Topographer Reveals Stochastic Dynamics of Cell Fate Decisions from Single-Cell RNA-Seq Data

Jiajun Zhang, Tianshou Zhou
School of Mathematical Sciences, Sun Yat-Sen University, Guangzhou 510275, China

Abstract: While single-cell measurement technologies provide an unprecedented opportunity to dissect developmental processes, revealing the mechanisms of cell fate decisions from single-cell RNA-seq data is challenging due to both cellular heterogeneity and transcriptional noise in the data. Here we developed *Topographer*, a bioinformatic pipeline, to construct an intuitive (i.e., every cell is equipped with both potential and pseudotime) developmental landscape, reveal stochastic dynamics of cell types, and infer both dynamic connections of marker gene networks and dynamic characteristics of transcriptional bursting kinetics across development. Applying this method to primary human myoblasts, we not only identified three known cell types but also estimated both their fate probabilities and transition probabilities among them. We found that the percent of the genes expressed in a bursty manner is significantly higher at the branch point than before or after branch, and there are apparent changes in both gene-gene and cell-cell correlations before and after branch. In general, single-cell transcriptome data with *Topographer* can well reveal the stochastic mechanisms of cell fate decisions from three different levels: cell lineage (macroscopic), gene network (mesoscopic) and gene expression (microscopic).

Key Words: cell fate decision, single-cell data, developmental landscape, cell-type dynamics, transition probability

1. Introduction

Multi-cell organisms start as a single cell that matures through a complex process involving multiple cell fate decision points, leading to functionally different cell types, many of which have yet to be defined. Transcriptional programs (in particular transcriptional networks) underlying cell fate decisions drive one cell type toward another often in a random manner due to both cellular heterogeneity[1] and transcriptional noise[2]. Since the structure of a multi-cell tissue is tightly linked with its function[3], elucidating the integrated (from gene to cell) mechanisms of cell fate decisions is crucial yet challenging.

Single-cell measurement technologies[4,5], which can measure a large number of parameters simultaneously in single cells and interrogate an entire tissue without perturbation, provide a great opportunity to elucidate developmental pathways and cell fate decisions. Several algorithms[7-15] developed for analysis of single-cell data have successfully ordered single cells based on their maturity. However, this pseudo-temporal ordering is only the first step towards understanding complex developmental processes involving multiple cell fate decision points. Many other important biological issues, e.g., in single-cell transcriptome data representing a complete development pathway, how many cell types there are, how one cell type transitions another, and how genotype determines phenotype, remain unsolved. The challenge is to devise new analysis approaches to reveal the dynamic mechanisms of cell fate decisions from single-cell data that lacks spatiotemporal information[6].

Here we developed *Topographer*, a bioinformatics pipeline that can construct an intuitive developmental landscape where by 'intuitive' we mean that every cell is equipped with both potential and pseudotime, quantify stochastic dynamics of cell types by estimating both their fate probabilities and transition probabilities among them, and infer dynamic characteristics of transcriptional bursting kinetics along the developmental trajectory. In addition, it can also identify various possible (e.g., bi- and tri-branching) cell trajectories with high resolution from single-cell data and infer dynamic connections of marker gene networks along the identified cell trajectories. In general, single-cell data with *Topographer* can overcome difficulties in constructing complex cell lineages, resolving intermediate stages of cell progress through development, and revealing the integrated mechanisms of cell fate decisions from three different levels: cell lineage, gene network and transcriptional burst (referring to schematic **Figure S1** in **Supplementary Information**).

We demonstrate the power of *Topographer* by analyzing single-cell RNA-seq data on the differentiation of primary human myoblasts[11]. We identified three known cell types: proliferating cells, differentiating myoblasts and interstitial mesenchymal cells, and constructed an intuitive Waddington developmental landscape. By estimating the fate probabilities of the identified cell types and transition probabilities among them, we found that the transition probability from proliferating cells to interstitial mesenchymal cells was approximately twice that from the proliferating cells to differentiating myoblasts, and that the fate probability of the differentiating myoblast type was approximately equal to that of the interstitial mesenchymal cell type. We also

found that the percent of the genes expressed in a burst manner was apparently higher at the branch point (~97%) than before or after branch (not beyond 80%). In addition, mean burst size (MBS) / mean burst frequency (MBF) monotonically decreased / increased before branch but monotonically increased / decreased after branch, with the pseudotime.

RESULT

In order to reveal the stochastic dynamics of cell fate decisions from single-cell transcriptome data, *Topographer* makes the following assumption about the data: the information on the entire development process is adequate, or a snapshot of primary tissue represents a complete development pathway. The overall *Topographer*, a multifunctional algorithm, comprises five functional modules: (1) the backbone module (**Figure 1B**); (2) the landscape module (**Figure 1C**); (3) the dynamics module (**Figure 1D**); (4) the network module (**Figure 1E**); and (5) the burst module (**Figure 1F**).

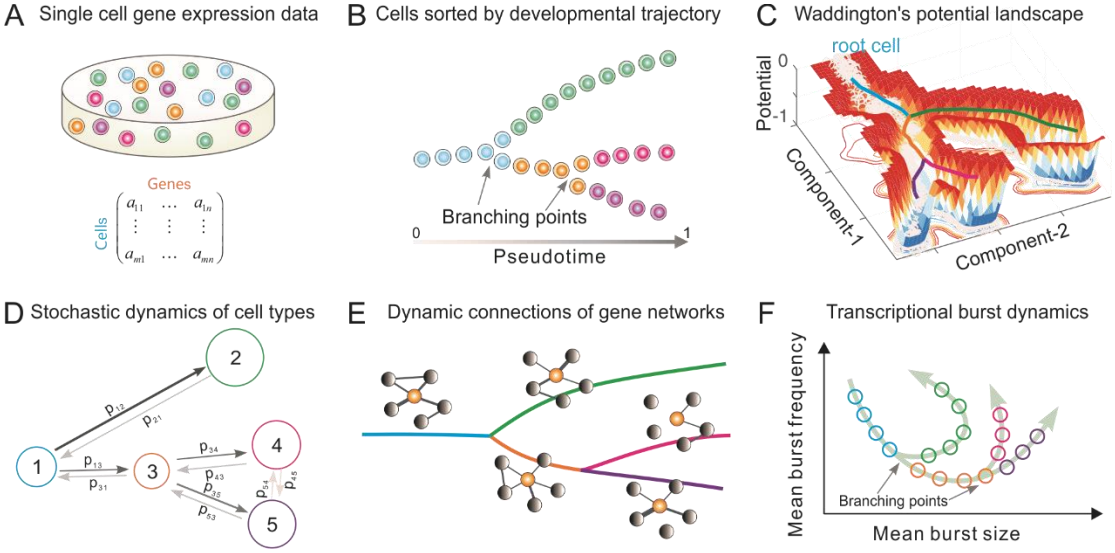

Figure 1 Overview of the five functions of *Topographer*, where every color represents the same meaning, and the only **(C)** subfigure is not schematic but is plotted using a set of data generated by a toy model (**Supplementary Information**).

(A) Single-cell data are represented by a matrix.

(B) *Topographer* identifies the backbone of cell trajectories from the data.

(C) *Topographer* constructs an intuitive Waddington potential landscape where every cell is equipped with both potential and pseudotime (**Online Methods**).

(D) *Topographer* quantifies stochastic dynamics of cell types by estimating the transition probabilities (indicated by symbols) between cell types and their fate probabilities (**Online Methods**), where numbers 1~5 represent cell types, the size of circle represents that of fate probability, and the thickness of the lines with arrows represents the size of transition probability.

(E) *Topographer* infers dynamic connections of marker gene networks along the identified cell

trajectories, where the orange ball represents the marker gene, and the thickness of connection line represents the strength of correlation.

(F) *Topographer* infers dynamic characteristics of transcriptional bursting kinetics (characterized by both burst size and burst frequency) along the pseudotime, where arrows represent the pseudotime direction.

We point out that: (1) different from the existing algorithms[7-15], *Topographer* identifies the backbone of cell trajectories by finding valley floor lines of a developmental landscape (**Online Methods**). The identified backbone is actually a projection of this landscape. (2) Previous algorithms[7-15] (partially) solved the question of ordering single cells in a dataset (**Figure 1B**) but not the others (**Figure 1C~1F**).

Below we simply introduce every functional module of *Topographer*. See **Online Methods** for more details and **Supplementary Information** for a complete description.

1. *Topographer* identifies the backbone of cell trajectories from single-cell data

The backbone module is a fast and local pseudo-potential-based algorithm (here pseudo-potential is defined as the negative of the logarithm of a local density function), which aims to identify the backbone of cell trajectories cross development from single-cell RNA-seq data. The essence of this module is to find valley floor lines in a developmental landscape.

Starting from an initial cell (**Figure 2A**) selected either based on the global minimal pseudo-potential or according to the prior knowledge, *Topographer* calculates an adaptive step (**Supplementary Information**) and searches for pseudo-potential wells on a super-ring (i.e., a high-dimensional circular tube) centered at this initial cell and with the radius equal to the step length (also **Figure 2A**). The search method, which is in essence to cluster cells on super-rings, is based on the idea that cluster centers are characterized by a lower pseudo-potential than their neighbors and by a relatively larger distance from points with lower pseudo-potentials (e.g., the only two 'large' pseudo-potential wells are desired in **Figure 2D**). This idea forms the basis of a procedure to find pseudo-potential wells on a super-ring. In this process, the number of pseudo-potential wells arises intuitively, outliers are automatically spotted and excluded from extra analysis, and pseudo-potential wells are recognized regardless of their shape and of the dimensionality of the space in which they are embedded. We stress that although there is analogy between our method and a density-based approach developed originally by Rodriguez and colleagues[17], the difference is that the former is

carried out on super-rings rather than in the full cell state space. Note that if the number of the found pseudo-potential wells (but not including the one found on the 'reverse' search direction) is more than 1, then this implies the existence of branching trajectories. Moreover, the greater number of the found pseudo-potential wells means more branching trajectories. The segments linking the center and the newly found pseudo-potential well/or wells on the super-ring can be taken as approximate part/or parts of the entire developmental trajectory. Similar processes are repeated recursively on sequential super-rings along a search direction until no new pseudo-potential wells are found (**Figure 2B**). Again by linking the centers and the pseudo-potential wells found on super-rings, *Topographer* thus builds a tree-like developmental backbone (**Figure 2C**). By projecting every cell onto this backbone (**Online Methods**) and by selecting a root node in the tree according to, e.g., the prior knowledge, *Topographer* thus orders all the single cells in the dataset, and equips every cell with pseudotime if this root node is set as an initial moment (without loss of generality, the full pseudotime may be set as the interval between 0 and 1).

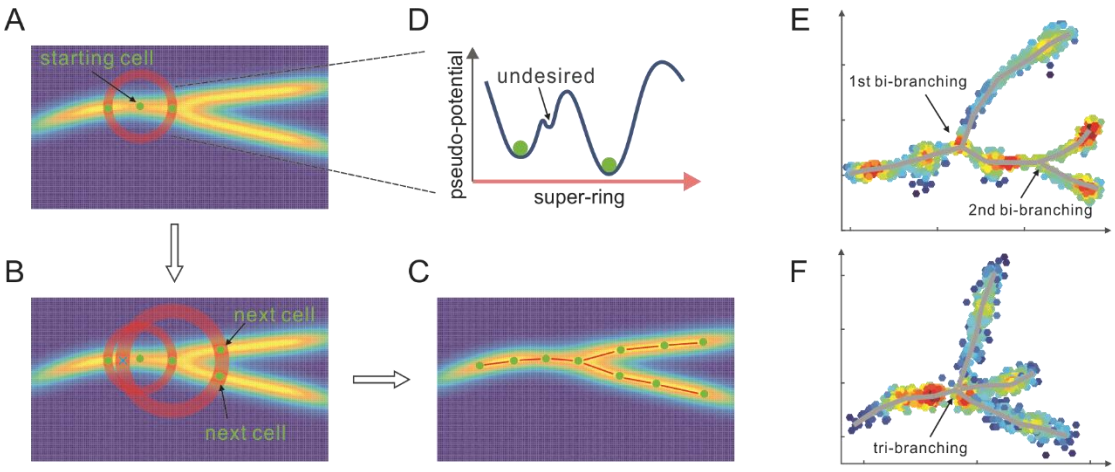

Figure 2 *Topographer* identifies the backbone of branching trajectories from a dataset.
(**A**, **B**, **C**) A flowchart (indicated by arrows): *Topographer* first selects an initial cell (**A**) as the center of a super-ring in the cell state space and searches for pseudo-potential wells on this ring (**A**), where a special case is shown with two desired pseudo-potential wells and the undesired one indicated in (**D**). Then, *Topographer* repeats recursively on every newly found pseudo-potential well (**B**), where symbol 'X' represents a pseudo-potential well found on a reverse search direction, which needs to be excluded in the search process) until no pseudo-potential wells are found, thus obtaining a tree-like backbone of cell trajectories (**C**). Finally, *Topographer* projects every cell onto the backbone, thus ordering all the cells in the dataset.
(E) Bi-branching trajectories identified from an artificial set of data.
(F) Tri-branching trajectories identified from another artificial set of data.

Figure 2E showed a doubly bi-branched trajectory identified from one artificial set of data

whereas **Figure 2F** showed a tri-branched trajectory identified from another artificial set of data. **Figure 3A** below showed a two-dimensional projection of the de novo cell trajectories identified from single-cell RNA-seq data on the development of primary human myoblasts[11] whereas **Figure 3B** showed the evolutions of five marker genes (MYOG, MYF5, MYH2, CDK1 and MEF2C) with branches along the identified developmental trajectory (or along the pseudotime). **Supplementary Information** demonstrated results obtained by analyzing other two examples (**Figure S12** and **Figure S14**), which further showed the power of *Topographer*.

We point out that mainly because of the ability to find pseudo-potential wells on super-rings, *Topographer* can identify developmental trajectories with non-, bi- and multi-branches (referring to **Figure 1E** and **1F**) (remark: the low resolution of experimentally sampling data may lead to, e.g., tri-branches). Thus, *Topographer* is advantageous over previous algorithms[7-15].

2. *Topographer* constructs a developmental landscape using single-cell data

The backbone module uses pseudo-potentials to construct the backbone of cell trajectories, which extracts the information on branches and cell ordering from single-cell data, but this kind of potential cannot correctly reflect transitions between different cell types since probability fluxes would exist between them due to cell division, cell death and/or other factors. For example, precursor cells should in principle have higher potentials (see **Online Methods** for definition) in a Waddington's developmental landscape in contrast to their generations, but if the precursor cells have higher densities, then they have lower pseudo-potentials. Both are apparently inconsistent. In addition, pseudo-potential lacks the time information on differentiation or development.

Because of the above shortcomings of pseudo-potential and since the Waddington's potential landscape has extensively been viewed as a powerful metaphor for how differentiated cell types emerge from a single and totipotent cell, the landscape module (an algorithm) is designed to construct a 'stereometric' developmental landscape where by 'stereometric' we mean that every cell is equipped with potential and pseudotime, in contrast to the 'planimetric' backbone constructed in the backbone module. The landscape module aims to provide a more intuitive understanding for the entire developmental process. The principle is simply stated as follows.

Since single-cell data are in general noisy due to both cellular heterogeneity and gene

expression noise, transitions among the cells scattered in the cell state space can be viewed as a random walker who randomly moves from a cell point to another[18]. *Topographer* first constructs a weighted directed graph based on the pseudotime information obtained in the backbone module, and then defines a conditional probability (**Online Methods**) that the random walker moves from one cell to another as the relative link weight. With these weights, *Topographer* calculates the visit probability for every cell by solving a master equation (also **Online Methods**). Furthermore, *Topographer* calculates the potential of every cell in the dataset, where potential is defined as the negative of the logarithm of visit probability (**Online Methods**). All these potentials are then used to construct a 'stereometric' developmental landscape in contrast to the 'planimetric' backbone constructed using pseudo-potentials. Note that when drawing such a landscape, dimension reduction[16] is used to visualization, the nearest neighbor interpolation is used to fit a landscape function of two variables in a 2-dimension space, and a Gaussian kernel is used to smooth interpolation (see **Online Methods** or **Supplementary Information** for details). Also note that in the constructed developmental landscape, every cell is equipped with both potential and pseudotime, two important attributes of a cell.

Since the constructed 'stereometric' developmental landscape considers the potential of every cell in the dataset whereas the constructed 'planimetric' backbone of cell trajectories considers pseudo-potential wells, the latter can in vision be viewed as an aerial photograph of the former (comparing **Figure 3C** with **Figure 3A)**.

In order to demonstrate developmental landscapes constructed by the landscape module, we analyzed two examples (see **Supplementary Information** for details): the one for the same set of artificial data as in **Figure 2E**, with the result shown in **Figure 1C**, and the other for a set of single-cell data on the differentiation of primary human myoblasts, with the constructed developmental landscape demonstrated shown in **Figure 3C** (where both the identified cell trajectories and the root cell selected according to the prior knowledge were indicated). It seemed to us that **Figure 3C** was the first Waddington's developmental landscape constructed from a realistic set of data (compared with Figure 5 in Ref. [19], which is a cartoon). **Supplementary Information** exhibited another Waddington's developmental landscape constructed using single-cell data on the development of somatic stem cells (**Figure S13**), which provides an intuitive understanding for the developmental process of this kind of cell.

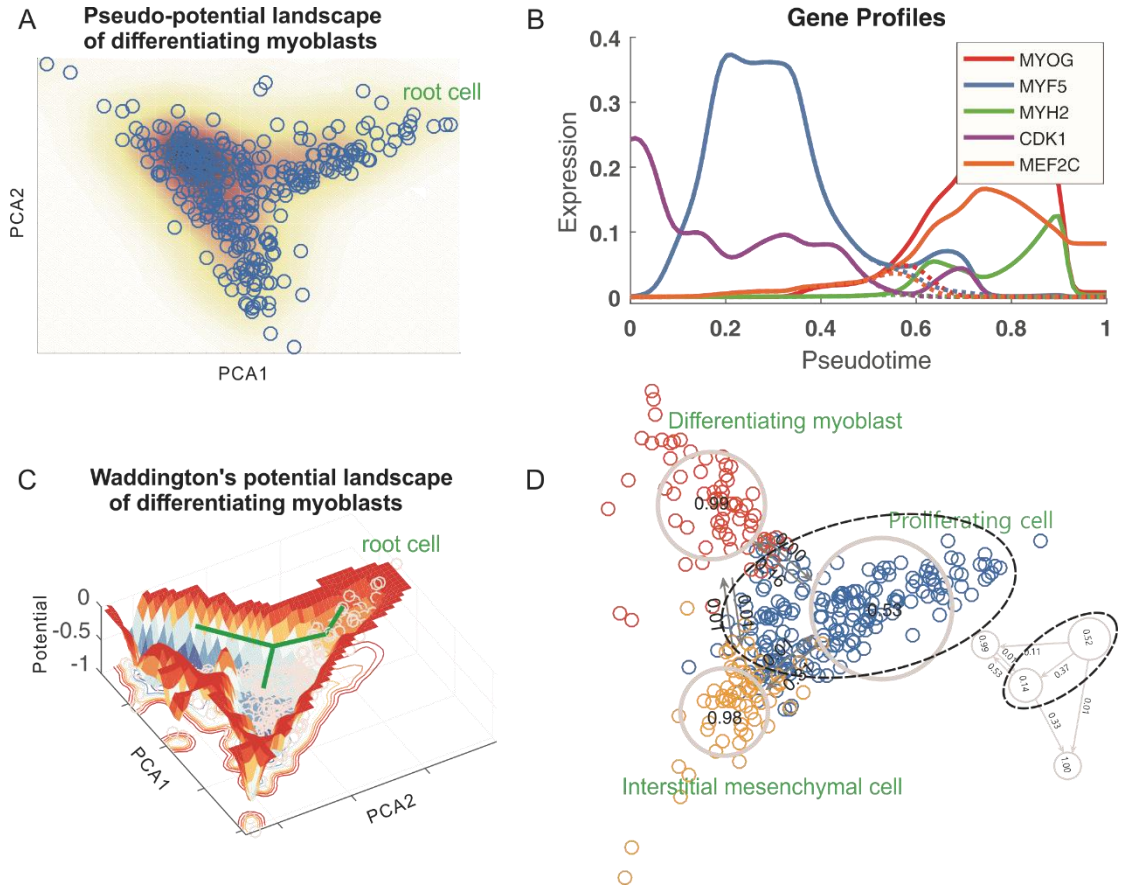

Figure 3. Analysis of single-cell data on the differentiation of primary human myoblasts.
(A) *Topographer* constructs a pseudo-potential landscape, where PCA1 and PCA2 represent components.
(B) Evolutions of the expression levels of five marker genes (indicated by different colors) associated to cell fate decisions along the pseudotime (i.e., along the identified cell trajectories), where dashed lines represent gene levels after branch.
(C) *Topographer* constructs a Waddington's potential landscape, where a thick green line with branches corresponds the backbone of cell trajectories identified by the backbone module, and each small, grey circle represents one cell. The normalized potential is shown with depth of color representing the size of potential.
(D) *Topographer* quantifies stochastic dynamics of cell types along branching trajectories by estimating both the survival probabilities of cell types (distinguished by colors) and transition probabilities among them. Three known cell types: proliferating cells, differentiating myoblast and interstitial mesenchymal cells, are indicated by dashed ellipse and circles. The dashed ellipse shows that the proliferating cell type (top panel) can further be classified into two subtypes (below panel). The survival probability of every cell subtype and transition probabilities between these subtypes are also indicated.

We point out that: (1) in contrast to the backbone module that aims to find a 'road' but ignores bumpiness of the road, the landscape module considers both the road (actually a valley floor of the constructed Waddington potential landscape) and its bumpiness (reflected by the potential of every cell). (2) Both the backbone module and the landscape module can identify cell trajectories from a

dataset, but the former uses pseudo-potentials that rely on neither pseudotime nor cell type whereas the latter use potentials depend on both pseudotime and cell type (see **Eq. (8)** with **Eq. (4)** in **Online Methods**). (3) Pseudo-potential cannot correctly reflect the motion of a 'ball' in the constructed Waddington potential landscape in which the ball has lower potential at the beginning than at the end, since a lower cell density means a higher pseudo-potential according to definition, referring to **Figure S5** in **Supplementary Information**.

3. *Topographer* estimates stochastic dynamics of cell types in single-cell data

Cellular heterogeneity and inherent noise in gene expression may result in stochastic transitions between cell types (even including those between cell subtypes). Quantifying such a transition using single-cell data is challenging but would be important for understanding the formation and functioning of cell types.

In order to quantify stochastic dynamics of cell types in single-cell RNA-seq data, it is first needed to determine types of the cells, which however is a fundamental issue in cell biology. *Topographer* determines cell type according to the following rule: (1) every branch of the identified developmental trajectory is viewed as a kind of cell type but a different branch as a different cell type; (2) for every branch, every found potential well is taken as a kind of cell subtype but a different potential well as a different cell subtype. Thus, the number of cell types is equal to that of branches whereas the total number of cell subtypes is equal to that of potential wells. In the following, we will not distinguish cell type and cell subtype unless confusion arises. We point out that the cell types determined in such a manner depend on the shapes of rugged potential wells (but the prior knowledge can provide guidelines in some cases). Therefore, the classification of cell types is not absolute but relative, referring to **Figure 3D** where the proliferating cell type indicated by a dashed ellipse is further divided into two subtypes. We also point out that in some situations, a potential well in the constructed Waddington potential landscape would not be apparent but would represent a small cell subtype or an intermediate cell state, which however may have important biological implications.

The dynamics module (also an algorithm) aims to reveal stochastic dynamics of the identified cell types in the dataset. For this, *Topographer* mainly calculates two kinds of probabilities: the fate

probability for every cell type and the transition probability between every two cell types. Importantly, in these calculations, *Topographer* makes use of the information on the cell trajectories identified in the backbone module.

Specifically, *Topographer* first calculates a weight of the directed edge from a cell to another based on the pseudotime (**Online Methods**), and then uses all the possible weights to calculate the so-called visit probability that the random walker visits a cell point in the cell state space, and further the conditional probability that is defined as a relative link weight. With these two kinds of probabilities, *Topographer* further calculates the probability that the random walker visits every cell type, and the transition probabilities between every two cell types (also **Online Methods**). These calculations indicate that transitions between cell types are in general not deterministic but random (referring to **Figure 3D**).

In order to demonstrate stochastic cell-type dynamics estimated by the dynamics module, we again analyzed an artificial set of data with results shown in **Figure S6** of **Supplementary Information** and a realistic set of single-cell RNA-seq data on the development of human myoblasts with results shown in **Figure 3D** (as well as another realistic set of single-cell RNA-seq data on the development of somatic stem cells with results shown **Figure S13** of **Supplementary Information**). From **Figure 3D**, we observed that the fate probability (~0.53) for the proliferating cell type is about the half of that for the differentiating cell or interstitial mesenchymal cell type (this is not strange since the proliferating cells are root cells) but the fate probabilities for the latter two (~0.99 and ~0.98) are approximately equal. In addition, proliferating cells differentiate into differentiating cells at the ~0.16 probability but the inverse differentiation probability is very small (~0.001). On the other hand, proliferating cells differentiate into interstitial mesenchymal cells at the ~0.31 probability but the inverse differentiation probability is also very small (~0.01), implying that proliferating cells tend to differentiate into interstitial mesenchymal cells. **Figure 3D** also showed the fate probabilities of cell subtypes and transition probabilities between them (low panel).

Apart from the above three main functional modules, *Topographer* can also infer both dynamic connections of marker gene networks and dynamic characteristics of transcriptional bursting kinetics along the pseudotime. We point out that the connections or characteristics to be inferred can in turn be used to infer whether and when (along the pseudotime) the branches of a developmental trajectory occur.

4. *Topographer* infers dynamic connections of marker gene networks along the pseudotime

A Waddington's potential landscape provides an intuitive understanding for developmental trajectories, but cell fate decisions mainly depend on the underlying regulatory networks. The network module, also an algorithm, aims to infer the trend of how marker gene networks dynamically change along the identified cell trajectories. For this, *Topographer* uses the network neighborhood analysis method[20] (or see **Online Methods**) to explore dynamic connections in a gene regulatory network (GRN) across development.

First, *Topographer* uses GENIE3[21] to generate a series of GRNs along the pseudotime. Then, based on these GRNs, it further analyzes the covariation partners of some particular gene (or genes) using a topological network analysis scheme[22] that can identify those genes that are most closely correlated with a given gene (or genes) of interest and most closely correlate to each other. See **Online Methods** for details.

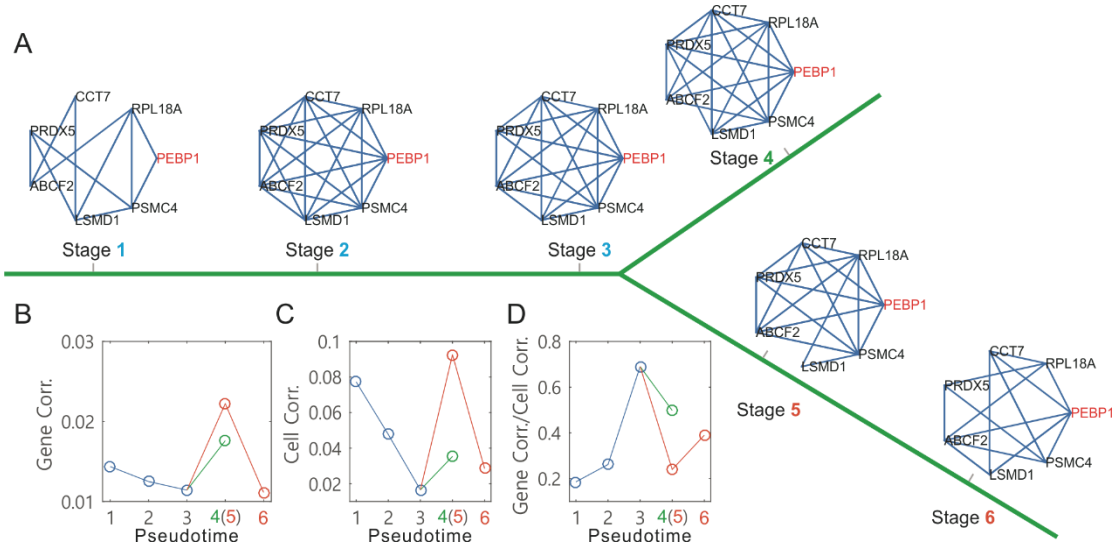

Figure 4 *Topographer* infers dynamic changes in the local connection network of a marker gene along the pseudotime from single-cell data on the differentiation of primary human myoblasts.
(A) Dynamic changes in a connection network of seven genes along the pseudotime, where the PEBP1 gene (orange) is taken as a core node of connection networks.
(B) Dynamic changes in gene-gene correlations along the pseudotime before and after branch (see different colors), where 6 empty circles correspond to the networks at 6 stages indicated in (A), respectively.
(C) Dynamic changes in cell-to-cell correlations along the pseudotime before and after branch.
(D) Dynamic changes in the ratio of the gene-to-gene correlation degree over the cell-to-cell correlation degree along the pseudotime before and after branch.

Here, we used the network module to analyze single-cell data on the differentiation of primary

human myoblasts[11], obtaining characteristics of the dynamic changes in connections of marker genes along the identified cell trajectory. **Figure 4A** demonstrated dynamic changes in the connection of a neighborhood network of the PEBP1 gene (as a marker) along the pseudotime, whereas **Figure 4B~4D)** displayed how mean gene-gene correlations (**Figure 4B**) and mean cell-cell correlations (**Figure 4C**) change also along the pseudotime. We observed that before branch, each of the mean gene-gene and cell-cell correlation degrees was a monotonically decreasing function in pseudotime (the blue line in **Figure 4B** or **4C**), but after branch, it becomes first monotonically increasing and then monotonically decreasing on one branch (the orange line in **Figure 4B** or **4C**), and it becomes monotonically increasing on the other branch (the green line in **Figure 4B** or **4C**). However, the change tendency for the ratio of the former degree over the latter degree was just opposite to the described trend (**Figure 4D**).

5. *Topographer* infers dynamic characteristics of transcriptional bursting kinetics along the pseudotime

Transcription occurs often in a bursty manner and single-cell experimental measurements have also provided evidence for transcriptional bursting both in bacteria and in eukaryotic cells[23]. By analyzing a simple stochastic model of gene expression, Xie, e al showed[24] that the number of mRNAs produced in the bursty fashion followed a Gamma distribution determined by two parameters: MBF (i.e., the mean number of mRNA production bursts per cell cycle), and MBS (i.e., the average size of the mRNA bursts). The burst module, also an algorithm, is designed to infer the trend of how transcriptional bursting kinetics dynamically changes across development. For this, *Topograph*er uses the maximum likelihood method[25] to infer the two parameters of MBF and MBS from single-cell RNA-seq data (see **Online Methods** for details), thus revealing dynamic characteristics of transcriptional bursting kinetics before branch, at the branching point and after branch of the developmental trajectory.

We used the burst module to analyze single-cell data on the differentiation of primary human myoblasts[11]. **Figure 5A~5E** showed how the cells at four pseudotime points (two before branch, one at branch point and one after branch) were distributed in the logarithmic plane of burst frequency (BF) and burst size (BS). A reference system (see two orthogonal blue lines: the horizontal line for MBF and the vertical line for MBS) was used to guide visual comparison between the rates (see the

indicated fractions) of cell numbers over the total cell number at a particular pseudotime point. The four quadrants of the reference system clearly showed how the genes in the dataset are expressed, e.g., the fourth quadrant showed that the genes were expressed in a manner of high frequency (i.e., the BF is more than 0.33) and small burst (i.e., the BS number is less than 200). We observed that the genes expressed in a bursty manner (i.e., the remaining three cases except for the case in the fourth quadrant) are more at the branch point (97%) than before or after branch (approximate or below 80%). In other words, the percent of the genes expressed with high frequency and small burst was apparently lower at the branch point. From these figures, we concluded that during the differentiation of primary human myoblasts, the percent of the genes expressed in a bursty manner was significantly greater at the branch point than before or after branch.

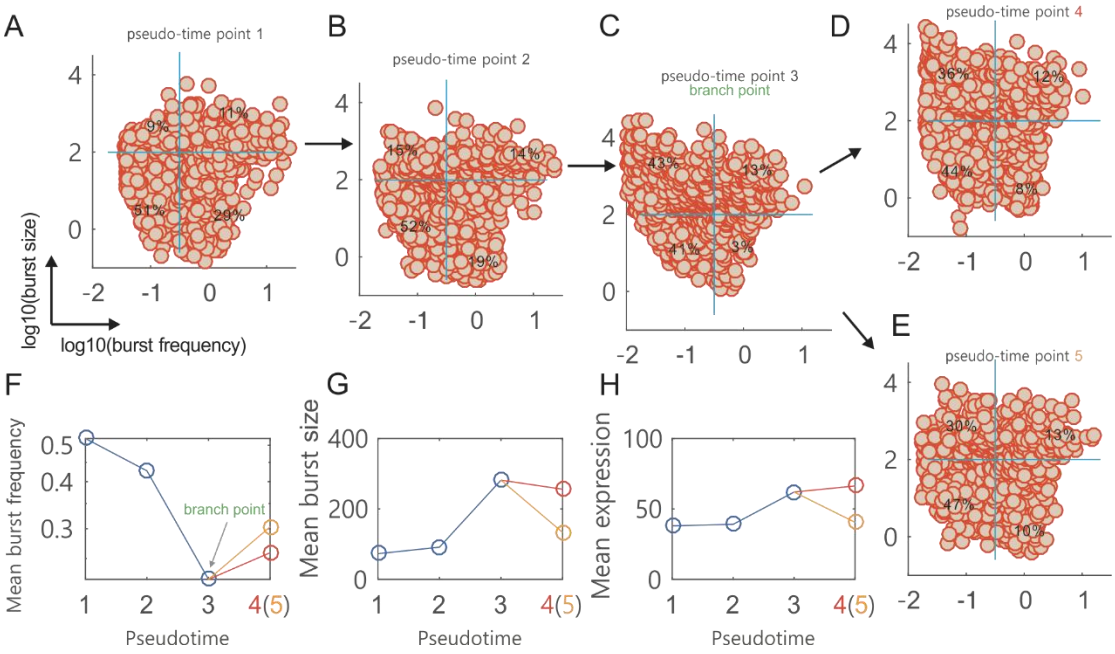

Figure 5 *Topographer* infers dynamic characteristics of transcriptional bursting kinetics along the pseudotime from single-cell data on the differentiation of primary human myoblasts.

(A-E) Scatter plot of the cells in the logarithmic plane of burst size (BS) and burst frequency (BF) at four pseudotime points respectively, where every circle represents a cell in the dataset. Four rates are indicated in a reference system (see two orthogonal blue lines at every subfigure, which correspond to mean BS and BF, respectively). Numbers 4 and 5 actually represent the same pseudotime point.

(F) Evolution of the mean BF along the pseudotime, where the branch point is indicated and two empty circles after the branch point correspond to (D) and (E), respectively.

(G) Evolution of the mean BS along the pseudotime.

(H) Evolution of the mean mRNA expression level along the pseudotime.

Figure 5F and **Figure 5G** showed the dependences of MBF and MBS on the pseudotime, respectively. We observed that there were apparently different trends before and after branch. **Figure**

5H showed the dependence of the mean mRNA expression level on the pseudotime, demonstrating the change tendency opposite to that of MBF. Although fundamentally similar to the trend of MBS on the whole, the mean mRNA level changed almost stably with the pseudotime for one branched trajectory (referring to the top line after branch in **Figure 5H**). These three subfigures implied that BF or BS can be taken as a better indicator of the branch occurrence than the mean mRNA expression level.

DISCUSSION

We have developed a bioinformatic pipeline -- *Topographer*, which enables the construction of developmental landscapes, the identification of de novo continuous developmental trajectories, and the uncovering of stochastic cell-type dynamics. The high resolution of *Topographer* can elaborately characterize both dynamic transcriptional bursting kinetics and dynamic connections of the networks of marker genes underlying cell fate decisions across development. When identifying the backbone of cell trajectories from single-cell data, *Topographer* is robust to the noise in the dataset (**Figure S8** - **Figure S10**). When applied to the differentiation of primary human myoblasts[11], *Topographer* first constructed an intuitive developmental landscape that provided an order and timing of events that closely recapitulated previous studies of this system, and then but more importantly, estimated the sizes of the fate probabilities for cell types and the transition probabilities between them. These estimations indicated not only that the transition from one kind of cell type to another during the differentiation of primary human myoblasts occurred in a probabilistic rather than deterministic manner, but also that the transitions between cell types may be unidirectional and bidirectional. Thus, our results challenged the traditional view that the development of primary human myoblasts was tree-like or that the process from a predecessor to its generations was both deterministic and unidirectional[3].

When applied to analysis of single-cell transcriptome data, *Topographer* (like similar methods in the literature) needs sufficiently many cells since it was established essentially based on the estimation of cell density. If the number of cells is too few (e.g., less than 100), this would lead to, e.g., inaccuracy of finding pseudo-potential wells on a super-ring in the backbone module. Fortunately, more and more cells currently can simultaneously be measured by single measurement technologies[22]. In addition, not limited to analysis of single-cell RNA-seq data, *Topographer* can

also be used to analysis of mass cytometry data[26] and single-cell PCR data[27]. This greatly extends the application ranges of *Topographer*.

Cell fate decisions would involve many factors or processes, which may lead to the hierarchy of cell types including intermediate cell states or cell subtypes. These cell subtypes, some of which have not been defined yet, may have important biological implications. Identifying elaborate cell types is a fundamental yet challenging problem in molecule biology. Apart from identifying known cell types (e.g., proliferating cells, differentiating cells and interstitial mesenchymal cells) from single-cell transcriptome data on the differentiation of primary human myoblasts, *Topographer* has the ability to identify cell subtypes, which in general correspond to shallow or small potential wells in the constructed developmental landscape (referring to the right below panel in **Figure 3D**). Moreover, *Topographer* can estimate the fate probability of every identified cell subtype and transition probabilities between every two identified cell subtypes (right below, **Figure 3D**). This is a main point of *Topographer* advantageous over the methods in the existing literature[7-15]. In addition, *Topographer* enables identification of non-, bi- and multi-branches (**Figure 2C** and **2D**), which is another advantage over the existing algorithms even including Wishbone[12].

It should be pointed out that *Topographer* provides only a framework for analyzing the stochastic mechanisms of cell fate decisions based on single-cell data from three different aspects: cell lineage committing dynamics (macroscopic), dynamic connections of gene networks (mesoscopic) and transcriptional bursting kinetics of genes (microscopic). First, these three aspects are inter-coupled and interplayed. *Topographer* provided useful information on their relationships that are implied by the pseudotime, but this kind of time only reflects the impact of the former on each of the latter two. The issues of how and in what degree the inferred gene connection networks or/and transcriptional bursting kinetics influence or/and determine cell fates in the underlying developmental process, remain unexplored. In order to study the relationship between the mesoscope or microscope and the macroscope, a possible way is to establish the so-called balance equation[27]. Second, in order to estimate the transition probabilities between cell types and their fate probabilities (see **Eq. (9)** and **Eq. (12)** in **Online Methods**, respectively), *Topographer* makes an assumption, that is, the transition from one cell to another along a developmental trajectory is linear (referring to **Eq. (4)** in **Online Methods** or **Eq. (6)** in **Supplementary Information**). In a realistic scenario, however, such a transition would be nonlinear since, e.g., cell-cell communication by

signal molecules would be nonlinear. Third, while cell-state dynamics are of particular significance in, e.g., tumor pathobiology[28], the traditional Waddington landscape figuratively describes a cell differentiation process as the trajectory of a ball into branching valleys, each of which represents a developmental state[29]. *Topographer* uses the potential of every cell to construct an intuitive developmental landscape helpful for understanding the underlying developmental pathway, and both the transition probabilities between cell types and their fate probabilities to characterize cell lineage committing dynamics. It is worth pointing out that these probabilities have definite physical meanings since they actually represent the Krammer escape rates[30] between potential wells in the constructed developmental landscape. However, how cell fate decisions including cell-state dynamics are related to Krammer escape rates is unclear. Fourth, based on the transition probabilities between cell types, one can establish a model of cell population dynamics, and further study stochastic state transitions from a dynamical-systems view (referring to an example analysis in **Supplementary Information**). Issues along these four directions are worth deep study.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Acknowledgments

We thank Prof. Qing Nie's constructive suggestions and Mr. Xiaoming Lai's partial computer program codes for the manuscript. This work was partially supported by National Natural Science Foundation of China (11775314, 91530320), and National Key Basic Research Program of China (2014CB964703).

AUTHOR CONTRIBUTIONS

J.J. and T. S. conceived and designed this work. J. J. developed a potential-based approach (*Topographer*) and applied it to several sets of single-cell RNA sequencing data. J.J. analyzed the data and T. S. wrote the manuscript.

COMPETING FINANCIALINTERESTS

The authors declare no competing financial interests.

References

1. Janes, K. A. Single-cell states versus single-cell atlases — two classes of heterogeneity that differ in meaning and method. Curr. Opin. Biotech. 39: 120–125(2016).
2. Paul, F., et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163: 1–15 (2015).
3. Perie L. et al. The Branching Point in Erythro-Myeloid Differentiation. Cell 163, 1655–1662 (2015).
4. Bendall, S.C., et al. Single cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science **332**, 687–696 (2011).
5. Jaitin, D.A., et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science **343**, 776–779(2014).
6. Trapnell, C. Defining cell types and states with single-cell genomics. Genome Res. 25: 1491–1498 (2015).
7. Riziv A. H. et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat. Biotechnol. 35, 551–560 (2017).
8. Bendall, S.C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell **157**, 714–725 (2014).
9. Shin, J. et al. Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell **17**, 360–372 (2015).
10. Marco, E. et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc. Natl. Acad. Sci. USA **111**, E5643–E5650 (2014).
11. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. **32**, 381–386 (2014).
12. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat. Biotechnol. **34**, 637–645 (2016).
13. Leng N. et al. Oscope identifies oscillatory genes in unsynchronized single cell RNA -seq experiments. Nat. Methods **12**, 947–950 (2015).
14. Kafri R. et al. Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. Nature **494**, 480-483 (2013).
15. Gut G. et al. Trajectories of cell-cycle progression from fixed cell populations. Nat. Methods **12**, 951–954 (2015).
16. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. J. Machin. Learn. Res. 9:2579-2605 (2008).
17. Rodriguez A. & Laio A. Clustering by fast search and find of density peaks. Science 344, 1492-1496 (2014).
18. Rosvall, M. & Bergstrom, C.T. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA 105, 15(4):1118–1123 (2008).
19. Olsson, A. et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature 537, (7622) : 698-702 (2016).
20. Li, A. & Horvath, S. Network neighborhood analysis with the multimode topological overlap measure. Bioinformatics 23, 222-231(2007).
21. Huynh-Thu,V. et al. (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS One, 5, e12776.

22. Klein A.M., Mazutis, …, Weitz D.A., Kirschner M.W., Droplet barcoding for single-cell transciptomics applied to embryonic stem cell. Cell 161, 1187-1201 (2015).
23. Larson D. R. What do expression dynamics tell us about the mechanism of transcription? Curr. Opin. Genet. Dev. 21:591-599 (2011).
24. Friedman, N., Cai, L. & Xie, X.S. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. Phys. Rev. Lett. 97, 168302 (2006).
25. Cam, L.L. Maximum likelihood — an introduction. Intern. Statist. Rev. 58 (2): 153–171 (1991).
26. Bendall, S.C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332, 687–696 (2011).
27. Wu J.C., Tzanakakis E.S. Contribution of stochastic partitioning at human embryonic stem cell division to NANOG heterogeneity. PLoS ONE 7(11), e50715 (2012).
28. Gupta P. B., Fillmore C.M.,…, Lander E.S. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell 146, 633–644 (2011).
29. Furusawa C. and Kaneko K. A dynamical-systems view of stem cell biology. Science 338, 215-217 (2012).
30. Gammaitoni L., Hänggi P., Jung P., and Marchesoni F. Stochastic resonance. Rev. Mod. Phys. 70, 223-287 (1998).

Online Methods

Cell differentiation or development is a complex process involving multiple cell fate decision points. Recent single-cell analysis technologies such as single-cell RNA-seq, which are enabling generation of data with high resolution, offer a great opportunity to reveal developmental processes and cell fate decisions, but require computational algorithms capable of exploiting this resolution. Although previously developed algorithms can successively order single cells in some single-cell data, this pseudo-temporal ordering is only the first step towards understanding developmental processes and cell fate decisions. Many other important yet fundamental issues, e.g., in single-cell transcriptome data representing a complete development process, e.g., how many cell types there are, how cell types are identified, how one cell type transitions another, and how genotype determines phenotype, remain unsolved. Using high-dimensional single-cell RNA-seq data, *Topographer* aims to reveal the integrated mechanisms of cell fate decisions across development from macroscope (e.g., cell lineage committing dynamics) to mesoscope (e.g., dynamic connections of gene networks) and to microscope (e.g., transcriptional bursting kinetics), referring to **Figure S1** in **Supporting Information**.

Topographer is a bioinformatic pipeline, comprising five functional modules: (1) identifying the backbone of cell trajectories from single-cell transcriptome data; (2) constructing a developmental landscape based on the data; (3) revealing stochastic dynamics of cell types in the data system; (4) inferring dynamic connections of marker gene networks along the identified developmental trajectory; and (5) inferring dynamic changes of transcriptional bursting kinetics along the identified cell trajectories. Main details of the method for every functional module are stated below and a complete detail is given in **Supplementary Information**. Because of irregularity, single-cell data needs pre-processing (**Supplementary Information**).

1. The method for identifying the backbone of cell trajectories from single-cell data

Assume that there are m cells and n genes in single-cell RNA-seq data of interest, which can in principle be represented as m points in the n-dimensional space (X) of gene expression

(called the cell state space for convenience).

The first functional module of *Topographer* (i.e., the backbone module) aims to identify the backbone of cell trajectories during development from the dataset. The essence is to find valley floors in a developmental landscape. Specifically, *Topographer* finds valleys with local minimal pseudo-potentials, where pseudo-potential is defined as

$$\tilde{E}(\boldsymbol{x}) = -\log \rho(\boldsymbol{x}) \tag{1}$$

with

$$\rho(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in X} \exp\left(-\frac{d(\boldsymbol{x},\boldsymbol{y})^2}{2\sigma^2}\right). \tag{2}$$

In **Eq. (2)**, d is the Euclidean distance between two state points $\boldsymbol{x}$ and $\boldsymbol{y}$ in the cell state space X (note: other kinds of distances are also suitable to *Topographer*). Note that ρ represents the local cell density, mostly accounting for the number of cells in a neighborhood defined by σ.

Roughly speaking, *Topographer* starts by cell state x_0 (i.e., an initial cell) and then searches for pseudo-potentials wells on super-rings (which are actually circular tubes in a high-dimensional space) by recursively applying a cluster algorithm[16]. Finally, all the centers of the super-rings are represented in a tree, T. Main details are stated below.

1.1 Constructing a developmental tree

Starting by an initial cell that has the global minimal pseudo-potential or by the cell that the user chooses according to the prior knowledge, *Topographer* calculates an adaptive radius or step (see Subsection 2.1.4 in **Supplementary Information**) and searches for potential wells on a super-ring centered at this cell and with the radius (referring to **Figure 2A**). The search method (called as the pseudo-potential well search algorithm) is based on the idea that cluster centers on the super-ring are characterized by a lower pseudo-potential than their neighbors and by a relatively larger distance from points with locally lower pseudo-potentials. Specifically, *Topographer* first defines

$$\delta(x) = \begin{cases} \max\limits_{y\neq x} d(x,y) & \text{if } \tilde{E}(x) = \min\limits_{y\in X} \tilde{E}(y) \\ \min\limits_{y:\tilde{E}(y)<\tilde{E}(x)} d(x,y) & \text{otherwise,} \end{cases} \tag{3}$$

and then finds local pseudo-potential wells on the super-ring, based on a combination of relatively smaller $\tilde{E}$ and relatively larger δ. Therefore, there is analogy between the pseudo-potential well search algorithm and a density-based approach developed originally by Rodriguez and colleagues[16].

The segments linking the center and the potential wells found on the super-ring can be taken as approximate parts of the entire developmental trajectory.

Then, taking every found pseudo-potential well as the center of a new super-ring with a new adaptive radius, *Topographer* performs similar calculations as at the previous step, thus finding pseudo-potential wells on this new super-ring. Again, the segments linking the new center and the newly found pseudo-potential wells on the new super-ring can be taken as other approximate parts of the entire developmental trajectory. This process is repeated until no new pseudo-potential wells are found. By linking the cluster centers, *Topographer* thus builds a tree-like developmental backbone, which is actually composed of valleys.

Note that for a super-ring' center rather than the starting point, the newly found valleys would include valleys on the "reverse direction" in the processes of searching local pseudo-potential wells on super-rings, which are not expected in our algorithm. To handle such an exception, *Topographer* excludes those valleys that are too close to the found valleys. In addition, any two newly found valleys with the distance smaller than the step length are merged by discarding the valleys with larger pseudo-potentials. Such a treatment may greatly improve the algorithm's robustness against the noise in the dataset.

Also note that a complete valley floor is constructed by terminating the recursive process for some super-ring on which there are no desired pseudo-potential wells to be found. Since no loops are assumed to exist in the developmental trajectory, there is definitely a boundary, implying that the search process necessarily stops within finite steps.

After the above search process is completed, all the found pseudo-potential valley floors are represented in an undirected acyclic graph (a tree with branches). To achieve better accuracy and coverage, *Topographer* refines a pseudo-potential valley tree by searching pseudo-potential wells on the line linking two centers on every edge of the tree. To that end, *Topographer* finishes construction of the backbone of a developmental tree from a given set of single-cell data.

1.2 Cell projection and pseudotime assignment

After a developmental tree has been constructed, *Topographer* then projects every cell point in the cell state space onto some edge of the tree according to the shortest distance principle (i.e., the perpendicular distance from the cell point to the edge is required to be shortest). Thus, every cell has its unique relative position in the identified backbone (or in the constructed tree).

Next, *Topographer* assigns a pseudotime for every cell in the dataset. Before that, however, it is needed to determine a root node in the constructed tree. *Topographer* chooses a root cell in such a manner that the distances between this cell and those cells that are initially set according to, e.g., the prior knowledge, are as short as possible. An initial pseudotime is assigned to this root node. Then, every other cell in the dataset is in order assigned with a pseudotime according to its position in the identified tree. Without loss of generality, the full pseudotime may be set as the interval between 0 and 1.

2. The method for constructing a developmental landscape based on single-cell data

2.1 Calculation of cell potential

After the backbone of a developmental trajectory has been identified and every cell has been equipped with a pseudo-moment, the second functional module of *Topographer* (i.e., the landscape module) is to calculate the potential of every cell in the dataset. All these potentials will then be used to construct a developmental landscape. It is expected that the potential to be introduced can be avoid shortcomings of the pseudo-potential as pointed out in the main text. For this calculation, *Topographer* will analogize transitions between cells at distinct stages of the differentiation process to a random walker who moves randomly between the data points scattered in the cell state space[17]. In addition, in order to construct a weighted directed graph $\mathbf{W}$, it is important that *Topographer* will use the pseudotime information.

Specifically, *Topographer* defines the weight of the directed edge from cell α to cell β as

$$W_{\alpha\to\beta} = W_0 e^{-\chi(\tau_\alpha - \tau_\beta)} . \quad (4)$$

(**Supplementary Information** gives a reason for this definition), where τ_α and τ_β represents the pseudotime points for cells α and β respectively, and positive constant χ represents a linearly changing rate that cell α transitions to cell β (this implies the assumption that the evolutional process from one cell to another along the pseudotime is linear). The setting of the χ value in general depends on the dataset under consideration (**Supplementary Information** gives a simple discussion) but the value of χ is set as 30 in our cases. It is worth pointing out that the

weight defined in such a manner has used the information on the pseudo-temporally ordered cell trajectories, and is a key for the full calculation.

Then, in order to determine cell visit probability on a random walk, *Topographer* defines a conditional probability that the random walker moves from cell β to cell α as the relative link weight, given by

$$p_{\beta\to\alpha}=\frac{W_{\beta\to\alpha}}{\sum_{\beta}W_{\beta\to\alpha}}\ , \tag{5}$$

which is apparently independent of initial W_0. If the stationary visit probability of cell α is denoted by p_α, then this probability can in principle be derived from a recursive system of the form

$$p_\alpha=\sum_{\beta}p_\beta p_{\beta\to\alpha}\ , \tag{6}$$

which represents the probability that the random walker visits the α cell from all the other cells. Note that **Eq. (6)** is actually a master equation[25] and can efficiently be solved with the power-iteration method[26]. However, to ensure that the unique solution of this equation is independent of the starting node in the directed network, the random walker instead teleports to a random node at a small rate τ. In addition, to obtain more robust results depending less on the teleportation parameter τ, it is most often to use teleportation to a node proportional to the total weight of the links to the node[27]. Because of these two points, the resulting stationary visit distribution for cell α is modified as

$$p_\alpha=(1-\tau)\sum_{\beta}p_\beta p_{\beta\to\alpha}+\tau\frac{\sum_{\beta}W_{\alpha\to\beta}}{\sum_{\alpha,\beta}W_{\beta\to\alpha}}\ . \tag{7}$$

Finally, *Topographer* calculates the potential of every cell in the dataset, according to

$$E_\alpha=-\log p_\alpha\ , \tag{8}$$

where p_α is given by **Eq. (7)**. Apparently, the potential defined in such a manner has made use of the information on the identified cell trajectories due to **Eq. (4)**. We point out that the potential of a cell depend on pseudotime but the pseudo-potential lacks the information on pseudotime.

2.2 Scatter plot of developmental landscape

After all the cells are equipped with potentials, all these potentials are then used to draw a developmental landscape. The method is stated as follows. First, dimension reduction is needed for visualization (the tSNE method[28] or the PCA method[29] may be used to achieve this purpose). In

general, dimension reduction cannot explicitly reflect the information on coordinates in a visualized landscape, e.g., PCA1 and PCA2 in **Figure 3C** do not actually represent components in the dataset. Second, *Topographer* uses the nearest neighbor interpolation method to perform interpolation on a 3-dimensional scattered data set. Specifically, *Topographer* uses ScatteredInterpolant (a function of the MATLAB software) to establish corresponding relationships between a set of points, (x, y), and a set of cell potentials, E. These relationships, denoted by $E = F(x, y)$, in principle define a curved surface in the 3-dimensional space for the developmental landscape, which in return passes through all the sampling points in the space under consideration. *Topographer* then uses the nearest neighbor interpolation to evaluate this surface at any query point (x_q, y_q), obtaining an interpolating value of every known potential given by **Eq. (8)**, i.e., obtaining $E_q = F(x_q, y_q)$. Third, a Gaussian kernel is used to smooth interpolation, and the identified developmental trajectory is drawn on the obtained developmental landscape (referring to the thick colored line in **Figure 1A** or the thick green line in **Figure 3C**).

We point out that pseudo-potential cannot correctly reflect the motion of a 'ball' in the constructed Waddington potential landscape in which the ball has lower potential at the beginning than at the end, since a lower cell density implies a higher pseudo-potential according to definition.

3. The method for quantifying dynamics of cell types in single-cell data

In order to quantify cell type dynamics, it is first needed to determine cell types. For this, *Topographer* adopts the following rule: Every branch in the identified developmental trajectory is defined as one kind of cell type, and a different branch is defined as one different kind of cell type. Furthermore, every potential well on every branch is defined as one kind of cell subtype, and a different potential well is defined as one different kind of cell subtype. Thus, the number of cell types is equal to that of branches, and the number of cell subtypes is equal to that of potential wells. It should be pointed out that the cell type determined in such a manner is not unique but depends on the choice of $\tilde{E}$ and δ (see their above respective definitions). In the following, we will not distinguish cell type and cell subtype unless confusion arises.

3.1 Estimating transition probabilities between cell types

Equation (5) has given the conditional probability ($p_{\beta\to\alpha}$) that the random walker moves from

cell β to cell α, whereas **Eq. (7)** has given the stationary visit probability of cell α, i.e., p_α. Then, *Topographer* calculates the transition probability at which a random walker visits the jth cell type from the ith cell type (denoted by $q_{i\curvearrowright j}$), according to

$$q_{i\curvearrowright j} = \sum_{\alpha\in i,\beta\in j} q_{\alpha\to\beta} \ , \tag{9}$$

and the transition probability at which the random walker exits the ith cell type (denoted by $q_{i\curvearrowright}$), according to

$$q_{i\curvearrowright} = \sum_{\alpha\in i,\beta\notin i} q_{\alpha\to\beta} \ , \tag{10}$$

where the unrecorded visit rate on a link, $q_{\beta\to\alpha}$ is given by

$$q_{\beta\to\alpha} = p_\beta p_{\beta\to\alpha} \ . \tag{11}$$

3.2 Estimating cell-type fate probabilities

The fate probability for cell type i, denoted by $fate_i$, is defined as

$$fate_i = 1 - q_{i\curvearrowright} \ , \tag{12}$$

which implies that a larger transition probability at which the random walker exits cell type i corresponds to a smaller fate probability for this cell type. This definition is in accordance with our intuition, so it is reasonable. Again, we emphasize that the above formulae for transition probability ($q_{i\curvearrowright j}$), and fate probability ($fate_i$) have all used the information on the pseudo-temporally ordered cell trajectories.

4. The method for inferring dynamic connections of gene networks along the pseudotime

In complex mixtures of cells, correlations of gene expression patterns would arise from differences between different cell lineages. To explore this correlation between the patterns of gene expression across cell development, *Topographer* constructs a series of gene regulatory networks (GRNs) along the pseudotime, which are directed networks for gene-gene interactions. Unsupervised GRNs are then created by GENIE3[20], which takes advantage of the random forest machine learning algorithm.

Based on the constructed GRNs, *Topographer* further explores the covariation partners of some particular gene (or genes) using a topological network analysis scheme[21]. The method is to identify

the set of those genes that are most closely correlated with a given gene (or genes) of interest and that correlate also most closely with each other, at a given pseudotime point (in fact, a window) (in **Figure 4** of the main text, however, we showed dynamic connections of gene networks at several representative pseudotime points). **Supplementary Information** provides more details of the method.

5. The method for inferring dynamic characteristics of transcriptional bursting dynamics along the pseudotime

Transcriptional bursting kinetics can be characterized by burst size and burst frequency. As is well known, Gamma distribution can well capture this bursty expression in some cases[23]. *Topographe*r uses a Gamma distribution to infer dynamic characteristics of transcriptional bursting kinetics along the cell trajectories identified from single-cell RNA-seq data. Assume that this distribution takes the form

$$p(x)=\frac{x^{a-1}}{b^{a}\Gamma(a)}e^{-\frac{x}{b}}, \tag{13}$$

where x represents the number of transcripts, a is the mean burst frequency (i.e., the mean number of mRNA production bursts per cell cycle) whereas b is the mean burst size (i.e., the average size of the mRNA bursts), and $\Gamma(\cdot)$ is the common Gamma function.

In order to infer dynamic characteristics of transcriptional bursting dynamics along the pseudotime, the key is to estimate two parameters a and b from the dataset at every pseudotime point. For this, *Topographer* makes use of the maximum likelihood method[24]. Since the number of cells at a single pseudotime point would be very few, *Topographer* uses the cell data in a window of this point to obtain more reliable estimations of a and b.

6. Data availability and software

Single-cell data on development of primary human myoblasts can be downloaded from doi:10.1038/nbt.2859[30]. The MATLAB codes used for data analysis and simulations are freely available on request from the corresponding author.

References

25. van Kampen N. G. Stochastic Process in Physics and Chemistry. North-Holland, Amsterdam, 1992.

26. Booth T. E. Power Iteration Method for the Several Largest Eigenvalues and Eigenfunctions. Nucl. Sci. & Eng. J. American Nucl. Soc. 154, (1):48-62 (2006).
27. Rosvall M. & Bergstrom C. T. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA **105**, (4):1118-1123 (2008).
28. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. J. Machin. Learn. Res. 9:2579-2605 (2008).
29. Wolfram Stacklies, Henning Redestig, Kevin Wright. A collection of PCA methods. Bioconductor version 3.6. DOI: 10.18129/B9.bioc.pcaMethods.
30. Trapnell C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotech. 32(4):381-386 (2014).