OXFORD

Briefings in Bioinformatics, 2022, **23(4)**, 1–12
https://doi.org/10.1093/bib/bbac225
Advance access publication date: 14 June 2022
Problem Solving Protocol

Entropy-based inference of transition states and cellular trajectory for single-cell transcriptomics

Yanglan Gan, Cheng Guo, Wenjing Guo, Guangwei Xu and Guobing Zou

Corresponding author: G. Zou, School of Computer Engineering and Science, Shanghai University, 200444 Shanghai, China.
E-mail: gbzou@shu.edu.cn

Abstract

The development of single-cell RNA-seq (scRNA-seq) technology allows researchers to characterize the cell types, states and transitions during dynamic biological processes at single-cell resolution. One of the critical tasks is to infer pseudo-time trajectory. However, the existence of transition cells in the intermediate state of complex biological processes poses a challenge for the trajectory inference. Here, we propose a new single-cell trajectory inference method based on transition entropy, named scTite, to identify transitional states and reconstruct cell trajectory from scRNA-seq data. Taking into account the continuity of cellular processes, we introduce a new metric called transition entropy to measure the uncertainty of a cell belonging to different cell clusters, and then identify cell states and transition cells. Specifically, we adopt different strategies to infer the trajectory for the identified cell states and transition cells, and combine them to obtain a detailed cell trajectory. For the identified cell clusters, we utilize the Wasserstein distance based on the probability distribution to calculate distance between clusters, and construct the minimum spanning tree. Meanwhile, we adopt the signaling entropy and partial correlation coefficient to determine transition paths, which contain a group of transition cells with the largest similarity. Then the transitional paths and the MST are combined to infer a refined cell trajectory. We apply scTite to four real scRNA-seq datasets and an integrated dataset, and conduct extensive performance comparison with nine existing trajectory inference methods. The experimental results demonstrate that the proposed method can reconstruct the cell trajectory more accurately than the compared algorithms. The scTite software package is available at https://github.com/dblab2022/scTite.

Keywords: trajectory inference, transition entropy, cell clusters, transition cells, Wasserstein distance

Introduction

The rapid development of scRNA-seq technology enables researchers to study the cell types, states and transitions along various biological processes at single-cell resolution [1]. Specifically, understanding cell state transitions can provide new perspectives into cell differentiation, development, diseases and other complex cellular processes [2]. An important task for describing the cell state changes over time is to infer pseudo-time cell trajectories [3], which generally reconstructs the trajectory of cells according to the degree of similarity between cells and further infer the pseudo-time ordering to sort the cells [4]. The inferred trajectory helps us to identify branches and instrumental genes at the branching points, as well as study gene expression dynamics during a biological process [5].

In recent years, many trajectory inference algorithms have been proposed to infer pseudo-lineages of cells. These algorithms can be divided into two major categories [2, 6]. One category of algorithms infers trajectory based on the tree structure. The minimum spanning tree (MST) is constructed as the main path, and subsequently various measures are adopted to obtain the pseudo-time ordering and lineage structure. TSCAN constructs MST based on cluster centroid and then infers pseudo-time ordering of cells, which effectively reduces the

Yanglan Gan is a professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. She received the Ph.D. degree in Computer Science from Tongji University in 2012, China. She has worked as a Visiting Scholar in the Department of Computer Science and Engineering at Washington University in St. Louis from 2009 to 2011, USA. Her research interests include bioinformatics, data mining, and Web services. She has published more than 40 papers on international journals and conferences, including Bioinformatics, IEEE/ACM TCBB, BMC Bioinformatics, BMC Genomics, Knowledge-based Systems, and Soft Computing. She served as a program committee member on BIBM 2021 and GIW 2018. She worked as a reviewer for a variety of international journals and conferences, such as BMC Bioinformatics,IEEE TCBB, Knowledge-based Systems.
Cheng Guo is currently a master student in the School of Computer Science and technology, Donghua University, China. Before that, he received a Bachelor degree in Taiyuan University of Technology, 2020. His research interests include data mining and bioinformatics.
Wenjing Guo is an assistant professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. She received the Ph.D. degree from East China Normal University, Shanghai, China, in 2014. Her research interests include data mining, routing of the wireless and sensor networks.
Guangwei Xu is a professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. He received the M.S. degree from Nanjing University, Nanjing, China, in 2000, and the Ph.D. from Tongji University, Shanghai, China, in 2003. His research interests include data secure storage, data integrity verification and privacy protection, secure computing and sharing of outsourced data, QoS and routing of the wireless and sensor networks.
Guobing Zou is an Associate Professor in the School of Computer Engineering and Science, Shanghai University, China. He received his PhD in Computer Science from Tongji University, Shanghai, China, 2012. His current research interests focus on data mining, intelligent algorithms and services computing. He has published around 70 papers on premier international journals and conferences,including Information Sciences, Expert Systems with Applications, Knowledge-Based Systems, IEEE Transactions on Services Computing, AAAI, ICWS and ICSOC.

Received: February 18, 2022. **Revised:** May 11, 2022. **Accepted:** May 12, 2022

complexity of the tree space [7]. Monocle2 iteratively determines the stable positions of cells projected in a low-dimensional space based on the DDRTree algorithm, and the generated MST represents the inferred lineage structure [8, 9]. Slingshot constructs a cluster-based MST to represent the global lineages, and fits smooth branching curves to these lineages using simultaneous principal curves algorithm [10]. TiC2D firstly adopts consensus clustering to identify cell clusters, and then constructs the MST based on these clusters [11, 12]. The other category of algorithms infers cell trajectories mainly based on graph structures. These algorithms represent data as an undirected or directed neighborhood graph of single cells, where nodes correspond to cells and edges represent the neighborhood relations among cells. Then a suitable path in the graph is chosen to trace the biological process from progenitor cells to different fates. Wanderlust constructs inter cellular graph based on a KNN graph, choosing the shortest path in the graph to infer a linear trajectory [13]. Wishbone improves the Wanderlust method to identify bifurcated trajectories [14]. SLICER first constructs a K-Nearest Neighbors (KNN) graph among cells, then finds the shortest path between the initial cell and the final cell as a pseudo-time trajectory [5]. GraphDDP starts from a user-defined cluster assignment and then utilizes a force-based graph layout approach to infer differentiation trajectories [15]. PAGA constructs a KNN graph and applies the Louvain algorithm to partition the graph with multiple resolutions [16]. Based on a cell–cell similarity matrix, SoptSC infers pseudo time from a cell-to-cell graph, and predicts cell-lineage relationships between clusters using the minimal spanning tree of a cluster-to-cluster graph [6]. These trajectory inference algorithms have made significant progress in reconstructing cell trajectories. Generally, these algorithms are effective in reconstructing trajectories for the phenotypically and molecularly distant cell states. However, they are less robust in identifying intermediate or transitional cell states and inferring the corresponding trajectories related with transitory states [17].

A transitional state is an intermediate state during complex cellular process in which a cell exhibits a mixed identity between two or more states [18]. To study the dynamics of cellular programs governing fate transitions, it is crucial to identify the transition cells and to trace the critical transitions. Till now, only a few methods focus on the possible existence of transition cells between cell subpopulations. As a measure of variability [19], different forms of entropy have been applied to analyze single-cell transcriptomic data. These entropy can be roughly divided into two categories, including IntRAcellular and IntERcellular entropy [20]. IntRAcellular entropy is used to quantify the heterogeneity of cell transcription state. IntERcellular is used to quantify the heterogeneity of gene transcription in a group of cells. There are several algorithms for trajectory inference using intracellular entropy. SLICE introduces single-cell entropy (scEntropy) to measure cell differentiation states and predict cell differentiation lineages via the construction of entropy directed cell trajectories [21]. On the basis of signaling entropy, SCENT quantifies the expression heterogeneity in single-cell populations, and reconstructs cell-lineage trajectories from time-course data [22, 23]. CellRouter identifies complex cell-state transition trajectories by using flow networks to explore the subpopulation structure of single-cells [24]. Due to the high similarity between transition cells and differentiated cells, identifying the cellular states and transitional states and inferring the trajectories based on scRNA-seq data profiles remain great challenges.

To overcome these challenges, we propose a new single-cell trajectory inference method based on transition entropy, named scTite, to identify transitional states and reconstructs cell trajectory from scRNA-seq data. Taking into account the continuity of cellular processes, we introduce transition entropy to measure the uncertainty of cells belonging to different cell states. Based on the transition entropy, we identify transition cells and partition the other cells into different cell clusters. Correspondingly, we adopt different strategies to infer the trajectory for the identified cell states and transition cells. We utilize the Wasserstein distance based on the probability distribution to calculate the distance between these clusters, and construct the minimum spanning tree. Meanwhile, we adopt the signaling entropy and partial correlation coefficient to determine transition paths. Then, the constructed transitional paths and MST are combined to obtain a more detailed cell trajectory. To evaluate the performance of scTite, we apply it to four real scRNA-seq datasets and an integrated dataset with a complex branching structure containing known lineage structures and developmental time information. The experimental results demonstrate that the proposed method can more accurately reconstruct the cell trajectory than the compared algorithms.

Materials and methods

The overview of scTite

To reconstruct cell trajectories from scRNA-seq data, we propose scTite, a new single-cell trajectory inference algorithm based on transition entropy. The metric transition entropy is introduced to identify transition cells and cell states. Based on the potency state of cells, the Wasserstein distance is used to construct MST for cell clusters. Partial correlation analysis is adopted to infer the transition path between clusters. The algorithm consists of four main steps, as illustrated in Figure 1. Firstly, we reduce the dimension of the pre-processed scRNA-seq dataset using the UMAP algorithm [25], and partition the cells into different subpopulations based on the Expectation Maximization clustering algorithm. Secondly, based on the probability of each cell belonging to these partitions, we define and measure a new metric transition entropy and identify the transition cells whose transition entropy is higher than a given threshold. Thirdly, the signaling entropy is used to measure the differentiation potential of the cells and identify different cell clusters

[23]. We estimate the probability densities of these cell clusters, calculate the Wasserstein distance based on the probability densities, and construct the minimum spanning tree which is the approximate path among the identified cell clusters. Finally, the transition paths containing transition cells are constructed, and cells are projected onto the paths to obtain detailed cell trajectories as well as pseudo-time ordering.

Step 1. Performing dimension reduction and single-cell clustering

The input to scTite is a single-cell gene expression matrix X of size m×n, with m genes and n cells. The rows correspond to genes and columns correspond to cells. Each element x_{ij} of X represents the expression of gene i in cell j. To reduce the effect of noises and high-dimensionality, we adopt the UMAP algorithm to reduce the dimension of the data. Specifically, UMAP algorithm is a widely-used dimension reduction method based on manifold learning. Here, we compare UMAP with other five dimensionality reduction algorithms, including a classical algorithm (t-distributed Stochastic Neighbor Embedding (t-SNE) [26]) and four newly-developed algorithms (IVIS [27], PHATE [28], DCA [29] and SAUCIE [30]). We respectively utilize these six dimensionality reduction algorithms to obtain the low-dimensional representation of data. Then, Expectation Maximization (EM) algorithm is used to identify cell clusters, which are evaluated by Adjusted Rand index (ARI). We run each method 20 times on two test datasets (the Fibroblast and ESC dataset) and calculate the average ARI. From the supplementary Table 1, we observe that UMAP outperforms all the five compared methods in the test datasets, implying that it can better preserve the global structure of data.

In order to infer accurate cell trajectories, we partition cells into different states, at the same time we want to identify transition cells across different but related states. Here, we adopt the EM clustering algorithm to cluster the cells. The EM algorithm is a soft clustering method, which does not compute actual assignments of cells to clusters, but classification probabilities. Put differently, each cell belongs to each cluster with a certain probability. Correspondingly, we obtain a probability matrix *P* of size n×k as follows:

$$P_{i,k} = P(x_i \epsilon C_k) \quad \text{with} \sum_{k=1}^{K} P_{i,k} = 1 \tag{1}$$

where x_i is the *ith* cell, $C_k \epsilon (C_1, \ldots C_K)$ is the *kth* cluster and $P_{i,k}$ represents the probability of cell *i* belonging to cluster *k*. The number of clusters *K* is determined using Bayesian Information Criterion (BIC).

Meanwhile, we can assign these cells into different clusters based on the largest probability belonging to the specific cluster and keep the distribution probabilities of cells belonging to different clusters.

Step 2. Identifying cell clusters and the transition cells based on transition entropy

Based on the EM clustering algorithm, each cell is assigned to each cluster with a certain probability. From the probability matrix *P*, we can observe that a small part of cells have similar probabilities belong to all cell clusters, which implies that these cells exhibit a mixed identity between two or more subpopulations and might be involved in several functional groups. Then, these cells are identified as transition cells. On the contrary, the low transition entropy indicates that the cell is dominantly belong to a certain cell cluster, and cell *i* is assigned to the unique cluster*k* if the probability $P_{i,k}$ is much higher than those in other clusters. Therefore, the probability matrix *P* provides an intuitive way to identify transition cells and cell clusters. To quantitatively assess the ability of a cell transiting to other cell states, we introduce a new metric single-cell transition entropy (transEntropy) as a measure of cell plasticity. Similar to the Shannon entropy, transition entropy is defined as:

$$H(i) = -\sum P_{i,k} log(P_{i,k}) \tag{2}$$

$$P_{i,k} = \frac{d_{i,k}}{\sum_{j \epsilon K} d_{i,j}} \tag{3}$$

where $P_{i,k}$ is the probability that cell *i* belongs to the *kth* cluster, *K* is the number of cell clusters and *H*(*i*) represents the transition entropy of cell *i*.

When the probabilities of a cell belonging to different clusters are similar, the transition entropy is high. Here, we select the top *m*% of cells with the highest transition entropy as transition cells, which may be in an intermediate state during complex cellular processes. Otherwise, the low transition entropy indicates that the cell is dominantly belong to a cell state. In order to determine the appropriate threshold of transition entropy, we calculate the transition entropy of each dataset and perform scTite to get the pseudo-time ordering under different *m* values. Using POS as evaluation metric, we compare the accuracy of the pseudo-time ordering under different *m* values and choose the most suitable value. The detailed experimental results are described in the Supplementary Note 4. For different datasets, *m* ranging from 5 to 20 can achieve good performance.

Step 3. Constructing the minimum spanning tree based on the Wasserstein distance

Based on the above step, we obtain a set of cell clusters and a set of transition cells. Specifically, these distinct cell clusters reveal the biologically relevant heterogeneity, which also implies that these cells are in states of different functional plasticity [23]. For example, pluripotent cells have the capability to differentiate into other major cell lineages, exhibiting high differentiation potential and functional plasticity. In contrast, differentiated cells activate lineage-specifying signaling pathways and therefore have lower functional plasticity. Therefore, quantification of the functional plasticity for each

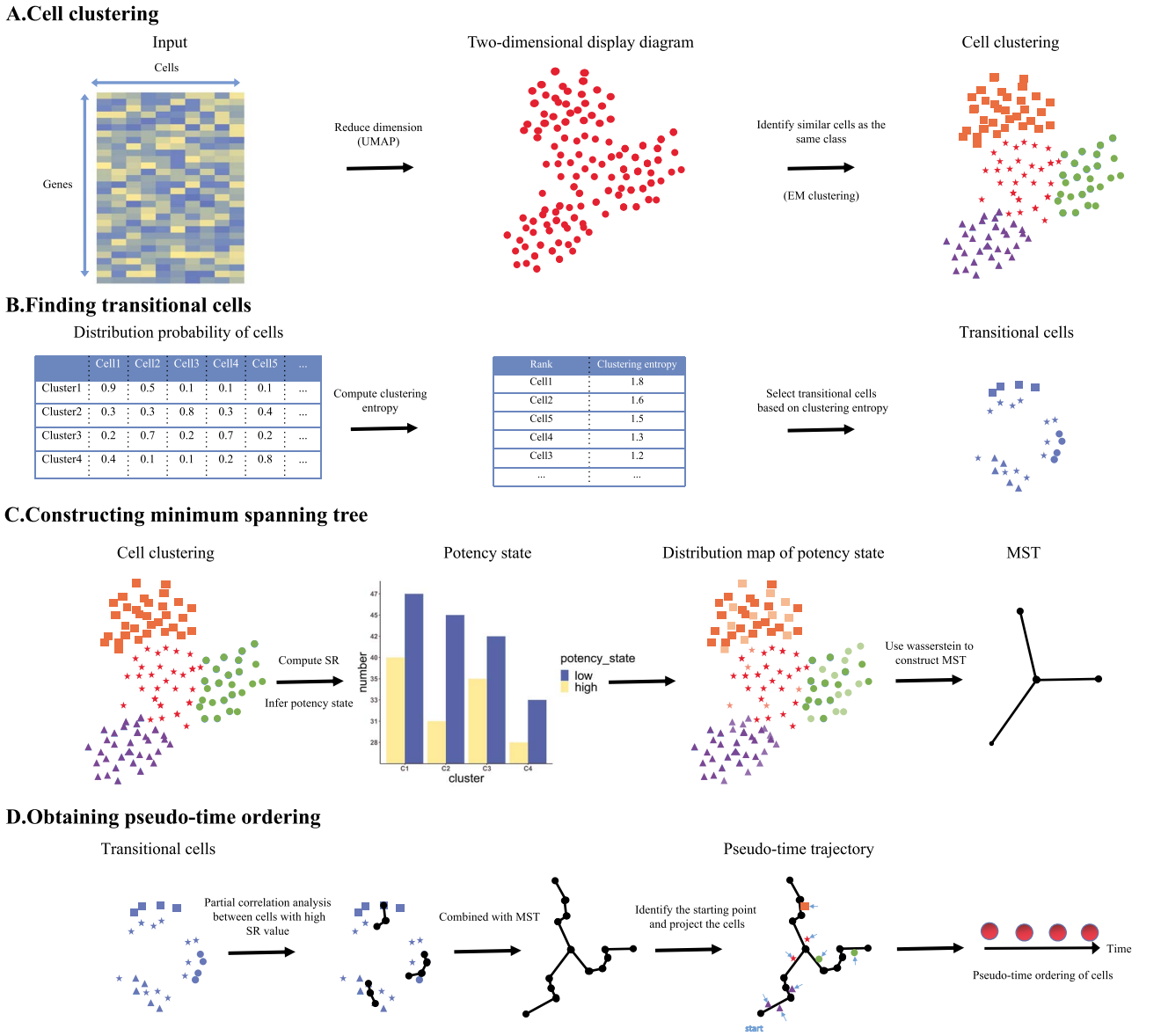

Figure 1. The overview of scTite. (A) Reducing the dimension of data and clustering cells. (B) Calculating the transition entropy of each cell, and the cells with the top 5 20 highest transition entropy are selected as transition cells. (C) Computing the signalling entropy (SR) of the cells, identifying the potency state, calculating the Wasserstein distance based on the probability density of the potency state in the cell clusters, and constructing the minimum spanning tree. (D) Constructing the network connection diagram between transition cells through partial correlation analysis. The transition trajectory is obtained by combining MST. Finally, the starting point is manually determined and the cells are projected onto the transition trajectory. Pseudo-time ordering is performed according to the projection length of the cells from the starting point.

identified cell cluster is critical to accurately infer cell trajectories. In previous studies, signaling entropy (SR) is introduced to measure the functional plasticity of cells. As in the method SCENT, the computation of signaling entropy need to estimate the interaction probabilities of proteins, which is assumed to be proportional to the normalized expression levels of the coding genes in the cell. Based on a curated protein-protein interaction (PPI) network, the edges between proteins are considered as signaling interactions, and the weights are interpreted as interaction probabilities. In detail, the weight of the edge between protein i and protein j is calculated as [22]:

$$w_{i,j} = \frac{n_{i,j}}{\sum_{k \epsilon N(i)} n_{i,k}} = \frac{n_{i,j}}{(An)_i} \tag{4}$$

where $N(i)$ represents the neighbors of node i in the PPI network, A is the adjacency matrix of the PPI network, $(An)_i$ represents the sum of interaction probabilities between the neighbors and protein i, $w_{i,j}$ represents the signaling probability between node i and node j.

Then, the entropy rate over the weighted network is defined as signaling entropy of the cell:

$$SE(\vec{X}) = -\sum_{i=1}^{m} \pi_i \sum_{j \epsilon N(i)} w_{i,j} log(w_{i,j}) \tag{5}$$

$$nSE(\vec{X}) = \frac{SE(\vec{X})}{maxSE} \tag{6}$$

where π_i denotes the invariant measure, satisfying $\pi_i P = \pi_i$ and the normalization constraint $\pi^t 1 = 1$. $nSE(\vec{X})$ is the normalized entropy rate, ranging from 0 to 1. *maxSE* indicates the maximum entropy in the cell.

According to the level of signal entropy, we can divide the cells into potency cell states. The optimal number of potency states is also determined by the Bayesian Information Criterion. For the identified potency states, we define the distribution probability of potency state j in cell cluster i as:

$$q_{i,j} = \frac{number(j)}{\sum_{j=1}^{h} number(j)} \tag{7}$$

where h represents the number of potency states, determined by Bayesian Information Criterion and $number(j)$ represents the number of cells belong to potency state j in the cell cluster.

For two cell clusters C_i and C_j, suppose there are S potency states. We can estimate the probability distributions of cell cluster C_i and C_j by Equation (7) as $C_i = \{q_{i1}, ..., q_{ih}\}$, $C_j = \{q_{j1}, ..., q_{jh}\}$. As the Wasserstein distance is usually used for measuring the distance between two probability distributions. Then we calculate the Wasserstein distance between different cell clusters as below [31, 32]:

$$W(C_i, C_j) = inf_{\gamma \sim \prod(C_i, C_j)} E_{(x,y)\sim\gamma}[\| x - y \|] \tag{8}$$

where $\prod(C_i, C_j)$ is the set of the joint distribution of C_i and C_j, the real sample x and the generated sample y are obtained by sampling from γ in the joint distribution, $\| x - y \|$ denotes the distance between samples, and $E_{(x,y)\sim\gamma}[\| x - y \|]$ is the expectation value of the samples for the distance under the joint distribution γ.

Finally, we use the Wasserstein distance as a distance metric to construct the minimum spanning tree, determining the main path of cell trajectory.

Step 4. Inferring transition trajectory and obtaining Pseudo-time ordering

Previous studies have shown that constructing MST at the cell cluster level can reduce the variability and complexity of the tree space [7]. However, when the identified cell clusters are close to each other, it is likely to ignore the influence of transition cells. In order to tackle these issues and more accurately infer cell trajectory, we further optimize the paths obtained from constructing MST by inferring the transition trajectory between clusters.

Partial correlation analysis refers to the analysis of the two variables only after excluding the influence of other variables when the two variables are related to other variables at the same time [33, 34]. When the number of cells is large, the calculation of the partial correlation coefficient may be time-consuming. To improve the time efficiency, we calculate partial correlation coefficient between cells based on the following strategy:

First, we project the transition cells onto the nearest edge of the constructed MST. For the transition cells on the same edge, we rank these cells in a descending order based on the signaling entropy.

Further, to determine the path along these transition cells, we calculate the partial correlation coefficient r_{ij} between the adjacent transition cell i and cell j. If the partial correlation coefficient $r_{ij} > 0$, cell i and cell j are connected, and the edge is assigned a weight r_{ij}. If $r_{ij} \leq 0$, the transition cell i and cell j are not connected. The cluster centroid is connected to the transition cells with the maximum and minimum signaling entropy, and construct a local directed graph for the transition cells on the path.

In the directed graph, we take the adjacent cluster centroid as the starting and ending point, and use the Dijkstra algorithm to find the path with the maximum sum of partial correlation coefficients, which is regarded as the transition path between clusters. Next, we adopt the principal curve algorithm to project the cells in the identified clusters to the transition path [35]. According to the projection value of the cells, we sort the cells and infer the pseudo-time ordering.

Datasets and data preprocessing

To demonstrate the effectiveness of scTite in inferring the cell trajectory, we apply the proposed algorithm to four real scRNA-seq datasets from Homo sapiens and Mus musculus, as well as an integrated dataset. The first data set is fibroblast-reprogramming-treutlein (Fibroblast) datasets, containing 355 cells directly reprogrammed from mouse embryonic fibroblasts (MEF) to induced neuronal (iN) cells on days 0, 2, 5, 20 and 22 [4]. The second dataset consists of 372 primary human skeletal muscle myoblasts (HSMM) single cells collected at 0, 24, 48, and 72 h [8]. The third dataset is germline-human-male weeks datasets (Germline), and contains 649 cells collected at 4, 9, 10, 19, 20, 21 and 25 weeks [4]. The fourth data set is mouse embryonic stem cells (ESC) and includes 2717 cells collected from days 0, 2, 4 and 7 [36]. The integrated dataset integrates snRNA-seq and scRNA-seq data, containing 11549 genes and 43791 cells reprogramming from adult dermal fibroblasts to primed and naive iPS cells [37]. The detailed information about the quantification and cell types of these datasets is described in the related articles. In the following, we respectively refer to these five datasets as Fibroblast, HSMM, Germline, ESC and Hrpi.

As previous studies did [38], we preprocess these four scRNA-seq datasets with three steps, including gene filtering, normalization, and log2-transformation. First, genes with zero gene expression values in all cells are filtered out [7]. The remaining gene expression values are subsequently normalized. To avoid a negative infinity result, when performing log2 transformation, we add a pseudo-count of 1 to the gene expression values. To further reduce the dimensions of the data, we choose the top υ% of genes with the greatest change in expression in all single cells. To determine an appropriate value for the filtering parameter υ, we

vary v from 5 to 20 in three datasets (Fibroblast, HSMM and ESC). Then, we use UMAP algorithm to obtain the low-dimensional representation of scRNA-seq data, and apply EM algorithm in mclust package to cluster cells. Adopting ARI as the evaluation metric, we compare the clustering accuracy under different v values, and select the most suitable value. The detailed experimental results are summarized in the Supplementary Note 3. For the Germline dataset, as the dimensionality has been reduced in the downloaded data, it is not further filtered in our analysis. As the dropout-events result in underestimating gene expression values and masking the correlation, dependency and other characteristics of the data, we need to further mitigate the effects of dropout-events. As MAGIC is an effective method to handle the dropout-events in single-cell datasets [39, 40], we apply it as an important data preprocessing step. To evaluate the effectiveness of MAGIC, we compare the clustering accuracy with and without MAGIC method (the Supplementary Note 5). The results demonstrate that the using of MAGIC can indeed improve the accuracy of clustering on the four datasets, which is important for further trajectory inference. For the Hrpi dataset, we directly utilize the processed Seruat objects from the original literature and select 10% of the cells for further analysis (the Supplementary Note 2).

Evaluation metrics

To evaluate the accuracy of the trajectory inference algorithms, we adopt the pseudo-temporal ordering score (POS) to measure how well the ordering of cells in the inferred trajectory matches the true one [7]. Here, it is assumed that external information is available to evaluate the pairwise order of cells. The POS is calculated as:

$$POS = \sum_{i=1}^{N-1} \sum_{j:j>i} g(i,j) \tag{9}$$

$$g(i,j) = \begin{cases} 0 & \text{if } T(i) = T(j) \\ (T(j) - T(i))/D & \text{otherwise} \end{cases} \tag{10}$$

where N represents the number of cells, D is the normalization parameter, allowing POS $\in$ [-1, 1], $g(i,j)$ is the score of each pair of cells $< i,j >$.

Assume that the real collection time for each cell is known in the dataset, the POS of the inferred pseudo-time ordering is defined as the degree to which the order of each pair of cells $< i,j >$ in the pseudo-time ordering matches the real collection time. If two cells are collected at the same time point, $g(i,j) = 0$; otherwise, $g(i,j)$ is positive if $T(i)$ represents an earlier time point, or negative if $T(i)$ represents a time point later than $T(j)$. $POS = 1$ indicates that the inferred pseudo-time ordering is completely consistent with the real order. $POS = -1$ represents that the inferred pseudo-time ordering is exactly the opposite of the true situation.

Meanwhile, we measure the robustness of the trajectory inference algorithm using Spearman's rank correlation coefficient [1], which is calculated as follows:

$$\rho = \frac{\sum_{i=1}^{N}(true_i - \overline{true})(rank_i - \overline{rank})}{\sqrt{\sum_{i=1}^{N}(true_i - \overline{true})^2}\sqrt{\sum_{i=1}^{N}(rank_i - \overline{rank})^2}} \tag{11}$$

where $rank(i)$ denotes the rank of cell i in the pseudo-time ordering, $true(i)$ denotes the rank of cell i in the real order, $\rho \in [-1, 1]$. The closer the ρ value is to 1, the more consistent the pseudo-time ordering is with the real order.

Results

Trajectory inference for real single-cell datasets

To assess the performance of scTite for inferring cell trajectories, we apply it to four real scRNA-seq datasets (Fibroblast, HSMM, Germline and ESC), which all containing pseudo-time information derived from the original study and labels for cell classification. For the four real datasets, we first perform soft clustering to identify the transitional cells and cell clusters. Then we construct the minimum spanning tree to connecting the identified clusters based on the Wasserstein distance, calculate the partial correlation coefficients between transition cells projected onto the same edge of the minimum spanning tree, and select the edge with the highest sum of similarities as the transition path. Finally the cells are projected onto the corresponding edges to obtain a pseudo-time ordering based on the projection length from the starting point. The reconstructed trajectories of these four datasets are respectively shown in Figure 2.

For the four dataset, the Bayesian Information Criterion is used to determine the optimal number of clusters. There are respectively 7, 4, 7 and 3 clusters in the Fibroblast, HSMM, Germline and ESC datasets, which is consistent with previous studies. Based on the defined transition entropy, we identify a set of transition cells in each dataset, which are labeled by heavy brown in the figure. The Fibroblast dataset mainly contains a range of cell subpopulations from mouse embryonic fibroblasts (MEF) to induced neuronal (iN) cells, and scTite identifies two branch trajectories, including $start \rightarrow 1$ and $start \rightarrow 2$ (Figure 2A). This finding is consistent with the tree-like topology for the differentiation trajectory of these cell subpopulations, as mentioned in previous study [4]. For the HSMM data set, previous studies have shown that the dataset contains two differentiation trajectories, one is affected by contaminated interstitial mesenchymal cells and the other is the main differentiation trajectory [7, 12]. As shown in Figure 2B, scTite also reconstruct two trajectories: $start \rightarrow 1$ and $start \rightarrow 2$, after identifying subpopulations of cells corresponding to the 0, 24, 48 and 72h, which is consistent with the differentiation process of the HSMM dataset. For the Germline data set, scTite identifies seven cell subpopulations according to the Bayesian Information Criterion, respectively correspondent to the cells collected at weeks 4, 9, 10, 19, 20, 21 and 25. Subsequently we identify two branch trajectories,

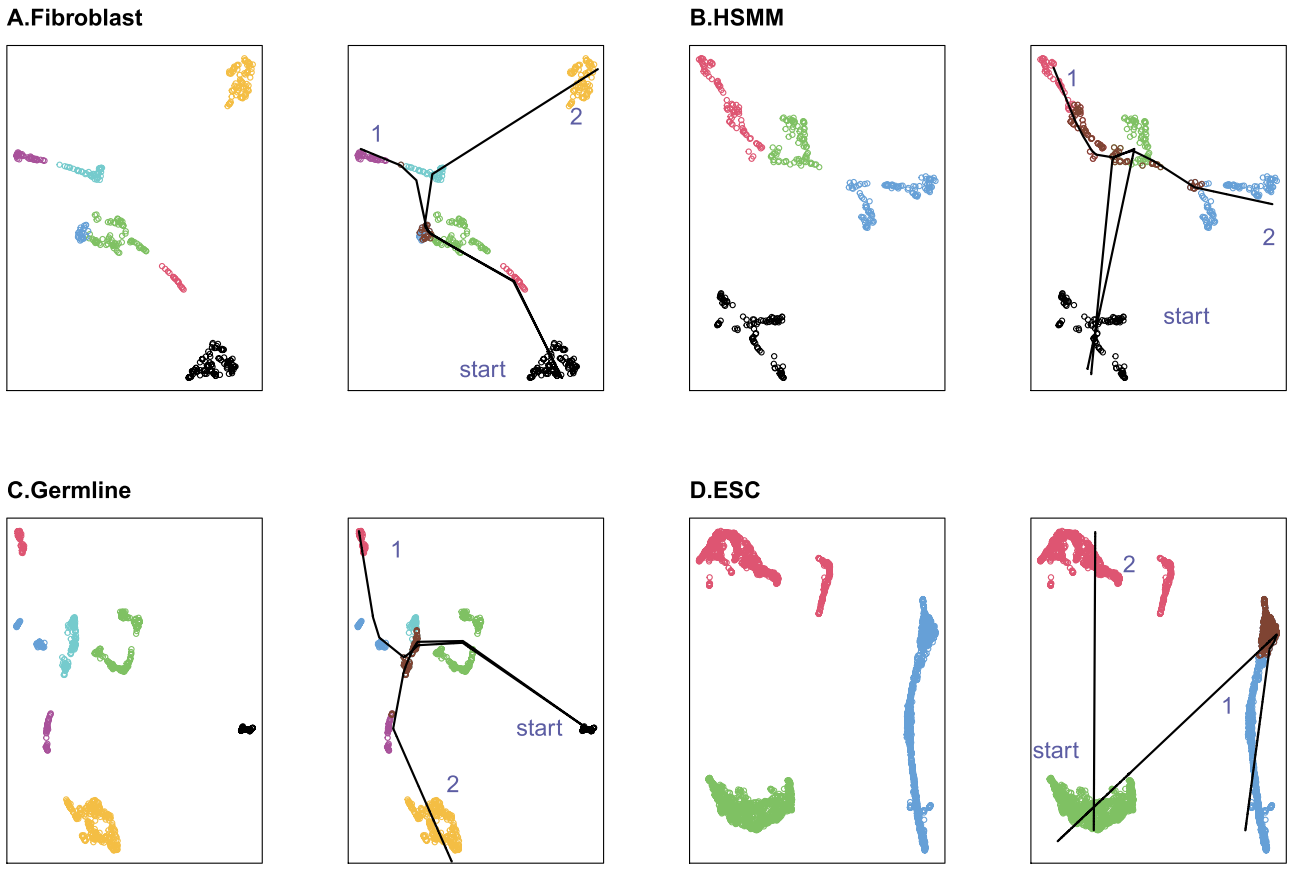

Figure 2. The proposed algorithm scTite reconstructs the pseudo-time trajectories of four scRNA-seq datasets (Fibroblast, HSMM, Germline and ESC). The heavy brown cells in the figure represent transition cells. (A) Pseudo-time trajectory of Fibroblast data set. (B) Pseudo-time trajectory of HSMM data set. (C) Pseudo-time trajectory of Germline data set. (D) Pseudo-time trajectory of ESC data set.

Table 1. The comparison results between Wasserstein and other five distance metrics

Distance metric	Fibroblast	HSMM	Germline	ESC
Wassertein	**0.971**	**0.734**	**0.626**	**0.742**
Euclidean	0.927	0.719	0.622	0.709
Manhattan	0.877	0.635	0.585	0.666
Chebyshev	0.926	0.716	0.586	0.719
Cosine	0.745	0.600	0.554	0.599
Jensen-shannon	0.745	0.602	0.588	0.591

including *start* → 1 and *start* → 2 (Figure 2C). For the ESC dataset, based on the constructed minimum spanning tree, scTite infers the differentiation trajectories containing two branches: *start* → 1 and *start* → 2 (Figure 2D), respectively.

To further validate the effectiveness of scTite on complex datasets, we apply it to the integrated dataset (the Hrpi dataset), which has been analyzed with Monocle3. The Hrpi dataset contains 43791 cells and 11549 genes reprogramming from adult human dermal fibroblasts into primed and naive iPS cells. We directly download the Seruat object processed by Liu et al. [37], including the preprocessed dataset and detailed experimental data. Then we extract about 10 % of the cells. Secondly, we use the BIC strategy to determine the number of clusters, which is set as 21. Finally, we apply scTite to infer the trajectory on the Hrpi dataset. As shown in Supplementary Figure 1, scTite identifies transition cells and reconstructs the branch trajectory of adult human dermal fibroblasts reprogramming into primed and naive iPS cells. The first branch trajectory *start* → 1 corresponds to the differentiation from fibroblasts into refractory cells; the second trajectory *start* → 2) corresponds to the differentiation from fibroblasts into primed iPS cells; the third trajectory *start* → 3 corresponds to the Trophectoderm branch during reprogramming; the fourth trajectory *start* → 4 corresponds to the differentiation from fibroblasts into naive iPS cells. This inferred trajectory is consistent with the branch topology found in the original literature. Subsequently, we project the cells onto the trajectory to get the pseudo-time ordering of the cells. This result demonstrates that scTite is capable of inferring complex trajectory from large datasets.

To validate the effectiveness of the Wasserstein distance for constructing MST, we replace it with five other widely used distance metrics in scTite and compare the corresponding performance. Here, we utilize POS to evaluate the performance of scTite with different distance metrics. The comparison results are shown in Table 1. On the Fibroblast data set, the POS of scTite using the Wasserstein distance is 0.971, which is 4.4% higher than that with the Euclidean distance. On the HSMM data set, the POS achieved by scTite using the Wasserstein distance is 0.734, which is 1.5% higher than that with the

Euclidean distance. On the Germline data set, the POS obtained by the algorithm is 0.626. On the ESC data set, the accuracy of the algorithm using the Wasserstein distance is 0.742, 2.3% higher than the Chebyshev distance with the second score. The results on these four datasets demonstrate that using the Wasserstein distance as a distance metric to construct the inter cluster MST outperforms the compared distance metrics.

Performance comparison

We further perform a comparative evaluation of scTite with state-of-the-art trajectory inference methods. As the previous comparison shows that TSCAN, Slingshot, MATCHER [41], PhenoPath [42] and CellTrails [43] are excellent trajectory inference methods, outperforming most existing methods. Here, we compared scTite with these five representative methods. Also, our comparison study includes four newly developed methods (TiC2D, SCOUT, Monocle3 [44], PAGA).

- TSCAN constructs MST to connect the cluster centroid after clustering similar cells, and further infers the lineage structure of cells.
- Slingshot also constructs the MST based on cluster centroid, utilizes simultaneous principal curves to merge the shared region of the smooth branching curves, and finally obtains the entire lineage structure.
- TiC2D adopts the Louvain algorithm to group genes and then performs ensemble clustering to obtain consensus subgroups of cells, construct the MST based on cluster centroid, and utilizes the principal curve algorithm to infer smooth differentiation trajectory.
- SCOUT uses the fixed-radius near neighbors algorithm based on cell density to find landmarks representing cell states, and constructs the MST based on landmarks. Then, the projection of Apollonian circle or a weighted distance is selected to determine the pseudo-time trajectory.
- PAGA allows graph partitions in low-dimensional space with different resolutions, and combines the highly confidence path with the distance metric based on random walk to sort cells.
- MATCHER uses Gaussian process latent variable model to infer the pseudo-time value of each type of data, and uses a non-linear function to map the pseudo-time value to the ”master time” value.
- PhenoPath utilizes the Bayesian statistical framework that integrates linear regression and latent variable modeling, as well as the covariants to learn the pseudo-time axis shared by different data objects.
- CellTrails exploits the k nearest neighbor information of cells in the low-dimensional space to estimate the geometric proximity of states, and then uses the trajectory fitted by straight lines passing through the geometric median of adjacent states.

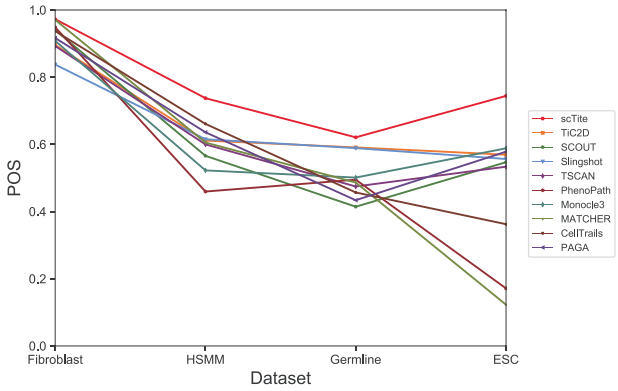

Figure 3. The performance comparison of scTite with nine state-of-the-art pseudo-time trajectory inference algorithms. The accuracy is measured by POS.

- Monocle3 is improved version of Monocle2, which can used for large scRNA-seq dataset. The main improvement strategy include using UMAP for dimension reduction and learning the principal graph based on PAGA graph.

In the comparative analysis, we utilize the widely used metric POS to evaluate the accuracy of the inferred pseudo-time trajectories. Figure 3 shows the performance of these trajectory inference algorithms on four real scRNA-seq datasets. On the whole, we observe that scTite performs better than the other nine competitive methods. Specifically, it is significantly better than the suboptimal method on HSMM, Germline and ESC datasets, and slightly better on Fibroblast datasets. In particular, on the Fibroblast data set, the POS of scTite is about 0.973, slightly higher than that of the suboptimal method MATCHER. For the dataset HSMM, the POS of scTite is about 0.738, which is about 7.7% higher than that of the suboptimal algorithm CellTrails. For the Germline dataset, the POS of scTite is about 0.621, while the highest POS of the other nine methods is about 0.591. For the ESC dataset, scTite significantly increases by 15.6% in the POS compared with the suboptimal method Monocle3.

Meanwhile, we evaluate the robustness of these compared algorithms based on the Spearman rank correlation coefficient. Figure 4 shows the comparison result of these ten trajectory inference algorithms on the four real scRNA-seq datasets. In detail, for the Fibroblast dataset, the Spearman coefficient of scTite is about 0.931, which is about 0.1% higher than the suboptimal algorithm MATCHER. For the HSMM data set, compared with the second method MATCHER, scTite significantly increased the Spearman coefficient by 3.7%. For the Germline dataset, the Spearman coefficient of scTite is about 0.771. It is about 4.6% higher than that of the suboptimal algorithm Slingshot. For the ESC dataset, the Spearman coefficient of scTite is about 0.654, which significantly increases by 10.8% compared with the second ranking method Slingshot. Overall, the result

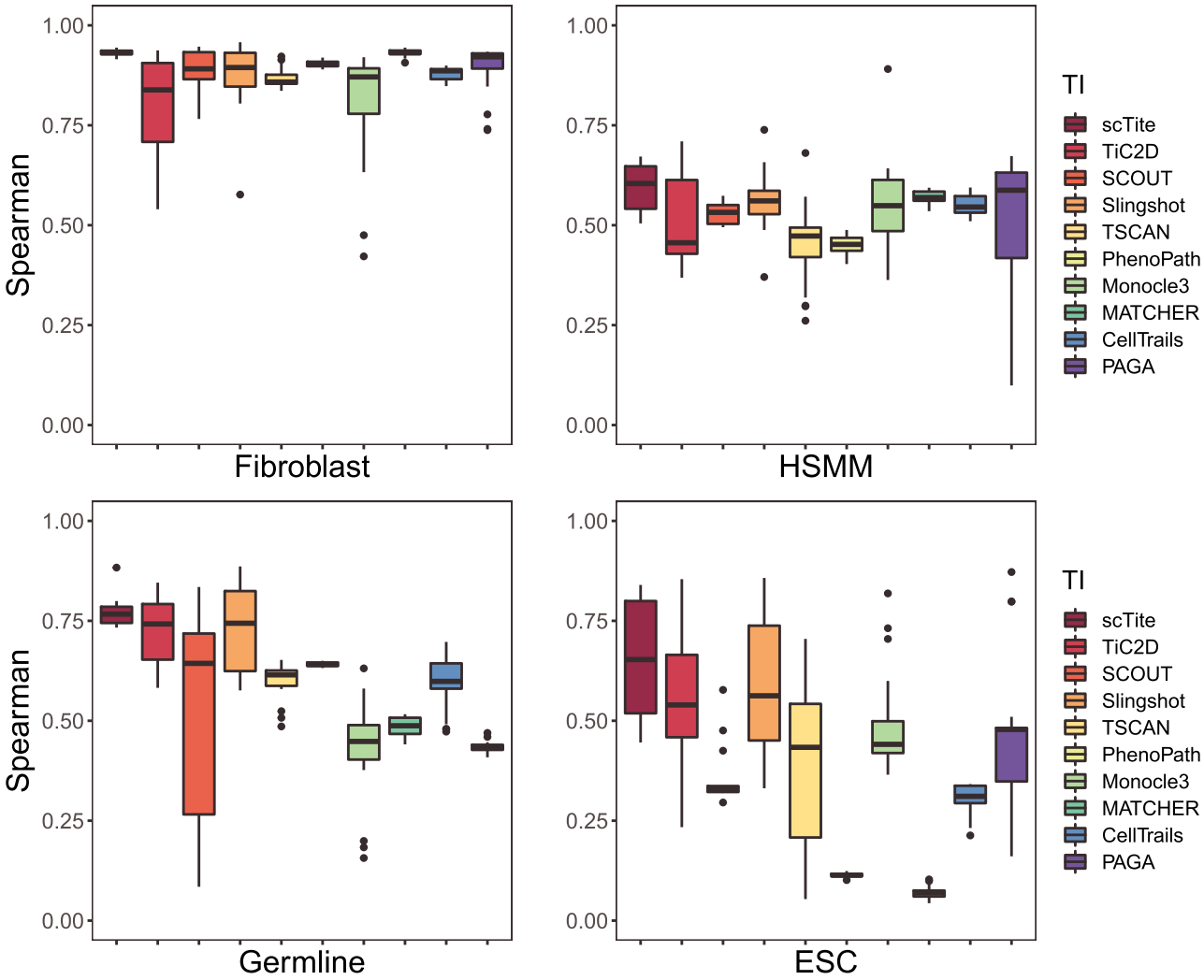

Figure 4. The performance comparison of scTite with nine state-of-the-art pseudo-time trajectory inference algorithms. The robustness is measured by the Spearman rank correlation coefficient.

demonstrates that scTite outperform the other nine algorithms with the respect of robustness.

To evaluate the efficiency of the proposed method, we compare the time cost of scTite with the other nine algorithms on four datasets. The experimental results demonstrate that scTite exhibits a high efficiency in inferring trajectory, which is similar with Slingshot and TSCAN. The detailed comparison is shown in the Supplementary Note 6.

Gene expression analysis

Further, to better understand complex biological process, we investigate to what extent the important genes are involved in the cell trajectories. Specifically, by ordering the expression of the important genes along the pseudo-time ordering, each part of the cell trajectory can be studied in terms of the behavior of certain genes.

Random Forest is a classifier composed of multiple decision trees, each of which can predict the category of input sample data independently. When classifying sample data, random forest can also calculate the importance score of each variable to evaluate the role of each variable in classification. Based on the input gene expression matrix, previous work [12] uses random forest to calculate the mean decrease in mean squared error (MSE) caused by genes, so as to obtain the importance score of genes for pseudo-time ordering. Subsequently, we perform descending order according to the importance score of genes. For each dataset, we select the top three genes with the highest importance for further analysis. We compare the spearman coefficients between the expression patterns of these marker genes along the true pseudo-time and the inferred orderings obtained by different methods.The comparison results are shown in Figure 5. On the whole, the gene expression profiles along the trajectories predicted by scTite are highly correlated with those along the real trajectories. In detail, for the Fibroblast dataset, we select Timp1, S100a6 and Tagln2 genes, which are related to positive regulation of fibroblast proliferation and epithelial cell differentiation. The spearman coefficients of scTite are 0.811, 0.848 and 0.790, respectively, which are slightly higher than the suboptimal algorithm SCOUT. For the HSMM dataset, we select TPD52L1, CCND3 and RNF41,

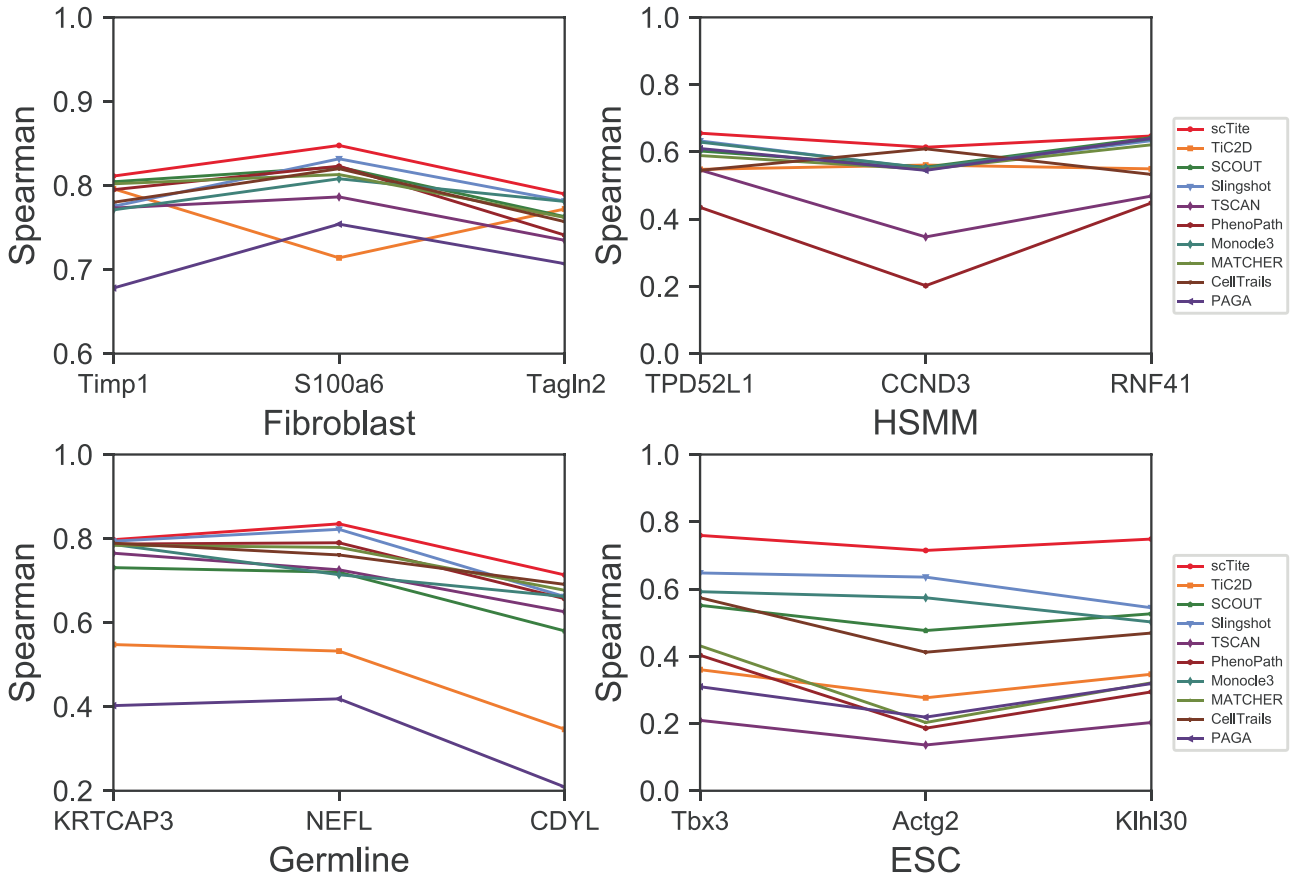

Figure 5. The performance comparison of scTite and other nine algorithms. The Spearman correlation coefficient is utilized to evaluate the consistence of the expression patterns of the marker genes along the true pseudo-time and those of the inferred orderings obtained by different methods.

which are associated with G2/M transition of mitotic cell cycle and T cell proliferation. The spearman coefficients of scTite are 0.656, 0.614 and 0.647, respectively, which are generally higher than the suboptimal algorithm Monocle3. On Germline data set, the three genes (KRTCAP3, NEFL and CDYL) exhibit the highest importance, which are associated with Neurotransmitter receptors and postsynaptic signal transmission. For these genes, the Spearman coefficients of scTite are 0.797, 0.835 and 0.713, respectively, which are generally higher than the suboptimal algorithm Slingshot. For the ESC data set, we select Tbx3, Actg2 and Klhl30, which are associated with regulating pluripotency of stem cells. The spearman coefficients of scTite are 0.760, 0.715 and 0.749, respectively, which are slightly higher than the suboptimal algorithm Slingshot. For the three important genes in each dataset, scTite can obtain higher spearman score, which means that the gene expression pattern of the important genes along the trajectory generated by scTite is more consistent with the real gene expression pattern than other algorithms, demonstrate that scTite can infer a more accurate pseudo-time ordering.

Conclusion

Recent studies have considered cell differentiation to be a continuous process of differentiation. The clustering approach assumes that cells are in a discrete state and cannot reflect the essence of cell differentiation. On the one hand, scRNA-seq datasets are noisy and heterogeneous, resulting in unstable trajectory inferred at the single cell level. On the other hand, the pseudo-time trajectory inferred in tree space does not capture the continuity of cell differentiation. To address the above issues, we propose a new trajectory inference algorithm scTite, designed to identify transition cells to describe the continuous differentiation process of cells. We first cluster single cells based on EM clustering, which can assign a cell to different clusters with a probability distribution. As transition state is an intermediate state in which a cell exhibits a mixed identity between two or more state. We define a new metric transition entropy to identify transition cells, adopt the Wasserstein distance to measure the distance between the identified clusters and construct the minimum spanning tree, and conduct partial correlation analysis to infer the detailed path between transition cells. Finally, we reconstruct cell differentiation trajectories on four real scRNA-seq datasets and compare the algorithm with nine state-of-the-art trajectory inference algorithms. The comparison results demonstrate that scTite can more accurately and robustly identify the pseudo-time trajectory of scRNA-seq data.

Key Points

- The development of scRNA-seq technology enables researchers to study the cell types, and state transitions along various biological processes. The existence of transition cells in the intermediate state of complex biological processes poses a challenge for the trajectory inference.
- To tackle this issue, we propose scTite, a new method to infer cell trajectory from scRNA-seq data based on transition entropy. Taking into account the continuity of the cell differentiation process, we define a new metric transition entropy to estimate the uncertainty of cells belonging to different cell clusters, and identify transition cells and cell states.
- Specifically, based on these identified cell clusters, we adopt the Wasserstein distance based on the probability distribution to calculate the distance between these clusters. Meanwhile, we utilize the signaling entropy and partial correlation coefficient to determine transition paths. Then, scTite combines the transitional paths and the MST to obtain a more detailed cell trajectory.
- The experimental results on four real scRNA-seq datasets show that scTite achieves superior performance compared with state-of-the-art methods.

Data availability

The Fibroblast dataset from GSE67310, The HSMM dataset from GSE52529, The Germline dataset from Germine, The ESC dataset from GSE65525.

Acknowledgements

We would like to thank MR. Xin Hu, MS. Xingyu Huang and MS. Huichun Zhu for fruitful discussions.

Funding

This work were sponsored in part by the National Natural Science Foundation of China (62172088, 61772128) and Shanghai Natural Science Foundation (21ZR1400400, 19ZR1402000).

References

1. Chen H, Albergante L, Hsu JY, *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. *Nat Commun* 2019;**10**(1):1–14.
2. Sagar and Dominic Grün. Deciphering cell fate decision by integrated single-cell sequencing analysis. *Ann Rev Biomed Data Sci* 2020;**3**:1–22.
3. Wei J, Zhou T, Zhang X, *et al.* Scout: a new algorithm for the inference of pseudo-time trajectory using single-cell data. *Comput Biol Chem* 2019;**80**:111–20.
4. Saelens W, Cannoodt R, Todorov H, *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;**37**(5):547–54.
5. Welch JD, Hartemink AJ, Prins JF. Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data. *Genome Biol* 2016;**17**(1):1–15.
6. Wang S, Karikomi M, MacLean AL, *et al.* Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res* 2019;**47**(11):e66–6.
7. Ji Z, Ji H. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic Acids Res* 2016;**44**(13):e117–7.
8. Trapnell C, Cacchiarelli D, Grimsby J, *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**(4):381–6.
9. Qiu X, Mao Q, Tang Y, *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 2017;**14**(10):979–82.
10. Street K, Risso D, Fletcher RB, *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018;**19**(1):1–16.
11. Gan Y, Li N, Zou G, *et al.* Identification of cancer subtypes from single-cell rna-seq data using a consensus clustering method. *BMC Med Genomics* 2018;**11**(6):65–72.
12. Gan Y, Li N, Guo C, *et al.* Tic2d: trajectory inference from single-cell rna-seq data based on consensus clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2021. https://doi.org/10.1109/TCBB.2021.3061720.
13. Bendall SC, Davis KL, Amir E-A D, *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell* 2014;**157**(3):714–25.
14. Setty M, Tadmor MD, Reich-Zeliger S, *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 2016;**34**(6):637–45.
15. Costa F, Grün D, Backofen R. Graphddp: a graph-embedding approach to detect differentiation pathways in single-cell-data using prior class knowledge. *Nat Commun* 2018;**9**(1):1–8.
16. Wolf F, Hamey FK, Plass M, *et al.* Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**(1):1–9.
17. Da Rocha EL, Rowe RG, Lundin V, *et al.* Reconstruction of complex single-cell trajectories using cellrouter. *Nat Commun* 2018;**9**(1):1–13.
18. Zheng X, Jin S, Nie Q, *et al.* scrcmf: Identification of cell subpopulations and transition states from single-cell transcriptomes. *IEEE Trans Biomed Eng* 2019;**67**(5):1418–28.
19. MacArthur BD, Lemischka IR. Statistical mechanics of pluripotency. *Cell* 2013;**154**(3):484–9.
20. Gandrillon O, Gaillard M, Espinasse T, *et al.* Entropy as a measure of variability and stemness in single-cell transcriptomics. *Curr Opin Syst Biol* 2021;**27**:100348.
21. Guo M, Bao EL, Wagner M, *et al.* Slice: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* 2017;**45**(7):e54–4.
22. Teschendorff AE, Sollich P, Kuehn R. Signalling entropy: a novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods* 2014;**67**(3):282–93.
23. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* 2017;**8**(1):1–15.
24. Lummertz E, da Rocha RG, Rowe VL, *et al.* Reconstruction of complex single-cell trajectories using cell-router. *Nat Commun* 2018;**9**(1):892.
25. Becht E, McInnes L, Healy J, *et al.* Dimensionality reduction for visualizing single-cell data using umap. *Nat Biotechnol* 2019;**37**(1):38–44.
26. Van der Maaten L, Hinton G. Visualizing data using t-sne. *J Mach Learn Res* 2008;**9**(11):2579–2605.

27. Szubert B, Cole JE, Monaco C, *et al.* Structure-preserving visualisation of high dimensional single-cell datasets. *Sci Rep* 2019;**9**(1):1–10.
28. Moon KR, van Dijk D, Wang Z, *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**(12):1482–92.
29. Eraslan G, Simon LM, Mircea M, *et al.* Single-cell rna-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**(1):1–14.
30. Amodio M, Van Dijk D, Srinivasan K, *et al.* Exploring single-cell data with deep multitasking neural networks. *Nat Methods* 2019;**16**(11):1139–45.
31. Panaretos VM, Zemel Y. Statistical aspects of wasserstein distances. *Ann Rev Stat Appl* 2019;**6**:405–31.
32. Gulrajani I, Ahmed F, Arjovsky M, *et al.* Improved training of wasserstein gans. *Adv Neural Inform Process Syst* 2017;**30**:5769–79.
33. Schäfer J, Strimmer K. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics* 2005;**21**(6):754–64.
34. Barzel B, Barabási A-L. Network link prediction by global silencing of indirect correlations. *Nat Biotechnol* 2013;**31**(8):720–5.
35. Hastie T, Stuetzle W. Principal curves. *J Am Stat Assoc* 1989;**84**(406):502–16.
36. Klein AM, Mazutis L, Akartuna I, *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**(5):1187–201.
37. Liu X, Ouyang JF, Rossello FJ, *et al.* Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature* 2020;**586**(7827):101–7.
38. Satija R, Farrell JA, Gennert D, *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**(5):495–502.
39. van Dijk D, Nainys J, Sharma R, *et al.* Magic: a diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *BioRxiv*. 2018;**174**(3):716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.
40. Moon KR, Stanley III JS, Burkhardt D, *et al.* Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Curr Opin Syst Biol* 2018;**7**:36–46.
41. Welch JD, Hartemink AJ, Prins JF. Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol* 2017;**18**(1):1–19.
42. Campbell KR, Yau C. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat Commun* 2018;**9**(1):1–12.
43. Ellwanger DC, Scheibinger M, Dumont RA, *et al.* Transcriptional dynamics of hair-bundle morphogenesis revealed with celltrails. *Cell Rep* 2018;**23**(10):2901–14.
44. Cao J, Spielmann M, Qiu X, *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**(7745):496–502.