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 $\log P(\boldsymbol{x}|\theta) = \sum_{t=1}^{T}\log P(\boldsymbol{x}_t|\theta_t)$ 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 $\boldsymbol{x}_t$ is used for denoting all of the data collected at time t, whereas the parameter θ_t 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 $\log P(\boldsymbol{x}_t|\theta_t)$. From a Bayesian perspective, the maximization of $\log P(\boldsymbol{x}_t|\theta_t)$, when $T=1$, is equivalent to the maximum a posteriori estimator (MAP) of θ_t with a flat prior. The second component of the likelihood function $\lambda\sum_c\|\boldsymbol{\mu}_c-\boldsymbol{\mu}_{a(c)}\|^2$ also has a simple Bayesian interpretation. To see this, we notice that, if the vector $\boldsymbol{\mu}_c-\boldsymbol{\mu}_{a(c)}$ has a prior covariance matrix λI where I is the identity matrix, then the maximization of $\log P(\boldsymbol{x}|\theta)-\lambda\sum_c\|\boldsymbol{\mu}_c-\boldsymbol{\mu}_{a(c)}\|^2$ 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: $P(\boldsymbol{x}_i\,|\,s_i=c)\sim N(\boldsymbol{\mu}_c,\sum)$, where s_i be 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

$$L(\theta)=-\frac{KN}{2}\log 2\pi-\frac{N}{2}\log\left|\sum\right| -\frac{1}{2}\sum_{i=1}^{N}\sum_{c}\delta_{s_ic}(\boldsymbol{x}_i-\boldsymbol{\mu}_c)'\sum^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_c)-\lambda\sum_{c}\|\boldsymbol{\mu}_c-\boldsymbol{\mu}_{a(c)}\|^2, \qquad \mathbf{[S1]}$$

where N is the number of cells and δ_{s_ic} is the Kronecker delta. We use the following iterative procedure maximizing each component at each time:

i) Update s_i by minimizing $(\boldsymbol{x}_i-\boldsymbol{\mu}_c)'\sum^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_c)$.

ii) Update $\boldsymbol{\mu}_c$ by setting $\boldsymbol{\mu}_c=\frac{2\lambda\boldsymbol{\mu}_{a(c)}+\sum_i\delta_{s_ic}\boldsymbol{x}_i}{2\lambda+\sum_i\delta_{s_ic}}$.

iii) Update $a(c)$ by minimizing $\sum_c\|\boldsymbol{\mu}_c-\boldsymbol{\mu}_{a(c)}\|^2$, 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):

$$\frac{dx}{dt}=-\nabla U(x)=-x^3+xa+b, \qquad \mathbf{[S2]}$$

where $U(x)$ is the potential function

$$U(x)=\frac{x^4}{4}-\frac{ax^2}{2}-bx+c. \qquad \mathbf{[S3]}$$

The parameter b controls the asymmetry of the potential and biologically accounts for any bias toward specific lineages during cell differentiation, a models 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 $b=0$, then the system is reduced to the supercritical pitchfork bifurcation, which may occur when the system has the symmetry $x\to -x$ (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

$$\frac{dx}{dt}=-\nabla U(x)+\sigma\frac{dW(t)}{dt}. \qquad \mathbf{[S4]}$$

The magnitude of diffusion is parameterized by the constant σ. 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, ψ, which evolves in time according to the Fokker–Planck equation

$$\frac{\partial\psi(x,t)}{\partial t}=\frac{\partial}{\partial x}\left(\frac{\partial U(x)}{\partial x}\psi(x,t)\right)+\frac{\sigma^2}{2}\frac{\partial^2\psi(x,t)}{\partial x^2}. \qquad \mathbf{[S5]}$$

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):

$$\psi_S(x)=C\,e^{-2U(x)/\sigma^2}=C\,e^{-2V(x)}, \qquad \mathbf{[S6]}$$

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 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

$$L(\xi) = \sum_{t=T_0}^{T_b} \frac{1}{N_t} \log \psi_S(\boldsymbol{x}_t | \sigma, b, a_t), \qquad \textbf{[S7]}$$

where ψ_S is given by Eq. **S6**, with unknown parameters σ, 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 $t = T_b$. $\boldsymbol{x}_t$ are the expression data on the bifurcation direction from the parent clusters until the differentiated state at $t = T_b$, with population means centered at the origin (Fig. 3 *A* and *C*). For simplicity, we assumed that a_t is the only parameter that changes between time points, and we used common σ and b values for all fitted time points. We used the simplex search method (10) to maximize $L(\xi)$, 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).

Booststrap 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 coclustering frequency: A score of 1 indicates they are always assigned to the same cluster, whereas a score of 0 indicates that they are never in the same cluster. The results in Fig. S2 show clear blocks of values close to 1, indicating that our method is robust.

Analysis of the Effects of Decreasing the Number of Cells in the Detection of Bifurcations. To estimate how many cells are needed for our bifurcation analysis we took a series of subsamples of decreasing size of the RT-PCR dataset and analyzed them with SCUBA. For the 32-cell bifurcation, we excluded all cells at the 64-cell stage to remove any possible confounding effect and gradually down-sampled the cells at the 32-cell stage. For each selected sample size a total of 1,000 subsamples were analyzed. For the 64-cell bifurcation, we gradually down-sampled the cells at the 64 cell-stage and repeated the analysis for each subsample as described above. The results show that for the 32-cell bifurcation, where there are only two different cell types that are clearly separated by the data, subsampling as few as 20 cells still allows detection of the bifurcation event (Fig. S3*A*). However, for the 64-cell bifurcation, which has three final cell types, the detectability of the bifurcation is more compromised (Fig. S3*B*). Fifty cells are needed to detect the bifurcation at least 70% of the time. These results indicate that the minimum number of cells required depends both on the difference between the bifurcating lineages and on the complexity of the lineage structure itself.

SPADE Analysis. The SPADE software (Version 1.0) was downloaded from Peng Qiu's website (pengqiu.gatech.edu) and applied to analyze the data in ref. 11 by using default parameter values. Four genes (Nanog, Id2, Sox2, and Gata4) were selected as markers for down-sampling and tree construction.

Prediction of the Effect of Perturbing Key Regulators on Lineage Bias. To analyze the effect of perturbing the expression level of an individual regulator on the lineage bias, we make use of the information obtained in the two steps of the SCUBA analysis: step 1, the projections of each gene along the bifurcation axis, and step 2, the shape of the potential extracted from the experimental data. At the end of each bifurcation, two new stable cell states emerge, corresponding to the two local minima of the potential $V(x)$. We use L and R to denote the left and right minima of the potential (Fig. 5*A*). At the 32-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 at position x on the bifurcation axis, the probability that it ends up at a specific cell state can be estimated by the splitting probability $\prod_R(x)$[or $\prod_L(x)$, 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):

$$\prod\nolimits_R(x) = \frac{\int_L^x e^{2V(z)}dz}{\int_L^R e^{2V(z)}dz}, \quad \prod\nolimits_L(x) = \frac{\int_x^R e^{2V(z)}dz}{\int_L^R e^{2V(z)}dz}. \qquad \textbf{[S8]}$$

The blue curve in Fig. 5*A* shows the relationship between $\prod_R(x)$ and x. Under a perturbation where the expression level of one gene is forcedly changed, the initial effect can be modeled as a displacement Δ_{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

$$\text{Bias}(gene) = \prod\nolimits_R(C + \Delta_{gene}) - \prod\nolimits_R(C). \qquad \textbf{[S9]}$$

Fig. 5 *B* and *C* show the predicted effect due to a twofold depletion of each assayed transcription 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 β-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 and 60 °C for 4 min) using the TaqMan PreAmp Master Mix Kit (Applied Biosystems) and gene-specific primers. Products were then diluted fivefold for PCR (Applied Biosystems) in 48.48 Dynamic Arrays on a BioMark System (Fluidigm). Threshold cycle (Ct) values were calculated using the system's software (BioMark Real-time PCR Analysis).

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

$$G = p_E * G_E + p_P * G_P + p_T * G_T. \qquad [S10]$$

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. 5*E*.

SCUBA prediction. Using the cell (the leftmost point in Fig. 5*E*) 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. 5*E*.

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

$$\Delta Nanog = \Delta p_E * Nanog_E + \Delta p_P * Nanog_P, \qquad [S11]$$

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

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 λ) and rescaling by [λ − min(λ)]/[max(λ) − min(λ)].

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. 7*C*). 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/gcyuan/SCUBA.

1. Ott E (2002) *Chaos in Dynamical Systems* (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
2. Lu Y-C (1976) *Singularity Theory and an Introduction to Catastrophe Theory* (Springer, New York), p xii.
3. Irving RS (2004) *Integers, Polynomials, and Rings: A Course in Algebra* (Springer, New York), p xv.
4. Guckenheimer J, Holmes P (1997) *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer, New York), p xvi.
5. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297(5584):1183–1186.
6. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31(1):69–73.
7. Raser JM, O'Shea EK (2005) Noise in gene expression: Origins, consequences, and control. *Science* 309(5743):2010–2013.
8. Waddington CH (1959) Canalization of development and genetic assimilation of acquired characters. *Nature* 183(4676):1654–1655.
9. van Kampen NG (2007) *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam), 3rd Ed.
10. Press WH (2007) *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ Press, Cambridge, UK), 3rd Ed, p xxi.
11. Guo G, et al. (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* 18(4):675–685.
12. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7(1):1–26.
13. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605.
14. Hastie T, Stuetzle W (1989) Principal curves. *J Am Stat Assoc* 84(406):502–516.
15. Trapnell C, et al. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32(4):381–386.
16. Bendall SC, et al. (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157(3):714–725.

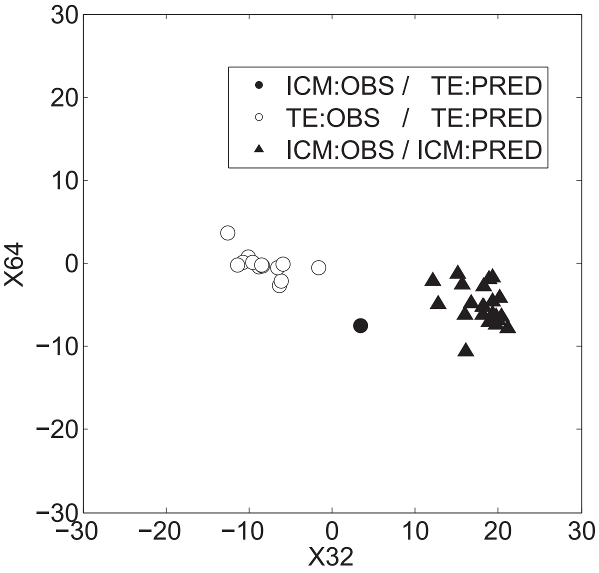

Fig. S1. Comparison between the predicted and observed cell lineages at the 32-cell stage. ICM:OBS, inner cell according to florescent marker; ICM:PRED, predicted as ICM cell; TE:OBS, outer cell according to florescent marker; TE:PRED, predicted as TE cell. X32 and X64 are the bifurcation directions for the 32- and 64-cell stages, respectively.

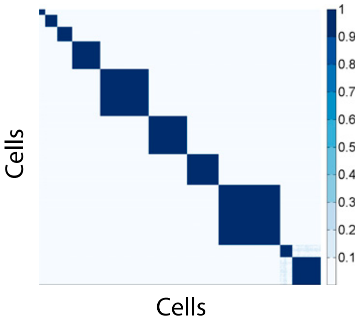

Fig. S2. Stability of the dynamic clustering step of SCUBA. Step 1 of SCUBA is repeated 1,000 times by bootstrapping. Each pixel in the grid represents the frequency of one specific pair of cells being assigned to the same cluster. The heat map shows that the cells are frequently associated within each of our 10 detected clusters. Only the last two clusters associated with EPI and PE states show a mild mixing.

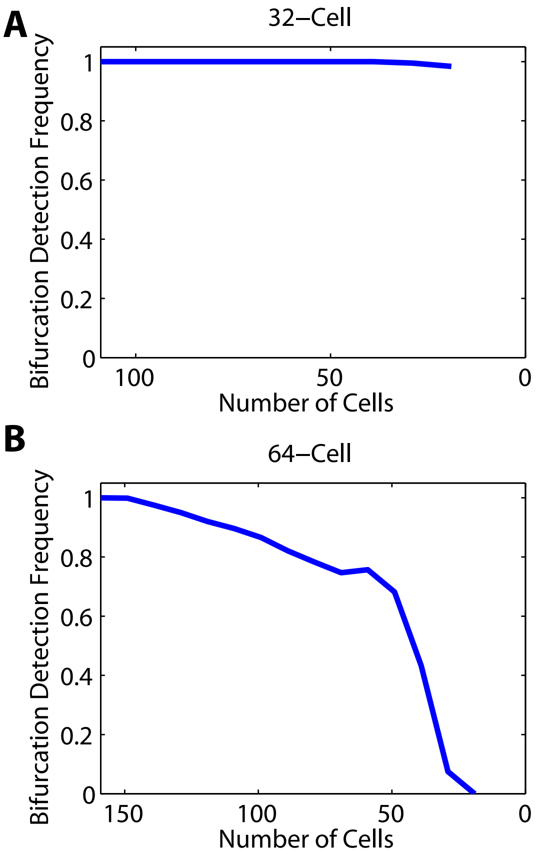

Fig. S3. Robustness of SCUBA with respect to the sample size. SCUBA is applied to randomly subsampled data. For each sample size, 1,000 subsamples were independently generated. The frequency at which the corresponding bifurcation event is detected is reported in the *y* axis. Shown are results for (*A*) the 32-cell stage and (*B*) the 64-cell stage.

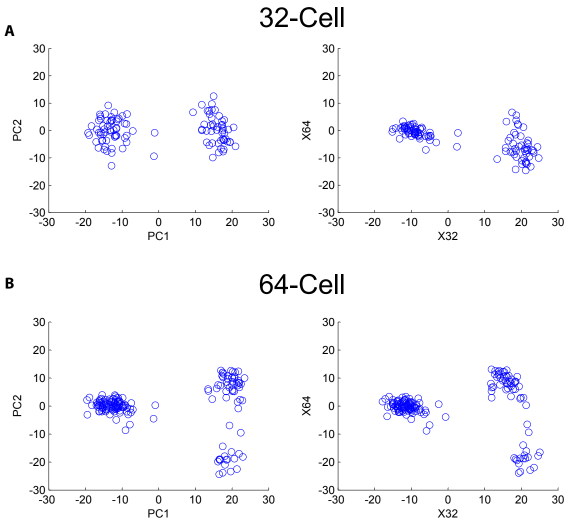

Fig. S4. Comparison between SCUBA and principal component analysis at (*A*) the 32-cell stage and (*B*) the 64-cell stage.

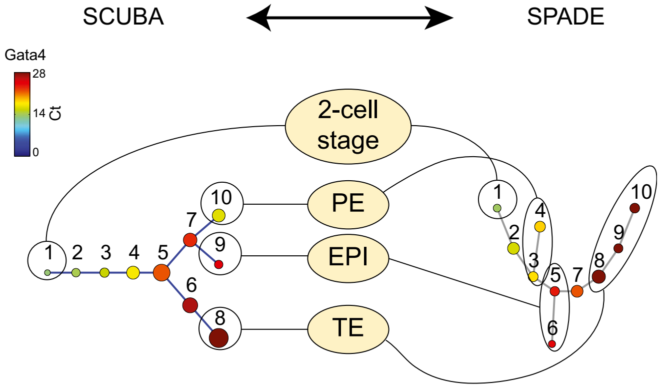

Fig. S5. Comparison between SCUBA and SPADE. The SCUBA- and SPADE-derived lineage trees are shown on the left and right, respectively. Each node represents a cluster of cells. The size of each node is proportional to the number of cells, and the color represents the level of Gata4. Note that the PE,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

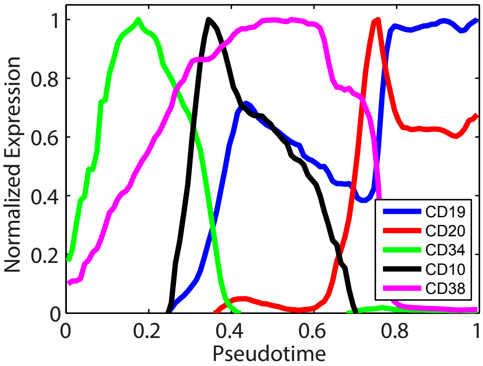

Fig. S6. Selected normalized gene expression profiles for cell sorted using Wanderlust pseudotime.

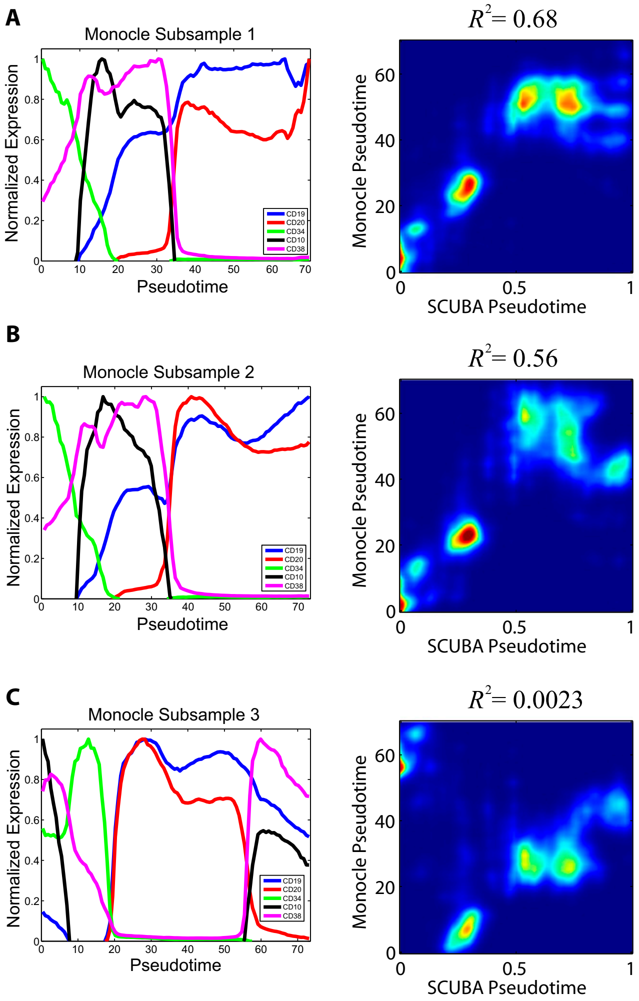

Fig. S7. Comparison between SCUBA and Monocle pseudotime. Monocle was applied to three independent subsamples (*A*–*C*), each including 900 cells. For each subsample, the left panel shows selected normalized gene expression profiles for cells sorted using Monocle pseudotime and the right panel shows the density plot for the distribution of SCUBA pseudotimes (*x* axis) against Monocle pseudotimes (*y* axis).

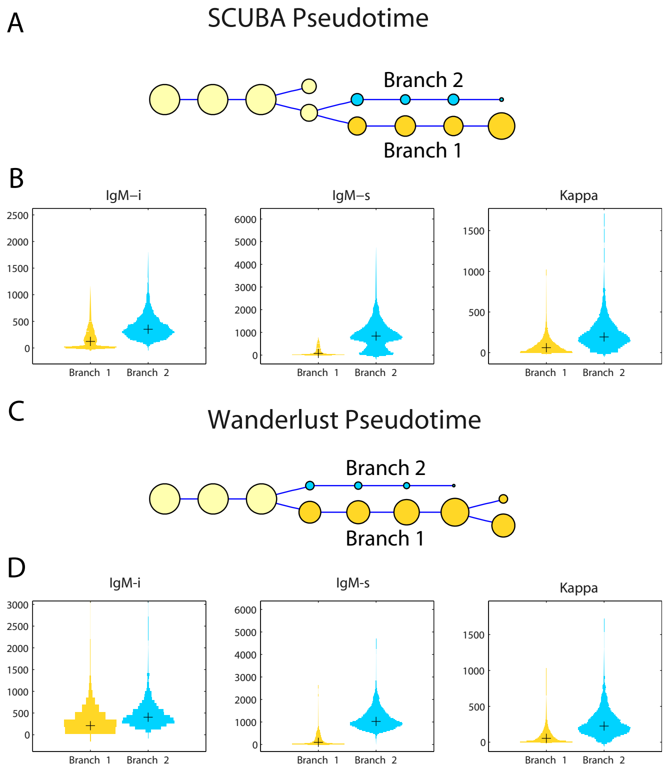

Fig. S8. Gene signature of the two branches found by SCUBA in the human B-cell development dataset. Developmental tree depicting two branches (orange and blue) found by SCUBA using SCUBA pseudotime (*A*) or Wanderlust pseudotime (*C*). Violin plots show the distribution of values for the most significantly 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