RESEARCH ARTICLE

Quantifying pluripotency landscape of cell differentiation from scRNA-seq data by continuous birth-death process

Jifan Shi[1], **Tiejun Li**[2]*, **Luonan Chen**[3,4,5,6]*, **Kazuyuki Aihara**[1,7]*

1 Institute of Industrial Science, The University of Tokyo, Tokyo, Japan, **2** LMAM and School of Mathematical Sciences, Peking University, Beijing, China, **3** Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China, **4** Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China, **5** School of Life Science and Technology, ShanghaiTech University, Shanghai, China, **6** Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China, **7** International Research Center for Neurointelligence, The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Tokyo, Japan

* tieli@pku.edu.cn (TL); lnchen@sibs.ac.cn (LC); aihara@sat.t.u-tokyo.ac.jp (KA)

Citation: Shi J, Li T, Chen L, Aihara K (2019) Quantifying pluripotency landscape of cell differentiation from scRNA-seq data by continuous birth-death process. PLoS Comput Biol 15(11): e1007488. https://doi.org/10.1371/journal.pcbi.1007488

Editor: Andrew E Teschendorff, University College London, UNITED KINGDOM

Received: April 13, 2019

Accepted: October 16, 2019

Published: November 13, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. The codes of LDD and the pre-processed datasets used in the paper can be downloaded from https://github.com/smsxiaomayi/LDD/blob/master/LDDcode.zip.

Funding: This work was supported by the National Key Research and Development Program of China (No.2017YFA0505500), National Natural Science Foundation of China (NSFC) (Nos.31930022, 31771476, 11421101, 91530322 and 11825102),

Abstract

Modeling cell differentiation from omics data is an essential problem in systems biology research. Although many algorithms have been established to analyze scRNA-seq data, approaches to infer the pseudo-time of cells or quantify their potency have not yet been satisfactorily solved. Here, we propose the Landscape of Differentiation Dynamics (LDD) method, which calculates cell potentials and constructs their differentiation landscape by a continuous birth-death process from scRNA-seq data. From the viewpoint of stochastic dynamics, we exploited the features of the differentiation process and quantified the differentiation landscape based on the source-sink diffusion process. In comparison with other scRNA-seq methods in seven benchmark datasets, we found that LDD could accurately and efficiently build the evolution tree of cells with pseudo-time, in particular quantifying their differentiation landscape in terms of potency. This study provides not only a computational tool to quantify cell potency or the Waddington potential landscape based on scRNA-seq data, but also novel insights to understand the cell differentiation process from a dynamic perspective.

Author summary

Quantifying the Waddington landscape of cell differentiation from high throughput data is a challenging problem in systems biology and biophysics. Here, we propose a theoretical method named LDD (Landscape of Differentiation Dynamics), which builds cell potentials and constructs their differentiation landscape by a continuous birth-death process from scRNA-seq data. This method well exploits the dynamical features of the differentiation process, thus quantifying the differentiation landscape in an accurate manner. We show that LDD can accurately and efficiently build the evolution tree of cells with pseudo-

Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), Japan Agency for Medical Research and Development (AMED) (No.JP18dm0307009), JSPS KAKENHI (No. JP15H05707), JST CREST (No.JPMJCR14D2), and World Premier International Research Center Initiative (WPI) MEXT, Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

time, in particular quantifying their differentiation landscape in terms of potency. Taken together, this study provides not only a computational tool to quantify cell potency based on scRNA-seq data, but also a theoretical approach to understand the cell differentiation process from a dynamic perspective.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Single-cell RNA sequencing (scRNA-seq) has become a rapidly developing technique since 2009 when Tang *et al*. [1–3] first proposed the sequencing method. SMART-seq2 [4], CELL-seq [5], Drop-seq [6], and 10X genomics [7] are the most popular protocols at present. They can measure gene expressions for individual cells rather than tissue-level bulk cells without high costs. By analyzing scRNA-seq data, we can determine tissue heterogeneity and capture various developing stages of cells.

Cell differentiation is a process in which several kinds of functional cells arise from one cell type called the pluripotent cell. It is considered that cell specification results from changes in gene expression patterns. As the expression data could be extracted from single cells, many mathematical models have been built, and statistical analysis has been applied to describe the differentiation process, such as RNA velocity [8] and pseudo-time. Pseudo-time is one of the most popular approaches, which attaches a number to each sample or cell as the evolution time from the pluripotent cell. There are mainly two approaches to estimate the pseudo-time of each cell. One is the distance-based method, including Wanderlust/Wishbone [9, 10], Diffusion maps/destiny [11–13], Monocle/Monocle2 [14, 15], scEpath [16] and others. This type of method defines the pseudo-time as the distance from a root cell based on a graph structure. The other type is the entropy-based method, including StemID [17], SLICE [18], SCENT [19], and Markov-chain entropy [20]. This type of method computes some predefined entropy of a cell-cell graph or a gene interaction network as the pseudo-time. We can refer to many comprehensive reviews [21–23] and comparison papers [24–26] for a survey of those works. Although many algorithms have been developed to analyze scRNA-seq data, how to accurately infer the pseudo-time of cells or quantify their potency has not yet been satisfactorily solved. In particular, most of the existing methods are based on statistical measures depending heavily on the samples, or based on the approximation of an equilibrium process, without a dynamical description which is essential for elucidating the differentiation process.

From a modeling viewpoint, cell differentiation is not an equilibrium, but a non-equilibrium, process due to frequent birth and death of cells, and thus can be well modeled by a continuous birth-death (or source-sink diffusion) process. In this paper, derived from such a dynamic process, we propose a new method named Landscape of Differentiation Dynamics (LDD) to analyze the cell differentiation process. Further, we use LDD to compute both the pseudo-time and directed differentiation paths, which are also known as the differentiation landscape. LDD not only quantifies the potency of cells, but also determines the pseudo-time derived from the continuous birth-death process, rather than from the geometric graphical distance used widely in traditional methods. Our method is based on the source-sink diffusion process, which exploits the dynamical features of the stochastic differentiation process, thus quantifying the differentiation landscape in an accurate manner from a dynamic perspective.

In this study, we constructed the potential landscape $V(\boldsymbol{x})$ by solving Eq (3) under the non-equilibrium steady state assumption of the differentiation dynamics. One key observation of our work was that we could obtain the net-flow rate $R(\boldsymbol{x})$ from the data without assuming its prior knowledge, which advances previous proposals on non-equilibrium dynamics of gene expression [27]. LDD is a two-level algorithm: one at the single cell level, which quantifies cell heterogeneity for each cell based on the diffusion process, and the other at the cluster level, which quantifies the transition between cell types (or clusters) based on the Markov process. Additionally, the reverse of the pseudo-time could be calculated for each cell type. Therefore, cells with the highest potential are deemed pluripotent and will evolve into differentiated cells with lower potential. Lineages or differentiation branches could be detected from a transition matrix between different clusters. From the differentiation landscape constructed by LDD, we could clarify the global landscape structure of a real biological process. Further, pluripotent cells with higher potential in our study were quantitatively shown to differentiate into down-stream cells with lower potential by LDD, similar to a ball rolling down a mountain as described by the Waddington landscape [28]. Taken together, this study provides not only a new dynamical model with a computational tool to quantify the cell potency based on scRNA-seq data, but also a new approach to understand the differentiation process from a dynamic and stochastic perspective.

Results

Modeling cell differentiation by continuous birth-death process

Cell differentiation is clearly a non-equilibrium process due to frequent birth and death of cells. Thus, to model the cell differentiation, we use the continuous birth-death process as the underlying dynamics of cell differentiation, which is also named as the source-sink Fokker-Planck equation in mathematics and the population balance equation in Klein *at al.*'s work [27, 29, 30]. The continuous birth-death process assumes that the probability density function $c(\boldsymbol{x}, t)$ of all sample cells develops as

$$\frac{\partial c(\boldsymbol{x}, t)}{\partial t} = \nabla \cdot (c(\boldsymbol{x}, t)\nabla F(\boldsymbol{x})) + D\Delta c(\boldsymbol{x}, t) + R(\boldsymbol{x})c(\boldsymbol{x}, t), \quad (1)$$

where $\boldsymbol{x}$ is a vector of gene expression, t is the time, $F(\boldsymbol{x})$ is a potential function, D is the noise amplitude, and $R(\boldsymbol{x})$ is the net-flow of cells at state $\boldsymbol{x}$. ∇, $\nabla\cdot$, and Δ denote the gradient, divergence, and Laplace operators, respectively. When the system reaches a non-equilibrium steady state, i.e. $\lim_{t\to\infty} c(\boldsymbol{x}, t) = p(\boldsymbol{x})$ or $\partial c(\boldsymbol{x}, t)/\partial t = 0$, the potential can be decomposed as $F(\boldsymbol{x}) = U(\boldsymbol{x}) + V(\boldsymbol{x})$ and calculated by

$$U(\boldsymbol{x}) = -D\log p(\boldsymbol{x}), \quad (2)$$

$$\mathcal{L}V(\boldsymbol{x}) = [\nabla \log p(\boldsymbol{x}) \cdot \nabla + \Delta]V(\boldsymbol{x}) = -R(\boldsymbol{x}), \quad (3)$$

according to [27]. We name the flow that is differentiation-oriented as "advection", while we use the term "diffusion" which is caused by random noise [31–33]. $U(\boldsymbol{x})$ is known as the equilibrium potential caused by diffusion without birth and death, and $V(\boldsymbol{x})$ is a new potential caused only by advection without diffusion. By definition, noise is known to only influence the diffusion process and generates meta-stable wells in $U(\boldsymbol{x})$. Therefore, $V(\boldsymbol{x})$ can be taken as the cell differentiation potential to describe the differentiation direction. The cells will evolve from pluripotent cells with high $V(\boldsymbol{x})$ to differentiated cells with low $V(\boldsymbol{x})$. $V(\boldsymbol{x})$ can represent the

Waddington landscape [28] of the cell at state $\boldsymbol{x}$, and the additive inverse of $V(\boldsymbol{x})$ can be considered as a reflection of the pseudo-time (see Materials and methods).

Estimating cell potential V(x) and constructing differentiation paths

To obtain $V(\boldsymbol{x})$ from Eq (3), the net-flow $R(\boldsymbol{x})$ and backward operator $\mathcal{L} = \nabla \log p(\boldsymbol{x}) \cdot \nabla + \Delta$ need to be estimated from scRNA-seq data. Due to the limitation of sample size, it is difficult to accurately measure net-flow $R(\boldsymbol{x})$ for every cell. In contrast to [27], which required additional information, we first clustered samples into different cell types/clusters, and then computed the net-flow $\hat{R}_s$ for each cluster s from the gene expression matrix by the divergence theorem and marginal decomposition. On the other hand, benefiting from the diffusion map theory [34, 35] and model reduction [36], the backward operator $\mathcal{L}$ was approximated by $\hat{L}$, which was the coarse-grained discrete matrix representation of $\mathcal{L}$ between cell clusters, obtained from the cell-to-cell transition matrix. Thus, with the approximated net-flow $\hat{R}$ for every cluster/cell type, and the approximated transition operator $\hat{L}$, $\hat{V}$ could be obtained numerically, generating a concrete value for the potency of each cell type. Additional details can be found in Materials and methods. We remark that in [27], the net-flow rate was set as the prior knowledge for each cell. However, we were able to obtain the value from the gene expression matrix if cells were clustered into different metastable states. This is one key point of our work.

To illustrate the entire differentiation process, we constructed its landscape, in which nodes were cell types with potential $\hat{V}$, paths were determined by the transition matrix between clusters, and directions were from high to low potential. We named this procedure as Landscape of Differentiation Dynamics (LDD). Fig 1 provides a flowchart of LDD, which is described in details in Materials and methods. An algorithmic description is provided in S1 Text section S1 and Fig I in S1 Text.

Computing differentiation landscape of simulated models

We used three simulated models to construct potential and verify our LDD method. The first is a drift-diffusion process, in which particles evolve by

$$\boldsymbol{x}(t + \Delta t) = \boldsymbol{x}(t) - \nabla F(\boldsymbol{x}(t)) \cdot \Delta t + \sqrt{D\Delta t} \cdot \xi(t), \tag{4}$$

where $\boldsymbol{x}$ is the position of a particle in 50 dimensions, F is a potential function, D is the noise amplitude, and $\boldsymbol{\xi}$ is a normal random vector standing for noise. A bifurcation from one branch to two branches occurred in the system, representing cell differentiation. Particles/cells were generated from a source, and two sinks indicated places for particle/cell removal or death. It imitated the cell's lifespan from birth to death. Details about the model are shown in S1 Text. In the example, 400 samples/cells were simulated. After reducing to two dimensions using principal component analysis (PCA) and clustering the 400 samples into four groups/clusters by k-means, we computed the potential $\hat{V}$ of each cluster. The landscape in a three-dimensional view is illustrated in Fig 2A. Fig A(a) in S1 Text shows the potential in a two-dimensional space. In Fig 2B, the differentiation paths between the four clusters were constructed. The cluster with the highest potential corresponded to the source region, and the two sinks with low potential values were at the end of two lineages.

The next two simulated examples are two-gene and six-gene regulatory networks. Their gene interactions are shown in Fig 2C and 2D, respectively. A region for cell birth is shown, as well as several regions for cell death. The detailed simulation method can be found in S1 Text. For the two-gene network, as the genes inhibit each other, two lineages formed. In each

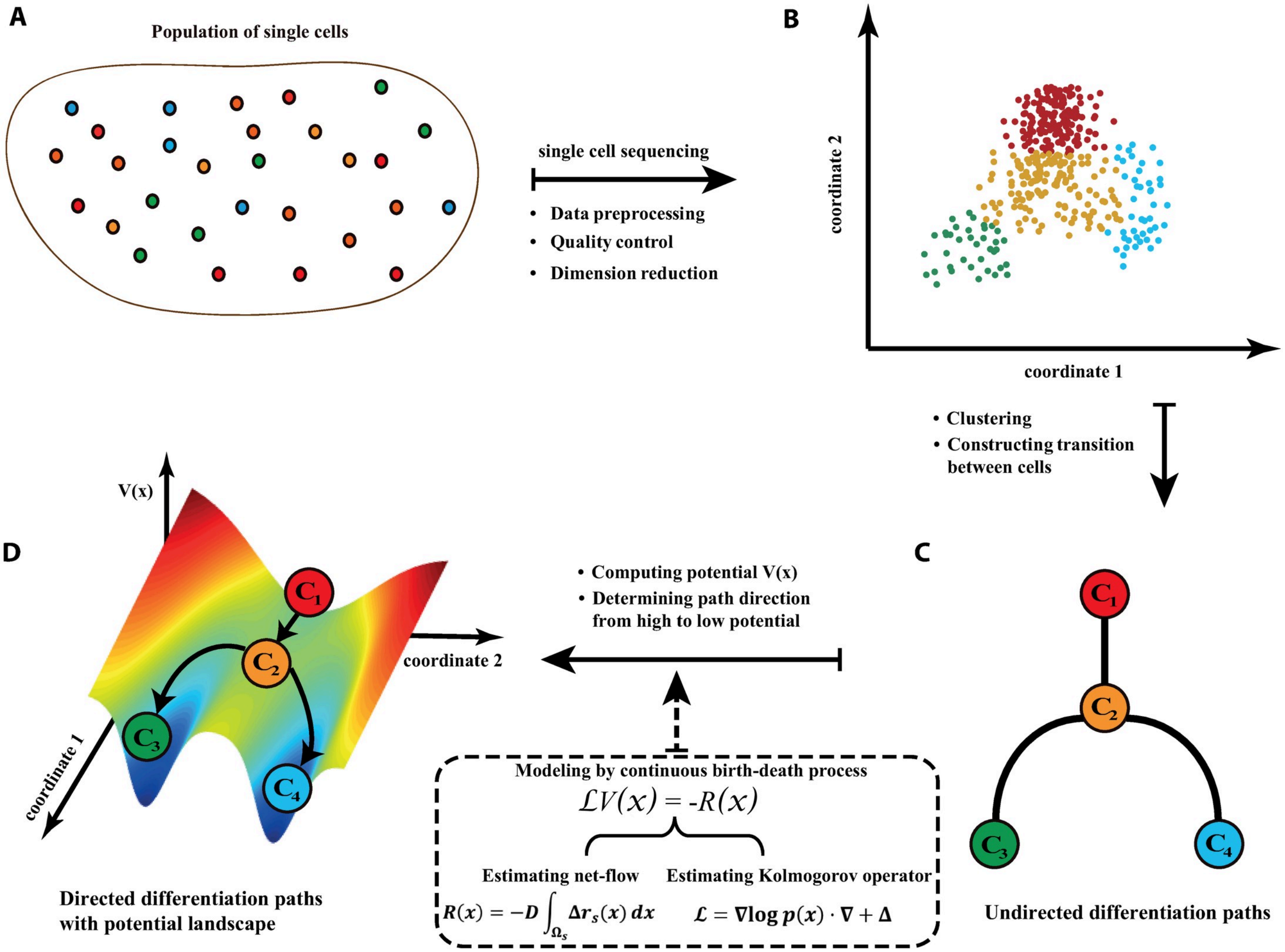
A
Population of single cells
single cell sequencing
• Data preprocessing
• Quality control
• Dimension reduction
B
coordinate 2
coordinate 1
• Clustering
• Constructing transition between cells
C
C_1
C_2
C_3
C_4
Undirected differentiation paths
• Computing potential V(x)
• Determining path direction from high to low potential
Modeling by continuous birth-death process
$\mathcal{L}V(x) = -R(x)$
Estimating net-flow
Estimating Kolmogorov operator
$R(x) = -D\int_{\Omega_s} \Delta r_s(x)\,dx$
$\mathcal{L} = \nabla \log p(x) \cdot \nabla + \Delta$
D
V(x)
coordinate 2
coordinate 1
Directed differentiation paths with potential landscape

Fig 1. Flowchart of the Landscape of Differentiation Dynamics (LDD) method. **A**: A pool of single cells, from which we can obtain the gene expression matrix by single cell sequencing. **B**: After preprocessing, quality control, and dimension reduction, a low-dimensional data matrix is obtained. **C**: The samples are clustered into different types. Undirected differentiation paths are determined by a transition matrix between clusters. **D**: After applying the continuous birth-death process to model the whole differentiation process, the potential $V(\boldsymbol{x})$, differentiation directions, and landscape can be constructed.

https://doi.org/10.1371/journal.pcbi.1007488.g001

branch, one gene had high expression, while expression of the other gene was very low. The illustrative landscape for the two-gene network is in Fig 2E, and the two-dimensional plot is in Fig A(b) in S1 Text. Fig 2F shows the differentiation paths between seven groups. LDD generated the potential in a correct order, i.e. pluripotent cells with higher values, and differentiated cells with lower values. Similar results of LDD were obtained in the six-gene network, which had four branches for the samples. Only two genes had high expression, while the others were low in each branch, which are shown in Fig A(c) and A(d) in S1 Text.

The three examples demonstrate the ability of LDD to quantitatively characterize the stochastic process of cell differentiation from a dynamic perspective, due to our model based on the continuous birth-death process. In particular, only using the observed data, LDD could quantitatively show that pluripotent cells with higher potential differentiated into downstream cells with lower potential, similar to a ball rolling down a mountain as described by Waddington [28].

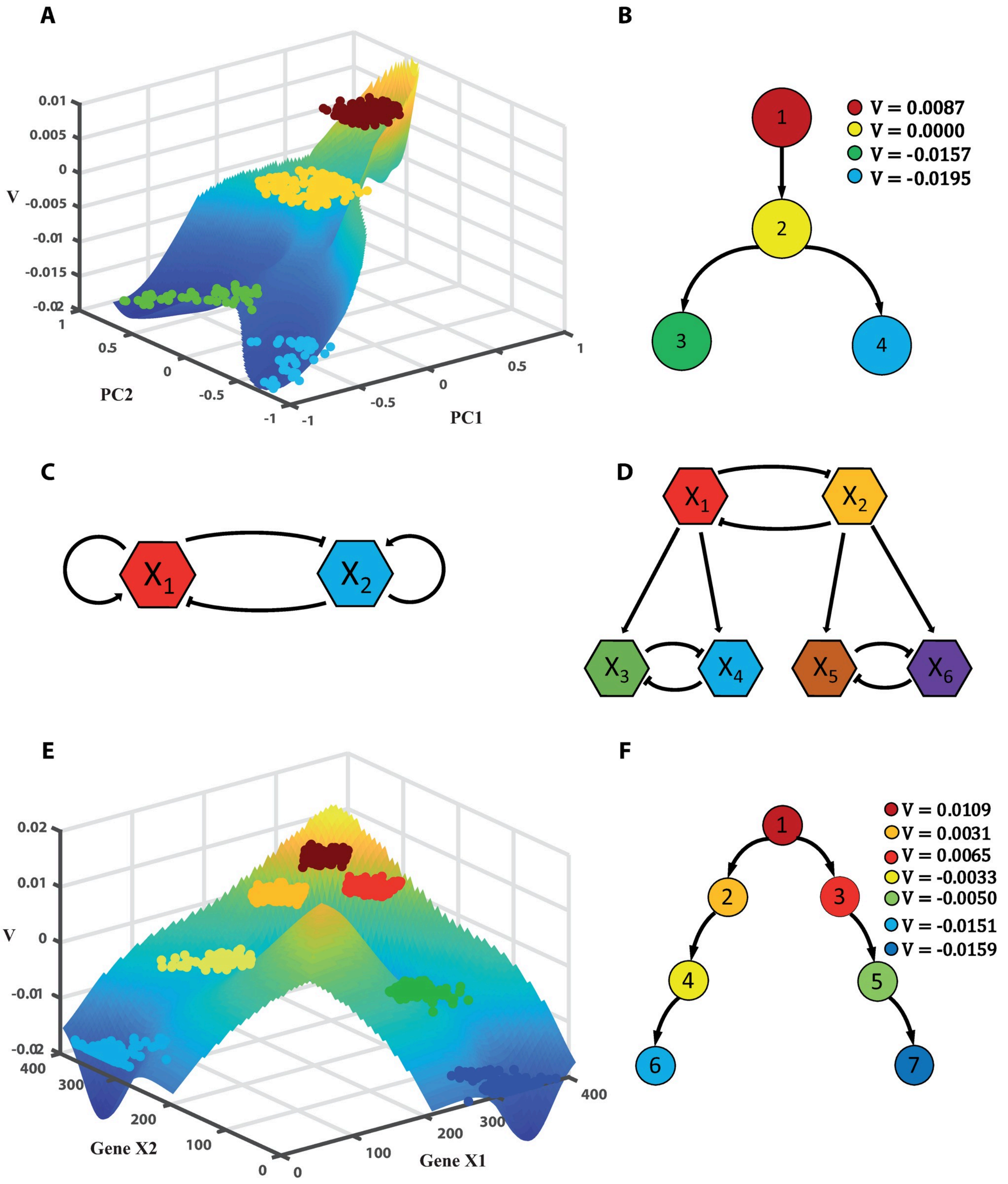

Fig 2. Differentiation landscape, differentiation paths, and gene networks for simulated models. **A** and **B** are the LDD potential landscape and differentiation paths using data from the simulated drift-diffusion process, in which samples/cells were clustered into four groups. **C** and **D** are the two-gene and six-gene regulatory networks for simulation, respectively. For the two-gene network, the potential landscape and differentiation paths for clusters are shown in **E** and **F**, respectively, where seven clusters were detected. Constant potential *V* for each cluster was computed by LDD, while the landscape is an illustration constructed by the method in Materials and methods.

https://doi.org/10.1371/journal.pcbi.1007488.g002

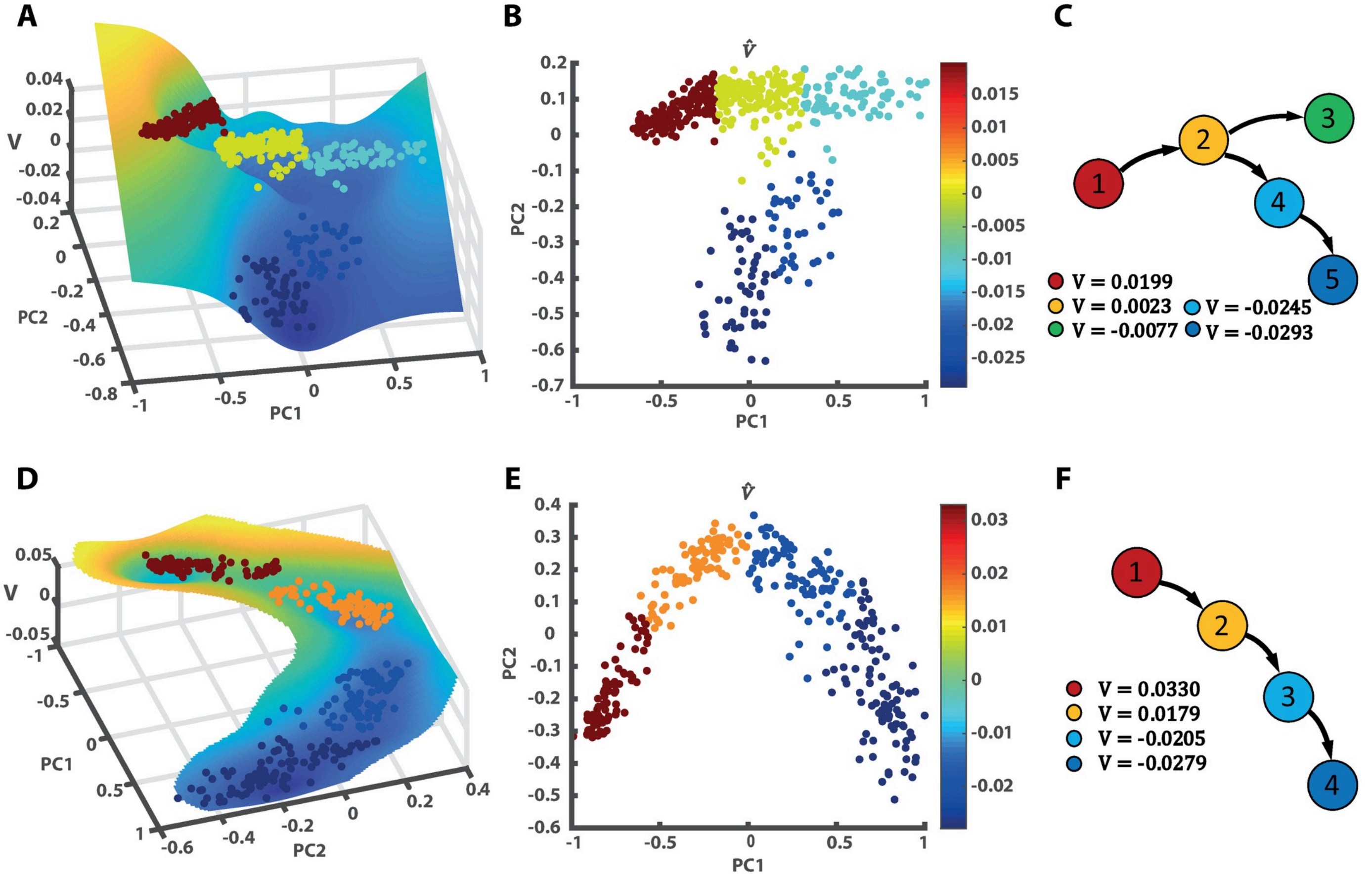

Fig 3. Differentiation landscape and differentiation paths for real datasets. A, **D** are the LDD potential landscapes, **B**, **E** are the potential values plotted in the two-dimensional reduction space, and **C**, **F** are the differentiation paths. **A**-**C** use Xu's dataset, which describes that hepatoblasts (cluster 1) differentiate into hepatocytes (cluster 3) and cholangiocytes (cluster 5). **D**-**F** use Furlan's dataset and show that chromaffin cells (cluster 4) are generated from SCPs (cluster 1).

https://doi.org/10.1371/journal.pcbi.1007488.g003

Computing differentiation landscape of real datasets

To further verify the efficiency of LDD, we applied it to four real datasets, i.e. Guo's dataset, Nef's dataset, Xu's dataset, and Furlan's dataset.

Guo's dataset [37] describes cells developing from zygote to blastocyst, through oocyte, 2-cell, 4-cell, 8-cell, morula, E3.5 blastocyst, and E4.25 blastocyst stages. At the end term of morula, the cells could differentiate into trophectoderm (TE) and inner cell mass (ICM). Nef's dataset [38] focused on the determination of mouse's sex. 400 samples were selected from E10.5, E11.5, E12.5, E13.5, and E16.5 stages. Two branches named the interstitial progenitor cell lineage and Sertoli cell lineage appeared during the observation time. Xu's dataset [39] showed that the progenitor hepatoblasts had bipotency to divide into hepatocytes and cholangiocytes. Furlan's dataset [40] found that a large number of chromaffin cells in adrenal medulla arose from a kind of peripheral glial stem cell called Schwann cell precursors (SCPs). The description and the preprocessing of these datasets can be found in S1 Text.

Using LDD, we computed the potential and differentiation paths of each dataset. The illustrative landscapes, potential values, and lineages of Xu's dataset are displayed in Fig 3A–3C, which show that the progenitor hepatoblasts (cluster 1) with high potency differentiated into hepatocytes (cluster 3) and cholangiocytes (cluster 5) with low potency. Fig 3D–3F are the corresponding results of Furlan's dataset. There was a single path through which the SCPs became chromaffin cells. For Guo's dataset and Nef's dataset, the results are given in Fig B in S1 Text. For Nef's dataset, we also computed Pearson correlations between LDD potential and three character genes *Pbx1*, *Gpc3*, and *Sfrp1* as 0.9015, 0.8582, and 0.8000. These genes decreased during the differentiation process, which indicated that the pseudo-time conformed to the differentiation direction.

Comparison with other pseudo-time algorithms

From both the simulated and real datasets, we showed that LDD could obtain the correct differentiation paths by quantifying potential values of cells. The pseudo-time is another important topic in single cell research, which provides a time label to each cell. We set the additive inverse of potential $\hat{V}$ as our pseudo-time by LDD, and compare it with several traditional methods. Because entropy-based methods usually need additional information, such as gene interactions or function annotations, we only compared LDD with six distance-based methods, i.e. TSCAN [41], Monocle2 [15], Diffusion Map [11, 12], DPT [13], SLICER [42], and Slingshot [43]. Their properties are listed in Table 1, and the details are given in S1 Text.

For the simulated datasets, we chose the Pearson correlation between the pseudo-times and the true-time labels as the measurement. The second to the fourth columns in Table 2 show the results for the Simu1 (the drift-diffusion process), Simu2 (the two-gene regulatory network), and Simu3 (the six-gene regulatory network) datasets. The closer the Pearson correlation approximated to 1, the more reliable the algorithm was. LDD and DPT outperformed the others; however, DPT required additional information, i.e. a started root sample to represent the stem cell. Hence, it implies that LDD effectively uses the information in the samples, which leads to correct differentiation paths.

For the real datasets, the accurate cell time is unavailable; however, several stage-level time labels, such as E10.5 and E11.5, are given. Under one stage label, there are several cell types. Therefore, instead of the Pearson correlation, we used the one-sided Wilcoxon ranksum test to determine whether the stem cell stage had earlier pseudo-time than the differentiated stage. The p-value was chosen as the measurement [19, 20]. The fifth to the eighth columns in Table 2 list the p-values for the four real datasets (Guo's, Nef's, Xu's, and Furlan's). The alternative hypothesis was that pluripotent stem cells had earlier pseudo-time than differentiated

Table 1. Properties of different pseudo-time methods.

	LDD	TSCAN	Monocle2	Diffusion map	DPT	SLICER	Slingshot
Basic Model	**Birth-death Dynamics**	**Distance on tree/ graph**	**Distance on tree/ graph**	**Distance on tree/ graph**	**Distance on tree/ graph**	**Distance on tree/ graph**	**Distance on tree/ graph**
Not use a root cell as input	Yes	No	No	No	No	No	No
Dimension Reduction	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Yes	Yes	No	No	No	No	Yes
Detecting ⩾ 1 branches	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reference	—	[41]	[15]	[11, 12]	[13]	[42]	[43]

Each column is a method. For the rows, "Basic Model" indicates the theoretical model used in the method. "Not use a root cell as input" indicates whether the algorithm requires a start cell as input. For some algorithms if giving a wrong root cell, the pseudo-time could be reverse of the true time direction or even be totally messed. "Dimension reduction" indicates whether dimensional reduction is applied on the data matrix. "Clustering" states whether the algorithm clusters cells during the process. "Detecting ⩾ 1 branches" indicates whether the algorithm could find more than one different branches/lineages. The corresponding papers are listed in the "Reference" row.

https://doi.org/10.1371/journal.pcbi.1007488.t001

Table 2. Comparison between seven pseudo-time methods.

	Pearson Correlation ρ			Wilcoxon Ranksum Test p-value			
Dataset	**Simu1**	**Simu2**	**Simu3**	**Guo**	**Nef**	**Xu**	**Furlan**
Measure	**Correlation ρ between pseudo-time and true time**			**Oocyte < E4.25 blastocyst**	**E10.5< E16.5**	**E10.5 < E17.5**	**SCPs < chromaffin**
LDD	**0.9135**	0.9520	0.8229	**3.842e-09**	9.432e-20	**3.557e-24**	**2.104e-39**
TSCAN	0.1343	0.0404	0.0055	1.344e-01	2.588e-01	6.145e-16	2.067e-34
Monocle2	0.0668	0.0121	0.0131	2.370e-07	3.084e-01	8.479e-22	2.067e-34
Diffusion map	0.7707	0.0398	0.0069	2.370e-07	6.277e-09	8.479e-22	2.067e-34
DPT	0.9079	**0.9603**	**0.8797**	2.370e-07	1.053e-15	8.479e-22	2.067e-34
SLICER	0.7348	0.8807	0.4514	3.672e-07	**4.921e-22**	8.479e-22	2.067e-34
Slingshot	0.0971	0.0147	0.0866	2.370e-07	3.725e-19	6.781e-22	2.067e-34

Seven pseudo-time methods, including LDD, were tested on three simulated datasets and four real datasets. For the simulated datasets, as we had true time labels for every cell, the Pearson correlation between true times and pseudo-times was calculated. For real datasets, we used the one-sided Wilcoxon ranksum test to determine whether pluripotent cells had earlier pseudo-time than differentiated cells. The alternative hypothesis is listed in the third row in the table. Bold numbers are the best method in its column. LDD performs among the best, while the other six methods require a root cell as prior information.

https://doi.org/10.1371/journal.pcbi.1007488.t002

cells. Thus, the smaller the p-value was, the better the algorithm performed. From Table 2, LDD was always shown to produce good results among the seven methods (the second place in Nef's dataset and the best in other three datasets).

The conclusion from the comparison between different pseudo-time methods is as follows: LDD and DPT performed well in determining the pseudo-time but DPT required additional information, i.e. a root cell. Thus, LDD is an efficient approach to model cell differentiation and quantitatively characterize the Waddington potential landscape or differentiation paths by exploiting dynamical and stochastic features of the differentiation process from the measured data.

Discussion

In this paper, in contrast to the approximation of the equilibrium process widely used in previous methods, we used a stochastic non-equilibrium steady process, i.e. a continuous birth-death process model, to describe the differentiation dynamics of cells, which well captures the dynamical and stochastic features of the cell differentiation process. LDD based on this model was proposed to compute the cell's potency value and construct the differentiation paths/landscape from scRNA-seq data, derived from the diffusion map theory and divergence theorem. Using the landscape, we showed that cells developed from high to low potential. Different lineages were also detected by the transition matrix obtained by LDD. As the LDD's pseudo-time (the additive inverse of potential) originates from a dynamic system, it gets rid of the limitation of distance measures on graphs. Comparison studies with traditional methods also showed that it was a powerful and effective method on both simulated and real datasets.

There are still some features and issues to be discussed: (1) The dynamical model used in LDD requires that the sample cells are in a non-equilibrium steady state, i.e., the cells' birth rate should be equal to the death rate. In a short time period when the environment does not change much, the model works, but for a long-term study, this condition may not hold. (2) As the dynamical model describes a continuous flow of differentiation, the sampled cells need to represent the whole differentiation process. When gathering data, this requirement should be considered. (3) Cell clustering is another important issue. Clustering cells in our method requires that different types are separated into different metastable states. However, our samples/cells also need to keep the continuity of differentiation. Thus, there is an appropriate balance between separability and continuity. In our examples, PCA and k-means clustering can maintain the orthogonal invariance for the landscape, which may not occur with other nonlinear approaches. Therefore, the method must be chosen carefully to ensure the system's properties (see Materials and methods). (4) The requirement on the samples is unclear and ambiguous for the distance-based or entropy-based methods, which limits their applications. In contrast, LDD provides a general mechanism for the differentiation process from a stochastically dynamic viewpoint, and its results are guaranteed by the diffusion model for the measured samples. (5) The birth-death process modeling is inspired by [27], but we further improve its theoretical result. In particular, we remove the requirement for measuring net-flow, i.e. we show that net-flow $\hat{R}_s$ can be directly computed from the gene expression matrix by the divergence theorem and marginal decomposition. In addition, rather than the graph Laplacian, we applied diffusion map theory, which considers the weights on edges and is a generalization of the graph Laplacian. (6) Saelens et al. [26] gave a comprehensive comparison over 45 scRNA packages, in aspects of their accuracy, scalability, stability, and usability. Our study focused on the theoretical framework and algorithm construction. Further improvements and tests of our algorithm on multiple and larger datasets will be performed as our future work.

In summary, by the diffusion map theory and divergence theorem, we provide a new approach to quantify cell differentiation from measured scRNA-seq data based on the continuous birth-death process, which well exploits the dynamical features of the cell differentiation process from a dynamic landscape viewpoint. The essential evolution laws of cells need much more efforts through joint experimental and theoretical studies in the future.

Materials and methods

Cell differentiation dynamics as continuous birth-death process

We use a stochastic non-equilibrium process Eq (1) to model the cell differentiation, which is also named the source-sink Fokker-Planck equation or the population balance equation [27, 29, 30]. In Eq (1), the function $c(\boldsymbol{x}, t)$ represents the probability density function (pdf) of cells at $\boldsymbol{x}$, which will be estimated from scRNA-seq data. The term $R(\boldsymbol{x})$ is crucial to understand the birth and death of the cells involved in the system. If there is a source and cells are increasing locally, $R(\boldsymbol{x})$ will be positive. Conversely, if there is a sink and cells are removed or die, $R(\boldsymbol{x})$ will be negative.

Deriving cell differentiation potential V(x)

Assume that the system reaches a non-equilibrium steady state, which means cells keep being born and dying but the whole population distribution is invariant. Let $t \to \infty$, then we will obtain from Eq (1)

$$\nabla \cdot (p(\boldsymbol{x})\nabla F(\boldsymbol{x})) + D\Delta p(\boldsymbol{x}) + R(\boldsymbol{x})p(\boldsymbol{x}) = 0, \tag{5}$$

where $p(\boldsymbol{x}) = \lim_{t\to\infty} c(\boldsymbol{x}, t)$. Only depending on $p(\boldsymbol{x})$, $F(\boldsymbol{x})$ could have many solutions, one of which can be written as

$$\nabla \cdot (p(\boldsymbol{x})\nabla U(\boldsymbol{x})) + D\Delta p(\boldsymbol{x}) = 0, \tag{6}$$

$$\nabla \cdot (p(\boldsymbol{x})\nabla V(\boldsymbol{x})) + R(\boldsymbol{x})p(\boldsymbol{x}) = 0, \tag{7}$$

and

$$F(\boldsymbol{x}) = U(\boldsymbol{x}) + V(\boldsymbol{x}). \tag{8}$$

The explicit forms for Eqs (6) and (7) are

$$U(\boldsymbol{x}) = -D \log p(\boldsymbol{x}), \tag{9}$$

$$\mathcal{L} V(\boldsymbol{x}) = -R(\boldsymbol{x}), \tag{10}$$

where $\mathcal{L}$ is the backward Kolmogorov operator

$$\mathcal{L} = \nabla \log p(\boldsymbol{x}) \cdot \nabla + \Delta. \tag{11}$$

$U(\boldsymbol{x})$ can be considered as the equilibrium potential caused by diffusion without samples' birth and death, while $V(\boldsymbol{x})$ can be taken as the potential caused by a birth-death flow without diffusion. If the noise amplitude D approaches zero, the diffusion vanishes and thus we can take $V(\boldsymbol{x})$ as the potential to describe cell pluripotency, or as the reverse pseudo-time of cell differentiation.

To compute $V(\boldsymbol{x})$, instead of setting $R(\boldsymbol{x})$ as the given values at each point $\boldsymbol{x}$ like [27], we cluster samples into groups/types to obtain $\hat{R}$ for each group only from the expression matrix. The backward operator $\mathcal{L}$ is also approximated by a coarse-grained $\hat{L}$ defined on the cell clusters. The cell differentiation potential $\hat{V}$ is then computed for each cluster based on a discrete version of Eq (10). The following subsections will show both theoretical and numerical details, and the overall algorithm is described in S1 Text.

Constructing Kolmogorov operator $\mathcal{L}$ by diffusion map from cell samples

To approximate the backward Kolmogorov operator $\mathcal{L}$ in Eq (11), we utilize the diffusion map theory [34, 35]. Denote the kernel function by

$$K_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{(4\pi\varepsilon)^{m/2}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{4\varepsilon}}, \tag{12}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two samples in m-dimensional space, and ε is a parameter adjusting the kernel width. If there are N samples/cells $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ in total obtained from probability density $r(\boldsymbol{x})$, we can define

$$q_\varepsilon(\boldsymbol{x}_i) = \sum_{j=1}^{N} K_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{13}$$

$$K_{\varepsilon,\alpha}(\boldsymbol{x}, \boldsymbol{y}) = \frac{K_\varepsilon(\boldsymbol{x}, \boldsymbol{y})}{q_\varepsilon^\alpha(\boldsymbol{x}) q_\varepsilon^\alpha(\boldsymbol{y})}, \tag{14}$$

$$d_{\varepsilon,\alpha}(\boldsymbol{x}_i) = \sum_{j=1}^{N} K_{\varepsilon,\alpha}(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{15}$$

and the transition matrix between samples i and j as

$$P_{\varepsilon,\alpha}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{K_{\varepsilon,\alpha}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{d_{\varepsilon,\alpha}(\boldsymbol{x}_i)}. \tag{16}$$

The discrete backward Kolmogorov operator is constructed as

$$L_{\varepsilon,\alpha} = \frac{P_{\varepsilon,\alpha} - I}{\varepsilon}. \tag{17}$$

When the sample size $N \to \infty$ and the kernel width ε goes to 0, the discrete operator $L_{\varepsilon,\alpha}$ tends to be an operator $\mathcal{L}_\alpha$ in the continuous space. When $\alpha = 1/2$, $\mathcal{L}_{1/2} = \nabla \log r(\boldsymbol{x}) \cdot \nabla + \Delta$ is exactly the backward operator in Eq (11) (see S1 Text for details). In our LDD algorithm, we first construct a weighted undirected k-nearest-neighbor (kNN) network by the similarity measure Eq (14) with the samples as nodes. After symmetrization, we obtain a Markov chain with transition matrix $P = (p_{ij})_{N\times N}$ through Eq (16) as

$$p_{ij} = P_{\varepsilon,\frac{1}{2}}(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad i, j = 1, 2, \ldots, N. \tag{18}$$

The stationary distribution of the Markov chain is defined as $\boldsymbol{\mu} = (\mu_i)_{1\times N}$, where

$$\mu_i = \frac{d_{\varepsilon,1/2}(\boldsymbol{x}_i)}{\sum_{i=1}^{N} d_{\varepsilon,1/2}(\boldsymbol{x}_i)}, \qquad i = 1, 2, \ldots, N, \tag{19}$$

which satisfies $\boldsymbol{\mu} = \boldsymbol{\mu} P$. P and $\boldsymbol{\mu}$ are used in the procedures below. Usually for the Gaussian kernel, the bandwidth $\sqrt{2\varepsilon}$ is set as the median value of the distances between all samples.

Clustering cells and computing the coarse-grained operator between clusters

There are lots of well-known clustering methods, such as k-means, spectral cluster, dynamical reduction [36], and new packages designed for single cell datasets, such as SC3 [44] and Seurat [45]. It is still a challenge to make a choice among the diverse approaches [46]. However, for our model, we have several limits when choosing a suitable clustering method. One important point is that different cell types should be separated into different groups. Each cell type holds a metastable state in the dynamical system. The other important point is that we need to maintain continuity of differentiation paths. Some nonlinear clustering approaches may fail, as they may scatter different clusters far away or eliminate the transition point that connects different branches. A suitable clustering is one fully using information from the data and coincident with the biological background. In our simulated datasets and real datasets, as the model is built on the original euclidean space, we used the linear principal component analysis (PCA) to reduce dimension and k-means to cluster the data, which has orthogonal invariance and keeps both the separation and continuity perfectly. When choosing the number of clusters in k-means, the following constrains should be considered. (1) From the expression of Eq (S7) in the S1 Text, we can conclude that the best number of the clusters is exactly the number of cell subtypes (the metastable states). The metastable wells are separated by the ridges of system potential $F(\boldsymbol{x})$. Hence, there is an upper limit for the cluster number, which is decided by the number of metastable wells. (2) For the differentiation process with one lineage, the lower limit of cluster number is two, and for two lineages, it is four; one is for stem cells, one is around the bifurcation point, and the other two clusters stand for two branches. (3) If two neighboring clusters, which do not include the cluster near the bifurcation point, merge into one, Eq (S7) will still hold, and the pseudo-time order of the cells will not change. Under these constrains, we can ensure the LDD robustness for different cluster numbers. The results are shown in Fig J in S1 Text.

If we get K clusters of all cells by one of the above methods, we can define a coarse-grained matrix $\hat{P} \in \mathbb{R}^{K\times K}$ between clusters s and t as

$$\hat{P}_{st} = \hat{P}(\Omega_s, \Omega_t) = \frac{\hat{\mu}_t}{n_s n_t}\sum_{i\in\Omega_s}\sum_{j\in\Omega_t} p_{ij}, \qquad s, t = 1, 2, \ldots, K, \tag{20}$$

where $\hat{\mu}_t = \sum_{i\in\Omega_t}\mu_i$ and n_s is the number of samples in Ω_s (see S1 Text for details). Correspondingly, we can get the approximated operator $\hat{L}$ through $\hat{P}$ by Eq (17) as

$$\hat{L} = \frac{\hat{P} - I}{\varepsilon}. \tag{21}$$

Theoretical derivation of net-flow for each cluster by divergence theorem

One of the most important advantages in the model is that we can compute the net-flow rate $\hat{R}$ of each cluster only based on scRNA-seq data. From Eq (1), if we only focus on the samples/cells in one cluster (or cell type) Ω_s, we can define the conditional probability density function (cpdf) as

$$r_s(\boldsymbol{x}, t) = \frac{c(\boldsymbol{x}, t)\chi_{\Omega_s}(\boldsymbol{x})}{\int_{\Omega_s} c(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}}, \tag{22}$$

where $\chi_{\Omega_s}(\boldsymbol{x})$ is the indicator function of Ω_s, i.e. $\chi_{\Omega_s}(\boldsymbol{x}) = 1$ when $\boldsymbol{x}$ is in Ω_s and 0 otherwise. In the long time limit, $c(\boldsymbol{x}, t)$ and $r_s(\boldsymbol{x}, t)$ will converge to the steady distribution $p(\boldsymbol{x})$ and $r_s(\boldsymbol{x})$, respectively. By applying the divergence theorem to the equation satisfied by $r_s(\boldsymbol{x})$ and eliminating equal terms, we can derive an equation satisfied by the net-flow rate $\hat{R}_s$ of cluster s as

$$\hat{R}_s \triangleq \int_{\Omega_s} R(\boldsymbol{x}) r_s(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = -D\int_{\Omega_s} \Delta r_s(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{23}$$

The derivation details are shown in S1 Text.

Numerical computation of net-flow for each cluster

The net-flow rate formula can be simplified by marginal density functions as

$$\begin{aligned}\hat{R}_s &= -D\int_{\Omega_s} \Delta r_s(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = -D\sum_{j=1}^{m}\int_{\Omega_s^{(j)}=[a_s^{(j)}, b_s^{(j)}]} \partial_x^2 r_s^{(j)}(x)\, \mathrm{d}x \\ &= -D\sum_{j=1}^{m}\left[\partial_x r_s^{(j)}(b_s^{(j)}) - \partial_x r_s^{(j)}(a_s^{(j)})\right], \qquad s = 1, 2, \ldots, K,\end{aligned} \tag{24}$$

where $\boldsymbol{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(m)})^T \in \Omega_s$, $r_s^{(j)}(x)$ is the marginal density of $r_s(\boldsymbol{x})$ on the $x^{(j)}$-axis, and $\Omega_s^{(j)} = [a_s^{(j)}, b_s^{(j)}]$ is the interval that $r_s^{(j)}(x)$ lies in. By approximating $r_s^{(j)}(x)$ in one-dimensional space through the kernel method and summation of the boundary derivatives, we can compute $\hat{R}_s$ conveniently.

When computing $\hat{R}_s$ for each cluster separately by Eq (24), the steady state Eq (5) of the original system may break down due to finite sample size effect and the numerical error in discrete computation. A post-processing of $\hat{R}_s$ is needed to ensure that the system is steady in discrete setting. By integrating Eq (1) in the whole space, letting $t \to \infty$ and using Eqs (22) and

(23), we obtain the equation for $\hat{R}_s$ as follows:

$$\sum_{s=1}^{K}(\hat{R}_s \cdot \int_{\Omega_s} c(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}) = 0. \tag{25}$$

If there are n_s samples in cluster s and $N = \sum_{s=1}^{K} n_s$, the constraint Eq (25) can be written as

$$\sum_{s=1}^{K} n_s \hat{R}_s = 0. \tag{26}$$

Then we post-process the results obtained from Eq (24) as

$$\tilde{R}_s = \hat{R}_s - \frac{\sum_{s=1}^{K} n_s \hat{R}_s}{N}, \tag{27}$$

in order to satisfy Eq (26). In other parts of this paper, for notational convenience, we still use $\hat{R}_s$ instead of $\tilde{R}_s$, but note that it is a post-processed value.

Computing potential and constructing differentiation paths of cells

By Eqs (10) and (21), the potential $\hat{V}$ in the cluster level satisfies $\hat{L}\hat{V} = -\hat{R}$, i.e.

$$(\hat{P} - I)\hat{V} = -\hat{R}\varepsilon, \tag{28}$$

where $\hat{P}$ is the coarse-grained matrix between clusters, vector $\hat{V} = (\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_K)^T$ is the LDD potential of different clusters, and vector $\hat{R} = (\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_K)^T$ represents the net-flow of each cluster. As the matrix on the left hand side of Eq (28) is degenerate, we need to compute its pseudo-inverse and get the least square solution of $\hat{V}$ as

$$\hat{V} = -(\hat{P} - I)^{\dagger}\hat{R}\varepsilon. \tag{29}$$

On the other hand, the structure of differentiation branches can be inferred from the weight matrix $\tilde{P} \in \mathbb{R}^{K\times K}$ between clusters, whose element is given by

$$\tilde{p}_{st} = \sum_{i\in\Omega_s}\sum_{j\in\Omega_t} \mu_i p_{ij}, \qquad s, t = 1, 2, \ldots, K, \tag{30}$$

and $\tilde{P}$ will become the transition matrix between clusters after row normalization [36]. Non-zero elements in $\tilde{P}$ represent differentiation paths, while in some cases we use a threshold to eliminate small values. The direction of differentiation is determined by $\hat{V}$ from high to low potential. Thus, using $\tilde{P}$ and $\hat{V}$, we can construct the whole differentiation landscape or picture only by the gene expression matrix obtained from an scRNA-seq dataset. Note that the noise D is not required for evaluating the potential landscape V of cells.

Drawing the illustrative landscape

For most of the datasets in this paper, we constructed an illustrative 3D landscape. It is plotted by fitting functions

$$\begin{aligned} f(x, y) &= \frac{1}{N}\sum_{i=1}^{N}\left(-e^{-\frac{(x-x_i)^2+(y-y_i)^2}{(\sigma/\tilde{p}_i)^2}}\right), \\ V(x, y) &= af(x, y) + bg(x, y), \end{aligned} \tag{31}$$

where (x_i, y_i) are the position of N samples in the two-dimensional reduced space, $\tilde{p}_i$ equals to $\tilde{P}_{ss}$ if sample i belongs to cluster s, $g(x, y)$ is a function (usually linear) positively correlated with $\hat{V}$, and a, b, σ are adjustable parameters. Some high values could be set as *NaN* when plotting.

Supporting information

S1 Text. Supplementary information of this paper. The supplementary document provides one algorithm, six notes, and ten supplementary figures for the main text.
(PDF)

Acknowledgments

The authors thank Peijie Zhou and Zhaoyuan Fang for helpful discussions and comments.

Author Contributions

Conceptualization: Jifan Shi, Tiejun Li, Luonan Chen.

Formal analysis: Jifan Shi.

Funding acquisition: Tiejun Li, Luonan Chen, Kazuyuki Aihara.

Methodology: Jifan Shi, Tiejun Li, Luonan Chen, Kazuyuki Aihara.

Supervision: Tiejun Li, Luonan Chen, Kazuyuki Aihara.

Validation: Jifan Shi.

Writing – original draft: Jifan Shi.

Writing – review & editing: Jifan Shi, Tiejun Li, Luonan Chen, Kazuyuki Aihara.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6(5):377–382. https://doi.org/10.1038/nmeth.1315 PMID: 19349980
2. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014; 11(1):41–46. https://doi.org/10.1038/nmeth.2694 PMID: 24141493
3. Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014; 42(14):8845–8860. https://doi.org/10.1093/nar/gku555 PMID: 25053837
4. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013; 10(11):1096–1098. https://doi.org/10.1038/nmeth.2639 PMID: 24056875
5. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2013; 2(3):666–673. https://doi.org/10.1016/j.celrep.2012.08.003
6. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161(5):1202–1214. https://doi.org/10.1016/j.cell.2015.05.002 PMID: 26000488
7. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8:14049. https://doi.org/10.1038/ncomms14049 PMID: 28091601
8. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018; 560(7719):494. https://doi.org/10.1038/s41586-018-0414-6 PMID: 30089906
9. Bendall SC, Davis KL, Amir EaD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014; 157(3):714–725. https://doi.org/10.1016/j.cell.2014.04.005 PMID: 24766814

10. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016; 34(6):637–645. https://doi.org/10.1038/nbt.3569 PMID: 27136076

11. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics. 2015; 31(18):2989–2998. https://doi.org/10.1093/bioinformatics/btv325 PMID: 26002886

12. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F. destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics. 2015; 32(8):1241–1243. https://doi.org/10.1093/bioinformatics/btv715 PMID: 26668002

13. Haghverdi L, Buettner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016; 13(10):845–848. https://doi.org/10.1038/nmeth.3971 PMID: 27571553

14. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32(4):381–386. https://doi.org/10.1038/nbt.2859 PMID: 24658644

15. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017; 14(10):979–982. https://doi.org/10.1038/nmeth.4402 PMID: 28825705

16. Jin S, MacLean AL, Peng T, Nie Q. scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. Bioinformatics. 2018; 34(12):2077–2086. https://doi.org/10.1093/bioinformatics/bty058 PMID: 29415263

17. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, et al. De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell. 2016; 19(2):266–277. https://doi.org/10.1016/j.stem.2016.05.010 PMID: 27345837

18. Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y. SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res. 2017; 45(7):e54. https://doi.org/10.1093/nar/gkw1278 PMID: 27998929

19. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat Commun. 2017; 8:15599. https://doi.org/10.1038/ncomms15599 PMID: 28569836

20. Shi J, Teschendorff AE, Chen W, Chen L, Li T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. Brief Bioinformatics. 2018; p. bby093. https://doi.org/10.1093/bib/bby093

21. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015; 25(10):1491–1498. https://doi.org/10.1101/gr.190595.115 PMID: 26430159

22. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016; 17(1):63. https://doi.org/10.1186/s13059-016-0927-y PMID: 27052890

23. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature. 2017; 541(7637):331–338. https://doi.org/10.1038/nature21350 PMID: 28102262

24. Cannoodt R, Saelens W, Saeys Y. Computational methods for trajectory inference from single-cell transcriptomics. Eur J Immunol. 2016; 46(11):2496–2506. https://doi.org/10.1002/eji.201646347 PMID: 27682842

25. Moon KR, Stanley J, Burkhardt D, van Dijk D, Wolf G, Krishnaswamy S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr Opin Syst Biol. 2017; 7:36–46. https://doi.org/10.1016/j.coisb.2017.12.008

26. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019; 37(5):547. https://doi.org/10.1038/s41587-019-0071-9 PMID: 30936559

27. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. Fundamental limits on dynamic inference from single-cell snapshots. Proc Natl Acad Sci USA. 2018; 115:E2467–E2476. https://doi.org/10.1073/pnas.1714723115 PMID: 29463712

28. Waddington CH. Principles of development and differentiation. Macmillan, New York; 1966.

29. Tusi BK, Wolock SL, Weinreb C, Hwang Y, Hidalgo D, Zilionis R, et al. Population snapshots predict early haematopoietic and erythroid hierarchies. Nature. 2018; 555(7694):54–60. https://doi.org/10.1038/nature25741 PMID: 29466336

30. Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science. 2018; 360(6392):eaar5780. https://doi.org/10.1126/science.aar5780 PMID: 29700227

31. Ferrell JE Jr. Bistability, bifurcations, and Waddington's epigenetic landscape. Curr Biol. 2012; 22(11):R458–R466. https://doi.org/10.1016/j.cub.2012.03.045

32. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002; 297(5584):1183–1186. https://doi.org/10.1126/science.1070919 PMID: 12183631

33. Wang J, Zhang K, Xu L, Wang E. Quantifying the Waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci USA. 2011; 108(20):8257–8262. https://doi.org/10.1073/pnas.1017017108 PMID: 21536909

34. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc Natl Acad Sci USA. 2005; 102(21):7426–7431. https://doi.org/10.1073/pnas.0500334102 PMID: 15899970

35. Coifman RR, Lafon S. Diffusion maps. Appl Comput Harmon Anal. 2006; 21(1):5–30. https://doi.org/10.1016/j.acha.2006.04.006

36. E W, Li T, Vanden-Eijnden E. Optimal partition and effective dynamics of complex networks. Proc Natl Acad Sci USA. 2008; 105(23):7907–7912. https://doi.org/10.1073/pnas.0707563105 PMID: 18303119

37. Guo G, Huss M, Tong GQ, Wang C, Sun LL, Clarke ND, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell. 2010; 18(4):675–685. https://doi.org/10.1016/j.devcel.2010.02.012 PMID: 20412781

38. Stévant I, Neirijnck Y, Borel C, Escoffier J, Smith LB, Antonarakis SE, et al. Deciphering cell lineage specification during male sex determination with single-cell RNA sequencing. Cell Rep. 2018; 22(6):1589–1599. https://doi.org/10.1016/j.celrep.2018.01.043 PMID: 29425512

39. Yang L, Wang WH, Qiu WL, Guo Z, Bi E, Xu CR. A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. Hepatology. 2017; 66(5):1387–1401. https://doi.org/10.1002/hep.29353 PMID: 28681484

40. Furlan A, Dyachuk V, Kastriti ME, Calvo-Enrique L, Abdo H, Hadjab S, et al. Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. Science. 2017; 357(6346):eaal3753. https://doi.org/10.1126/science.aal3753 PMID: 28684471

41. Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016; 44(13):e117. https://doi.org/10.1093/nar/gkw430 PMID: 27179027

42. Welch JD, Hartemink AJ, Prins JF. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol. 2016; 17(1):106. https://doi.org/10.1186/s13059-016-0975-3 PMID: 27215581

43. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics. 2018; 19(1):477. https://doi.org/10.1186/s12864-018-4772-0 PMID: 29914354

44. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017; 14(5):483–486. https://doi.org/10.1038/nmeth.4236 PMID: 28346451

45. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018; 36(5):411–420. https://doi.org/10.1038/nbt.4096 PMID: 29608179

46. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019; p. 1.