We present single-cell clustering using bifurcation analysis (SCUBA), a novel computational method for extracting lineage relationships from single-cell gene expression data and modeling the dynamic changes associated with cell differentiation. SCUBA draws techniques from nonlinear dynamics and stochastic differential equation theories, providing a systematic framework for modeling complex processes involving multilineage specifications. By applying SCUBA to analyze two complementary, publicly available datasets we successfully reconstructed the cellular hierarchy during early development of mouse embryos, modeled the dynamic changes in gene expression patterns, and predicted the effects of perturbing key transcriptional regulators on inducing lineage biases. The results were robust with respect to experimental platform differences between RT-PCR and RNA sequencing. We selectively tested our predictions in Nanog mutants and found good agreement between SCUBA predictions and the experimental data. We further extended the utility of SCUBA by developing a method to reconstruct missing temporal-order information from a typical single-cell dataset. Analysis of a hematopoietic dataset suggests that our method is effective for reconstructing gene expression dynamics during human B-cell development. In summary, SCUBA provides a useful single-cell data analysis tool that is well-suited for the investigation of developmental processes.

Significance

Characterization of cellular heterogeneity and hierarchy are important tasks in developmental biology and may help overcome drug resistance in treatment of cancer and other diseases. Single-cell technologies provide a powerful tool for detecting rare cell types and cell-fate transition events, whereas traditional gene expression profiling methods can be used only to measure the average behavior of a cell population. However, the lack of suitable computational methods for single-cell data analysis has become a bottleneck. Here we present a method with the focuses on automatically detecting multilineage transitions and on modeling the associated changes in gene expression patterns. We show that our method is generally applicable and that its applications provide biological insights into developmental processes.
identify bifurcation events directly from single-cell data without prior biological knowledge.

We have successfully applied SCUBA to three distinct data-types: RT-PCR (9), RNA sequencing (RNA-seq) (19), and mass cytometry data (33). Using single-cell RT-PCR data (9), we have correctly identified two bifurcation events during early development of mouse embryos, reconstructed the dynamic landscape of changes in gene expression patterns, and experimentally validated our model by testing its prediction for the effect of Nanog perturbation on cell lineage biases. Analysis of RNA-seq data gave similar results, indicating that our method is robust with respect to experimental platform differences. We have further developed an approach based on principal curve analysis (34) to infer temporal order, thereby extending the applicability of SCUBA to datasets with no temporal information. Taken together, SCUBA provides a useful and robust tool for characterizing cellular heterogeneity and gene expression dynamics from single-cell gene expression data.

Results and Discussion

General Framework of SCUBA. Consider an experimental study in which, to investigate cell differentiation events during development, multiple cells are subjected to single-cell measurements and grouped according to developmental time (Fig. 1, Top), which might be either known a priori or inferred indirectly. Our goal is to automatically identify gene expression patterns associated with cell differentiation from single-cell data. We model the developmental process using a stochastic dynamical system that has the following properties: First, at each developmental time, single-cell gene expression changes are determined by a stochastic dynamical system, containing both deterministic and stochastic components; second, each cell is randomly sampled from the equilibrium distribution of the stochastic dynamical system; and third, the changes of the stochastic dynamical system across time can be parameterized. An immediate consequence is that most cells reside in states that are close to the attractors, whereas only a small number of cells may undergo transitions from one attractor to another. The appearance of multiple new cell types is modeled as a bifurcation process, corresponding to the emergence of new attractors. The major goals of SCUBA are to recover the cellular hierarchy and to quantify the dynamics along the bifuraction events.

Specifically, our method uses a two-step approach, as illustrated in Fig. 1. In the first step, we estimate the locations of the stage-specific attractors and their relationships, using a binary tree model. For simplicity, we only consider steady-state attractors. In the second step, we quantitatively model the dynamics in the reduced state space along each bifurcation direction, using a potential $V(x)$ to characterize gene expression dynamics associated with each bifurcation event (Fig. 1, Bottom). Of note, the parameter space is divided into two regions, corresponding to one or two attractor states, respectively, and their boundary is given by $4a^2 - 27b^2 = 0$. The details are explained in Materials and Methods.

Bifurcation Events During Early Embryonic Development. We first applied SCUBA to analyze a published dataset (9) where the developmental stage for each cell is known. In that study, the authors used high-throughput RT-PCR to quantify the expression levels of 48 selected genes, including 27 key developmental transcription factors, in 438 individual cells isolated from early-stage mouse embryos. Cells were extracted at seven distinct time points, each corresponding to a cell-doubling event, from the 1-cell zygote to the 64-cell blastocyst. There are two well-characterized cell differentiation events during this process (35). The first one occurs at the 32-cell stage, where totipotent cells differentiate into trophectoderm (TE) and inner cell mass (ICM), whereas the second event occurs at the 64-cell stage, where ICM further differentiates into primitive endoderm (PE) and epiblast (EPI). At the end of this period, the embryo contains three distinct cell types: TE, PE, and EPI.

By applying the first step of SCUBA we identified two bifurcation events, at the 32-cell and 64-cell stages, respectively (Fig. 24). The timing of these events matched exactly the occurrence of the aforementioned cell-differentiation events. To test whether our clustering results indeed reflected true lineage differences, we used our results as the basis to predict cell lineages in an independent fluorescently labeled cell population studied in ref. 9. Out of the 37 cells that could be compared in this manner, we found only one misclassification error, indicating that our predictions were highly accurate (Fig. S1) (see SI Materials and Methods for details). To further test the robustness of our clustering results, we simulated and analyzed 1,000 datasets by subsampling the data to test how many cells were needed to reliably detect bifurcations. Whereas the 32-cell bifurcation was detected with as few as 20 cells (Fig. S3), at least 50 cells were required to detect the 64-cell stage with...
We then focused on the local dynamic change of gene expression patterns associated with each bifurcation event. As expected, the overall variance of gene expression increased dramatically during both bifurcation events (see total bar lengths in Fig. 2C). Interestingly, the increase was almost entirely contributed by the bifurcation direction (red portion in bars in Fig. 2C), suggesting that insights can be gained by focusing on the reduced dynamics along the bifurcation directions.

Modeling Dynamic Changes in Gene Expression Patterns Associated with Bifurcations. Next we investigated the gene expression dynamics associated with each bifurcation event by using step 2 of SCUBA. Specifically, we projected the high-dimensional gene expression pattern on the bifurcation direction and then inferred the potential function $V(x)$ by fitting the projected data (see Eq. 3 in Materials and Methods). The fitted parameters values are shown in Table 1. As expected, the potential changed from single-well to double-well for both bifurcations (Fig. 3). Such catastrophic changes are characteristic of multilineage cell-fate transitions.

Prediction of the Effect of Biological Noise on the Maintenance of Lineage Diversity. Our analysis provides a systematic way to evaluate the contributions of deterministic and stochastic forces in establishing cell-fate selection. It is important to note that $4d^2 - 27b^2 > 0$ does not guarantee that the two states after the bifurcation will be clearly distinguishable in the data, because stochastic noise may mask the difference between these two states. Similarly, $4d^2 - 27b^2 < 0$ may not be sufficient to maintain the stability of a cell type, if its stabilizing effect can be countered by noise. Eq. 3 (Materials and Methods) provides a guide to quantitatively assess the balance between the deterministic and stochastic forces. In particular, for cases with small $b$ and approximately symmetric attractors, differences between the two attractors after bifurcation can only be detected when $a > \sigma V/2$. For both bifurcations, $b$ is small and the estimated value of $a$ is so that $a > \sigma V/2$, providing a theoretical explanation for why distinct cell types can be observed at these time points. However, the existence of noise provides a window of opportunity for manipulating cell fates, which may have interesting applications.

To investigate the effect of gene expression noise on the choice of cell fates during differentiation, we compared results from changing the noise level $\sigma$ (see Eq. 3 and Eq. S4 in SI Materials and Methods) to $K_x$. The steady-state distribution $\psi_f(x)$ now becomes $\psi_f(x) = C e^{-W/(\beta K_x)}$. Fig. 4A shows the peaks corresponding to the two attractors at the 32-cell stage become broader as $K$ increases, indicating each attractor state becomes less stable. Also, the areas under the peaks are more similar, indicating that the bias between these two states is reduced. For example, doubling the noise ($K = 2$) would result in an almost even distribution between the two states, whereas reducing the noise by a factor of 2 ($K = 1/2$) would lead to a stronger bias toward the TE lineage. The effect of noise is more dramatic at the 64-cell stage (Fig. 4B), where the potential $V(x)$ is more asymmetric. It is important to note that our calculations represent an upper-bound estimate of the effects of biological noise, because they do not take into account the technical variation in single-cell gene expression measurements. These results point out that noise may play an important role in the maintenance of cell-type diversity.

Table 1. Fitted model parameters for the 32- and 64-cell bifurcations of the RT-PCR dataset

<table>
<thead>
<tr>
<th>Bifurcation</th>
<th>$\sigma$</th>
<th>$b$</th>
<th>$a_{T0}$</th>
<th>$a_{T1}$</th>
<th>$a_{Tb}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>32-cell</td>
<td>75.6</td>
<td>-18.0</td>
<td>-1,156.9</td>
<td>-204.0</td>
<td>232.7</td>
</tr>
<tr>
<td>64-cell</td>
<td>81.8</td>
<td>84.8</td>
<td>-1,003.3</td>
<td>-68.5</td>
<td>205.7</td>
</tr>
</tbody>
</table>
Prediction and Experimental Validation of the Effects of Transcription Factor Expression Levels on Lineage Bias. SCUBA provides a venue to predict the effect of perturbing the expression level of a certain transcription factor on the differentiation process leading to two new cell types. We reasoned that if the perturbation size is sufficiently small its effect could be approximated by the change in the initial conditions without modifying the underlying epigenetic landscape. In a system that contains multiple attractor cell states, changes in initial conditions may alter the final population composition into different cell types. We defined the lineage bias introduced by a transcription factor perturbation as the change induced in the probability of reaching each attractor cell state. To predict the bias resulting from perturbing each transcription factor, we first calculated its effect in changing the initial conditions and then made use of these results for details). As expected, decreasing Nanog expression values (higher Ct) led to a bias toward PE in mutant embryos (Fig. 5E). However, looking at Nanog values provides only a partial explanation, because predictions of a null model based on the Nanog expression levels alone drastically overestimated the effect of the perturbation (Fig. 5E). A likely explanation is that the loss of Nanog was counterbalanced by other factors. To test whether such coordinated effects can be correctly predicted by our SCUBA analysis, we predicted the bias introduced by Nanog perturbation based on the perturbed gene expression dynamics as discussed above (also see Fig. 5C and SI Materials and Methods). This provides a much more accurate prediction (Fig. 5E). The remarkable agreement between our predictions and the experimental results strongly validates our method.

Analysis of Single-Cell RNA-seq Data Shows Robustness of SCUBA. Recent developments in single-cell RNA-seq technologies have enabled whole-transcriptome profiling. To test whether SCUBA is useful for analyzing RNA-seq data we reanalyzed a recently published dataset (19) covering the same time span in early mouse embryo development as the RT-PCR dataset analyzed here (9). The RNA-seq experiments detected a total of 22,958 genes in 294 single cells, but many genes were expressed at a low level and subject to considerable technical variation (19, 44). Therefore, we focused on a subset of genes that were likely to be discriminative, selecting the 1,000 most variable genes that were expressed (>1 reads per kilobase of transcript per million reads mapped) in at least 30% of the cells. SCUBA analysis of this filtered RNA-seq gene signature resulted in a binary tree structure similar to that for the RT-PCR data (Figs. 2C and 6D), both having two bifurcations at the same developmental stages. The slight difference of the timing of the second bifurcation is likely because the RNA-seq dataset also includes some 48-cell embryos, which were not profiled in the RT-PCR dataset.

To do a quantitative comparison we focused on the 32-cell bifurcation, because the other bifurcation was only supported by a small number of cells in the RNA-seq dataset. Among the 1,000 most variable genes, 13 were also present in the RT-PCR dataset.
Of note, the contributions of these genes to the bifurcation axis were remarkably reproducible despite the platform differences ($R^2 = 0.86$; Fig. 6B). In addition, the RNA-seq analysis uncovered additional genes known to be important for either embryonic development [such as Sox15 (45) or Id2/Id3 (46, 47)] or the establishment of tight junctions to form the placenta [such as claudins (48, 49)] that were also associated with high weights (Fig. 6C). We projected the expression profile of the 1,000 genes onto the bifurcation direction and fitted the potential landscape based on Eq. 3 (Materials and Methods and Fig. 6D and E). The resulting landscape had a shape similar to the one obtained for the RT-PCR dataset (Fig. 3B). Taken together, these analyses strongly suggest that SCUBA is also useful for RNA-seq data analysis and the results are robust with respect to experimental platform differences.

Analysis of Human B-cell Differentiation and Comparison with Other Methods. Whereas the bifurcation analysis in SCUBA requires temporal information, it has not escaped our notice that such information may be difficult to obtain experimentally. In some cases, it is feasible to infer the temporal order between the cells by inspecting the expression pattern of known lineage markers. More generally, computational methods [Wanderlust (33) and Monocle (50)] have been recently developed to infer “pseudotime” in silico. Therefore, one strategy is to combine these methods with SCUBA to analyze datasets with no temporal information. In addition, here we present an alternative strategy to infer pseudotime and compare its performance with existing methods.

As an example, we obtained a publicly available single-cell mass cytometry dataset, measuring 18 markers in ∼20,000 cells at
different stages of human B-cell development (33). Cells were extracted from a snap-shot of the bone marrow, therefore bearing no temporal information. The B-cell development is primarily a monocline differentiation process, serving as a new test for SCUBA. We inferred the pseudotime in two steps. First, we used t-SNE (51) to reduce the data into a 3D space. Second, we fitted a smooth curve passing through the reduced data using the principal curved analysis (34) (see SI Materials and Methods for details). Although the resulting curve had no direction, we were able to further distinguish the start and end positions based on the expected change of CD34 expression during hematopoiesis. For each cell, its corresponding pseudotime, called SCUBA pseudotime, was quantified by its relatively mapped position along the principal curve and the values were normalized between 0 and 1 (Fig. 7A). After sorting the cells based on pseudotime, we reconstructed the temporal gene expression profiles during B-cell development (Fig. 7B) and found that the pattern was in good agreement with the literature (52). Specifically, cells had initially high values of CD34, followed by CD38 and CD10, and finally high levels of CD19 and CD20, which are known landmarks of B-cell development.

Compared with two recently published methods, Wanderlust (33) and Monocle (50), our pseudotime inference strategy is conceptually simpler. Also, unlike Wanderlust, it is unnecessary to select an initialization cell, but the principal curve analysis automatically detects the start and end as part of the curve fitting procedure. Of note, the inferred pseudotime was highly correlated with Wanderlust ($R^2 = 0.70$; Fig. 7C). The temporal gene expression patterns inferred from SCUBA and Wanderlust were also similar (compare Fig. 7B and Fig. S6). In contrast, Monocle (50) seemed to have problems analyzing a large number of cells because it failed to run whenever we included more than ~900 cells in the analysis. We tried to overcome this limitation by random subsampling but found the results were highly sensitive to the sampling differences (see Fig. S7 and SI Materials and Methods for details).

Using the pseudotime inferred from SCUBA (or Wanderlust, respectively), we divided the cells into eight equally sized groups ordered by pseudotime and then applied our bifurcation analysis to infer cellular hierarchy. Most of the cells were aligned along a single branch of the binary tree, largely consistent with a monocline differentiation process view of B-cell development. However, analyses of the data ordered with both methods detected a bifurcation event, separating cells into two branches with about one-third and two-thirds of the population, respectively, for the SCUBA analysis (Fig. 7D). Comparison of the signatures of the two branches revealed that cells in the smaller subpopulation had higher IgM (intracellular and especially on the surface) and Kappa (Fig. S8), indicating that a fraction of the cells formed a more mature B-cell subpopulation. These results highlight the utility of SCUBA to detect cell populations with distinct gene signatures.

**Discussion**

We have presented a method, SCUBA, for analyzing single-cell gene expression data. Our method is suitable for the analysis of time-course data sampled with sufficient temporal resolution, and it can detect bifurcations reliably with as few as 20 cells. We have shown that SCUBA is applicable to RT-PCR, RNA-seq, and mass cytometry data and its results are robust with respect to experimental platform differences. SCUBA uses bifurcation theory to focally investigate the dynamic changes of gene expression patterns during development. The major strengths are to automatically detect critical multilineage cell-fate transitions without using prior biological knowledge and to model the gene expression dynamics associated with bifurcation events. SCUBA may also be used to test whether the progression of a developmental process is along a monocline trajectory, however, in that case the second step of SCUBA is not applicable. We have applied SCUBA to analyze...
three different datasets and shown that it provides a useful tool for reconstructing cellular hierarchy and dynamics in complex systems. Through analysis of public datasets we correctly identified two bifurcation events during early development of mouse embryos and quantified the major initiation events during cell differentiation. Our model exquisitely explained the gene expression dynamics around each bifurcation event and predicted the effect of perturbing key regulators in inducing lineage bias. We experimentally tested these predictions by gene expression profiling of Nanog mutant embryos and found excellent agreement between our predictions and the experimental data. Although it requires additional experimental validation, our method also provides a promising framework to systematically evaluate the function of stochastic noise in development. The agreement between RT-PCR and RNA-seq analysis suggests that our method is robust with respect to the experimental platform differences.

One of the limitations of SCUBA is its requirement of data with temporal information for bifurcation analysis. Such information may be difficult to obtain experimentally owing to technical challenges. In certain situations one might be able to infer missing temporal information by applying existing computational methods (33, 50, 53) or the principal curve analysis approach presented here. However, it remains difficult to infer temporal information in general, especially if the cellular hierarchy is complex. During the preparation of this paper we were aware of a recent study (54) that also used Fokker–Planck equations as a model to study the epigenetic landscape during cell reprogramming. In their model, a constant energy function was used to model the entire epigenetic landscape, and cell-fate transition was modeled as moving from one local minimum to another. This is very different from our current approach, where we use a series of energy functions to model the epigenetic landscape. Our strategy is essential here to identify the bifurcation events, where a local change of energy function leads to the emergence of new minima. Comparing these two approaches, SCUBA provides a more natural framework for modeling multilineage differentiations.

A major goal of systematic characterization of cellular heterogeneity is to provide insights into disease processes, which in turn may lead to novel disease-treatment methods. For example, it is well known that each cancer constitutes a highly heterogeneous set of cells and tissues. A fundamental task is to understand the origin of each cell type in tumor genesis and maintenance. In particular, increasing experimental evidence suggests the dominant role of a small set of specialized cells known as cancer stem cells (55, 56). Single-cell gene expression analysis, powered by both technological and computational advances, will likely play an important role in addressing these issues.

Materials and Methods

SCUBA uses a two-step approach, as illustrated in Fig. 1. The mathematical details of the two steps are explained below.

Step 1: Inference of Cellular Hierarchy Using Dynamic Clustering. We infer the cellular hierarchy by iteratively clustering and mapping between cells at the different developmental time points. We assume that the gene expression patterns change smoothly in time, so that a parental cell and its immediate progeny have similar gene expression profiles. During development, a cell may differentiate in a monolineage manner or may differentiate into multiple cell lineages, which we refer to as a bifurcation event. In such an event, we assume that it only gives rise to two new lineages and that the temporal resolution of the data is sufficiently high to capture every bifurcation. Although these assumptions are not universally applicable, they are likely to be valid in many situations and in practice may not be a severe limitation.

At the initial time point we divide cells into clusters with similar gene expression pattern using k-means clustering and use the gap statistic (57) to determine the number of clusters. At each of the following time points, each cell is assigned to a parental cluster based on its gene expression profile. To determine whether a bifurcation event occurs, the progeny of each parental cluster is further divided into two distinct clusters by k-means, and the gap statistic is used to select either the single-cluster or two-cluster model. This procedure is repeated until the final time point. In this way we create a binary tree (Fig. 1, Middle) as an initial estimate of the cellular hierarchy. Of note, if the process only involves monolineage differentiation, then the resulting tree simply has no bifurcations.

Next, we refine the binary tree structure to optimally describe the global gene expression pattern. To this end, we evaluate the performance of each parameterization by using the following penalized likelihood function:

\[ L(\theta) = \log P(x|\theta) - \frac{1}{2} \sum_{i=1}^{n} |x_i - \mu_{i}\theta_{i}|^2, \]

where \( \theta \) indicates all of the parameters involved in defining the tree structure, \( x \) is the observed data, \( \mu_i \) and \( \theta_i \) are the centers of clusters \( c \) and \( a \) (c), respectively, \( a(c) \) is the parent cluster of \( c \), and \( \lambda \) is a predefined constant, set to \( \lambda = 1 \) in this paper. During this refinement process, the overall tree structure might change as some clusters become empty, but it may not create additional bifurcations. Further details and certain generalizations are described in SI Materials and Methods.

Step 2: Modeling Gene Expression Dynamics Using Bifurcation Theory. Our next goal is to model the dynamic changes of gene expression patterns along the cellular hierarchy reconstructed in step 1. We focus on the bifurcation events identified in step 1 and simplify the dynamics to one dimension by projecting the high-dimensional gene expression patterns onto the bifurcation direction, which is defined as the line connecting the centers of the two clusters obtained from a common parental cluster. In the applications discussed in the main text we found that such a dramatic reduction of dimensionality still preserved significant information, allowing us to gain key mechanistic insights about the developmental process.

We begin by considering an idealized scenario where the underlying dynamics is deterministic. In this case, for all initial conditions the system will eventually approach one of the attractor states. Therefore, each observable cell state should correspond to an attractor, and two new cell types arise as a result of a change in the attractor landscape, namely, one attractor loses stability and is replaced by two new attractors. The general bifurcation that can describe the appearance of new attractors in one-dimensional dynamical systems is the two-parameter cusp bifurcation (see the equation in Fig. 1, Bottom and ref. 58), one of the seven irreducible unfoldings according to Thom’s Classification Theorem (25).

Mathematically, a cusp bifurcation is represented by the following first-order ordinary differential equation (ODE) (24, 58):

\[ \frac{dx}{dt} = -x^3 + xa + b \]

with control parameters \( a \) and \( b \). Depending on the values of these parameters, Eq. 2 may have either one or two attractor states (Fig. 1, Bottom). To further take into account the intrinsic stochasticity of gene expression (30, 31, 59), we modify Eq. 2 by addition of a stochastic term and model the ensemble distribution of differentiation trajectories by the Fokker–Planck equation (see details in SI Materials and Methods). The equilibrium distribution is given by ref. 41:

\[ \psi(x) = C e^{-2V(x)}, \]

with \( C \) a normalization constant and \( V(x) \) our potential (see step 2 in Fig. 1). In this form, this potential \( V(x) \) is analogous to the epigenetic landscape schematically described by Waddington (60), represented by a marble rolling down a hill with rugged topology. By fitting Eq. 3 to single-cell gene expression data, the model parameters can be estimated (see details in SI Materials and Methods). Of note, in this step we do not make any assumption about the mechanisms controlling the potential landscape.

ACKNOWLEDGMENTS. We thank Drs. Kimberly Glass, Sidinh Luc, and Junhyong Kim for helpful discussions. G.-C.Y.’s research was supported by National Institutes of Health Grant R21HG006778 and grants from Harvard Stem Cell Institute and the Army Research Office.


Supporting Information

Marco et al. 10.1073/pnas.1408993111

SI Materials and Methods

Mathematical Details of SCUBA.
Additional details for step 1. Inference of cellular hierarchy using dynamic clustering: Refinement of the lineage tree. In the final calculation in step 1 of SCUBA we used a penalized likelihood function to refine the binary tree structure (see Eq. 1 in the main text). The first component \( \log P(x|\theta) = \sum_{i=1}^{N} \log P(x_i|\theta_i) \) can be decomposed into \( T \) terms, one for each time point. In our case \( t = 1, \ldots, T \) corresponds to the 1, 2, ..., 64 cell stages, respectively. The notation \( x_i \) is used for denoting all of the data collected at time \( t \), whereas the parameter \( \theta_i \) includes \( i \) the number of clusters and \( (ii) \) the cluster-specific centers and dispersions. If the cluster-specific dispersions are spherical and identical across clusters, then \( k \)-means clustering is the routinely used method to identify maxima of \( \log P(x|\theta) \). From a Bayesian perspective, the maximization of \( \log P(x|\theta) \), when \( T = 1 \), is equivalent to the maximum a posteriori estimator (MAP) of \( \theta_i \) with a flat prior. The second component of the likelihood function \( \lambda \sum_i \| \mu_i - \mu_{a(i)} \|^2 \) also has a simple Bayesian interpretation. To see this, we notice that, if the vector \( \mu_i - \mu_{a(i)} \) has a prior covariance matrix \( I \) where \( I \) is the identity matrix, then the maximization of \( \log P(x|\theta) - \lambda \sum_i \| \mu_i - \mu_{a(i)} \|^2 \) is equivalent to the MAP estimator. In this case the parameters include not only cluster center locations but also the tree structure.

We assume that within each cluster the gene expression data follow a multivariate normal distribution: \( P(x_i|s_i = c) \sim N(\mu_c, \Sigma) \), where \( s_i \) is the missing data that indicate the cluster from which the \( i \)-th cell is drawn. Eq. 1 in the main text can be expanded as

\[
L(\theta) = -\frac{KN}{2} \log 2\pi - \frac{N}{2} \log \left| \Sigma \right| - \frac{1}{2} \sum_{i=1}^{N} \left( x_i - \mu_i \right)' \Sigma^{-1} \left( x_i - \mu_i \right) - \lambda \sum_i \| \mu_i - \mu_{a(i)} \|^2,
\]

where \( N \) is the number of cells and \( \delta_{ic} \) is the Kronecker delta. We use the following iterative procedure maximizing each component at each time:

i) Update \( s_i \) by minimizing \( (x_i - \mu_i)' \Sigma^{-1} (x_i - \mu_i) \).

ii) Update \( \mu_c \) by setting \( \mu_c = \frac{2x_i \delta_{ic} + \sum_1^{N} \delta_{ic} x_i}{2I + \sum_1^{N} \delta_{ic}} \).

iii) Update \( a(c) \) by minimizing \( \sum_1^{N} \| \mu_i - \mu_{a(i)} \|^2 \), with the constraint that each parent cluster can only have one or two progenies.

iv) Repeat i–iii until convergence.

Convergence is guaranteed here because the likelihood function is increased at each individual step. In general, the procedure only identifies a local maximum, as is the case for \( k \)-means clustering. Notice that the refinement process can change the model parameters as well as the tree structure; for example, certain clusters may become empty after a few iterations, resulting in a truncated tree. Because the refinement process does not involve further partitioning of existing clusters it cannot add new bifurcations. Taken together, the above iterative procedure allows us to identify the optimal partition of the gene expression data into coherent dynamic clusters, whereas each bifurcation captures a cell-differentiation event.

It is possible to generalize our current model by allowing a bifurcation to give rise to more than two cell lineages. For this purpose, we use the gap statistic to determine the optimal number of clusters at each time step without imposing additional constraint. However, this “constraint-free” version of clustering may introduce spurious bifurcations and is not considered in our analysis. Additional details for step 2. Modeling gene expression dynamics using bifurcation theory: Formulation of the mathematical equations. Mathematically, a cusp bifurcation can be represented by the following first-order ODE (1, 2):

\[
\frac{dx}{dt} = -\nabla U(x) = -x^3 + xa + b,
\]

where \( U(x) \) is the potential function

\[
U(x) = \frac{x^4}{4} - \frac{ax^2}{2} - bx + c.
\]

The parameter \( b \) controls the asymmetry of the potential and biologically accounts for any bias toward specific lineages during cell differentiation, a model the dynamic changes during development, and \( c \) is a constant that does not affect the dynamics and is set to zero here. For combinations of parameters such that \( 4a^3 - 27b^2 < 0 \) (green area in step 2 in Fig. 1), Eq. S2 has a single steady-state solution (3), which is the only attractor (see blue marble in step 2 in Fig. 1). However, for \( 4a^3 - 27b^2 > 0 \) (blue area in step 2 in Fig. 1), Eq. S2 has three real roots, corresponding to three steady states, of which one is unstable (red marble in step 2 in Fig. 1) and the other two are stable (purple marbles in step 2 in Fig. 1). If we assume the value of \( b \) does not change between developmental stages, then a bifurcation occurs as \( a \) passes through the critical value, where \( 4a^3 - 27b^2 = 0 \). A special case is when \( b = 0 \), and the system is reduced to the supercritical pitchfork bifurcation, which may occur when the system has the symmetry \( x \rightarrow -x \). (4). In the context of gene expression analyses, gene expression levels are intrinsically stochastic (5–7). Therefore, we modified Eq. S2 by adding a Brownian diffusion term, \( dW(t)/dt \), to incorporate the stochastic deviations. Therefore, our model becomes

\[
\frac{dx}{dt} = -\nabla U(x) + \sigma \frac{dW(t)}{dt}.
\]

The magnitude of diffusion is parameterized by the constant \( \sigma \). In this form, this potential \( U(x) \) is analogous to the epigenetic landscape schematically described by Waddington (8), represented by a marble rolling down a hill with rugged topology. Because each cell is measured only once, it is infeasible to identify the unknown parameters by fitting Eq. S4 directly. Instead, we turned our attention to the distribution of cell-states, \( \psi(t) \), which evolves in time according to the Fokker–Planck equation

\[
\frac{\partial \psi(x,t)}{\partial t} = \frac{\partial}{\partial x} \left( \nabla U(x) \psi(x,t) \right) + \frac{\sigma^2}{2} \frac{\partial^2 \psi(x,t)}{\partial x^2}.
\]

To get the equilibrium distribution, we set \( \partial \psi/\partial t = 0 \), resulting in an equation that can be solved analytically, with the following form (9):

\[
\psi_s(x) = C e^{-2U(x)/\sigma^2} = C e^{-2V(x)},\]

with \( C \) a normalization constant and \( V(x) \equiv U(x)/\sigma^2 \) a rescaled potential (see step 2 in Fig. 1). Assuming that the equilibrium

Marco et al. www.pnas.org/cgi/content/short/1408993111
distribution can be approximated by the observed single-cell data, we estimated the model parameters by fitting Eq. S6 to single-cell gene expression data. The details are described in the next section.

In step 2, the bifurcation direction is used as the basis for dimensional reduction and subsequent dynamical system analysis. Therefore, it can only be used to study bilineage differentiation processes.

Additional details for step 2. Modeling gene expression dynamics using bifurcation theory: Inference of model parameters. The model parameters are inferred by using a maximum likelihood procedure. Specifically, the log-likelihood of a given set of parameters is given by

$$L(z) = \frac{1}{T_n} \log \psi_S(x_t | \sigma, b, a),$$

where $\psi_S$ is given by Eq. S6, with unknown parameters $\sigma, b,$ and $a$ (see Eqs. S3 and S4). In Eq. S7 the sum starts at an undifferentiated state at $t = T_0$ and ends at the bifurcation event at $t = T_n$. $x_t$ are the expression data on the bifurcation direction from the parent clusters until the differentiated state at $t = T_n$, with population means centered at the origin (Fig. 3A and C). For simplicity, we assumed that $a$ is the only parameter that changes between time points, and we used common $\sigma$ and $b$ values for all fitted time points. We used the simplex search method (10) to maximize $L(z)$, as implemented by the fminsearch function in MATLAB.

Validation of Clustering Results by Comparison with Cell-Position Labels. To test whether our clustering results in step 1 of SCUBA indeed reflected true lineage differences, we used our clusters as the basis to predict cell lineages in an independent cell population studied in ref. 11. These authors applied the same procedure and generated an additional dataset containing 134 cells. In addition to the gene expression levels, the location of each cell was labeled by using a fluorescent marker (PKH26). At the 32-cell stage, the cell lineage can be uniquely determined by its location, with the ICM cells located at the inner embryo and the TE cells located at the outer embryo. Therefore, we focused on the 32-cell stage and selected the 37 cells whose locations were unambiguously determined. Using our previously obtained clustering results, we assigned each cell to the closest cluster and evaluated the prediction accuracy by comparing the cluster assignments with the experimentally determined cell lineage. Out of the 37 cells that could be compared in this manner, we found only one misclassification error, indicating that our predictions are highly accurate (Fig. S1).

Bootstrap Analysis of Clustering Results. To test the robustness of our clustering results, we simulated 1,000 datasets by resampling the data using bootstrap (12) and repeated our analysis pipeline for each simulated sample. For each pair of cells we enumerated their co-clustering frequency: A score of 1 indicates they are always assigned to the same cluster, whereas at the 64-cell stage L and R represent the TE and ICM, respectively, whereas at the 64-cell stage L and R represent the EPI and PE, respectively.

We first estimate the probability of a cell differentiating into a certain lineage in normal conditions. For a cell with starting position $x$ on the bifurcation axis, the probability that it ends up at a specific cell state can be estimated by the splitting probability $\Pi_R(x)$ or $\Pi_L(x)$, respectively, which is the probability that a cell first reaches the attractor state at $R$ (or $L$, respectively). The splitting probability is related to the potential $V(x)$ via the following (9):

$$\Pi_R(x) = \int_L^R e^{2V(z)dz}, \quad \Pi_L(x) = \int_L^R e^{2V(z)dz}.\quad [S8]$$

The blue curve in Fig. S4 shows the relationship between $\Pi_R(x)$ and $x$. Under a perturbation where the expression level of one gene is forcedly changed, the initial effect can be modeled as a displacement $\Delta_{gene}$ along the bifurcation axis, and this displacement of this initial condition leads to altered splitting probabilities. Therefore, the lineage bias due to such a perturbation can be estimated by

$$\text{Bias}_{\text{gene}} = \Pi_R(C + \Delta_{gene}) - \Pi_R(C).\quad [S9]$$

Fig. 5B and C show the predicted effect due to a twofold depletion of each assayed splitting factor at the 32-cell and 64-cell stage bifurcation, respectively.

Experimental Procedure for Blastocyst Generation and Single-Blastocyst RT-PCR. A null mutation of the mouse Nanog gene was generated by homologous recombination in embryonic stem cells. The Nanog allele was modified to produce a fusion between Nanog amino acid 60 (Leucine) and the $\beta$-galactosidase (LacZ) reporter gene. NanoglacZ/LacZ homozygotes die shortly after implantation (embryonic day 5.5), whereas NanoglacZ/+ heterozygotes are phenotypically normal.

For single-blastocyst quantitative PCR, total RNA was extracted from individual blastocysts at approximately the 64-cell stage using the PicoPure RNA Isolation Kit (Arcturus Bioscience) and cDNA synthesized at 37 °C for 2 h using the high-capacity cDNA Archive Kit (Applied Biosystems). One-eighth of each cDNA preparation was preamplified for 16 cycles (95 °C for 15 s
Estimates of Cell-Type Composition for Mutant Embryos and Comparison with SCUBA Prediction. To validate our SCUBA predictions for the effects of Nanog perturbations, we first used the whole-embryo dataset to estimate its decomposition into fractions of the different cell types present at the 64-cell stage, TE, PE, and EPI. Next, we calculated the lineage bias in embryos with decreasing Nanog expression and compared with our SCUBA predictions. Finally, we also compared the lineage bias data with a null model that uses only the values of Nanog. The details are described as follows.

**Estimation of the cell-type composition in each embryo.** At the 64-cell stage there are three different cell types present: EPI, PE, and TE. Assuming that their related fractions are \(p_E\), \(p_P\), and \(p_T\), respectively, with \(1=p_E+p_P+p_T\), the expression level for each gene, \(G\), can be decomposed into contributions from the three cell types, \(G_E\), \(G_P\), and \(G_T\), as

\[
G = p_E \cdot G_E + p_P \cdot G_P + p_T \cdot G_T.
\]  

The values of \(G_E\), \(G_P\), and \(G_T\), were estimated from single-cell data by taking the mean value for each cell type. The values of \(p_E\), \(p_P\), and \(p_T\) were then obtained by linear regression, based on the expression levels of all 48 genes. To estimate the lineage bias, we first calculated the fractions of PE cells at the 64-cell stage, \(p_P/(p_P+p_E)\), and then calculated the bias as the change in these fractions with respect to wild-type (taken as the embryo with highest Nanog gene expression). The biases for 25 embryos with decreasing values of Nanog are shown in Fig. 5E.

**SCUBA prediction.** Using the cell (the leftmost point in Fig. 5E) with highest Nanog gene expression as reference, we predicted the lineage bias in each embryo by projecting the observed Nanog expression level onto the bifurcation axis, followed by calculating the bias as in Eq. S9. The predicted effect is shown as the magenta curve in Fig. 5E.

**Prediction using only Nanog values.** As a null model, we also estimated the different cell-type composition for each embryo using only the expression level of Nanog in Eq. S10. Assuming that the Nanog perturbation has no impact on the TE lineage, the change of Nanog expression in the embryo is given by

\[
\Delta \text{Nanog} = \Delta p_E \cdot \text{Nanog}_E + \Delta p_P \cdot \text{Nanog}_P.
\]  

with \(0 = \Delta p_E + \Delta p_P\). We solve Eq. S11 for \(\Delta p_P\) as the null model and used it to calculate the lineage bias as defined above (blue line in Fig. 5E).

---


---

**Method to Infer Pseudotime from Non-Time-Ordered Datasets.** Our method to infer pseudotime from non-time-ordered datasets consisted of two steps. In the first step, we used t-SNE (13) to reduce the data into a three-dimensional space. We used the MATLAB implementation of t-SNE [Matlab Toolbox for Dimensionality Reduction (v0.8.1b)]. In the second step, we fitted a smooth curve passing through the reduced data using the principal curved analysis (14). We used the R package “princurve” version 1.1-12. For each cell, the pseudotime is estimated by taking the corresponding projection index along the principal curve (parameter \(\lambda\)) and rescaling by \([\lambda - \min(\lambda)]/[\max(\lambda) - \min(\lambda)]\).

**Comparison with Wanderlust and Monocle.** Recently, two new methods have been developed to estimate the temporal order of single-cell data from non-time-ordered datasets (15, 16). To compare the performance of each method, we applied each method to analyze a human B-cell development datasets (16). The dataset consisted of 19,486 cells with 17 lineage markers. One of them (IgM) was measured in two cell locations, surface and intracellular, resulting in a set of 18 signatures per cell.

For Wanderlust we used their published pseudotime estimation, which showed a high correlation with our pseudotime estimates (Fig. 7C). For Monocle, with this dataset and markers, analysis of more than 1,000 cells resulted in a run-time error, probably owing to its limited capacity for large-scale data. To overcome this difficulty, we applied Monocle to analyze a randomly selected subsample containing 900 cells. The inferred pseudotime was then extrapolated to all other cells by using \(k\)-nearest neighbor (\(k = 3\)) method. To evaluate the variation due to subsampling, we repeated the analysis three times, each using an independently selected subsample. For each pseudotime reconstruction method we calculated the temporal expression profiles of several signature genes using 100 equally spaced time windows and normalizing it to a maximum value of 1, as in ref. 16.

We applied SCUBA to infer the lineage tree and compared the results using two different estimates of pseudotime, obtained from principal curve analysis and Wanderlust, respectively. Specifically, we sorted cells based on the inferred pseudotime and divided them into eight equally sized groups. We then sequentially constructed the lineage tree using step 1 of SCUBA, treating each group of cells as representing a single time point. In both cases, SCUBA detected two branches, indicating that there is significant cell heterogeneity. The three markers with most significant differences (using Mann–Whitney U test) are shown in Fig. S8.

**Software.** A MATLAB implementation of the SCUBA algorithm is available at github.com/gyeyuan/SCUBA.
Fig. S1. Comparison between the predicted and observed cell lineages at the 32-cell stage. ICM:OBS, inner cell according to fluorescent marker; ICM:PRED, predicted as ICM cell; TE:OBS, outer cell according to fluorescent marker; TE:PRED, predicted as TE cell. X32 and X64 are the bifurcation directions for the 32- and 64-cell stages, respectively.

Fig. S2. Stability of the dynamic clustering step of SCUBA. Step 1 of SCUBA is repeated 1,000 times by bootstrapping. Each pixel in the grid represents the frequency of one specific pair of cells being assigned to the same cluster. The heat map shows that the cells are frequently associated within each of our 10 detected clusters. Only the last two clusters associated with EPI and PE states show a mild mixing.
Fig. S3. Robustness of SCUBA with respect to the sample size. SCUBA is applied to randomly subsampled data. For each sample size, 1,000 subsamples were independently generated. The frequency at which the corresponding bifurcation event is detected is reported in the y axis. Shown are results for (A) the 32-cell stage and (B) the 64-cell stage.

Fig. S4. Comparison between SCUBA and principal component analysis at (A) the 32-cell stage and (B) the 64-cell stage.
Fig. S5. Comparison between SCUBA and SPADE. The SCUBA- and SPADE-derived lineage trees are shown on the left and right, respectively. Each node represents a cluster of cells. The size of each node is proportional to the number of cells, and the color represents the level of Gata4. Note that the PE, EPI, and TE cells are placed in a sequential order in the SPADE analysis, which is incorrect.

Fig. S6. Selected normalized gene expression profiles for cell sorted using Wanderlust pseudotime.
Fig. S7. Comparison between SCUBA and Monocle pseudotime. Monocle was applied to three independent subsamples (A–C), each including 900 cells. For each subsample, the left panel shows selected normalized gene expression profiles for cells sorted using Monocle pseudotime and the right panel shows the density plot for the distribution of SCUBA pseudotimes (x axis) against Monocle pseudotimes (y axis).
Fig. S8. Gene signature of the two branches found by SCUBA in the human B-cell development dataset. Developmental tree depicting two branches (orange and blue) found by SCUBA using SCUBA pseudotime (A) or Wanderlust pseudotime (C). Violin plots show the distribution of values for the most significantly different markers in branch 1 (orange) and branch 2 (blue), analyzed using SCUBA pseudotime (B) or Wanderlust pseudotime (D).

Dataset S1. Gene weights associated with the 32-cell and 64-cell bifurcation directions for the single-cell RT-PCR dataset

Dataset S1