culture, followed by their transfer into a 3D culture medium as aggregates with the inclusion of BMP4 and other signals. The dataset for scRNA-seq was generated using the 10X Genomics platform and consists of five-time points ($I = 5$); we refer to iMeLC as day 0, whereas 3D-cultured aggregates are denoted as day 0.5, day 1, day 1.5, and day 2, respectively (Fig. 4A).

3.2 Preprocessing and clustering

As a preprocessing of scEGOT, we applied a noise reduction method RECODE [13] and selected genes with the top 2,000 higher normalized variances (variances divided by means) of non-mitochondrial genes, followed by log scaling. We then performed PCA and applied scEGOT to the top 150 principal components (their cumulative contribution rate was 93.67%). We also put the regularization parameter of EGOT to $\varepsilon = 0.01$.

We set the number of clusters K_i as 1, 2, 4, 5, and 5 for day 0, day 0.5, day 1, day 1.5, and day 2, respectively (Figs. 4 and S1A). In particular, by examining the expression of the signature genes, we annotated the clusters at day 2 with four cell types: PGCLC ($NANOG^+/SOX17^+/TFAP2C^+/PRDM1^+$), amnion-like cell (AMLC, $TFAP2A^+/TFAP2C^+/GATA3^+$), endoderm-like cell (EDLC, $GATA6^+/SOX17^+/FOXA2^+$), and extra-embryonic mesenchyme-like cell (EXMCLC, $HAND1^+/FOXF1^+$) (Figs. 4B and S1B–C). This annotation of the clusters is consistent with [6].

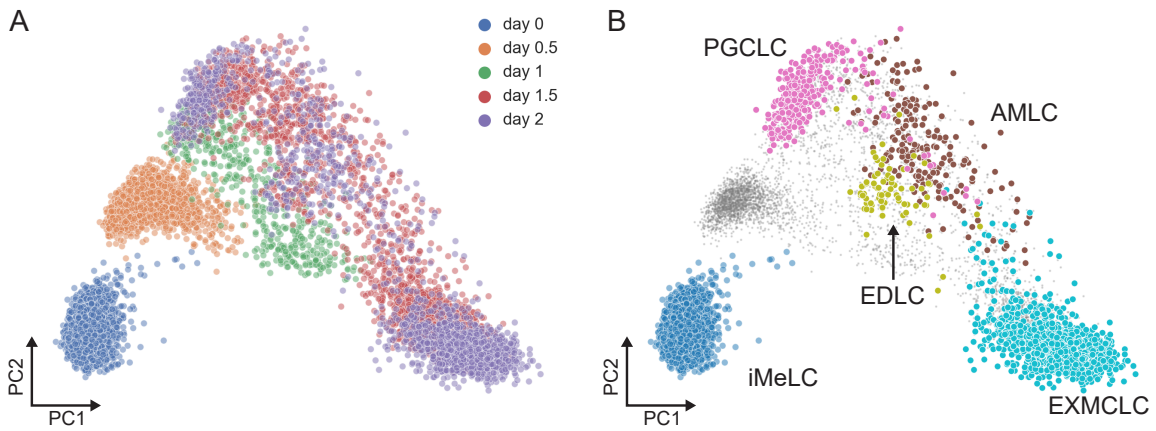


Figure 4: PCA plots of PGCLC induction dataset. The cells are colored according to (A) experimental day and (B) cell type. The gray points in (B) are cells at the middle stages (days 0.5–1.5).

3.3 Verification of scEGOT interpolation

We verify the accuracy of the entropic displacement interpolation of scEGOT (12). We set the data at day 1 as the reference distribution and generated the interpolation distribution using the data at day 0.5 (two clusters) and 1.5 (five clusters). The analysis shows that the interpolation distribution (1,000 cells) properly reproduces the reference distribution (Fig 5A). This was quantitatively verified using the silhouette score (Fig 5B), showing that the interpolation by scEGOT is well overlapped with the reference data rather than the source and target data. Here, the silhouette score indicates that the clusters are separated (overlapped, resp.) if the value is close to one (zero, resp.).

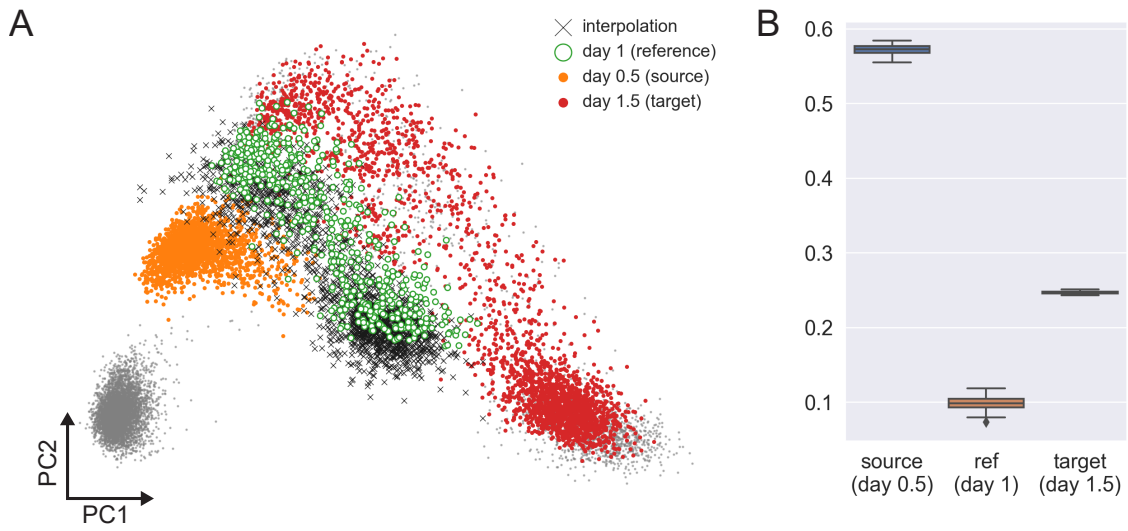


Figure 5: Verification of scEGOT interpolation. (A) The reference distribution (day 1) and scEGOT interpolation distribution generated by datasets at days 0.5 and 1.5 on PCA coordinates. (B) Box plot of the silhouette scores over 100 trials of the scEGOT interpolation versus the source (day 0.5)/reference (day 1)/target (day 1.5).

3.4 Trajectory inference with cell state graph

In this section, we analyze the trajectory of cell differentiation in the human PGCLC induction system using the cell state graph (Fig. 6) generated by scEGOT and gain molecular insights into the mechanism that segregates PGCLCs from somatic lineages. The cell state graph shows four primary differentiation pathways leading to PGCLC, AMLC, EDLC, and EXMCLC as follows:

PGCLC: day0-1 → day0.5-1 → day1-1 → day1.5-1 → day2-1;
AMLC: day0-1 → day0.5-1 → day1-1 → day1.5-2 → day2-2;
EDLC: day0-1 → day0.5-1 → day1-1 → day1.5-3 → day2-3;
EXMCLC: day0-1 → day0.5-2 → day1-3 → day1.5-5 → day2-5.

The top path in the cell state graph (Fig. 6B) represents the PGCLC differentiation pathway. This path features key PGCLC markers such as *NANOG*, *SOX17*, *TFAP2C* and *PRDM1* [14, 18, 37, 40] (Fig. S1C). The expression levels of these markers increase from day 0.5 to day 1.5. It is of note that major EXMCLC progenitors are segregated as early as day 0.5 (see *Space-time continuous gene expression analysis* section for further analysis). Additionally, EXMCLC can be generated through alternative pathways, most typically, day0-1 → day0.5-1 → day1-2 → day1.5-4 → day2-4, suggesting that day0.5-1 cells retain a competence to differentiate into EXMCLC. Indeed, while there are pathways from the PGCLC progenitors until day 1 to the somatic cell populations, there is no pathway in the opposite direction. This result is reminiscent of Weismann's barrier [52].

To gain a deeper understanding of PGCLC/EXMCLC segregation, we examine the difference in the gene expression value between the clusters. The volcano plots (Fig. 7A–B) show the comparison between two branched clusters at day 0.5 (day0.5-1 and day0.5-2) and day 1 (day1-1 and day1-2) of the PGCLC and EXMCLC lineages. Notably, we found that *TFAP2A* is expressed in the PGCLC progenitor cell population (day0.5-1). Previous studies have shown that this gene plays an important role in PGCLC differentiation [5, 6, 57], suggesting that the cell state graph generated from scEGOT has the ability to capture the progenitor cell population of PGCLC differentiation.

We also discovered that *NKX1-2* is highly expressed in the PGCLC progenitor compared to the somatic cells. This gene is known to be expressed during mesoderm development [44].

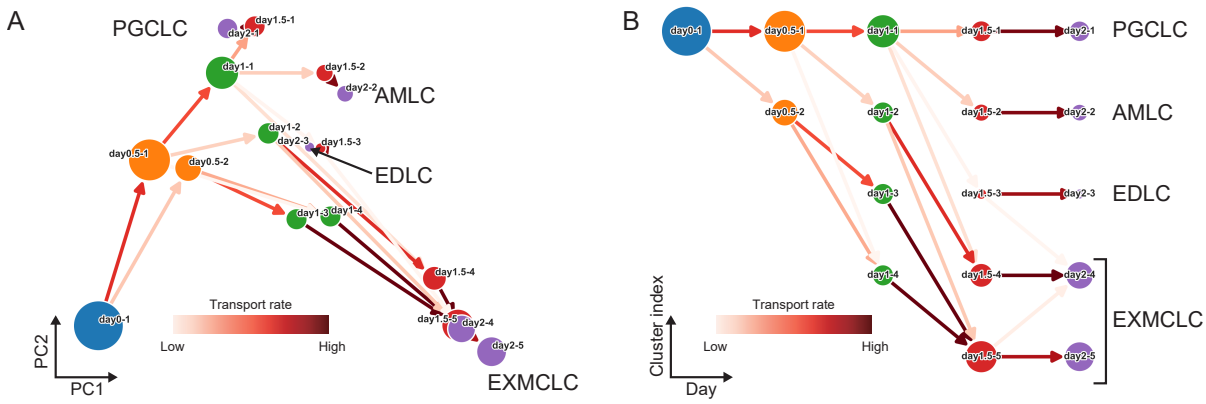


Figure 6: Cell state graphs (A) on PCA coordinates and (B) by a hierarchical layout. The colors of the edges denote the transport rates ($w_{k,l}/\pi_i^k$).

However, its role in PGCLC differentiation has been unknown. This finding will be further investigated in later sections using other functions of scEGOT.

Conversely, genes that are highly expressed in the EXMCLC pathway may also play a critical role in PGCLC/EXMCLC segregation. The early mesoderm markers (*MESP1*, *MESP2*) and the later mesoderm markers (*GATA6*, *FOXF1*) are significantly upregulated on day0.5-2. This shows that these somatic genes might act as repressors of PGCLC during the segregation.

Furthermore, when comparing the clusters day1-1 and day1-2, the former, which is on the way to the PGCLC pathway, shows enrichment of key PGCLC specification genes, namely *NANOG*, *TFAP2C*, *SOX17*, *KLF4*, *SOX15* and *PRDM1* (Fig. 7B). This indicates that the PGCLC path is more specified on day 1.

In mouse PGC, genes such as *stella* and *fragilis* are actively expressed, while HOX genes such as *Hoxa1*, *Hoxb1* are repressed [20, 29, 35, 54]. On the other hand, we did not observe any HOX gene expression in human PGCLC. Interestingly, in our induction system, *HOXA1*, *HOXB2*, *HOXB3*, *HOXB5* and *HOXB6* are highly expressed in the EXMCLC lineage.

From the above observation, the cell state graph well captures the trajectories of human PGCLC induction and can potentially find key genes by analyzing differentially expressed genes among the clusters. In particular, the cell state graph succeeded in tracing PGCLC/somatic cell segregation and identified key genes that may enhance or suppress PGCLC differentiation, such as *NKX1-2*, *MESP1*, *MESP2*, and *GATA6*.

3.5 Velocity analysis

We investigated the dynamics of the developmental process in human PGCLC induction using the cell velocity (11), and compared our method with the RNA velocity [24, 1]. RNA velocity is a standard method to estimate the velocity field in the gene expression space from scRNA-seq data, and it is also widely used to infer cell trajectories [33].

The cell velocity properly indicated that the lineage fate is determined, and the transfer to other lineages (between EDLC, EXMCLC, AMLC, and PGCLC) terminates when the flow becomes stationary (Fig. 8A). On the other hand, the RNA velocity showed that the flow passes through the AMLC, EDLC, and PGCLC populations without stopping, and finally moves towards the EXMCLC lineage on day 2. However, this behavior is not consistent with biological observations (Fig. 8B).

Furthermore, scEGOT can provide the velocities for all the genes, whereas the RNA velocity cannot calculate the velocities for genes without a sufficient amount of detection of unspliced RNA [1, 24] (Fig. 8C). We also emphasize that since the cell velocity is a data-driven method, it can be applied not only to scRNA-seq data but also to any other time-series single-cell data, such as scATAC-seq data. We apply it to time-series scATAC-seq data for innate immune cells

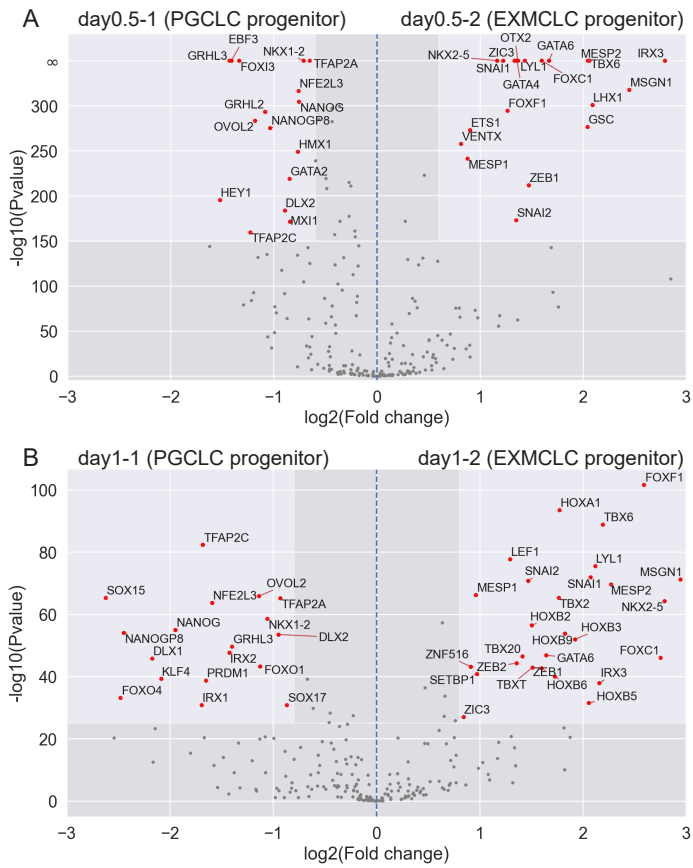


Figure 7: Volcano plots for (A) day 0.5 (day0.5-1 and day0.5-2) and (B) day 1 (day1-1 and day1-2) clusters. The horizontal and vertical lines show the log2 fold change, where the fold change indicates the rate of variation of gene expression and the negative common logarithm of p-values calculated from the independent samples t-test, respectively. The annotated points show cluster-specific genes with (A) $|\log_2(\text{Fold change})| > 0.6$ and $-\log_{10}(\text{Pvalue}) > 150$ and (B) $|\log_2(\text{Fold change})| > 0.8$ and $-\log_{10}(\text{Pvalue}) > 25$.

from mouse-draining lymph nodes (Fig. 8D). The flow of NK cells and monocytes from inactive to active states can be observed. Overall, the cell velocity allows us to perform a comprehensive velocity analysis.

3.6 Space-time continuous gene expression analysis

To study the cell differentiation dynamics, such as the bifurcation time of PGCLC/somatic cell segregation, we constructed the time interpolations of cell populations and the time-continuous gene expression dynamics (animation) using the entropic displacement interpolation (12) (Fig. 9 and Movies S1–S4). The result clearly shows the temporal evolution of cell differentiation and gene expression patterns for the marker genes of PGCLC. In particular, as early as 0.25 and more clearly at 0.75 days, the clusters *NKX1-2*$^+$ and *NKX1-2*$^-$ are separated, and the former cluster exactly moves to the PGCLC population (*TFAP2C*$^+$/*PRDM1*$^+$) (Movies S1, S3).

On the other hand, although the cluster *TFAP2A*$^+$, which has been reported as the progenitor cell population of the PGCLC in previous studies [5, 6, 57], shows a similar tendency to the cluster *NKX1-2*$^+$ until day 1, *TFAP2A* is also highly expressed in the AMLC population after this day, implying that *TFAP2A* alone cannot identify the PGCLC progenitor cell population. This analysis suggests that *NKX1-2* is one of the earliest marker genes of the PGCLC progenitor cell population, and its expression guides the cell population to the PGCLC pathway. It also shows that the bifurcation of PGCLC/somatic cell segregation occurs much earlier, as early as 0.25 days.

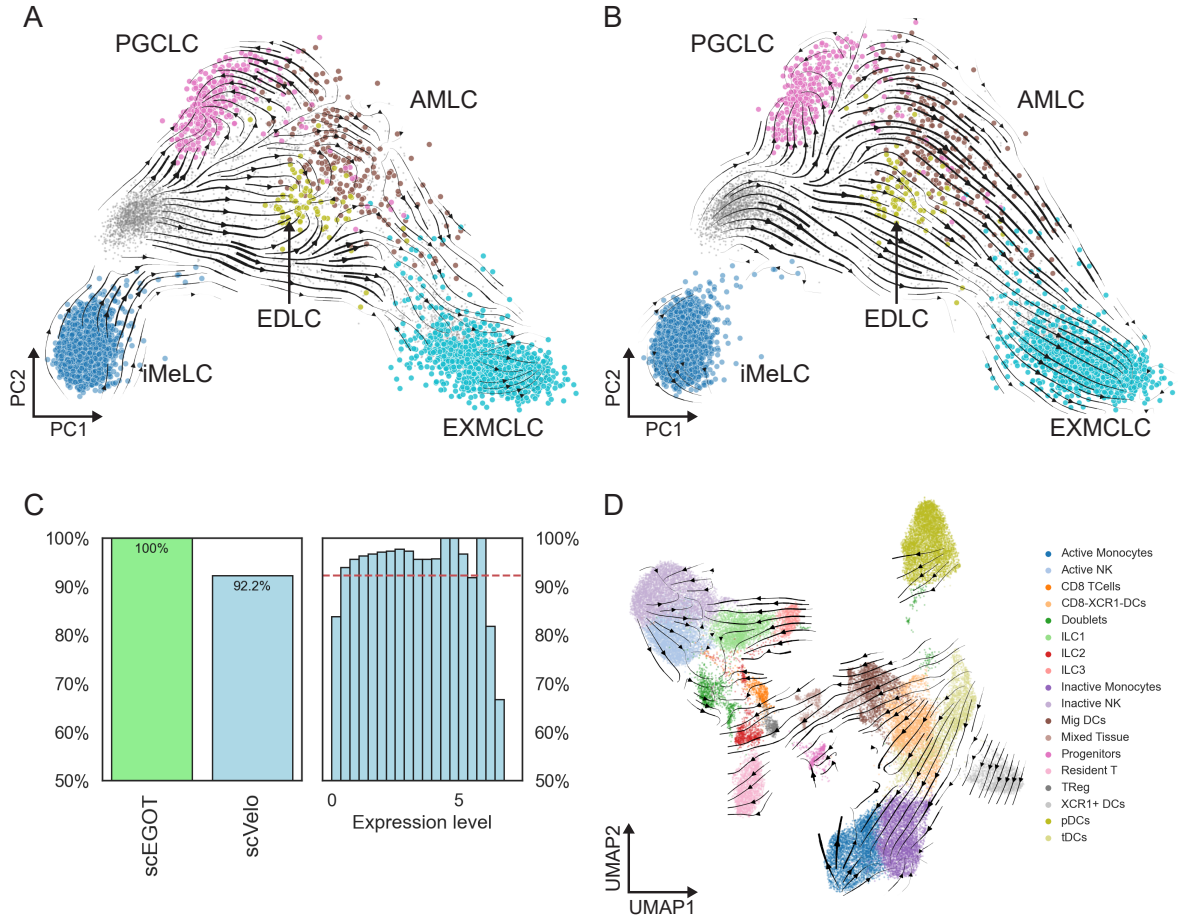


Figure 8: Comparison of velocities between scEGOT (cell velocity) and scVelo (RNA velocity). (A) Stream plot of cell velocity generated by scEGOT. (B) Stream plot of RNA velocity generated by scVelo (stochastic mode). (C) Left: Percentage of genes for which velocities can be computed by scEGOT and scVelo. Right: Histogram of the coverage of scVelo for mean expression levels. The red dashed line is the total coverage of scVelo (92.3%). (D) Cell velocity for scATAC-seq data of mouse innate immune cells at three time points (days 0, 1, and 28) [22].

3.7 GRN analysis

We inferred the GRNs for PGCLC induction. We computed the GRN matrix A^* using the cells in the clusters on the PGCLC pathway, which is determined by the cell state graph analysis. We then extracted a subnetwork of the GRN specified by the key genes of the human PGCLC specification *PRDM1*, *SOX17*, *TFAP2C* and two candidates *TFAP2A* and *NKX1-2* of PGCLC progenitor marker genes (Fig. 10). We cut the small regulatory edges with weight $|A^*_{ij}|$ less than 0.02. The regularization parameter $\lambda > 0$ was automatically determined by cross-validation.

On days 0–0.5, we found that both *TFAP2A* and *NKX1-2* can activate *SOX17* which is essential for PGCLC differentiation [14, 18]. However, on days 0.5–1, *NKX1-2* continues to be able to activate the other PGCLC regulators, whereas *TFAP2A* does not interact with them. This result suggests that *NKX1-2* may play a more essential role than *TFAP2A* during human PGCLC differentiation.

3.8 Reconstruction of Waddington's landscape

Finally, we reconstructed Waddington's landscape to validate the differentiation ability of cells in the PGCLC induction system. We constructed the k-nearest neighbor graph of cells ($k = 15$) on the top two principal components to compute the graph Laplacian. Fig. 11 shows the

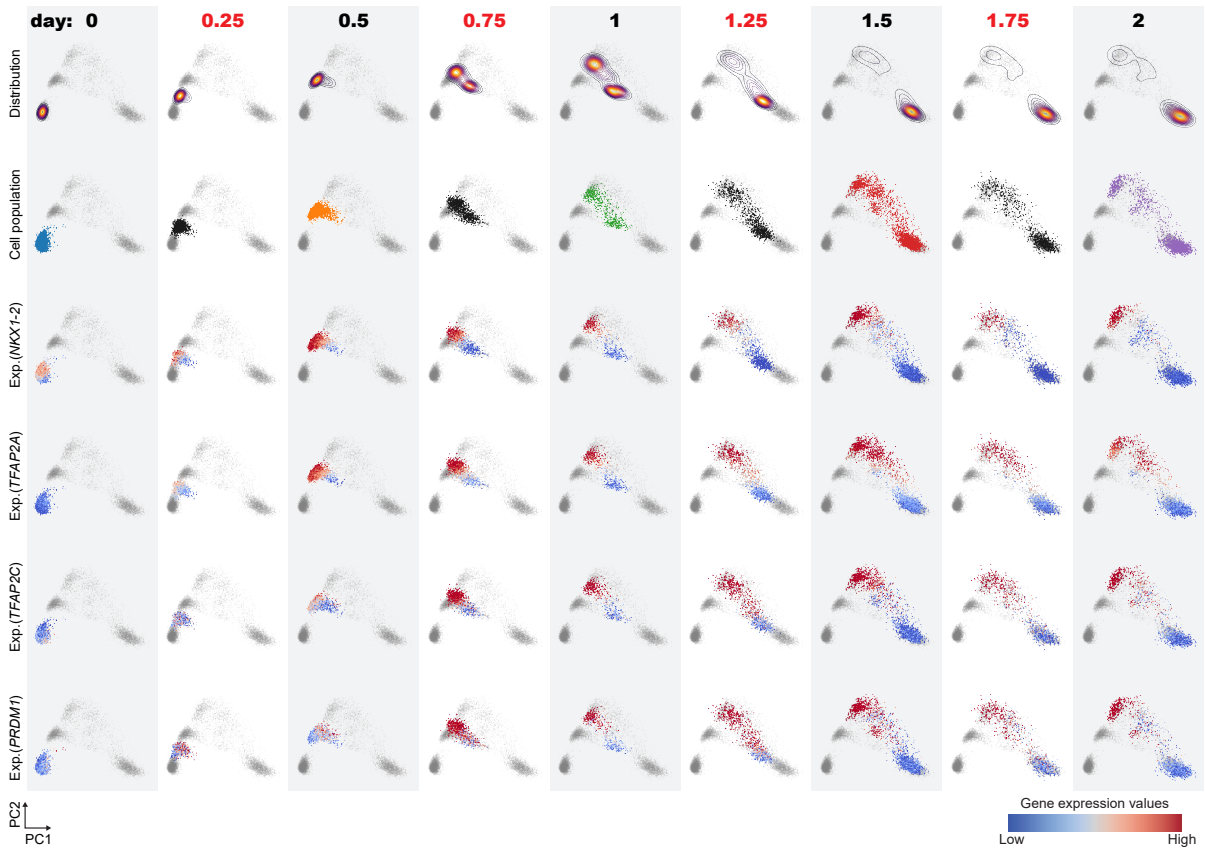


Figure 9: Input scRNA-seq data (gray columns) and interpolated data (white columns) on the top two principal components. The first row shows the contour plots of the Gaussian mixture distributions. The second row denotes the cell populations of the real scRNA-seq data (gray columns) and those generated by the interpolated Gaussian mixture distributions (white columns). The third to sixth rows show the gene expression values of *NKX1-2*, *TFAP2A*, *TFAP2C*, and *PRDM1* in the cell populations.

reconstructed Waddington's landscape together with the expression values of the genes *NKX1-2*, *MESP1*, and *GATA6*.

The landscape (Fig. 11A) shows that the iMeLCs, which are characterized by high levels of pluripotent genes (*SOX2*, *NANOG*, *POU5F1*) and relatively low levels of early mesoderm genes (*TBXT*, *MIXL1*, *EOMES*) (Fig. S1B), are located at the top of the potential. This suggests that these cells have high plasticity, which is consistent with biological knowledge.

In addition, viewing the gene expression values on the landscape allows us to understand the contribution of the genes to cell differentiation. For instance, it can be seen that *NKX1-2* contributes to the early stage of PGCLC (Fig. 11B). In contrast, *MESP1* and *GATA6* contribute to the early stage of EXMCLC and EDLC differentiation, respectively (Figs. 11C–D). Importantly, we emphasize that their expression levels are complementary at days 0.5-1, suggesting that these genes may play the role of landscape pegs, forming a barrier between the PGCLC and the somatic cell pathways.

3.9 Conclusion

We have formulated an entropic Gaussian mixture optimal transport (EGOT) and proved the mathematical properties of EGOT. Based on EGOT, we have also proposed a novel single-cell trajectory inference framework (scEGOT) that infers cell differentiation pathways and gene expression dynamics from time-series scRNA-seq data. Furthermore, we applied scEGOT to the human PGCLC induction system and validated it biologically.

ScEGOT provided us with the following outputs in biological applications. The cell state

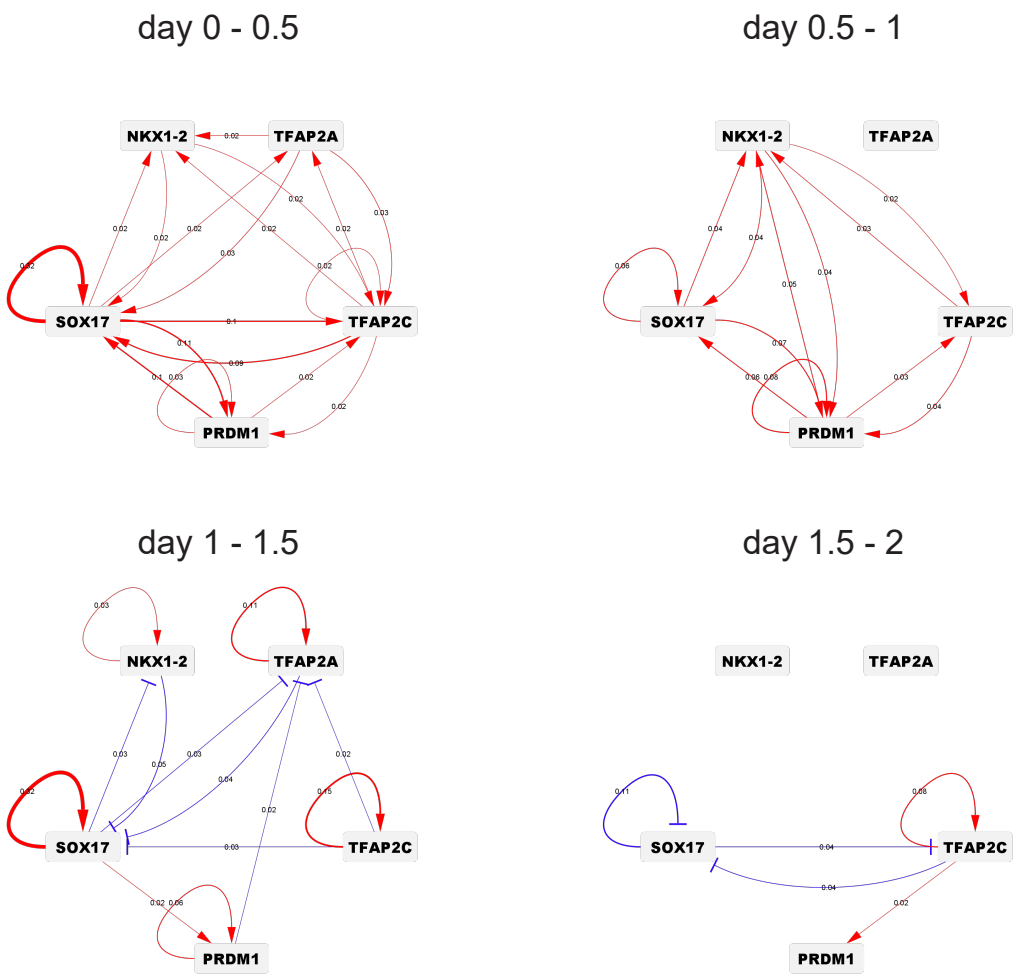


Figure 10: Gene regulatory networks of the human PGCLC induction system generated from scRNA-seq data at days 0–0.5 (top left), days 0.5–1 (top right), days 1–1.5 (bottom left), and days 1.5–2 (bottom right).

graph, which represents the cell clusters and their transport rates as a graph model, showed the trajectories between the cell populations. The cell velocity (11) generated by the entropic barycentric projection map (9) has expressed the cell differentiation flow, which describes more detailed and consistent structures in the case of human PGCLC than the conventional method. The scEGOT interpolation (12), which can generate a pseudo cell population at any intermediate time, has simulated the temporal evolution of gene expression dynamics in single-cell resolution and has certainly identified the bifurcation time. The GRN analysis, derived by solving the regression problem (15) between the gene expression and cell velocity, has contributed to the discovery of candidates of upstream genes that induce a cell differentiation system. The reconstruction of Waddington's landscape, constructed based on the potential of the gradient flow associated with the cell velocity, has summarized the pathway of cell differentiation and represented the cell potency of each cell type. We note that all the outputs represented by the PCA coordinates can be replaced by those with other coordinates generated by other dimensionality reduction methods, such as uniform manifold approximation and projection (UMAP) [25], which provides biological interpretations from different aspects (Fig. S2).

Through the abovementioned scEGOT analysis, we have identified genes such as *NKX1-2*, *MESP1*, and *GATA6* as potential regulators for human PGCLC differentiation or its divergence towards somatic lineages. The biological functions of these genes in human PGCLC are still unknown and are expected to be revealed through biological experiments.

Looking back on this paper, we have revisited an epigenetic world that was once envisioned

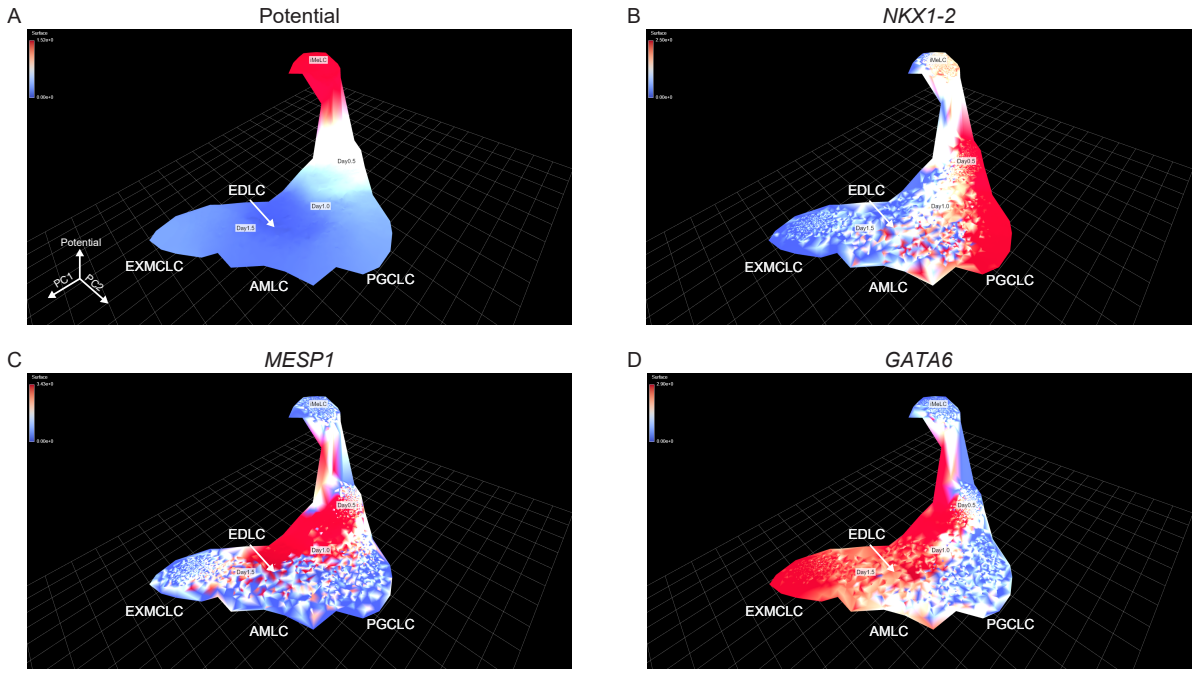


Figure 11: Reconstruction of Waddington's landscape of human PGCLC induction data. The x-, y-, and z-axes denote the PC1, PC2 coordinates, and the potential φ, respectively. The visualization was prepared using CellMapViewer https://github.com/yusuke-imoto-lab/CellMapViewer. The colors indicate (A) the magnitude of potential, (B) *NKX1-2*, (C) *MESP1*, and (D) *GATA6* expression values.

by Waddington. He viewed the process of cell development as a canal and stated that a cell within a region denoting a particular state would be carried through the canals to a single specified cell, which is the corresponding steady state. He modeled such a time transition as a phase-space diagram of development (Fig. 3 in [47]), which is a similar concept to the cell state graph. He then modeled the epigenetic landscape, which is a well-known potential model, by simply describing the phase-space diagram in a three-dimensional picture (Fig. 4 in [47]). He further explained that the chemical states of the genes and their regulatory system underlie the landscape model (Fig. 5 in [47]).

Overall, scEGOT has realized the images envisioned by Waddington through a comprehensive mathematical framework called EGOT and has succeeded in capturing complex cell differentiation systems and their dynamics by analyzing the cell state graph, cell velocity, interpolation analysis, gene regulatory network, and reconstruction of the epigenetic landscape.

We believe that scEGOT is useful for a broader range of biological data because it is a data-driven method. In other words, all scEGOT settings are never limited to scRNA-seq data. For example, scATAC-seq data containing open chromatin information could be applied in scEGOT to provide a deeper explanation of cell differentiation and lineage fate determination at epigenome resolution (Fig. 8D). Indeed, the epigenetic regulations during cell differentiation, such as the regulations of cis-regulatory elements and transposon elements, are fundamental questions, and our approach will be instructive in addressing such questions.

Acknowledgements

This work was supported by the World Premier International Research Center Initiative (WPI) and AMED-CREST Grant (JP19gm1310002h, Y.H.). This work was partially supported by JSPS Grant-in-Aid for Early-Career Scientists (21K13822, T.Y.), JST PRESTO (JPMJPR2021,

Y.I.), Grants-in-Aid for Specially Promoted Research from JSPS (17H06098, 22H04920, M.S.), Grants from the Open Philanthropy Project (2018-193685, M.S.), JSPS Grant-in-Aid for Transformative Research Areas (A) (22H05107, Y.H.), and JST MIRAI Program Grant (22682401, Y.H.).

Conflicts of Interest

The authors declare no conflict of interest.

Data availability

ScRNA-seq data of the human PGCLC induction system are available in the Gene Expression Omnibus (GEO) under identification number GSE241287.

Code availability

The Python software scEGOT is available through an open-source package at https://github.com/yachimura-lab/scEGOT.

References

[1] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38:1408–1414, 12 2020.

[2] C. Bunne, L. Papaxanthos, A. Krause, and M. Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.

[3] D. Bures. An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

[4] R. Cannoodt, W. Saelens, and Y. Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506, 2016.

[5] A. Castillo-Venzor, C. A. Penfold, M. D. Morgan, W. W. C. Tang, T. Kobayashi, F. C. K. Wong, S. Bergmann, E. Slatery, T. E. Boroviak, J. C. Marioni, and M. A. Surani. Origin and segregation of the human germline. preprint, Developmental Biology, July 2022.

[6] D. Chen, N. Sun, L. Hou, R. Kim, J. Faith, M. Aslanyan, Y. Tao, Y. Zheng, J. Fu, W. Liu, M. Kellis, and A. Clark. Human primordial germ cells are specified from lineage-primed progenitors. *Cell Reports*, 29:4568–4582.e5, 12 2019.

[7] Y. Chen, T. T. Georgiou, and A. Tannenbaum. Optimal transport for gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019.

[8] J. Delon and A. Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13:936–970, 2020.

[9] D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[10] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[11] C. R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

[12] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848, 2016.

[13] Y. Imoto, T. Nakamura, E. G. Escolar, M. Yoshiwaki, Y. Kojima, Y. Yabuta, Y. Katou, T. Yamamoto, Y. Hiraoka, and M. Saitou. Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. *Life Science Alliance*, 5(12):e202201591, Dec. 2022.

[14] N. Irie, L. Weinberger, W. Tang, T. Kobayashi, S. Viukov, Y. S. Manor, S. Dietmann, J. Hanna, and M. Surani. SOX17 Is a Critical Specifier of Human Primordial Germ Cell Fate. *Cell*, 160(1-2):253–268, Jan. 2015.

[15] Q. Jiang, S. Zhang, and L. Wan. Dynamic inference of cell developmental complex energy landscape from time series single-cell transcriptomic data. *PLOS Computational Biology*, 18(1):e1009821, 2022.

[16] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[17] D. Klein, G. Palla, M. Lange, Z. Piran, M. Gander, L. Meng-Papaxanthos, M. Sterr, A. Bastidas-Ponce, M. Tarquis-Medina, H. Lickert, M. Bakhti, M. Nitzan, M. Cuturi, and F. Theis. Mapping cells through time and space with moscot, 05 2023.

[18] Y. Kojima, K. Sasaki, S. Yokobayashi, Y. Sakai, T. Nakamura, Y. Yabuta, F. Nakaki, S. Nagaoka, K. Woltjen, A. Hotta, T. Yamamoto, and M. Saitou. Evolutionarily distinctive transcriptional and signaling programs drive human germ cell lineage specification from pluripotent stem cells. *Cell Stem Cell*, 21:517–532.e5, 10 2017.

[19] Y. Kojima, C. Yamashiro, Y. Murase, Y. Yabuta, I. Okamoto, C. Iwatani, H. Tsuchiya, M. Nakaya, T. Tsukiyama, T. Nakamura, T. Yamamoto, and M. Saitou. GATA transcription factors, SOX17 and TFAP2C, drive the human germ-cell specification program. *Life Science Alliance*, 4(5):e202000974, May 2021.

[20] K. Kurimoto, Y. Yabuta, Y. Ohinata, M. Shigeta, K. Yamanaka, and M. Saitou. Complex genome-wide transcription dynamics orchestrated by blimp1 for the specification of the germ cell lineage in mice. *Genes & Development*, 22(12):1617–1635, 2008.

[21] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[22] A. Lee, M. K. Scott, F. Wimmers, P. S. Arunachalam, W. Luo, C. B. Fox, M. Tomai, P. Khatri, and B. Pulendran. A molecular atlas of innate immunity to adjuvanted and live attenuated vaccines, in mice. *Nature communications*, 13(1):549, 2022.

[23] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[24] G. L. Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. Rna velocity of single cells. *Nature*, 560:494–498, 8 2018.

[25] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[26] G. McLachlan and D. Peel. *Finite mixture models*. Wiley Ser. Probab. Math. Stat. Chichester: Wiley, 2000.

[27] N. Moriel, E. Senel, N. Friedman, N. Rajewsky, N. Karaiskos, and M. Nitzan. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature protocols*, 16(9):4177–4200, 2021.

[28] M. Nitzan, N. Karaiskos, N. Friedman, and N. Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, 2019.

[29] Y. Ohinata, B. Payer, D. O'Carroll, K. Ancelin, Y. Ono, M. Sano, S. C. Barton, T. Obukhanych, M. Nussenzweig, A. Tarakhovsky, M. Saitou, and M. A. Surani. Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature*, 436(7048):207–213, 2005.

[30] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[31] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[32] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979–982, 2017.

[33] X. Qiu, Y. Zhang, J. D. Martin-Rufino, C. Weng, S. Hosseinzadeh, D. Yang, A. N. Pogson, M. Y. Hein, K. H. J. Min, L. Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

[34] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[35] M. Saitou, S. C. Barton, and M. A. Surani. A molecular programme for the specification of germ cell fate in mice. *Nature*, 418(6895):293–300, 2002.

[36] F. Santambrogio. *Optimal transport for applied mathematicians*, volume 55. Springer, 2015.

[37] K. Sasaki, S. Yokobayashi, T. Nakamura, I. Okamoto, Y. Yabuta, K. Kurimoto, H. Ohta, Y. Moritoki, C. Iwatani, H. Tsuchiya, S. Nakamura, K. Sekiguchi, T. Sakuma, T. Yamamoto, T. Mori, K. Woltjen, M. Nakagawa, T. Yamamoto, K. Takahashi, S. Yamanaka, and M. Saitou. Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell*, 17(2):178–194, Aug. 2015.

[38] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176:928–943.e22, 2 2019.

[39] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16, 2018.

[40] W. W. C. Tang, A. Castillo-Venzor, W. H. Gruhn, T. Kobayashi, C. A. Penfold, M. D. Morgan, D. Sun, N. Irie, and M. A. Surani. Sequential enhancer state remodelling defines human germline competence and specification. *Nature Cell Biology*, 24(4):448–460, Apr. 2022.

[41] A. E. Teschendorff and A. P. Feinberg. Statistical mechanics meets single-cell biology. *Nature Reviews Genetics*, 22(7):459–476, 2021.

[42] A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pages 9526–9536. PMLR, 2020.

[43] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

[44] L. Verrier, L. Davidson, M. Gierliński, A. Dady, and K. G. Storey. Neural differentiation, selection and transcriptomic profiling of human neuromesodermal progenitors-like cells in vitro. *Development*, page dev.166215, Jan. 2018.

[45] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[46] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[47] C. Waddington. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. Allen & Unwin, 1957.

[48] D. E. Wagner and A. M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(7):410–427, 2020.

[49] J. Weed. An explicit analysis of the entropic penalty in linear programming. In *Conference On Learning Theory*, pages 1841–1855. PMLR, 2018.

[50] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019.

[51] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E2467–E2476, 3 2018.

[52] A. Weismann. *The germ-plasm: a theory of heredity*. Scribner's, 1893.

[53] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20:1–9, 2019.

[54] Y. Yabuta, K. Kurimoto, Y. Ohinata, Y. Seki, and M. Saitou. Gene expression dynamics during germline specification in mice identified by quantitative single-cell gene expression profiling1. *Biology of Reproduction*, 75(5):705–716, 2006.

[55] K. D. Yang, K. Damodaran, S. Venkatachalapathy, A. C. Soylemezoglu, G. Shivashankar, and C. Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

[56] G. H. T. Yeo, S. D. Saksena, and D. K. Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):3222, 2021.

[57] Y. Zheng, R. Z. Yan, S. Sun, M. Kobayashi, L. Xiang, R. Yang, A. Goedel, Y. Kang, X. Xue, S. N. Esfahani, Y. Liu, A. M. Resto Irizarry, W. Wu, Y. Li, W. Ji, Y. Niu, K. R. Chien, T. Li, T. Shioda, and J. Fu. Single-cell analysis of embryoids reveals lineage diversification roadmaps of early human development. *Cell Stem Cell*, 29(9):1402–1419.e8, Sept. 2022.

Supporting information for scEGOT: Single-cell trajectory inference framework based on entropic Gaussian mixture optimal transport

Toshiaki Yachimura, Hanbo Wang, Yusuke Imoto, Momoko Yoshida, Sohei Tasaki, Yoji Kojima, Yukihiro Yabuta, Mitinori Saitou and Yasuaki Hiraoka

In the following, proofs of propositions and theorems in this paper are shown. In addition, figures and animations that could not be included in the paper are also given.

Proofs of propositions and theorems

Proof of Proposition 1

Since the discrete entropy H is a strongly concave function, the objective function of (2) is a strongly convex function. Therefore, the optimization problem (2) has a unique optimal solution. Also, we consider the following Lagrangian associated with the problem (2):

$$\mathcal{L}(w, f, g) = \sum_{k,l} w_{k,l} W_2^2(\mu_i^k, \mu_{i+1}^l) - \varepsilon H(w) - (f, w\mathbf{1}_{K_{i+1}} - \pi_i) - (g, w^T \mathbf{1}_{K_i} - \pi_{i+1}), \tag{S1}$$

where $f \in \mathbb{R}^{K_i}$ and $g \in \mathbb{R}^{K_{i+1}}$ are the Lagrange multipliers and $(\cdot, \cdot)$ denotes the standard Euclidean inner product of vectors. The first-order condition of (S1) yields that

$$\frac{\partial \mathcal{L}}{\partial w_{k,l}} = W_2^2(\mu_i^k, \mu_{i+1}^l) + \varepsilon \log(w_{k,l}) - f_k - g_l = 0.$$

Thus the unique solution $w = w^\varepsilon$ of (2) is given by

$$w_{k,l}^\varepsilon = e^{\frac{f_k}{\varepsilon}} e^{-\frac{W_2^2(\mu_i^k, \mu_{i+1}^l)}{\varepsilon}} e^{\frac{g_l}{\varepsilon}}. \tag{S2}$$

Putting $u_k = e^{\frac{f_k}{\varepsilon}}$ and $v_l = e^{\frac{g_l}{\varepsilon}}$ in (S2), then we obtain

$$w_{k,l}^\varepsilon = u_k \, e^{-\frac{W_2^2(\mu_i^k, \mu_{i+1}^l)}{\varepsilon}} \, v_l, \tag{S3}$$

which is the desired conclusion.

Proof of Proposition 2

We consider the convergence of EGOT and the solution (S3) as $\varepsilon \to 0$. By using an estimate similar to that for discrete entropic regularized optimal transports with Shannon entropy by Weed [2], we can show that the solution (S3) of EGOT exponentially converges to an optimal transport plan of d_G with the maximal entropy as $\varepsilon \to 0$. That is, there exists $M > 0$ independent of $\varepsilon > 0$ such that

$$\|w^\varepsilon - w^*\|_1 \leq e^{-\frac{M}{\varepsilon}}. \tag{S4}$$

By using the estimate (S4), we have

$$\begin{aligned}
\left|d_G^\varepsilon(\mu_{t_i},\mu_{t_{i+1}}) - d_G(\mu_{t_i},\mu_{t_{i+1}})\right| &= \left|\sum_{k,l} w_{k,l}^\varepsilon W_2^2(\mu_i^k,\mu_{i+1}^l) - \varepsilon H(w^\varepsilon) - \sum_{k,l} w_{k,l}^* W_2^2(\mu_i^k,\mu_{i+1}^l)\right| \\
&\le \sum_{k,l} \left|w_{k,l}^\varepsilon - w_{k,l}^*\right| W_2^2(\mu_i^k,\mu_{i+1}^l) + \varepsilon H(w^\varepsilon) \\
&\le \max_{k,l} W_2^2(\mu_i^k,\mu_{i+1}^l) \sum_{k,l} \left|w_{k,l}^\varepsilon - w_{k,l}^*\right| + \varepsilon H(w^\varepsilon) \\
&\le \max_{k,l} W_2^2(\mu_i^k,\mu_{i+1}^l) \|w^\varepsilon - w^*\|_1 + \varepsilon H(w^\varepsilon) \\
&\le \max_{k,l} W_2^2(\mu_i^k,\mu_{i+1}^l) e^{-\frac{M}{\varepsilon}} + \varepsilon H(w^\varepsilon).
\end{aligned}$$

From Jensen's inequality and the fact that the function $-\log t$ is convex, we obtain

$$\begin{aligned}
H(w^\varepsilon) &= -\sum_{k,l} w_{k,l}^\varepsilon (\log w_{k,l}^\varepsilon - 1) \\
&= 1 - \sum_{k,l} w_{k,l}^\varepsilon \log w_{k,l}^\varepsilon \\
&= 1 + \sum_{k,l} w_{k,l}^\varepsilon \log\left(\frac{1}{w_{k,l}^\varepsilon}\right) \\
&\le 1 + \log \sum_{k,l} w_{k,l}^\varepsilon \left(\frac{1}{w_{k,l}^\varepsilon}\right) \\
&= 1 + \log \sum_{k,l} 1 \\
&= 1 + \log K_i K_{i+1}.
\end{aligned}$$

This leads to the following estimate

$$\left|d_G^\varepsilon(\mu_{t_i},\mu_{t_{i+1}}) - d_G(\mu_{t_i},\mu_{t_{i+1}})\right| \le \max_{k,l} W_2^2(\mu_i^k,\mu_{i+1}^l) e^{-\frac{M}{\varepsilon}} + (1+\log K_i K_{i+1})\varepsilon,$$

which is the desired conclusion.

Proof of Theorem 3

As shown in [1, Proposition 4, p.944], the discrete optimal transport problem

$$d_G(\mu_{t_i},\mu_{t_{i+1}}) = \min_{w\in\Pi(\pi_i,\pi_{i+1})} \sum_{k,l} w_{k,l} W_2^2(\mu_i^k,\mu_{i+1}^l)$$

coincides with the continuous optimal transport problem (4). Therefore, by Proposition 2, we obtain the following estimate:

$$|d_G^\varepsilon(\mu_{t_i},\mu_{t_{i+1}}) - M W_2^2(\mu_{t_i},\mu_{t_{i+1}})| \le \max_{k,l} W_2^2(\mu_i^k,\mu_{i+1}^l) e^{-\frac{M}{\varepsilon}} + (1+\log K_i K_{i+1})\varepsilon.$$

Moreover, for any bounded continuous functions $\phi \in C_b(\mathbb{R}^n \times \mathbb{R}^n)$,

$$\begin{aligned}
&\left|\int_{\mathbb{R}^n\times\mathbb{R}^n} \phi(x,y)\, d\gamma^\varepsilon(x,y) - \int_{\mathbb{R}^n\times\mathbb{R}^n} \phi(x,y)\, d\gamma^*(x,y)\right| \\
&= \left|\int_{\mathbb{R}^n} \sum_{k,l} w_{k,l}^\varepsilon \phi(x, T_{k,l}(x)) g_{m_i^k,\Sigma_i^k}(x) dx - \int_{\mathbb{R}^n} \sum_{k,l} w_{k,l}^* \phi(x, T_{k,l}(x))\, g_{m_i^k,\Sigma_i^k}(x) dx\right| \\
&\le \|w^\varepsilon - w^*\|_1 \|\phi\|_\infty.
\end{aligned}$$

By using (S4), we obtain

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} \phi(x,y)\, d\gamma^\varepsilon(x,y) \to \int_{\mathbb{R}^n\times\mathbb{R}^n} \phi(x,y)\, d\gamma^*(x,y).$$

Thus, the entropic transport plan γ^ε converges to γ^* in the sense of the narrow convergence.

Proof of Proposition 4

Recall that the definition of the entropic barycentric projection map is

$$T^\varepsilon(x) := \int_{\mathbb{R}^n} y\, d\gamma^\varepsilon_x(y), \tag{S5}$$

where $\gamma^\varepsilon_x(y)$ is the disintegration of the entropic transport plan γ^ε with respect to the first marginal μ_{t_i}, i.e., $d\gamma^\varepsilon(x,y) = d\gamma^\varepsilon_x(y) d\mu_{t_i}(x)$. Then, by (8), the probability density of $\gamma^\varepsilon_x(y)$ is given by

$$\begin{aligned}\frac{d\gamma^\varepsilon(x,y)}{d\mu_{t_i}(x)} &= \frac{\sum_{k,l} w^\varepsilon_{k,l} g_{m^k_i,\Sigma^k_i}(x)\delta_{y=T_{k,l}(x)}}{\sum_j \pi^j_i g_{m^j_i,\Sigma^j_i}(x)} \\ &= \sum_{k,l} P^\varepsilon_{k,l}(x)\delta_{y=T_{k,l}(x)},\end{aligned}$$

where $P^\varepsilon_{k,l}$ denotes

$$P^\varepsilon_{k,l}(x) = \frac{w^\varepsilon_{k,l} g_{m^k_i,\Sigma^k_i}(x)}{\sum_j \pi^j_i g_{m^j_i,\Sigma^j_i}(x)}.$$

Thus, the entropic barycentric projection map (S5) can be computed as follows:

$$\begin{aligned}T^\varepsilon(x) &= \int_{\mathbb{R}^n} y\, d\gamma^\varepsilon_x(y) \\ &= \int_{\mathbb{R}^n} y \left(\sum_{k,l} P^\varepsilon_{k,l}(x)\delta_{y=T_{k,l}(x)}\right) dy \\ &= \sum_{k,l} P^\varepsilon_{k,l}(x) T_{k,l}(x),\end{aligned}$$

which is the desired conclusion.

Proof of Theorem 5

By the results of [1, Proposition 5 and Corollary 2, pp.946–948], the probability distribution

$$\mu^*_{t_{i+s}} = \sum_{k,l} w^*_{k,l}\mu^{k,l}_{t_{i+s}}$$

is the geodesic with the maximum entropy between μ_{t_i} and $\mu_{t_{i+1}}$ in the space $G_{2n}(\infty)$ equipped with the distance MW_2, where w^* is the solution of (6). Also, the proof that the entropic displacement interpolation $\mu^\varepsilon_{t_{i+s}}$ converges to $\mu^*_{t_{i+s}}$ narrowly can be shown in the same way as Theorem 3. Indeed, for any bounded continuous functions $\phi \in C_b(\mathbb{R}^n)$,

$$\begin{aligned}\left|\int_{\mathbb{R}^n} \phi\, d\mu^\varepsilon_{t_{i+s}} - \int_{\mathbb{R}^n} \phi\, d\mu^*_{t_{i+s}}\right| &= \left|\sum_{k,l}(w^\varepsilon_{k,l} - w^*_{k,l}) \int_{\mathbb{R}^n} \phi\, d\mu^{k,l}_{t_{i+s}}\right| \\ &\le \|w^\varepsilon - w^*\|_1 \|\phi\|_\infty \to 0.\end{aligned}$$

Here, we used (S4).

Proof of Theorem 6

Theorem 6 can be shown by direct calculation as follows:

$$\begin{aligned}
\operatorname{div} v &= \operatorname{div}\ (T^{\varepsilon} - x) \\
&= \operatorname{div}\ (T^{\varepsilon}) - n \\
&= \operatorname{div}\ \left(\sum_{k,l} P^{\varepsilon}_{k,l} T_{k,l} \right) - n \\
&= \sum_{k,l} \left((\nabla P^{\varepsilon}_{k,l}, T_{k,l}) + P^{\varepsilon}_{k,l} \operatorname{div} T_{k,l} \right) - n.
\end{aligned}$$

By the definition of $P^{\varepsilon}_{k,l}$, we have

$$\log P^{\varepsilon}_{k,l}(x) = \log g_{m^k_i, \Sigma^k_i}(x) - \log \sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x) + \log w^{\varepsilon}_{k,l}.$$

Then, we get

$$\frac{\nabla P^{\varepsilon}_{k,l}(x)}{P^{\varepsilon}_{k,l}(x)} = \frac{\nabla g_{m^k_i, \Sigma^k_i}(x)}{g_{m^k_i, \Sigma^k_i}(x)} - \frac{\sum_j \pi^j_i \nabla g_{m^j_i, \Sigma^j_i}(x)}{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x)}. \tag{S6}$$

Also, in the same manner as above, by the definition of $g_{m^k_i, \Sigma^k_i}$, we have

$$\log g_{m^k_i, \Sigma^k_i}(x) = -\frac{1}{2}(x - m^k_i)^T (\Sigma^k_i)^{-1} (x - m^k_i) + (\text{const.}).$$

Then, we obtain

$$\nabla g_{m^k_i, \Sigma^k_i}(x) = -g_{m^k_i, \Sigma^k_i}(x) (\Sigma^k_i)^{-1} (x - m^k_i) \tag{S7}$$

By substituting (S7) into (S6), we have

$$\begin{aligned}
\frac{\nabla P^{\varepsilon}_{k,l}(x)}{P^{\varepsilon}_{k,l}(x)} &= \frac{\nabla g_{m^k_i, \Sigma^k_i}(x)}{g_{m^k_i, \Sigma^k_i}(x)} - \frac{\sum_j \pi^j_i \nabla g_{m^j_i, \Sigma^j_i}(x)}{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x)} \\
&= \frac{-g_{m^k_i, \Sigma^k_i}(x) (\Sigma^k_i)^{-1} (x - m^k_i)}{g_{m^k_i, \Sigma^k_i}(x)} + \frac{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x) (\Sigma^j_i)^{-1} (x - m^j_i)}{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x)} \\
&= -(\Sigma^k_i)^{-1} (x - m^k_i) + \frac{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x) (\Sigma^j_i)^{-1} (x - m^j_i)}{\sum_j \pi^j_i g_{m^j_i, \Sigma^j_i}(x)} \\
&= C_k,
\end{aligned}$$

where we set C_k as the final expression. Therefore, we obtain

$$\nabla P^{\varepsilon}_{k,l} = P^{\varepsilon}_{k,l} C_k. \tag{S8}$$

Also, let us recall that the optimal transport map between μ^k_i and μ^l_{i+1} is given by

$$T_{k,l}(x) = m^l_{i+1} + A_{k,l}(x - m^k_i),$$

where $A_{k,l}$ denotes

$$A_{k,l} = (\Sigma^k_i)^{-\frac{1}{2}} \left((\Sigma^k_i)^{\frac{1}{2}} \Sigma^l_{i+1} (\Sigma^k_i)^{\frac{1}{2}} \right)^{\frac{1}{2}} (\Sigma^k_i)^{-\frac{1}{2}}.$$

The direct computation implies that the following equality holds:

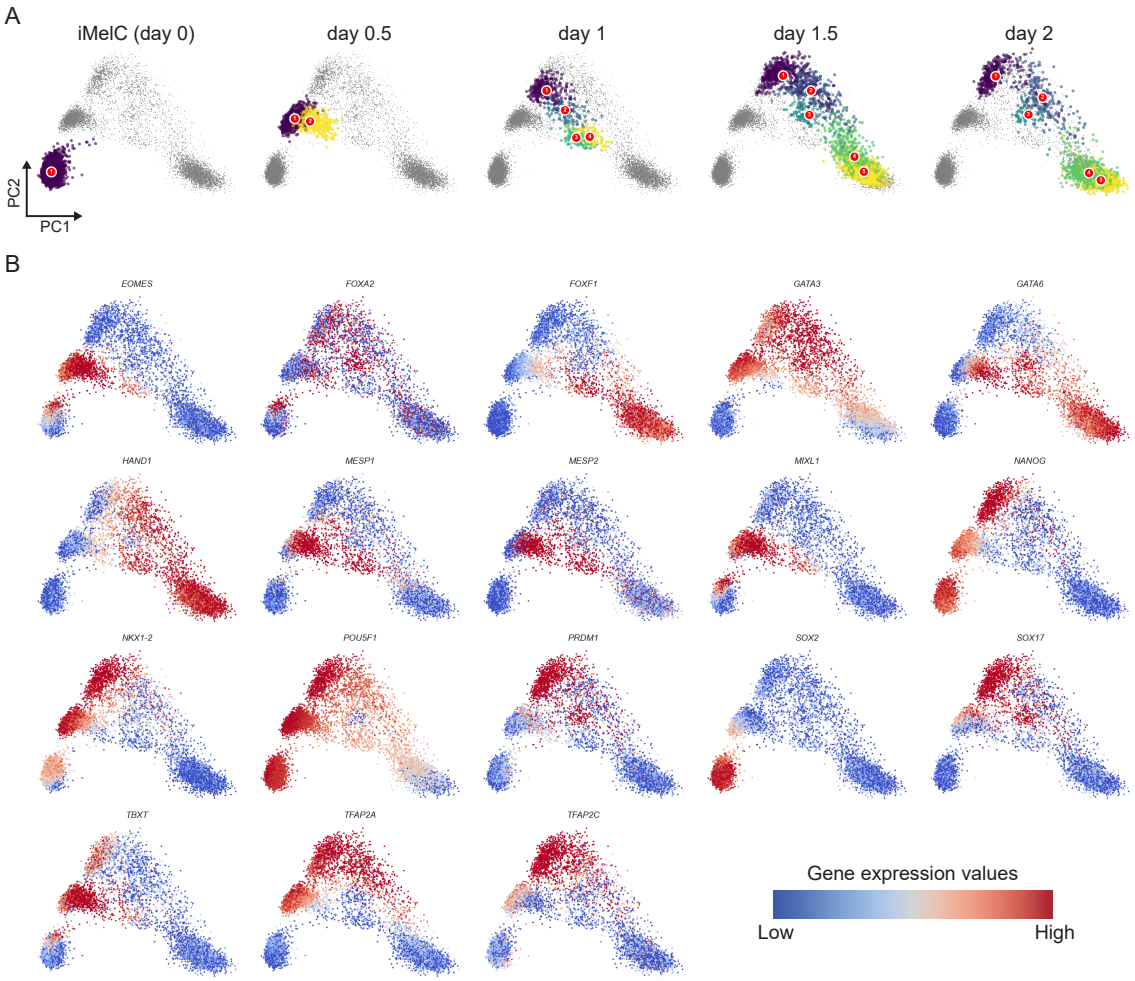
$$\operatorname{div} T_{k,l} = \operatorname{tr}(A_{k,l}). \tag{S9}$$

By (S8) and (S9), we obtain

$$\begin{aligned}
\operatorname{div} v &= \sum_{k,l} \left((\nabla P^{\varepsilon}_{k,l}, T_{k,l}) + P^{\varepsilon}_{k,l} \operatorname{div} T_{k,l} \right) - n \\
&= \sum_{k,l} \left((P^{\varepsilon}_{k,l} C_k, T_{k,l}) + \operatorname{tr}(A_{k,l}) P^{\varepsilon}_{k,l} \right) - n \\
&= \sum_{k,l} \left((C_k, T_{k,l}) + \operatorname{tr}(A_{k,l}) \right) P^{\varepsilon}_{k,l} - n,
\end{aligned}$$

which is the desired conclusion.

Figures



C

	PGCLC marker	AMLC marker	EDLC marker	EXMCLC marker	day 0	day 0.5		day 1				day 1.5					day 2					
					1	1	2	1	2	3	4	1	2	3	4	5	1	2	3	4	5	
NANOG	✓				2.77	0.74	-0.40	0.16	-1.13	-0.78	-0.75	0.47	-0.72	-1.01	-0.66	-0.63	1.08	-0.59	-0.61	-0.52	-0.49	*NANOG*
SOX17	✓		✓		-0.29	0.31	-0.14	1.64	0.37	-0.55	-0.27	1.35	-0.43	1.66	-0.52	-0.48	1.48	-0.28	1.65	-0.38	-0.38	*SOX17*
TFAP2C	✓	✓			0.23	-0.02	-1.12	1.18	-0.45	-0.74	-0.67	1.26	0.33	-0.81	-0.55	-0.54	1.45	0.05	-0.62	-0.38	-0.37	*TFAP2C*
PRDM1	✓				-0.42	-1.03	-0.91	0.14	-0.91	-0.84	-0.98	0.50	-0.62	0.05	-0.62	-0.64	0.67	-0.67	0.20	-0.61	-0.58	*PRDM1*
TFAP2A		✓			-0.45	1.38	0.11	0.69	-0.17	-0.88	-0.81	0.85	1.27	-0.58	-0.20	-0.46	0.59	1.21	-0.90	-0.38	-0.49	*TFAP2A*
GATA3		✓			-0.42	1.25	0.32	0.65	0.58	0.18	0.40	0.41	1.35	1.24	0.31	0.27	0.07	1.23	1.19	0.13	-0.04	*GATA3*
GATA6			✓		-0.53	-0.59	1.21	-1.26	0.08	0.83	0.54	-1.01	-0.41	1.22	1.00	1.26	-1.04	-0.21	1.21	1.06	1.23	*GATA6*
FOXA2			✓		-0.74	-1.47	-1.37	-1.17	-1.33	-0.93	-1.01	-1.13	-0.84	0.85	-0.79	-0.73	-1.16	-0.79	1.07	-0.71	-0.65	*FOXA2*
HAND1				✓	0.22	-0.47	0.32	0.61	2.14	2.14	2.35	0.54	2.43	0.94	2.97	2.86	-0.26	2.60	0.78	3.00	3.00	*HAND1*
FOXF1				✓	-0.79	-1.17	-0.76	-1.52	-0.32	0.12	0.27	-1.35	-0.66	-0.97	0.43	0.56	-1.21	-0.72	-1.05	0.53	0.33	*FOXF1*
NKX1-2					1.56	1.24	-0.04	0.36	-0.45	-0.85	-0.86	0.68	-0.28	-0.40	-0.62	-0.62	0.56	-0.44	-0.91	-0.66	-0.61	*NKX1-2*
MESP1					-0.35	0.96	2.56	0.00	1.85	1.65	1.55	-1.13	-0.56	-0.90	-0.05	-0.25	-0.98	-0.52	-0.81	-0.44	-0.39	*MESP1*
MESP2					-0.77	-1.13	0.21	-1.48	-0.26	0.64	0.24	-1.44	-0.87	-1.30	-0.69	-0.59	-1.24	-0.86	-1.20	-0.64	-0.55	*MESP2*

Figure 1: Clusters by the Gaussian mixture model (GMM) and marker gene's expression values. (A) Clusters on PCA coordinate for each day. (B) Expression values of 18 marker genes. (C) Mean z-score of marker gene expression values for each cluster.

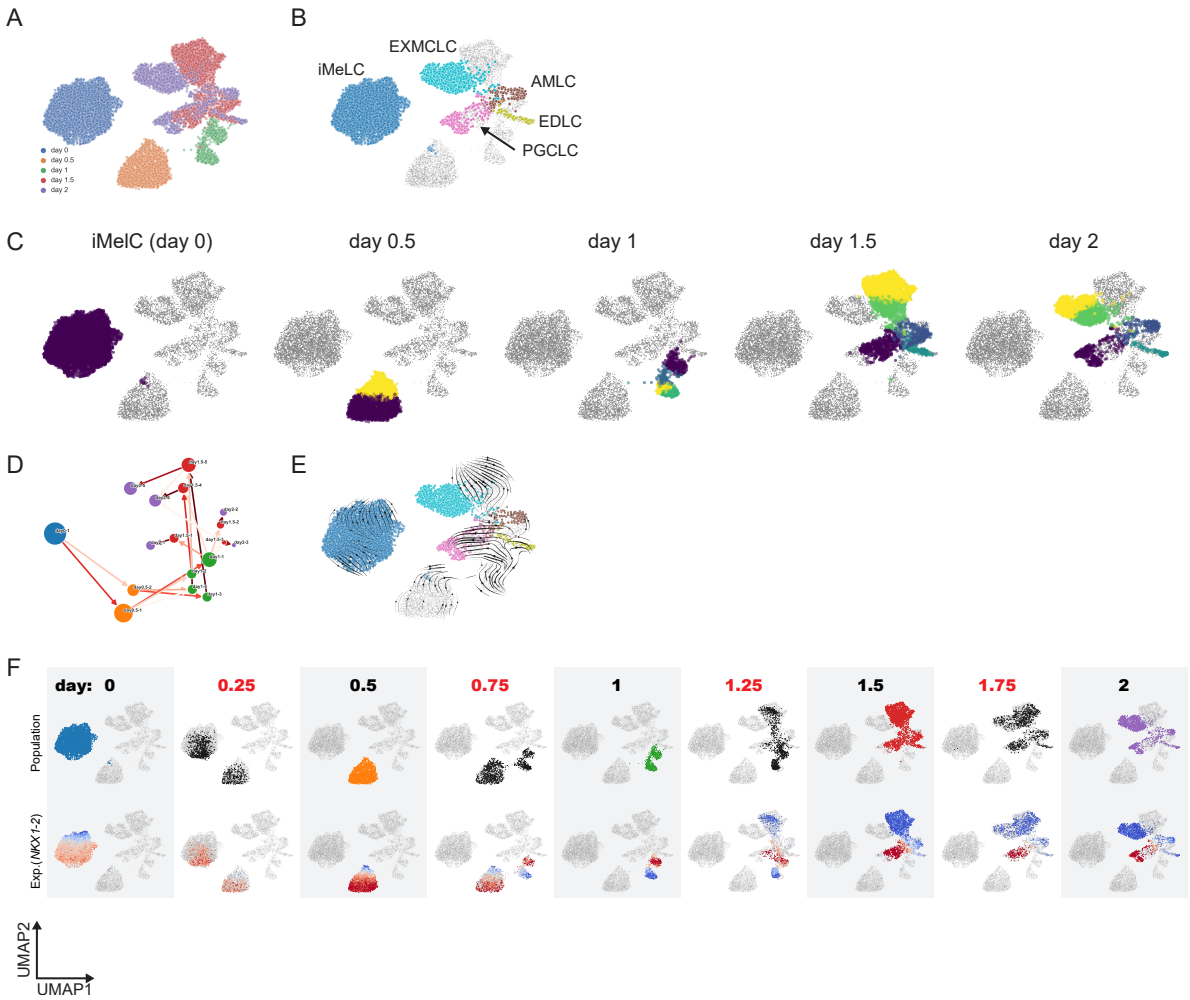


Figure 2: Visualization on the UMAP coordinate. UMAP plots are colored by (A) experimental days, (B) cell types, and (C) clusters by the Gaussian mixture model (GMM). (D), (E), (F) ScEGOT's outputs, cell state graph, cell velocity, and interpolation (top; cell population, bottom; *NKX1-2* expression value).

Movies

Movie S1. Time-continuous gene expression dynamics of *NKX1-2* on the PCA coordinates.

Movie S2. Time-continuous gene expression dynamics of *TFAP2A* on the PCA coordinates.

Movie S3. Time-continuous gene expression dynamics of *NKX1-2* on the UMAP coordinates.

Movie S4. Time-continuous gene expression dynamics of *TFAP2A* on the UMAP coordinates.

References

[1] J. Delon and A. Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13:936–970, 2020.

[2] J. Weed. An explicit analysis of the entropic penalty in linear programming. In *Conference On Learning Theory*, pages 1841–1855. PMLR, 2018.