Supplementary Data

scEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data

Suoqin Jin[1], Adam L MacLean[1], Tao Peng[1] and Qing Nie[1,2,*]

[1]Department of Mathematics and Center for Complex Biological Systems, and [2]Department of Developmental and Cell Biology, University of California, Irvine, CA 92697, USA.

*To whom correspondence should be addressed (qnie@uci.edu).

Table of Contents

Supplementary Methods

Full details of the scEpath algorithm

scEpath calculates the energy landscape, cell state transition probabilities, pseudotime, and pseudotime-dependent maker genes by the following steps: (i) preprocessing of scRNA-seq data; (ii) construction of a gene-gene interaction network; (iii) calculation of the single cell energy; (iv) energy landscape visualization via principal component analysis and structural clustering; (v) Inference of transition probabilities; (vi) Inference of cell lineage hierarchy via probabilistic directed graph construction; (vii) reconstruction of pseudotime. In addition, scEpath, via downstream analyses, can reveal molecular, dynamical, and functional mechanisms that control or regulate cell fate decisions. We now present full details to describing each of the steps that underlie scEpath algorithm:

Preprocessing of scRNA-seq data. scEpath allows user to easily explore basic quality control (QC) of scRNA-seq data based on any user-defined criteria. First, scEpath filters low-quality cells in which the number of expressed genes is less than 100; Second, scEpath filters genes that are expressed in less than 3 cells; Third, scEpath excludes 'uninteresting' sources such as ERCC spike-ins and Ribosomal genes. At last, the filtered expression matrix X is log2-transformed after adding a pseudocount 1: $\log2(X+1)$. For convenience, we still denote the preprocessed expression matrix by X.

***Construction of a gene-gene interaction network*.** Most subcellular components function through interactions with other subcellular components. These functionally relevant interactions imply that the impact of the perturbations of a specific gene is not restricted to the activity of the gene product, but can spread along the links of the network and alter the activity of other gene products (Barabasi, et al., 2011; Jin, et al., 2014; Jin, et al., 2017). Therefore, an understanding of a gene's network context is essential in determining the phenotypic variability. When trying to infer regulatory interactions between genes one of the most obvious things to look for is correlation in gene expression levels. If there is a strong correlation between two genes, this may indicate that they are highly co-expressed. A network with *n* nodes (genes) is fully specified by its adjacency matrix $A = (a_{ik})$, where a_{ik} takes value 1 or 0 depending on the presence whether nodes *i* and *k* are linked or not. Specifically, for two genes with gene expression profiles across *m* individual cells $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{im}\right)$ and $x_k = \left(x_{k1}, x_{k2}, \ldots, x_{km}\right)$, a_{ik} can be defined by

$$a_{ik} = \begin{cases} 1, & \text{if } \left|cor(x_i, x_k)\right| > \tau \\ 0, & \text{otherwise,} \end{cases} \qquad (1)$$

where τ is the threshold parameter. Thus, two genes are linked (a_{ik} = 1) if the absolute value of Spearman correlation between their expression profiles exceeds the threshold τ. Different threshold will lead to the different number of nodes (genes) in the constructed networks. In this study, we investigate the relationship between the number of nodes, edges, and the threshold τ, and choose the highest threshold without a significant reduction in the total number of genes of the constructed network. We go on to study network properties (e.g. scale-freeness); see below. This approach enables some network pruning (to be more conservative in determining edges) while retaining a large portion of the transcriptome for energy calculations (i.e. reducing information loss). We also systematically explore levels of the threshold to assess its impact (via the resulting networks) on the performance of scEpath and find that within a certain range, the results derived using scEpath are most consistent.

Most biological networks are close to being scale-free, meaning that their node connectivities follow a power law (Albert, 2005), i.e., the probability that a node is connected with k other nodes (the degree distribution $p(k)$ of a network) decays as a power law $p(k) \sim k^{-\gamma}$. To visually inspect the topological properties of the constructed network, we plot log10($p(k)$) versus log10(k). A straight line is indicative of scale-free topology. To measure how well a network satisfies a scale-free topology, previous study proposed a signed version of the model fitting index R^2 (i.e., the square of the correlation between log($p(k)$) and log(k)) (Zhang and Horvath, 2005). R^2 is multiplied by −1 if the slope of the regression line between log($p(k)$) and log(k)) is positive. If R^2 of the model approaches 1, then there is a straight line relationship between log($p(k)$) and log(k)).

***Calculation of single cell energy (scEnergy)*.** Waddington's epigenetic landscape is an abstract metaphor frequently used to describe lineage specification and cell fate decisions (Li, et al., 2016; Moris, et al., 2016). The development of cells from pluripotent to committed states is often compared to balls rolling down hill through several valleys. In the Waddington's epigenetic landscape, a cell, reimagined as a ball, begins at the top of a hill (indicating higher cellular potential energy) and follows existing paths in the landscape into one of several possible fates represented as valleys. Once a cell makes a decision, it is restricted in its subsequent decisions by the route it has taken, representing decreased cellular potential energy and fate restriction. However, the question of

whether such a landscape can be mapped out quantitatively to provide means to infer transition probabilities between cell states and cellular trajectories remains largely unanswered.

To address this question and better understand the relationship between gene expression stochasticity and phenotypic variability, we employed a statistical physics-based approach to quantitatively measure developmental states of single cells by deriving energy landscapes from single cell transcriptome data. The scEnergy of a cell j with the gene expression pattern y was given by

$$E_j(\mathrm{y}) = \sum_{i=1}^{n} E_{ij}(\mathrm{y}) = -\sum_{i=1}^{n} y_{ij} \ln \frac{y_{ij}}{\sum_{k \in N(i)} y_{kj}}, \qquad (2)$$

where y_{ij} represents the normalized expression level (rescaled between 0 and 1) of gene *i* in cell *j* and $N(i)$ is the neighborhood of node *i* (including *i*) in the network. We define $E_{ij}(y) = 0$ when $y_{ij} = 0$. This model shows that not only each cell has an associated value representing its scEnergy E_j but also each gene is assigned a local energy state E_{ij}. The rescaling of the expression is done as follows: $y_{ij} = (x_{ij} - x_{\cdot j}^{\min}) / (x_{\cdot j}^{\max} - x_{\cdot j}^{\min})$ where $x_{\cdot j}^{\min}$ and $x_{\cdot j}^{\max}$ are the minimum and maximum of the expression in cell j across all the genes, respectively. Moreover, we define the normalized scEnergy (taking values between 0 and 1) as

$$\hat{E}_j(\mathrm{y}) = \frac{\left(E_j(\mathrm{y}) / \bar{E}(\mathrm{y})\right)^2}{1 + \left(E_j(\mathrm{y}) / \bar{E}(\mathrm{y})\right)^2}, \qquad (3)$$

where $\bar{E}(y)$ is the average scEnergy across all the cells. (See details in Methods section in main text)

Energy landscape visualization via principal component analysis and structural clustering. Since it is not possible to visualize the potential energy landscape as a function of extremely high dimensional coordinates, it is essential to choose an appropriate set of reaction coordinates that allows to distinguishing among different cell states. To retain global information for the whole gene networks, scEpath then performs Principal Component Analysis (PCA) on the energy matrix $\mathrm{E} = (E_{ij})$ and uses the first two components as the reaction coordinates, leading to two dimensional representations of high dimensional energy landscapes and single cell data.

In addition, scEpath performs structural clustering of cells using an unsupervised framework called single-cell interpretation via multikernel learning (SIMLR), which learns an appropriate cell-to-cell similarity metric from single-cell RNA-seq data (Wang, et al., 2017). How to determine the number of clusters is a challenging problem in the identification of subpopulations from single cell data. scEpath uses gap properties of the eigenvalue spectrum (von Luxburg, 2007) to determine the number of desired clusters by analyzing the Laplacian matrix derived from the cell-to-cell similarity matrix learned by SIMLR. Based on perturbation theory and spectral graph theory, it has been theoretically proven that the number of clusters *N* equals the multiplicity of the eigenvalue 0 of the laplacian matrix of a symmetric weighted matrix (von Luxburg, 2007). Therefore, in the ideal case of *N* completely disconnected clusters, the eigenvalue 0 has multiplicity *N*, and then there is a gap to the (*N* + 1)th eigenvalue λ_{N+1}. More generally, the number of clusters *N* is usually given by the value of *N* that maximizes the eigengap (difference between consecutive eigenvalues). i.e., choose the number *N* such that all eigenvalues $\lambda_1,\ldots,\lambda_N$ are very small, but λ_{N+1} is relatively large. Thus, the optimal number of clusters *N* is determined by

$$N = \max_{i>1} eigengap(i) \triangleq \max_{i>1} (\lambda_{i+1} - \lambda_i), \qquad (4)$$

where λ_i is the ordered *i*-th eigenvalue of the laplacian matrix such that $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$.

Once we obtain the cell clusters, we first removed the outliers of cells in each cluster using the interquartile range (IQR) rule. Specifically, in each cluster, we removed the cells with energy less than $Q_1 - 1.5*(Q_3 - Q_1)$ or greater than $Q_3 + 1.5*(Q_3 - Q_1)$ based on the energy distribution (Q1: first quartile; Q3: third quartile). Then we fit a surface using piecewise linear interpolation, where the x-axis and y-axis correspond to the first two PCA components, and the z-axis is reserved for the calculated energy of each cell.

Inference of transition probabilities. scEpath defines the metacell as the set of cells that occupies θ_1 percent of the total energy in each cluster. In our current analysis, scEpath sets θ_1 to be 80% by default. The probability that a given system will be in metacell *k* with energy E_k^M can be calculated from the Boltzmann–Gibbs distribution:

$$p_k^M = \exp(-E_k^M) / \sum_{j=1}^{N} \exp(-E_j^M), \quad (5)$$

where *N* is the number of metacells. The probability q_k^M that the system leaves this metacell is thus $1 - p_k^M$. Next we assume that the probability of entering a state *l* from state *k* is inversely proportional to the pair-wise distance in reduced dimensional space. The cell adjacency relations based on the pair-wise distance can be quantified by

$$G_{kl} = \exp\left(-\left\|z_k - z_l\right\|_2^2 / (4\varepsilon^2)\right), \quad (6)$$

where z_k is the centroid of metacell *k* in the reduced dimensional space and ε is the maximum of the Euclidean distances among all paired metacell.

Using the row normalization, one can define a transition matrix

$$\tilde{G}_{kl}^{\text{asym}} = G_{kl} / \sum_{j=1}^{N} G_{kj}.$$

Here, $\tilde{G}^{\text{asym}}$ is asymmetrical. All one learns about data using Euclidean distances is about the adjacency relations and not about directions. According to the Markov chains theory (Aldous and Fill, 2002), one can compute a stationary distribution π for $\tilde{G}^{\text{asym}}$ as

$$\pi_k = \sum_{j=1}^{N} G_{kj} / \sum_{k=1}^{N} \sum_{j=1}^{N} G_{kj},$$

which gives $\pi_k \tilde{G}_{kl}^{\text{asym}} = \pi_l \tilde{G}_{lk}^{\text{asym}}$. Then a symmetrical transition matrix $\tilde{G}^{\text{sym}}$ based on the pair-wise distance between metacells is defined as

$$\tilde{G}_{kl}^{\text{sym}} = \pi_k^{1/2} \tilde{G}_{kl}^{\text{asym}} \pi_l^{-1/2}. \quad (7)$$

Combining Eqs. (5) and (7) together, scEpath defines the transition probability T_{kl} between metacell *k* and metacell *l* as follows:

$$T_{kl} = \begin{cases} (1 - p_k^M)\tilde{G}_{kl}^{\text{sym}}, k \neq l \\ p_k^M \qquad\qquad\quad , k = l. \end{cases} \quad (8)$$

Inference of cell lineage hierarchy via probabilistic directed graph construction. To infer cell lineages, scEpath first constructs a probabilistic directed graph in which nodes represent metacells, edges connecting metacells are weighted by transition probabilities between these metacells and directions are determined by the energy flow with significant changes from high to low. scEpath then removes the connections with low transition probabilities by learning a maximum probability flow in the probabilistic directed graph defined by a weighted matrix *W*. This problem is equivalent to find a minimum directed spanning tree (MDST) by setting the edge weights to be 1-*W*. A MDST rooted at *r* is a directed spanning tree rooted at *r* of minimum weights. scEpath

determines the root node (initial state) as the metacell with highest scEnergy. However, one special case is that there is no significant difference between the two metacells with the two highest average scEnergies. In this case, scEpath will need the prior knowledge to distinguishing these two metacells; otherwise the default root node is the metacell with highest scEnergy. scEpath implements Edmonds' algorithm to find a MDST by omitting self-loops of the probabilistic directed graph because we only need to find the shortest path that connects all the matacells (Gibbons, 1985). The resulting trimmed graph is used to present a candidate cell lineages of cellular development (See details in Methods section in main text).

Reconstruction of pseudotime. Once the cell lineage structure has been determined, scEpath reconstructs pseudotime by ordering individual cells along developmental trajectories. scEpath re-orders cells separately for each lineage branch via a principal curve-based approach. In this approach, a smooth one-dimension curve that passes through the middle of data in reduced dimensional space is fit. To reduce the noise effect, scEpath first identifies the core cells in each cluster. The core cell is one that the distance from the centroid of its corresponding metacell is less than θ_2 quantile of the distances of all cells from this centroid in the cluster (θ_2 = 0.75 in our current analyses). Then scEpath uses the R "princurve" package to fit a principal curve of these core cells. To assign each cell a projection index, each cell including the cells that are not core cells is projected onto the principal curve such that the projected points is closest to the cell in an orthogonal sense, leading to a location of each cell on the curve. The projection index is a value that quantifies the distance between the projected point and one end of the curve by calculating the arc length between them on the principal curve. In this way, all cells can be placed in order according to the projection indexes. Once cells are ordered, pseudotime is computed for each ordered path. For a given lineage path, the projection index of a cell on the path is set to be its pseudotime. Then scEpath rescales the pseudotime such that it is bounded in [0, 1].

Identification of pseudotime-dependent genes and key transcription factors. Identification of pseudotime-dependent genes and transcription factors (TF) that are significantly changed as the cells make progress is critical for understanding the molecular mechanisms of state transitions (Wagner, et al., 2016). To model gene expression changes in dependency of pseudotime, scEpath first divides the pseudotime into 10 equally spaced bins. Then the expression of individual gene in each bin is estimated by the trimean of the expressions of this gene across all the cells located into this bin. The density of cells is non-uniform along the lineage path and binning the pseudotime for estimating average expression rather than moving average captures the density differences. Furthermore, scEpath smoothens the average expression of each gene using cubic regression splines, leading to a smoothed version of gene expression along a lineage path. To determine the pseudotime-dependent genes that are significantly changed along pseudotime, we compared the standard deviation of the observed smoothed expressions with a set of similarly permuted expressions by randomly permuting the cell order (1000 permutations in our current analyses). We considered all genes with a standard deviation greater than 0.5 and a Bonferroni-corrected p-value below a significance level $\alpha = 0.01$ to be pseudotime dependent.

Transcriptional regulators can act at different stages, and in different combinations, through the path of cell development and differentiation. TFs are major drivers of cellular identity, which create a unique 'code' for a particular cell fate and, if possible, for cell fate decisions (Moris, et al., 2016). To discover critical transcription factor programs responsible for cell states and state transitions during development, we first collected the TFs that are annotated in the Animal Transcription Factor Database (AnimalTFDB 2.0,

http://bioinfo.life.hust.edu.cn/AnimalTFDB/) (Zhang, et al., 2015) among the identified pseudotime dependent genes. scEpath then declares potentially important TFs for directing cell fate choices if these TFs are differentially expressed between consecutive clusters on one lineage path. Specially, a Bonferroni-corrected p-value from the two-sample t-test of the TF expression is below a significance level 0.01, and fold-change of the average TF expression is greater than a threshold (e.g. log2(fold-change) > 1).

Cellular decision-making is mediated by a complex interplay between external stimuli and the intracellular environment, in particular transcription factor regulatory networks. Therefore, transcription factor regulatory networks are also inferred to study cell-state transitions at the network level. The key TFs derived by scEpath were used to construct a transcriptional network based on Spearman correlations of their expression profiles across all the cells. To infer robust and significant TF associations at a single cell resolution, we first computed the significance of the correlation and used a p-value below 10^{-5} as a cutoff. Two TFs that exhibited a significant correlation were connected by an edge. Then, we subsampled 90% cells for 1,000 iterations. Finally, the significant correlations that survived in all of the 1,000 subsampling were kept as the candidate network.

Discovery of temporal expression patterns and functional signatures. To explore expression patterns and functional signatures of pseudotime-dependent genes, we created a heatmap of pseudotime-dependent genes organized according to a hierarchical clustering of their smoothed gene expression profiles using Pearson's correlation with Ward's method. The heatmap is ordered such that nearest neighbors have similar expression profiles. Genes within each cluster were ordered according to expression peak. The average expression pattern of each cluster is calculated by the trimean of smoothed expressions of all the genes in that cluster. Then functional enrichment analysis of genes in each cluster was performed using the curated GO gene sets from Molecular Signatures Database (MSigDB, http://software.broadinstitute.org/gsea/msigdb/) (Subramanian, et al., 2005). Biological processes enriched with multiple hypothesis testing-corrected p-value (FDR) below 0.01 were considered to be significant. The top five terms of each cluster were shown in this study.

Data sources. The first dataset (GSE36552) consists of 88 cells from human early embryos (HEE) (Yan, et al., 2013). The second dataset (GSE52583) consists of 198 samples from mouse lung epithelial cells (Treutlein, et al., 2014). According to the annotated metadata available from Supplementary Data 5 in Treutlein et al (Treutlein, et al., 2014), all cells annotated as ciliated cells, clara cells or bulk sample were excluded, yielding 155 cells for downstream analysis. The third dataset (GSE52529) consists of 271 cells from human skeletal muscle myoblasts (HSMM), which were collected at 0, 24, 48 and 72h after switching human myoblasts from high-mitogen conditions (GM) to low-serum medium (Trapnell, et al., 2014).

Supplementary Results

Supplementary Note 1: Application of scEpath to a simulated data

We first demonstrate the effectiveness of scEpath using a simulated data. The simulated data contains 40 genes and 250 cells. We simulated four clusters of cells that represent the dynamical changes of developmental process. For the first cluster C1, the first 30 genes have a high expression level 1 and the last 10 genes have a low expression level 0; for the second cluster C2, the first 20 genes have a high expression level 1 and the last 20 genes have a low expression level 0; for the third cluster C3, the first 10 genes have a high expression level 1 and

the remaining 30 genes have a low expression level 0; for the fourth cluster C4, the first 20 genes have a low expression level 0, the next 10 genes have a high expression level 1 and the last 10 genes have a low expression level 0. Then we added a Gaussian noise to the data in each cluster with a standard deviation σ. Therefore, there is a branched lineage underlying this simulated data: one lineage is from C1, through C2, to C3; another lineage is from C1 to C4. To show the performance of scEpath with respect to different noise levels, we simulated the single cell data with three different standard deviations (i.e., $\sigma = 0.2$, $\sigma = 0.4$ and $\sigma = 0.6$). The negative data were set to be zero.

First, we show the performance of scEpath on the simulated data with a standard deviation 0.2. Because there are only 40 genes, so we constructed the gene-gene network by setting the threshold τ to be 0.1. Next, the scEnergies E_{ij} of all these 40 genes and 250 cells were computed. Performing a principal component analysis of the calculated energy matrix E = (E_{ij}) allows a low dimensional representation of high dimensional energy landscapes, showing how the cells transition along the "valley". (Supplementary Figures S1B, D and E). In addition, through performing the unsupervised clustering of the simulated data, scEpath identified four clusters (Supplementary Figures S1B). The number of clusters was determined by maximizing eigengap (Supplementary Figure S1A). Energy landscapes show that the scEnergy consistently decreased along each lineage (Supplementary Figure S1D and E). This result was further confirmed by the two-sided Wilcoxon rank-sum test of the energy distribution of consecutive clusters, suggesting the significantly reduced energies along the developmental process (Supplementary Figure S1C). scEpath further infers the cell lineages by trimming the energy-directed probabilistic graph between clusters. scEpath de novo predicted a branched lineage. One is from C1, the state of which had the highest scEnergy and the smallest probability that the system will be in, through C2, to C3, the state of which had the lowest scEnergy and the largest probability that the system will be in; the other is from C1, the state of which had the highest scEnergy and the smallest probability that the system will be in, to C4, the state of which had the lowest scEnergy and the largest probability that the system will be in (Supplementary Figure S1F). Expectedly, the probability that the system will be in the state C3 or C4 is almost the same. Interestingly, the contour plot of the energy landscape also indicates that there are two unique paths along the "valleys" (Supplementary Figure S1E).

Similarly, we also successfully inferred the branched lineage of the simulated data with a standard deviation 0.4 or 0.6 (Supplementary Figure S1). In addition to the expected two paths, we also observed another "valley" from C1 to C3 in both cases. However, compared to the probability in transition from C1 to C2, the probability in transition from C1 to C3 is much lower (0.42 vs. 0.28 for $\sigma = 0.4$; 0.42 vs. 0.31 for $\sigma = 0.6$). These results indicate that the energy landscape implies potential transitions among clusters, but it is not sufficient for the accurate path. In contrast, the combination of energy landscape and transition probabilities enables an accurate inference of the transitional path.

Supplementary Note 2: Application of scEpath to the mouse lung alveolar type 2 (AT2) data

To demonstrate the effectiveness of scEpath, we studied cells comprising the mouse lung alveolar type 2 (AT2) branch in isolation, to investigate the development of this lineage branch, which enables us to investigate the full life cycle of the alveolar type2 cell lineage during maturation of progenitors. According to the annotated metadata

available from Treutlein et al (Treutlein, et al., 2014), all cells annotated as “AT2” or “Sftpc+” cells were selected for scEpath analysis, yielding 101 individuals cells at four different stages (i.e., E14.5, E16.5, E18.5 and Adult). We investigated the relationship between the number of nodes (genes), and the threshold tau, and chose the highest threshold without a significant reduction in the total number of genes of the constructed network (Supplementary Figure S2). We thus set the threshold τ to be 0.5, leading to a network consisting of 11,226 nodes (genes) (Supplementary Figures S2). Next, the scEnergies E_{ij} of all these 11,226 genes and 101 cells were computed. Then a principal component analysis was performed on the calculated energy matrix E = (E_{ij}), leading to a two dimensional representation of high dimensional energy landscapes (Supplementary Figures S12B). In addition, through performing the unsupervised clustering of the scRNA-seq data, scEpath identified five metacells including one distinct metacell at every stage of E14.5, E16.5 and E18.5 and early and late metacells at adult stage (Supplementary Figures S12A and B). The number of metacells was determined by maximizing eigengap (Supplementary Figure S8B). Energy landscapes show that the scEnergy consistently decreased along the identified metacells from C1 to C5 (Supplementary Figures S12C, D and E). This result was further confirmed by the two-sided Wilcoxon rank-sum test of the energy distribution of consecutive metacells, suggesting the significantly reduced energies along the four developmental stages from E14.5 to adult (Supplementary Figure S12F, inset). Moreover, scEnergy distance--scEnergy plot shows a clear linear trajectory (Supplementary Figure S12F). Notably, from the energy landscape, we observed that some cells are on the barrier between two “wells”, which we called as “edge” cells for example the red cells in domain ①. In addition, energy landscape also indicates that metacell C3 is an intermediate state which are extremely not stable as shown by the green cells in domain ②. Although potential transitions among metacells can be observed in the energy landscapes, scEpath further infers the cell lineages by trimming the energy-directed probabilistic graph between metacells. scEpath *de novo* predicted a linear lineage from metacell C1, the stable state of which had the highest scEnergy and the smallest probability that the system will be in, through metacells C2, C3 and C4, to metacell C5, the stable state of which had the lowest scEnergy and the largest probability that the system will be in (Supplementary Figure S12G).

scEpath next re-ordered the cells along the lineage path, identified pseudotime-dependent genes that are with statistically significant variation in expression levels along the differentiation trajectory and grouped genes with similar trends in expression. This analysis revealed 2,068 pseudotime dependent genes and eight temporal “rolling wave” of transcriptional changes during differentiation (Supplementary Figures S12H and J, Supplementary Figure S13A). Among these pseudotime-dependent genes, the progenitor cell markers such as Sox11 and Sox9 and cell cycle genes such as Foxm1 was downregulated, while the AT2 cell differentiation markers such as Sftpb and Lyz2 was upregulated (Supplementary Figure 14A), indicating the resconstructed pseudotime represents the progression of AT2 cell differentiation. The top 20 pseudotime dependent genes exhibited significantly changed dynamics (Supplementary Figure 14B), suggesting the effectiveness of scEpath in uncovering pseudotime dependent genes. Genes in different clusters were regulated on different time scales: gene clusters I and II are immediately downregulated, then clusters III, IV and V exhibit gradually downregulated at different scales, cluster VI shows transition downregulation, cluster VII shows transition upregulation and cluster VIII is gradually upregulated although a slight downregulation happens at late stage (Supplementary Figure S12J). Combining all the time scales together gives a high-resolution continuous transcriptional spectrum of differentiation. scEpath further discovered pseudotime dependent transcription factors (TFs) responsible for cell

states and state transitions during differentiation. Supplementary Figures S12K and S13B demonstrated that the expression profiles of various transcription factors were also continuous throughout the cell differentiation, suggesting that the differentiation process is accurately regulated by cell state-specific TFs. For example, scEpath revealed several well known regulators such as Sox11, Hmga1-rs and Sox9 that exhibited immediate or gradual downregulation. In addition, scEpath also predicted several TFs (e.g., Hmgb2, Tead2, Sp1) that have not been previously described as relevant for AT2 differentiation. In sum, scEpath enabled the reconstruction of genetic programs during AT2 differentiation at a high resolution.

Supplementary Note 3: Application of scEpath to the mouse lung epithelial specification data

Applying scEpath to all the full dataset (155 epithelial cells), we inferred a gene co-expression network containing 12,293 genes (Supplementary Figures S2). scEpath identified six metacells within this developing epithelial tissue. Based on cluster-specific expression of known cell type marker genes (Figure 3B), we identified three previously reported epithelial cell types: early progenitors (EPs) (Ccnb2 and Cdk1), AT1 (Ager and Pdpn), and AT2 (Sftpb and Scd1).EPs further segregated into two subgroups: one showing highest expression of proliferation markers (early EPs; C1), and the other exhibiting lower expression of these proliferation markers (late EPs; C2). This result indicates the high level of heterogeneity present within the EP compartment. Cluster C3 was characterized by the decreased co-expression of AT1 and AT2 marker genes, suggesting a population of alveolar bipotential progenitors (BPs). Nascent AT2 (cluster C5) was characterized by a decreased expression of AT1 marker or AT2 markers, while mature AT2 (cluster C6) expressed AT2 markers. Thus scEpath enabled the identification of various progenitors and differentiated cell types in the alveolar maturation pathway, which is well in agreement with previous study (Treutlein, et al., 2014).

Furthermore, scEpath inferred AT1 and AT2 lineages emerging from a common BP. Specifically, AT1 lineage is in sequence from early EPs, through late EPs and BP, to AT1; AT2 lineage occurs in a progressive manner from early EPs, through late EPs, BP and nascent AT2,to mature AT2 (Figure 3G).

scEpath re-ordered the cells along the AT1 or AT2 branches, identified pseudotime-dependent genes and then grouped these genes with similar trends in expression. This analysis revealed eight temporal “rolling waves” of transcriptional changes during lung epithelial specification (Figure 3H, Supplementary Figure S15). Three clusters (V, VI and VIII) of genes showed distinct expression kinetics along the AT1 versus AT2 differentiation paths (Figure 3H). Gene cluster V, enriched in cell development and cell proliferation, was immediately upregulated only on the AT1 differentiation path (right panel in Figure 3H). Gene cluster VI, enriched in some common functions with cluster V such as cell development and cell morphogenesis, was gradually downregulated on AT2 path but transiently downregulated on AT1 path. Gene cluster VIII was significantly upregulated only on the AT2 path, and was enriched for lipid metabolic process and immune response. Other than clusters V, VI and VIII, the remaining clusters followed almost identical dynamic trends on both branches. Gene cluster I comprising mostly cell cycle genes was rapidly downregulated and enriched in cell cycle and chromosome organization. Genes from cluster II were gradually downregulated and were largely involved in RNA processing and splicing. Genes from cluster III were downregulated involving in protein localization. Genes from cluster IV were also gradually downregulated involving in metabolic process and oxidative phosphorylation. Supplementary Figure S16 shows the temporal

dynamics of the marker genes and the top 20 pseudotime-dependent genes along AT1 and AT2 lineages respectively, confirming the significantly distinct dynamics of the differentiation from BPs into AT1 lineage and AT2 lineage. Taken together, scEpath revealed common and specific temporal dynamics and function signatures between AT1 and AT2 branches, which gains new insights into the mouse lung epithelial specification.

Moreover, scEpath further uncovered 84 pseudotime-dependent TFs, among which there were 32 TFs important for state transitions during differentiation (Figure 3K and Supplementary Figure S15B). Several regulators (e.g., Runx1, Foxn2, Foxn3, Klf5, Id3) have been shown to play critical roles in differentiation, but no studies reported their roles in mouse lung epithelial specification. For example, previous studies stated that Runx1 promotes proliferation and neuronal differentiation in adult mouse neurosphere cultures and inhibition of Runx1 transcriptional regulation reduces cell proliferation (Logan, et al., 2015). The activity of transcription factors of multiple families including Fox and Klf families all plays important roles in the differentiation of the respiratory epithelium (Ustiyan, et al., 2012).

Supplementary Note 4: Application of scEpath to the human skeletal muscle myoblasts (HSMM) data

scEpath reordered the cells along myoblast differentiation path, allowing us to identifying total 1,116 pseudotime-dependent genes that were dynamically regulated along myoblast differentiation. Supplementary Figure S19 shows the temporal dynamics of the marker genes and the top pseudotime dependent genes along myoblast differentiation, suggesting the significantly changed dynamics. We then grouped these genes with similar trends in expression, leading to five temporal “rolling wave” of transcriptional changes during myoblast differentiation (Figure 4E, Supplementary Figure S20A). Two gene clusters (IV and V), that strongly mark for myoblast differentiation by comparing expression patterns of cluster C3 (containing mesenchymal cells) (Figure 4E), were enriched for muscle development process such as muscle organ development, muscle cell differentiation and muscle structure development (Supplementary Figure S18D). Gene clusters II and III, enriched in extracellular structure organization and tissue development, were highly expressed only in mesenchymal cells while transiently upregulated during myoblast differentiation (Figure 4E). Gene cluster I comprising mostly cell cycle genes was rapidly downregulated and enriched in cell cycle and chromosome organization. Taken together, scEpath revealed the specific-cluster temporal dynamics and dynamical function signatures along myoblast differentiation.

Supplementary Note 5: Comparison of scEpath with existing algorithms

To evaluate the performance of pseudotime/lineage reconstruction, we compared scEpath with four current pseudotime inference algorithms: Monocle 1/Monocle 2 (Qiu, et al., 2017; Trapnell, et al., 2014), TSCAN (Ji and Ji, 2016), and DPT (Haghverdi, et al., 2016) on the four experimental datasets. To evaluate the performance of measuring developmental states, we compared scEnergy with three entropy-based measures: signalling entropy (SCENT) (Teschendorff and Enver, 2017), SLICE (Guo, et al., 2017) and StemID (Grun, et al., 2016). Among pseudotime inference methods, TSCAN and Monocle 2 can automatically infer the number of branches, but Monocle 1 and DPT need user to specify the number of branches. There were two versions of Monocle, i.e., Monocle 1, a classical and popular method which employed Independent Component Analysis (ICA) to perform

dimension reduction (Trapnell, et al., 2014), and Monocle 2, a very recently released method which used reverse graph embedding, producing more accurate, robust trajectories than its previous version (Qiu, et al., 2017). TSCAN used a cluster-based minimum spanning tree (MST) approach to order cells, which has been demonstrated to be robust with respect to the selection of genes (Ji and Ji, 2016). DPT robustly estimated cell order according to diffusion pseudotime (DPT), which measures transitions between cells using diffusion-like random walks (Haghverdi, et al., 2016). Among the entropy-based measures, SCENT estimates differentiation states of a single cell by computing the signalling entropy of a cell's transcriptome in the context of an interaction network (Banerji, et al., 2013; Teschendorff and Enver, 2017); SLICE calculates the entropy of single cells based on their probability distributions of functional activation (Guo, et al., 2017); and StemID calculates the transcriptome entropy of single cells based on their probability distributions of the number of transcripts of each gene in each cell (Grun, et al., 2016). Both scEnergy and signalling entropy make use of network information, while SLICE and StemID do not.

5.1 Comparison of scEpath with existing pseudotime inference algorithms

To compare pseudotime inference algorithms, we evaluated the following three aspects. First, we calculated the similarity by pseudotime reconstruction score (PRS) between the inferred pesutotemporal ordering with the true time ordering (i.e., experimental stage/time), to access accuracy. Second, we evaluated the robustness of pesutotemporal ordering. In details, we used two measures, the PRS and the Pearson correlation coefficient, to calculate the similarity of each pair of pseudotemporal ordering runs under repeated subsampling of 90, 80, or 70% of the total number of cells in each dataset. The similarity of two runs was calculated on the intersection (i.e., common cells in two subsampled sets). In each scenario, the original data was independently subsampled 50 times, resulting in a 50 by 50 (upper triangular) similarity matrix for each method, which we call robustness analysis. Third, we qualitatively evaluated the performance of lineage reconstruction regarding the changes of the input gene set.

In the main text, we quantitatively compared the similarity with true time ordering and the robustness of pesutotemporal ordering. Generally, robustness alone is not sufficient to indicate good performance of pseudotemporal odering. Robustness of a method should be interpreted in the context whether it can improve cell ordering accuracy (e.g., increased PRS). Our results showed that that scEpath produces cellular trajectories that are – in combination – more accurate and robust than state-of-art methods including Monocle1/2, TSCAN, and DPT. In contrast, other methods were not able to simultaneously attain accuracy and robustness. Next we will show that scEpath produced more robust results when varying the number of selected genes for lineage reconstruction. To demonstrate this, we varied the threshold for network construction, leading to different number of genes for lineage inference (Supplementary Figure S2). For the HEE and AT2 data, all the methods can successfully reconstruct the linear development trajectory (Supplementary Figures S4 and S5). For LES data, all the methods except for DPT can exactly separate AT1 and AT2 branches (Supplementary Figure S6). DPT only succeeded when the number of genes was 967. For the HSMM data, when the number of genes was 14968, only scEpath can isolate mesenchymal cells from myoblast lineage (Supplementary Figure S7). When the number of genes was 8128, scEpath, TSCAN and DPT succeeded. However, when we used 12 marker genes of myoblast differentiation from Monocle package, Monocle 1/2 and scEpath can detect the important split between mesenchymal state and myoblast differentiation. TSCAN cannot isolate mesenchymal cells from myobalst

differentiation and DPT cannot isolate mesenchymal cells from proliferation cells. These results indicate that scEpath can robustly identify the developmental trajectories without extensive screening of genes, while other existing methods were not robust to the variation of the size of input gene set, suggesting that they may need to use the high variable genes (informative genes) for lineage inference. It should be mentioned that in most of the cases the biological ground truth of the cell lineages is not known, so it may be not sufficient to indicate good performance by comparing methods as to whether they can detect branching events. Here we showed that scEpath can robustly separate the biological distinct cell populations (e.g., AT1 cells and AT2 cells, muscle cells and mysenchymal cells) into different trajectories without feature selection.

5.2 Detailed implementation of existing algorithms

For the comparison compatibility, we use the same set of genes as an input of all the algorithms. Details on the implementation of algorithms used in this study are as follows.

(i) Robustness analysis of pseudotemporal ordering to the subsampling dataset with Monocle1/2, TSCAN, DPT and scEpath. For Monocle 1/2, we first convert relative expression values (e.g., FPKM) to transcript counts using “relative2abs” function, and then select the genes for dimension reduction and ordering (estimateSizeFactors,estimateDispersions, dispersionTable and subset functions in Monocle). Parameters (mean_expression and dispersion_empirical in subset function) are properly selected for each dataset such that the number of ordering genes is about 2000 and the original dataset can correctly infer the cell lineages. Finally, we run orderCells function with default parameters. But for Monocle 1, we set num paths = 1 (HEE data and AT2 data) and num paths = 2 (LES data and HSMM data) in the orderCells function. For TSCAN, we use the log-transformed expression value, set the parameter “clusternum” to be 5 percent of number of all genes in the preprocess function and the parameter clusternum = 3:9 in the exprmclust function, and then run TSCAN with default parameters. For DPT, we downloaded the DPT software in Matlab version from Supplementary Software of its original paper (Haghverdi, et al., 2016). We also use the log-transformed expression value and then run DPT with default parameters except for the parameters “branching” and “root”. For HEE data and AT2 data, we set branching = 0; for LES data and HSMM data, we set branching = 1 and properly select a root cell to ensure pseudotime calculation for all subsampling datasets starts from the same cell. For scEpath, we run it by default parameters, use the same gene-gene interaction network as the original dataset and set the number of clusters to be the same as the original dataset. For LES data, cells were ordered along separate branches with TSCAN and scEpath. We calculated the average similarity (measured by Pseudotemporal Ordering Score or Pearson’s correlation) of these two branches as the final value. For HSMM data, we calculated the similarity of pairwise runs by eliminating the branch involving mesenchymal cells for TSCAN and scEpath, or eliminating the mesenchymal cells for Monocle1/2 and DPT.

(ii) Robustness analysis of lineage inference to the variation in the size of the input gene set with Monocle1/2, TSCAN, and DPT. For Monocle 1/2, we use log-transformed expression value and use the input gene set for dimension reduction and ordering. For TSCAN, we use the log-transformed expression value and set the parameter “clusternum” to NULL in the preprocess function. For DPT, we run it as the previous one.

(iii) Comparison with entropy measures with SCENT (Teschendorff and Enver, 2017), SLICE (Guo, et al., 2017) and StemID (Grun, et al., 2016). For SCENT, we use the log transformed expression value and the interaction network we constructed (for HEE, AT2 and LES data) or network information provided in this package (hprdAsigH-13Jun12.Rd (https://github.com/aet21/SCENT) for HSMM data), and then run SCENT with default parameters. The entropy of each cell is computed using CompSRana function in SCENT. For SLICE, we use the raw expression value and then run SLICE with default parameters. The entropy of each cell is computed using getEntropy function in SLICE. For StemID, we download the StemID2 (https://github.com/dgrun/RaceID3_StemID2). We use the raw expression value and then run StemID with default parameters. The entropy of each cell is computed using compentropy function in StemID2.

Supplementary Note 6: Robustness analysis of the parameters in scEpath algorithm

We tested the robustness of scEpath on the four data used in this study by varying the following parameters:

Parameter	Description	Default
τ	Threshold for constructing gene-gene interaction network	
θ_1	Proportion of cells included in the metacell	0.8
θ_2	Proportion of cells used to fit the principal curve	0.75
nPC	Number of principal components used in the calculation of transition probability	

We first investigate the robustness of scEpath to different choices of τ. In this study, we investigate the relationship between the number of nodes (genes), edges, and the threshold τ, and choose the highest threshold without a significant reduction in the total number of genes of the constructed network. Different threshold will lead to different number of genes for downstream analysis. In the Supplementary Note 5, we have shown that scEpath produced more robust results when varying the number of selected genes (i.e., varying parameter τ) for lineage reconstruction. Here we used the Pearson correlation to quantify the similarity of estimated scEnergy of single cells among different runs under different choices of the threshold τ. Supplementary Figure S3 shows the comparisons of pairwise runs. For AT2 data and LES data, these runs are highly correlated (above 0.95 in all cases). For HEE data, most of the correlations are above 0.9 except for three cases (0.80, 0.83 and 0.88). For HSMM data, these runs also show high correlations (above 0.9) when the number of genes in the two runs are in the same scale. If the number of genes in the two runs are far away from each other, for example τ = 0.2 (#21022 genes) and τ = 0.4 (#8128 genes), they exhibited a relative lower correlation r = 0.79. Taken together, these results indicate that within a certain range of the threshold τ, the overall results (e.g. scEnergies and cell lineages) derived using scEpath were most consistent and robust.

Next, we investigated the robustness of scEpath to different choices of θ_1 and θ_2. These two parameters are related to the lineage inference and pseudotime reconstruction in scEpath algorithm. θ_1 and θ_2 define thresholds for the energy measure and the energy distance (here energy distance means the distance from each cell to its

cluster center in low dimensional space), respectively. The motivation behind both of these is to exclude significant outliers in a cluster from its corresponding metacell. Within a cluster, cells are ordered both by scEnergy and scEnergy distance. Cells are then selected such that: i) the sum of the scEnergy of the selected cells is θ_1 (default 80%) of the total scEnergy for that cluster; ii) cells lie within θ_2 (default 75%, the third quartile of the distance distribution) of the farthest distance from a cell to the cluster center in low dimensional (similarity) space. We varied θ_1 from 0.7 to 1, and θ_2 from 0.55 to 0.85 simultaneously. For all these runs in the four data, we can identify the correct lineage paths (Supplementary Figures 25-29). To quantify the similarity of pseudotime between different runs, we calculated the Kendall rank correlation, which is defined as

$$R = \frac{c - c'}{m(m-1)/2}, \quad (9)$$

where c is the number of concordant pair of cells, c' is the number of disconcordant pair cells and m is the total number of cells in one lineage path. The denominator is the total number of pair combinations, so the coefficient must be in the range [-1 1]. If the agreement between the two pseudotime is perfect (i.e., the order of cells are the same) the coefficient has value 1. If the disagreement between the two pseudotime is perfect (i.e., one order is the reverse of the other) the coefficient has value −1. If the two pseudotime is independent, then we would expect the coefficient to be approximately zero. Supplementary Figures 25-29 demonstrated that the pseudotime is robust across the different parameter settings. For HEE data, Kendall rank correlations of these runs are above 0.98. For AT2 data, these runs are highly correlated (above 0.95 in all cases and 0.97 when compared to the default parameter). For LES data, extremely high correlations were observed (above 0.98 for the AT1 path and 0.99 for the AT2 path). For HSMM data, these runs also show high correlations (above 0.96 for most cases, 0.88 for some cases) for the muscle path.

Finally, we investigated the robustness of scEpath to the number of principle components in the calculation of transition probability and the inference of cell lineages. By default, we calculated the transition probability based on reduced dimensional space given by the first two components. We calculated the standard deviations of the first 20 components and found that the standard deviations of the first two components were significant larger than the remaining components in the four experimental datasets we studied (Supplementary Figure S33A). However, the top two components may be not enough for complex scRNA-seq datasets in reality. Thus, we also adaptively selected the number of significant components using a method of inferring the optimal hard threshold for singular values (Gavish and Donoho, 2014). Significant components are selected such that their singular values are above the optimal threshold. We found that they yield the same cell lineages in the four datasets (Supplementary Figure S33B). Moreover, the correlations of transition matrix between the “optimal” one and the original one (using the first two components) are larger than 0.96 in all the four datasets. In details, the correlation is 0.999, 0.998, 0.999 and 0.962 in HEE, AT2, LES and HSMM data, respectively.

Supplementary Note 7: Performance of scEpath without network information or without scEnergy

We estimated the energy of single cells (scEnergy) using two inputs: the expression of gene i in cell j and the expressions of all the neighboring genes of gene i in cell j. The method does not consider genes in isolation, but estimates scEnergy in the context of a large gene network, making the resulting inferences much less prone to

variation due to perturbations (e.g. gene dropout and the changes in the input gene sets), as seen in our robustness analysis (Supplementary Note 5). If we ignored the denominator (i.e., the network information) and defined the scEnergy as simple as $E_j(y) = -\sum_{i=1}^{n} y_{ij} \ln y_{ij}$, we found that this simplified measure shows a clear correlation with the original measure that we use, but does not performs as well as our original definition at discriminating between developmental states (Supplementary Figure S32). For example, application to HEE data, this simplified measure cannot discriminate cells in 8-cell stage from cells in Morulae or 4-cell stage (a two tailed Wilcoxon rank sum test: p-value = 0.39 for the comparison with Morulae; p-value = 0.12 for the comparison with 4-cell stage); on the contrast, our original definition can significantly discriminate them (p-value = 0.01 and p-value = 0.004, respectively). Application to AT2 data, the significance level (P-value) of the case without network information increases compared to the case with network information. Particularly, there is a significant difference between C3 and C4 in the case with network information, while no significance was observed in the case without network information. Intuitively, the term involving the gene-gene network used in the denominator provides a gene's non-local information, which may provide a more robust prediction in general. To test this we have compared the similarity of pseudotemporal ordering under repeated subsampling of cells, to assess robustness. We used the Pseodotemporal Ordering Score (PRS) to calculate the similarity of each pair of pseudotemporal ordering runs under repeated subsampling of 90, 80, or 70% of the total number of cells. Compared to the case without network information (i.e., this simplified definition), we found that scEpath with network information produced more robust pseudotime calculation under repeated subsampling of cells from HEE data and AT2 data (Supplementary Figure S32). Moreover, compared to the current pseudotemporal ordering methods (e.g., Monocle1/2, TSCAN and DPT) which we note do not use any gene-gene network information, we found scEpath produced cellular trajectories more robust in most cases, as seen in our simulations (Figure 5 and Supplementary Figure S23). Taken together, integration of gene expression with a gene network improves the prediction power (including discriminative ability of developmental states and pseudotemporal ordering) of scEpath.

Inference of the transition probability between cell states is important in the analysis of single cell data. In this study, we estimated the transition probability by combining scEnergy (the energy state of metacells) with scEnergy distance (the distance from one metacell to another) based on the statistical physics. The symmetric distance matrix (e.g., Euclidean distance) does not permit asymmetric transitions. We expect such asymmetric transitions both from biological knowledge (stem cell differentiation is very different from reprogramming), and also from our motivation based on the Boltzmann-Gibbs distribution, from which the probability p of a cell remaining in its current state is estimated. Given p, the probability of a cell exiting the state is (1-p): this definition leads to an asymmetrical transition probability between states. It is plausible (even probable) that in most cases transition probabilities are asymmetric. Thus we argue that the transition probability derived in scEpath is more suitable for quantifying cell state transitions; we use it to estimate a weight for the probabilistic directed graph connecting metacells. However, it should be noted that, based on the definition of transition probability, the inverse of Euclidean distance has a high correlation with the transition probability and the two tree construction methods may yield similar cell lineages. To formally show this, we inferred cell lineages using the Euclidean distance of pairs of metacell centroids as a weight of the edge, and observed the same cell lineages as our original ones in the four datasets we studied (Supplementary Figure S34).

Supplementary Note 8: Performance of scEnergy in the discrimination between pluripotent/multipotent cells and the less potent cells

Given the observed high correlation of scEpath with other entropy measures (including signalling entropy, SLICE and StemID) and developmental stages (Supplementary Figure S30), we further attempted to test the discrimination power of scEnergy on pluripotent or multipotent cells versus the less potent cells in various scRNA-seq datasets.

First, to test the discrimination power of pluripotent versus non-pluripotent cells, we estimated scEnergy for 1,018 single-cell RNA-seq cells (Chu, et al., 2016), including pluripotent human embryonic stem cells (hESCs, n = 374, including 212 cells from H1 cell line and 162 cells from H9 cell line) and non-pluripotent cells (non-Pluri, n = 644, including 173 neural progenitor cells, 138 definite endoderm cells, 105 endothelial cells, 69 trophoblast-like cells and 159 human foreskin fibroblasts). We found that scEnergy of pluripotent hESCs was significant higher than the non-pluripotent cells (one tailed Wilcoxon rank sum test P = 1e-69) (Supplementary Figure S31A).

Second, to test the discriminative power of totipotent versus pluripotent cells, we used the single cell data from human early embryos development (Yan, et al., 2013). scEnergy provided a significant discriminator of totipotency (44 cells including Oocyte, Zygote,2-cell, 4-celland 8-cell) versus pluripotency (12 inner cell mass (ICM) cells of Late blastocyst) (P = 5e-7, Supplementary Figure S31B).

Third, to further test the discrimination power of scEnergy on multipotent versus less potent cells, we computed it for scRNA-seq profiles of 117 cells capturing haematopoietic stem cells (HSC) formation during mouse embryogenesis (GSE67120) (Zhou, et al., 2016), including the haematopoietic stem cells (HSC, n = 37), the precursors of HSC (pre-HSC, n = 50) and adult HSC (n = 30). Comparative results showed that scEnergy can significantly discriminate these populations (Supplementary Figure S31C).

Finally, to further test the general validity of scEnergy, we studied the hepatocyte-like lineage progression from the induced pluripotent stem (iPS) cells in two-dimensional culture during human liver development (GSE81252) (Camp, et al., 2017). We calculated the scEnergy on 425 single-cell transcriptomes from multiple time points during differentiation from iPS cells (day 0, n = 80) to DE (definitive endoderm, day 6, n = 70), HE (hepatic endoderm, day 8, n = 113), IH (immature hepatoblast-like, day 14, n =81) and MH (mature hepatocyte-like, day 21, n = 81). We found that scEnergy can significantly discriminate any pairs of these cell types and closely correlated with the hepatic cell differentiation from iPS cells (Supplementary Figure S31D, left panel). Moreover, scEnergy can significantly discriminate iPS cells (n = 80) from all other (non-pluripotent, n = 345) cells (P = 5e-42, Supplementary Figure S31D, right panel).

Taken together, these results indicated that scEnergy exhibits a discriminative ability of pluripotent/multipotent cells versus the less potent cells and can be as a measure of developmental states.

Supplementary Figures

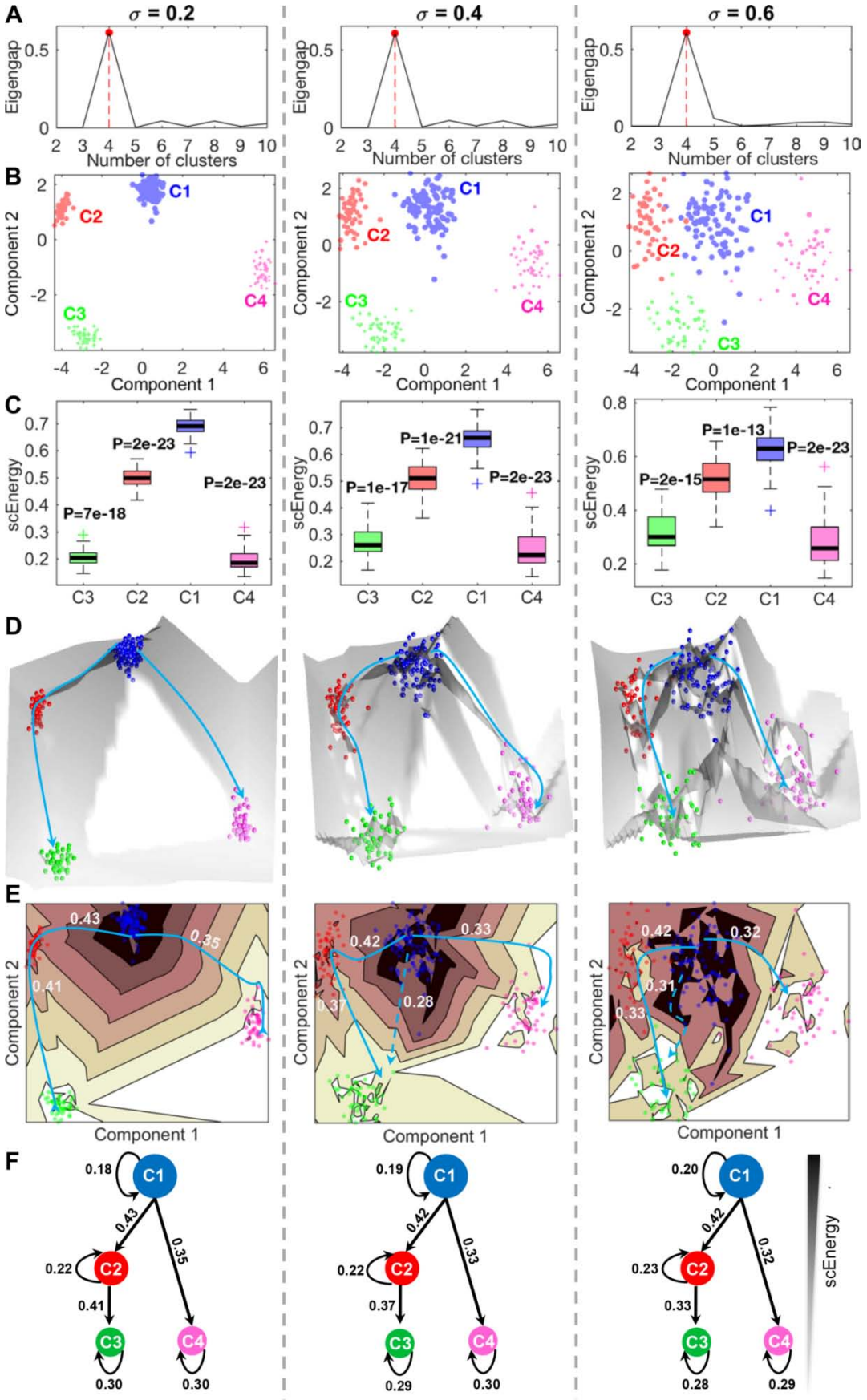

Figure S1. scEpath reconstructed the branched lineage on the simulated data with different standard deviations (left panel: σ = 0.2; middle panel: σ = 0.4; right panel: σ = 0.6). (A) Eigengap of recommending the best choice of the number of clusters with SIMLR. The higher the eigengap is, the more likely the number is close to the optimal number of clusters. The red dot and dashed line indicate the optimal number of clusters for each data set. (B) Energy landscape visualized on the first two components. Cells are colored according to the unsupervised clustering. Sizes of cells were proportional to the scEnergies of the cells. (C) Comparison of energy distribution among the identified cell clusters, suggesting a significant reduced energy during development. P-value from a two-sided Wilcoxon rank-sum test is indicated. (D) Overall view of energy landscape in three dimensions. The developmental trajectories are shown by a curve. (E) Contour plot of the energy landscape. The transition path is highlighted by a solid blue line. The dashed blue lines indicate two possible paths along the "valeys". White numbers represent the transition probability between two clusters. (F) scEpath revealed a branched lineage in which transition rates are shown.

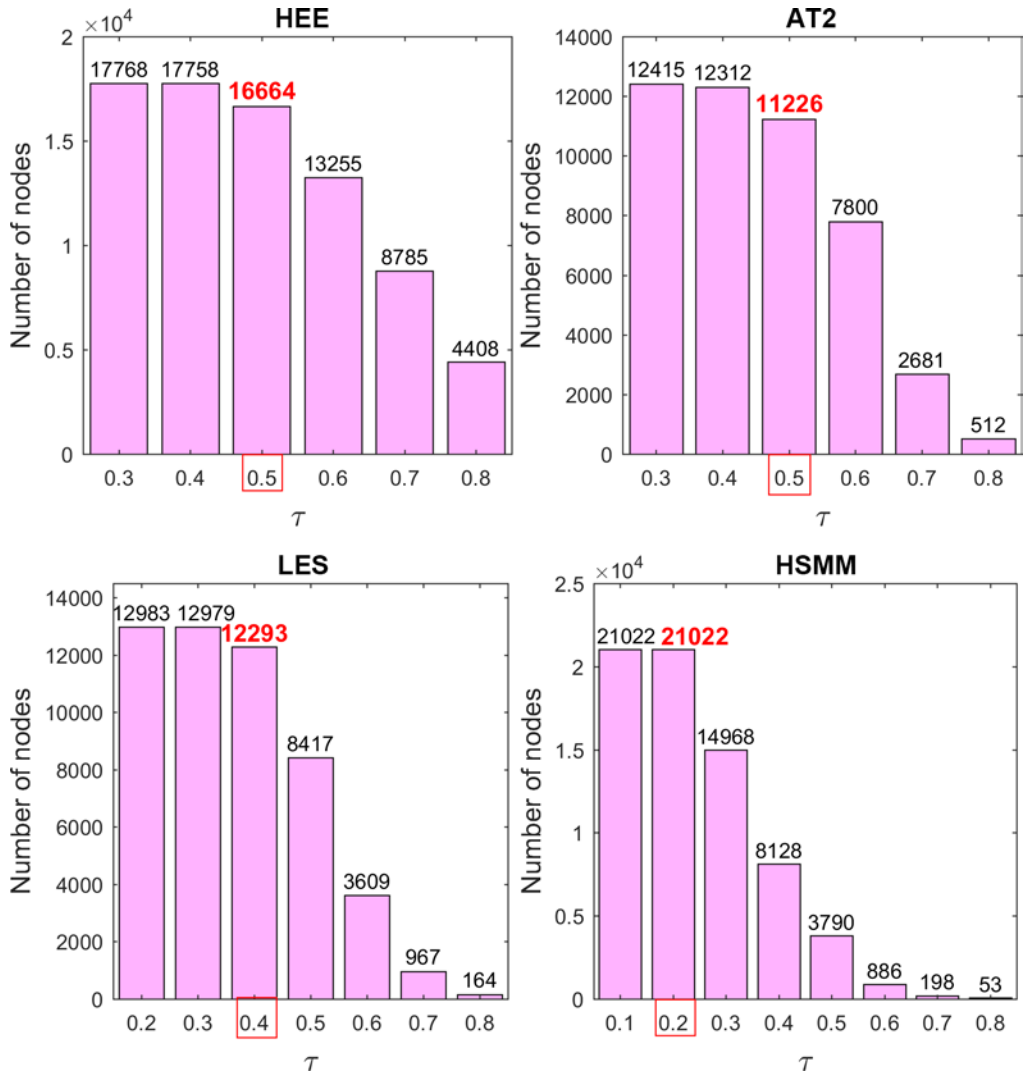

Figure S2. The number of nodes in each constructed network under different threshold τ. The selected threshold and the number of nodes are indicated by red box and red font color, respectively.

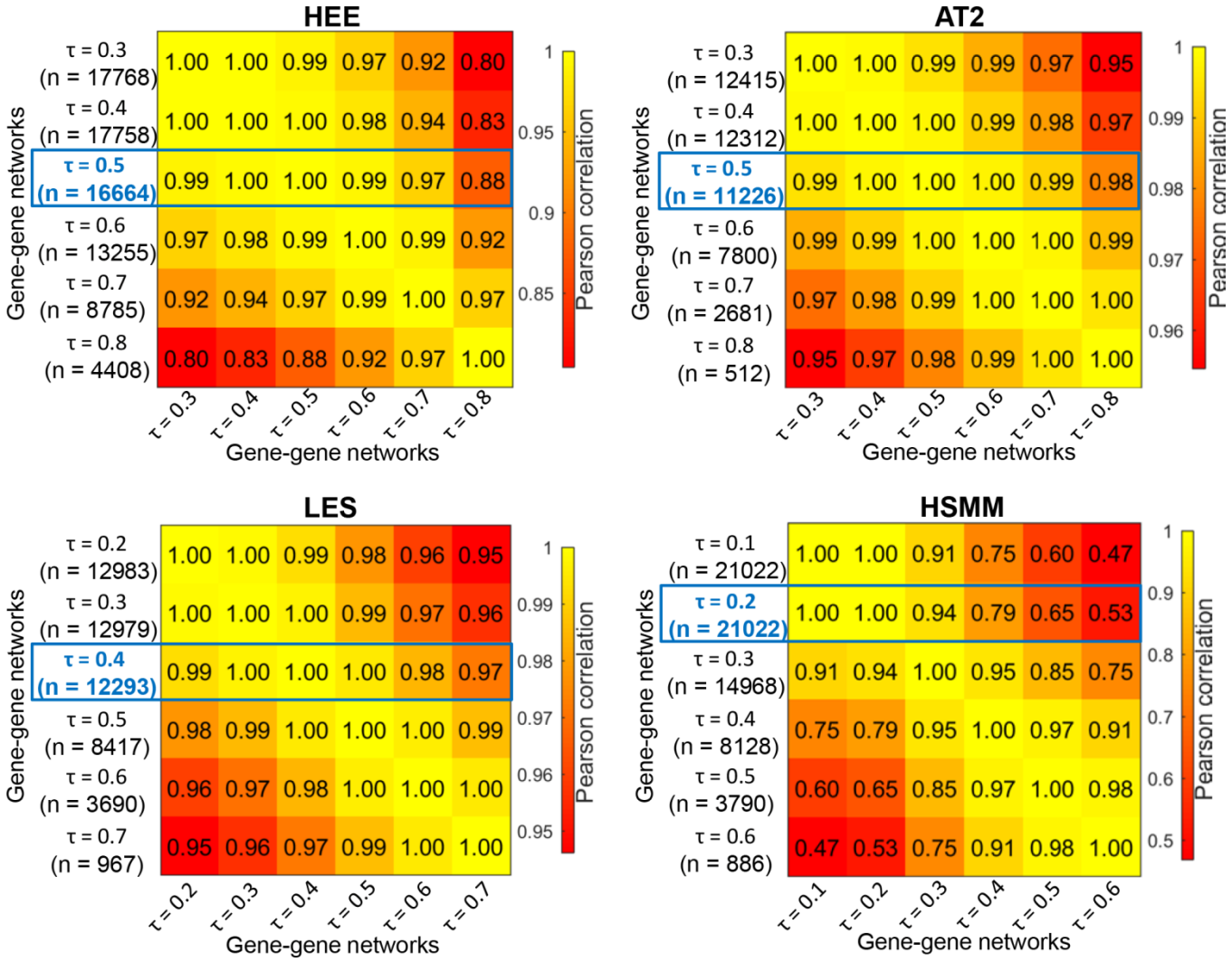

Figure S3. Pairwise correlation of the estimated scEnergy of single cells among different runs under different choices of the threshold τ. The threshold selected in this study is indicated by a blue box. The number of genes under each threshold is also depicted.

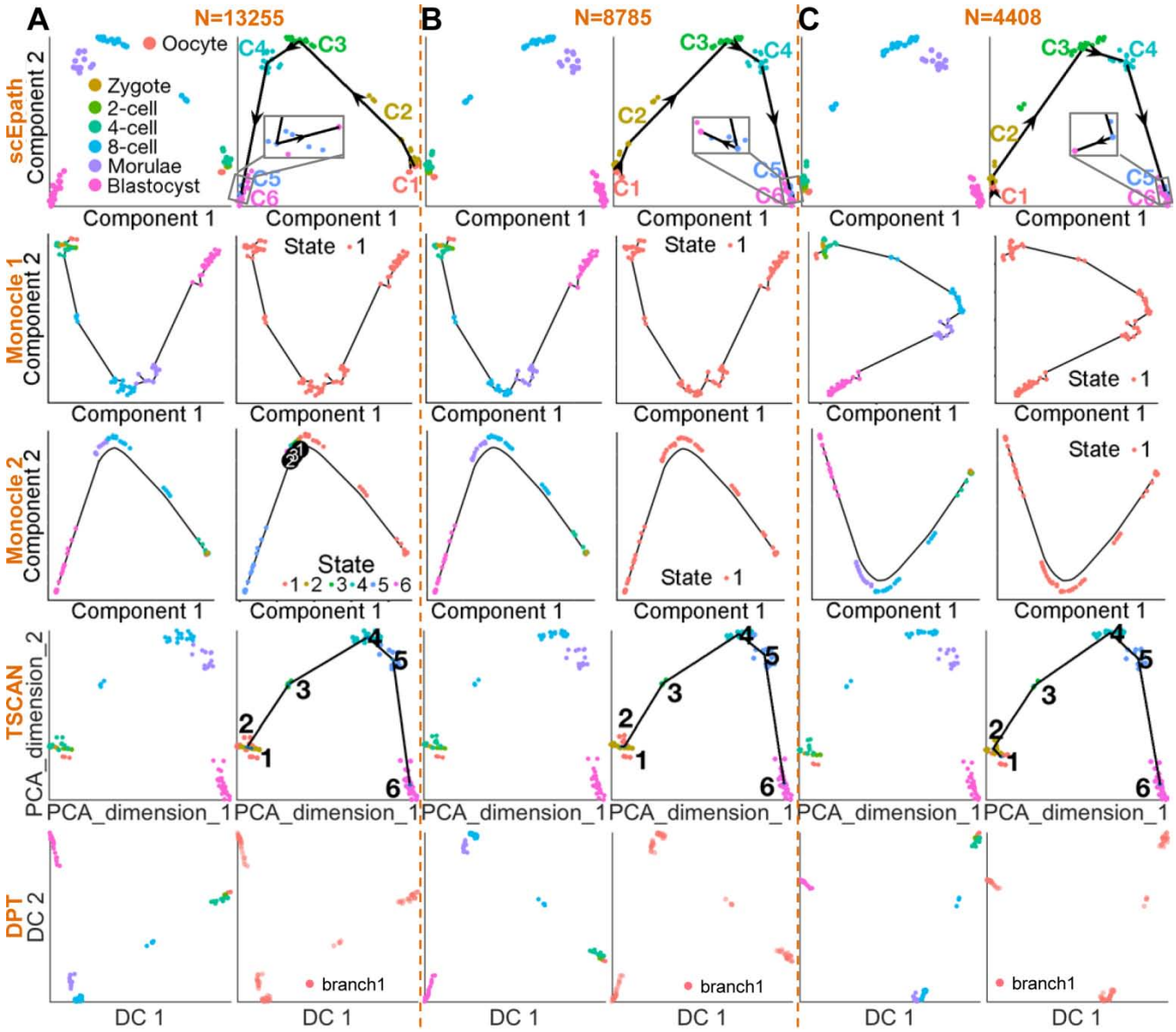

Figure S4. scEpath, Monocle 1, Monocle 2, TSCAN, and DPT were applied to the HEE single-cell RNA-seq data from our first example while varying the number of input genes for lineage inference. (A) The number of genes is 13255 when the threshold for network construction is 0.6. Left panel: Cells were colored based on the experimental stages ('Oocyte','Zygote','2-cell','4-cell','8-cell','Morulae' and 'Late blastocyst'). Right panel: Cells were colored based on the clusters (or states/branches) assigned by the algorithms. Inferred lineage trees were also depicted. (B) The number of genes is 8785 when the threshold for network construction is 0.7. (C) The number of genes is 4408 when the threshold for network construction is 0.8.

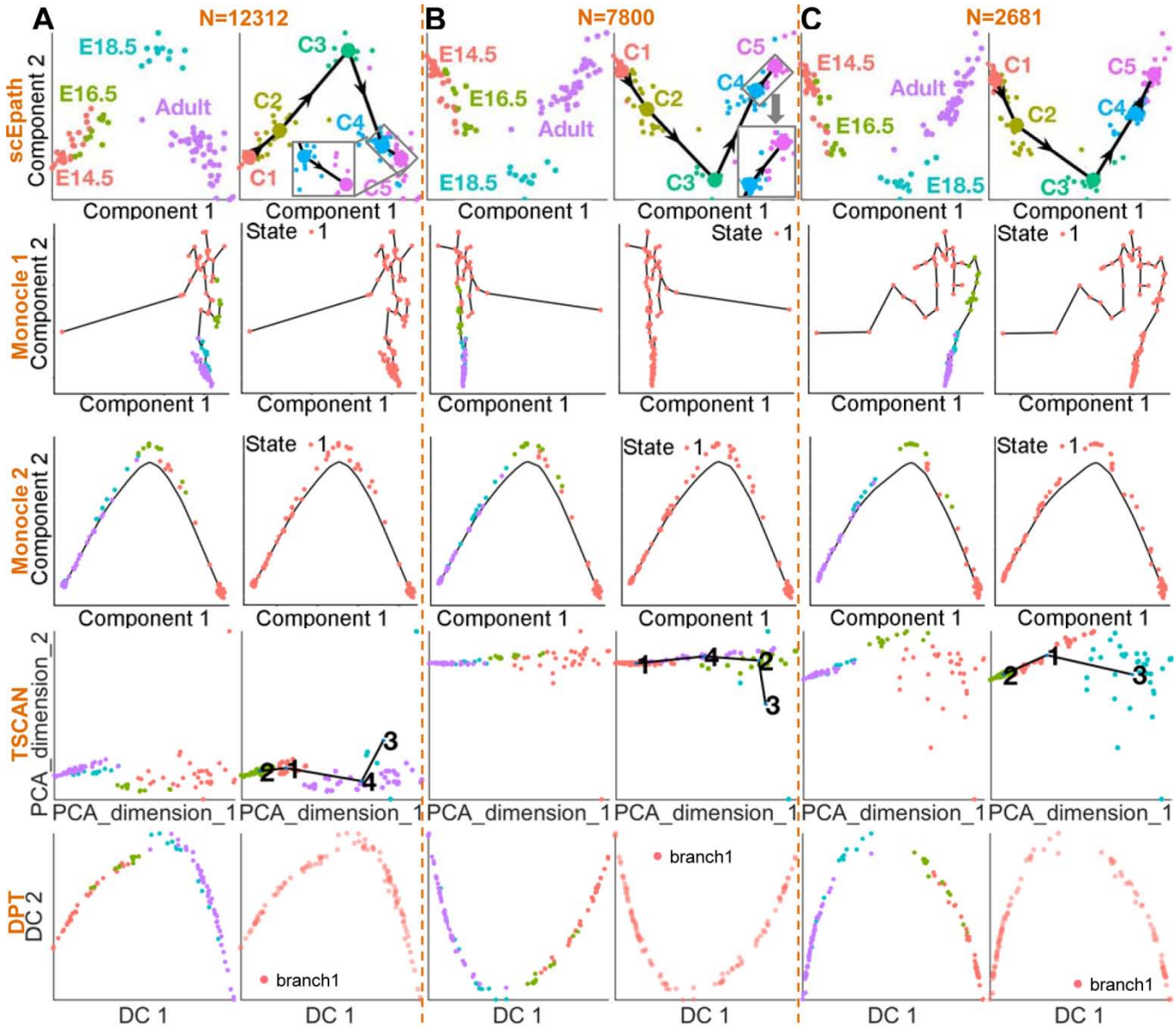

Figure S5. scEpath, Monocle 1, Monocle 2, TSCAN, and DPT were applied to the AT2 single-cell RNA-seq data while varying the number of selected genes for lineage inference. (A) The number of genes is 12312 when the threshold for network construction is 0.4. Left panel: Cells were colored based on the experimental stages (E14.5, E16.5, E18.5 and Adult). Right panel: Cells were colored based on the clusters (or states/branches) assigned by the algorithms. Inferred lineage trees were also depicted. (B) The number of genes is 7800 when the threshold for network construction is 0.6. (C) The number of genes is 2681 when the threshold for network construction is 0.7.

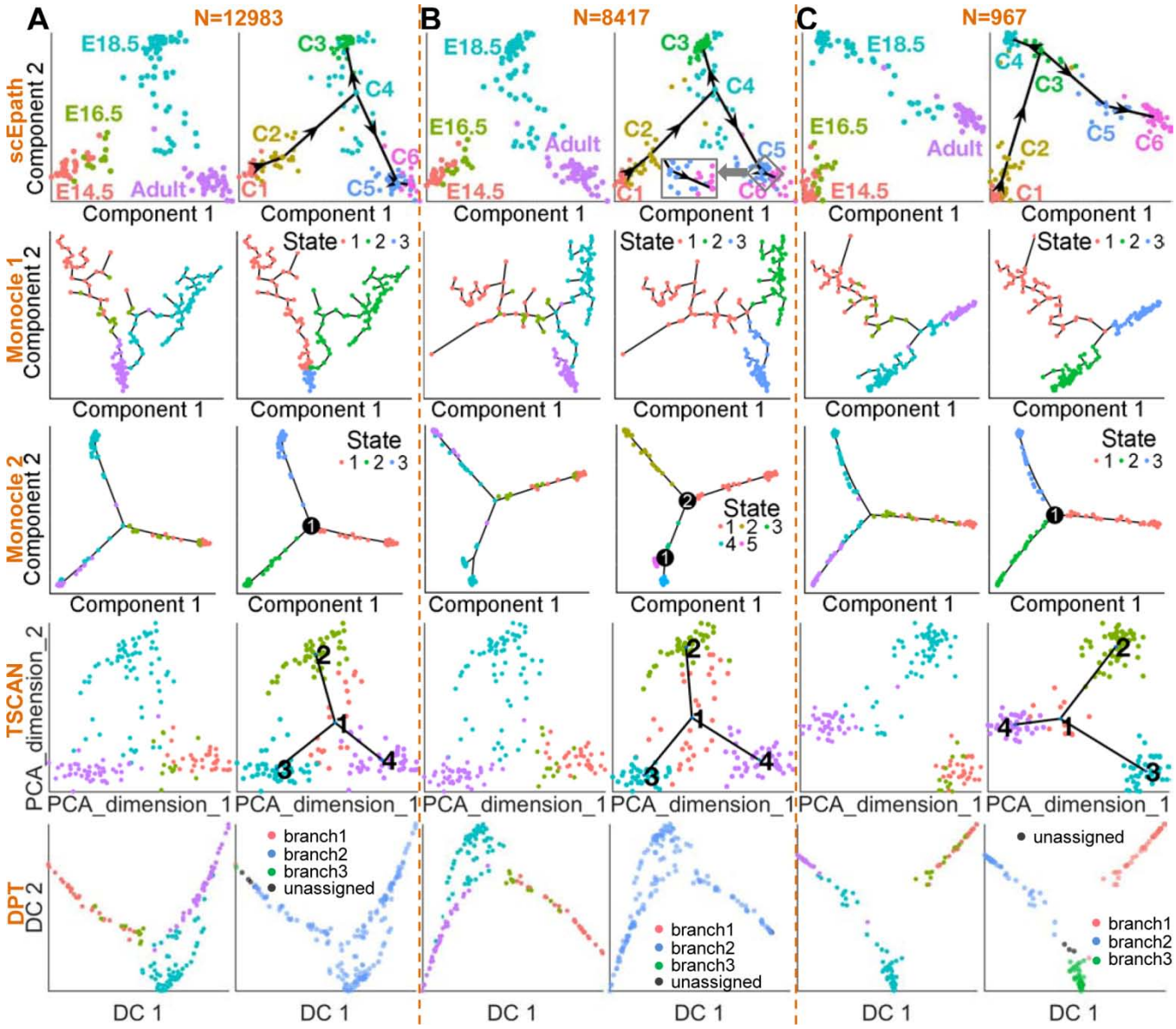

Figure S6. scEpath, Monocle 1, Monocle 2, TSCAN, and DPT were applied to the lung epithelial specification (LES) single-cell RNA-seq data while varying the number of selected genes for lineage inference. (A) The number of genes is 12983 when the threshold for network construction is 0.2. Left panel: Cells were colored based on the experimental stages (E14.5, E16.5, E18.5 and Adult). Right panel: Cells were colored based on the clusters (or states/branches) assigned by the algorithms. Inferred lineage trees were also depicted. (B) The number of genes is 8417 when the threshold for network construction is 0.5. (C) The number of genes is 967 when the threshold for network construction is 0.7.

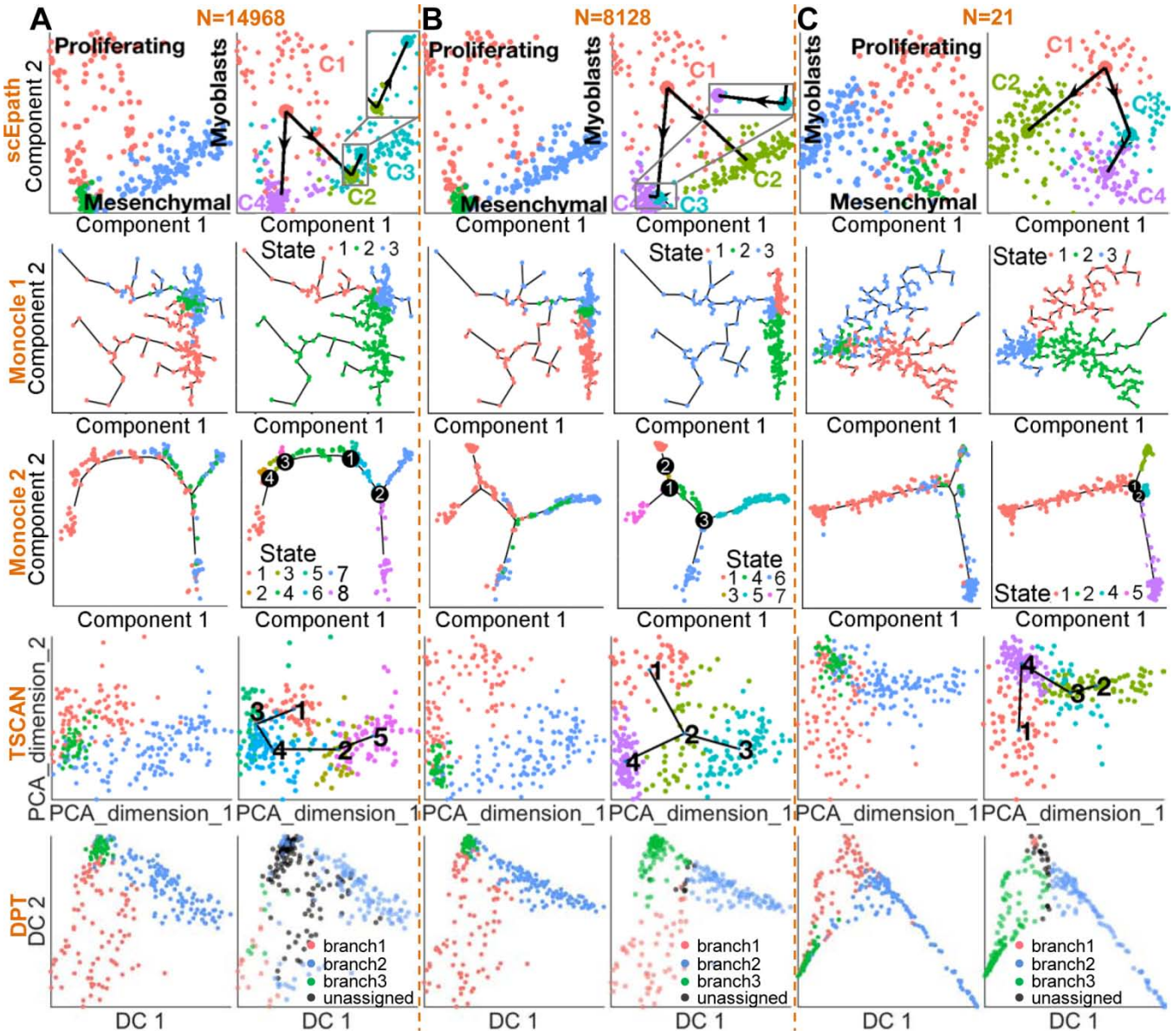

Figure S7. scEpath, Monocle 1, Monocle 2, TSCAN, and DPT were applied to the human skeletal muscle myoblasts (HSMM) single-cell RNA-seq data from our last example while varying the number of selected genes for lineage inference. (A) The number of genes is 14968 when the threshold for network construction is 0.3. Left panel: Cells were colored based on the cell identity assigned by Monocle (Trapnell, et al., 2014) (proliferating cells, differentiating myoblasts cells and Interstitial mesenchymal cells). Right panel: Cells were colored based on the clusters (or states/branches) assigned by the algorithms. Inferred lineage trees were also depicted. For the lineage trees reported by Monocle 2, each black circle with a white number represents a branched point. (B) The number of genes is 8128 when the threshold for network construction is 0.4. (C) The number of genes is 21 when we used the maker genes (from Monocle package) for differentiating myoblasts.

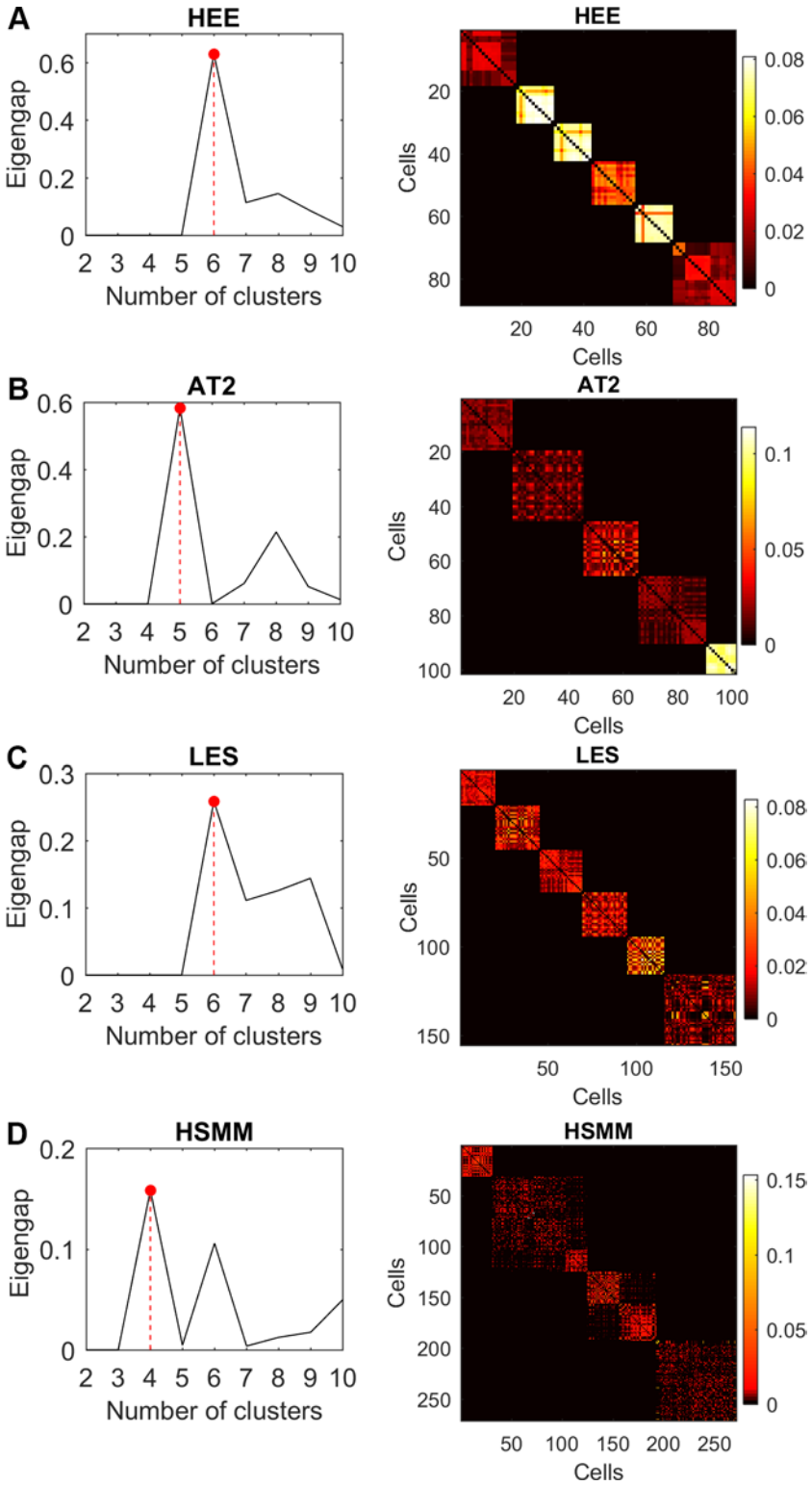

Figure S8. Determining the number of clusters using eigengap. (A) Eigengap of recommending the best choice of the number of clusters with SIMLR on the four real single cell data sets: HEE, AT2, LES and HSMM. The higher the eigengap is, the more likely the number is close to the optimal number of clusters. The red dot and dashed line indicate the optimal number of clusters for each data set. Right panel: With the optimal number of clusters, the learned cell-cell similarity matrix by SIMLR on the four real singlecell data sets. We can find that, the eigengap can provide a good estimate of the optimal number of clusters, which leads to a clear block-diagonal structure with *N* blocks whereby cells within the same subpopulation are more similar.

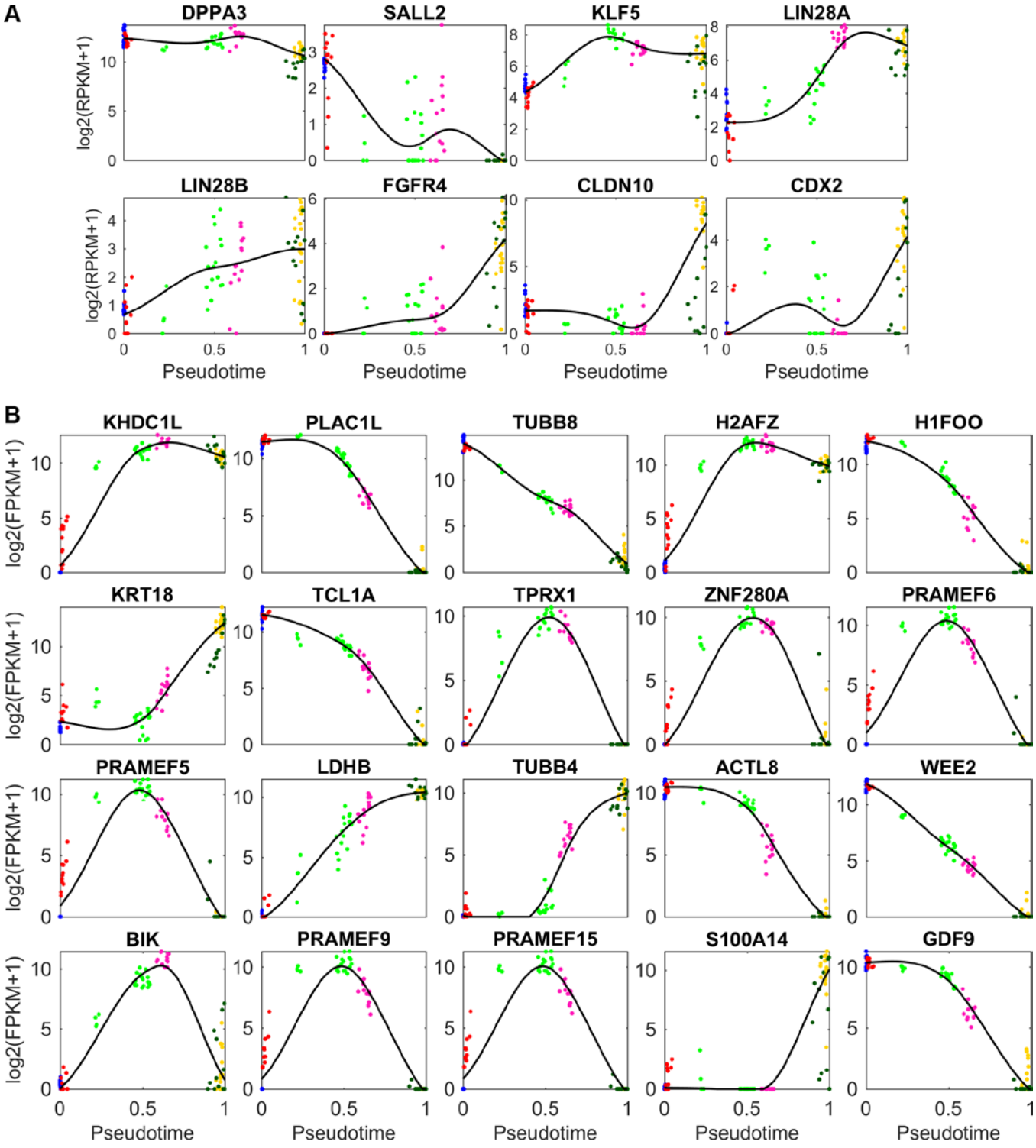

Figure S9. The expression patterns of genes along the reconstructed pseudotime during the development in human early embryos cells (HEE). (A) Validation of pseudotemporal ordering using the known cell stage markers. Gene expression in cells plotted along pseudotime and fitted with a cubic smoothing spline (black line). Cells are colored according to cell clusters defined by scEpath. (B) The expression patterns of top 20 genes identified by scEpath show significant changes along pseudotime.

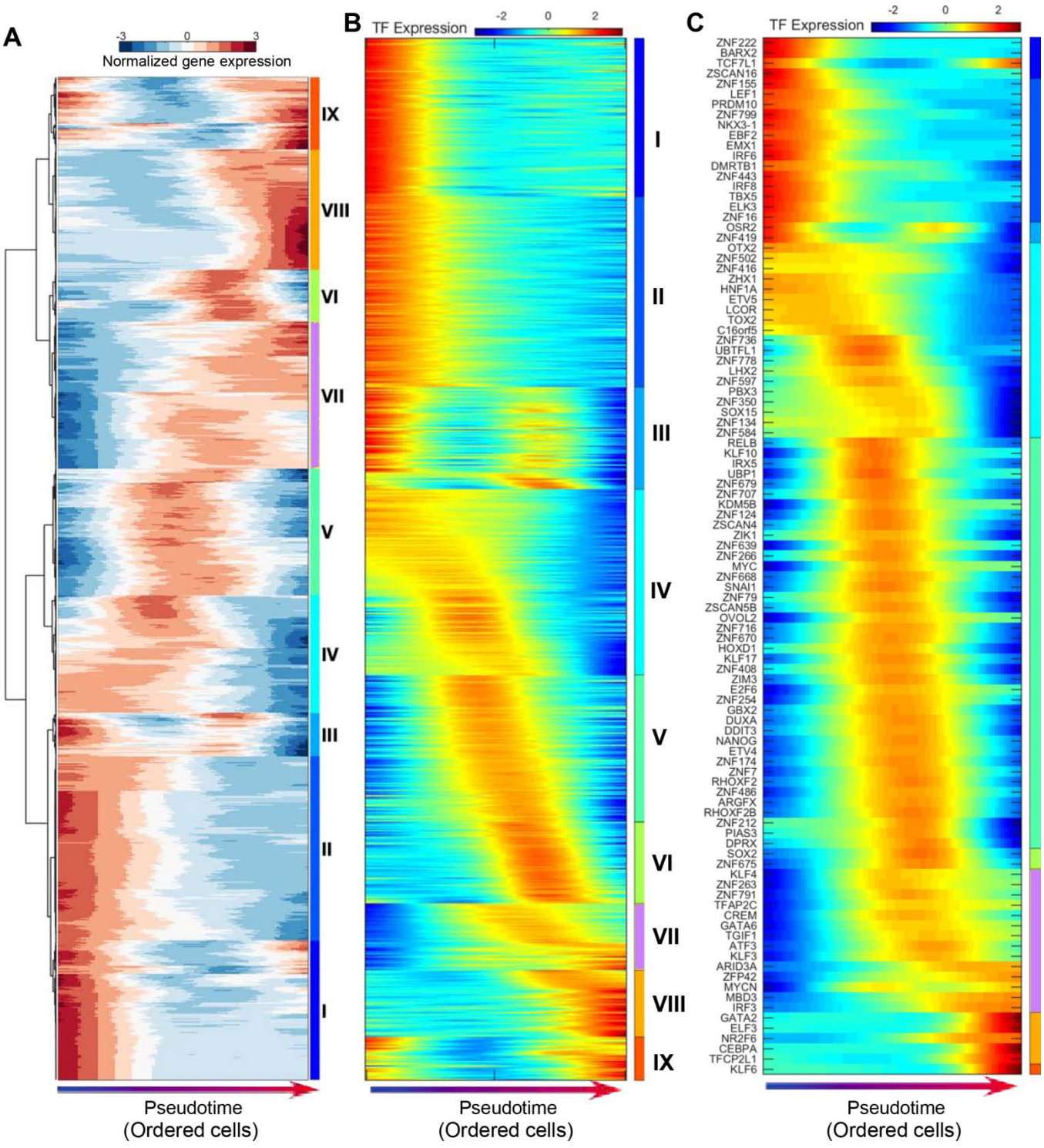

Figure S10. Heatmap of the expressions of pseudotime dependent genes and transcription factors along the reconstructed pseudotime during the development in human early embryos cells (HEE). (A) Heatmap of pseudotime dependent genes (n = 2068) organized according to a hierarchical clustering of their smoothed gene expression profiles using Pearson's correlation with Ward's method. Genes were clustered into nine groups (I–IX). (B) "Rolling wave" showing the smoothed expression pattern of all the significantly differentiation-related transcription factors (TFs). TFs were ordered according to cluster membership (right) and peak expression.

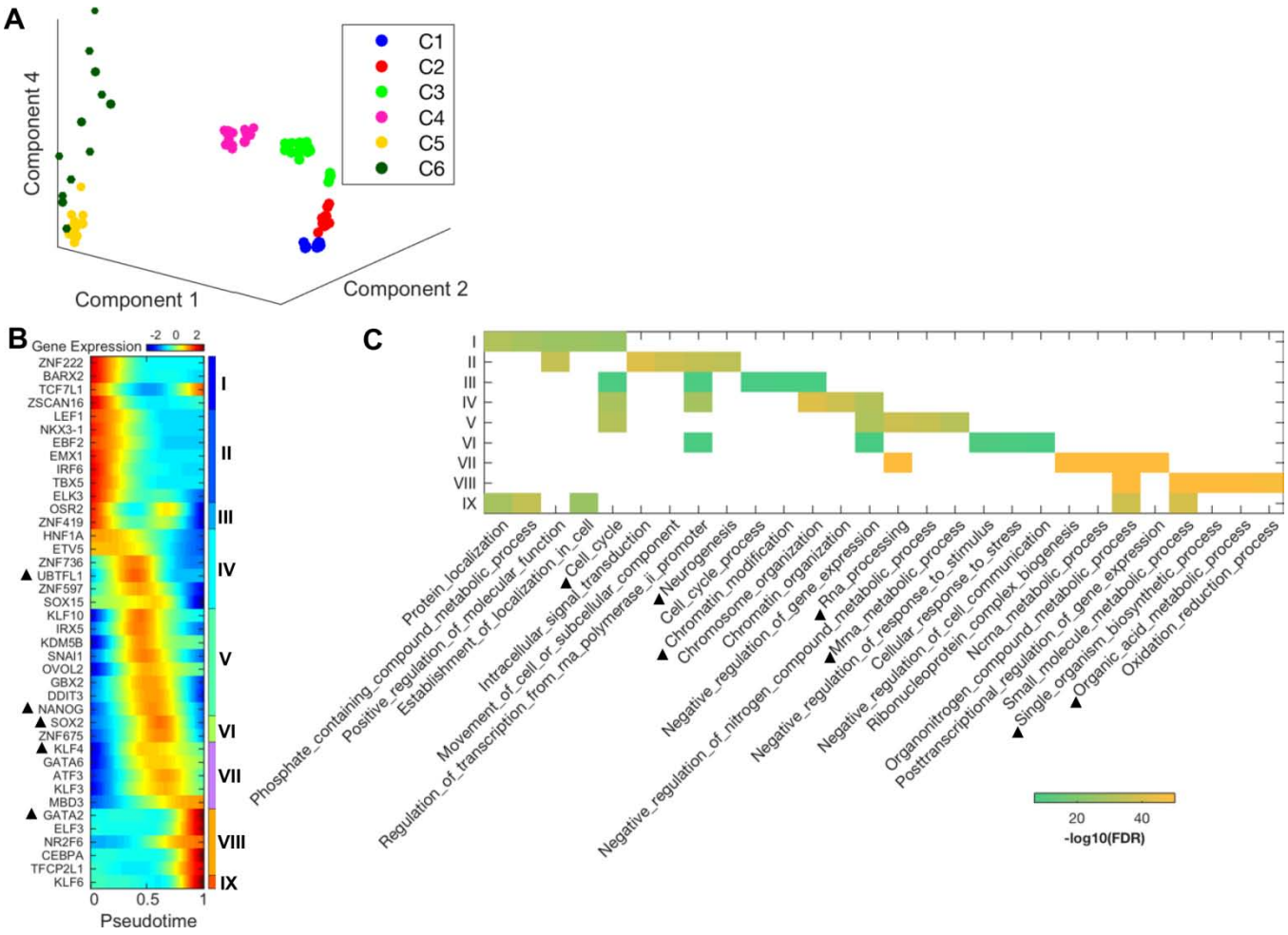

Figure S11. scEpath uncovered critical transcription factors and functional signatures during the development of human early embryos (HEE). (A) Cells visualized on the Component 1, Component 2 and Component 4. Clusters C5 and C6 show a mixed domain in the first two components. However, Component 4 clearly separates C6 from all the other clusters. (B) “Rolling wave” plot showing the smoothed expression pattern of the important TFs delineated by scEpath. TFs were ordered according to cluster membership (right) and peak expression as shown in Figure 2G in main text. TF indicated by a triangle have been previously described as relevant for human early embryo development. (C) Functional map shows the top five enriched GO biological processes of pesudotime-dependent genes in each gene cluster defined by Figure 2G in main text. Go terms indicated by a triangle shows some representative functional signatures.

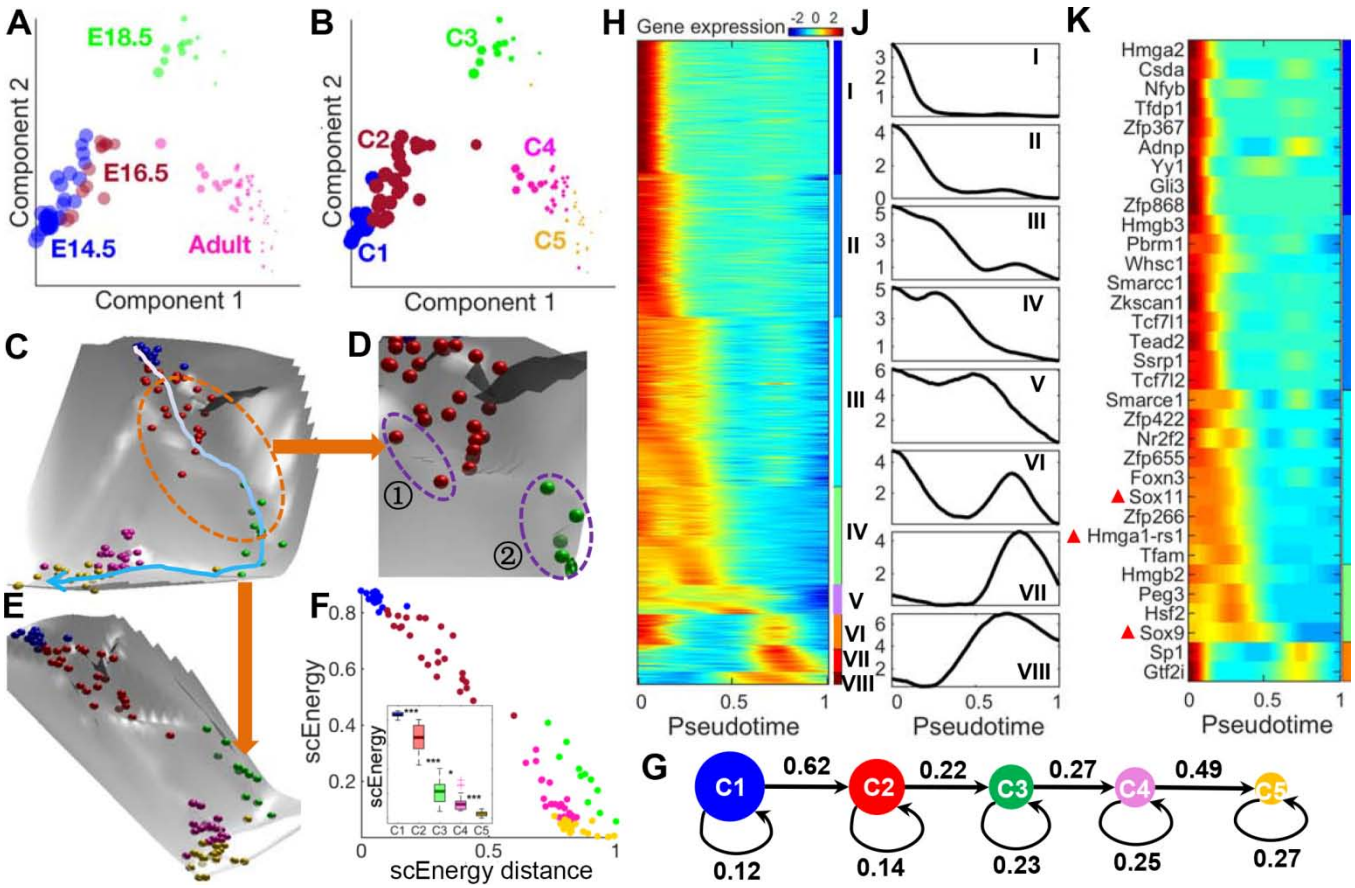

Figure S12. scEpath reconstructed the differentiation lineage and the high-resolution transcriptional programs of mouse lung alveolar type 2 (AT2) cells. (A) Energy landscape visualized on the first two components, colored by the experimental time points. (B) Cells are colored according to unsupervised clustering. Sizes of cells were proportional to the scEnergies of the cells, scEpath identified five clusters without feature selection. (C) Overall view of energy landscape in three dimensions. The developmental trajectories are shown by a curve in which white indicates initial stage and blue indicates late stages. (D) Zoom into the area indicted by light brown oval in (C), which shows the clear barrier between C2 and C3. Red cells indicated by ① show an example of "edge cells". Green cells indicated by ② show that cells in C3 are intermediate states which are not stable. (E) Different view of energy landscape shows the transition path in the late stages. (F) Main panel: Cells visualized on the scEnergy distance-scEnergy space, suggesting a linear trajectory. Inset: Comparison of energy distribution among the identified cell clusters, suggesting a significant reduced energy during differentiation. "***" represents the p-value (a two-sided Wilcoxon rank-sum test) is less than 0.001, "*" represents the p-value is less than 0.05 but greater than 0.01. (G) scEpath revealed a linear lineage path where transition rates are shown. (H) "Rolling wave" plot showing the normalized-smoothed expression pattern of pseudotime-dependent genes (n = 2068) clustered into eight groups (I–VIII) and ordered according to their peak expression. (J) Average expressions (log2(FPKM+1)) of the eight gene clusters along pseudotime. (K) "Rolling wave" plot showing the smoothed expression pattern of the most significantly differentiation-related transcription factors (TFs). TFs were ordered according to cluster membership (right) and peak expression as shown in (H). TF indicated by a triangle have been previously described as relevant for AT2 cell differentiation.

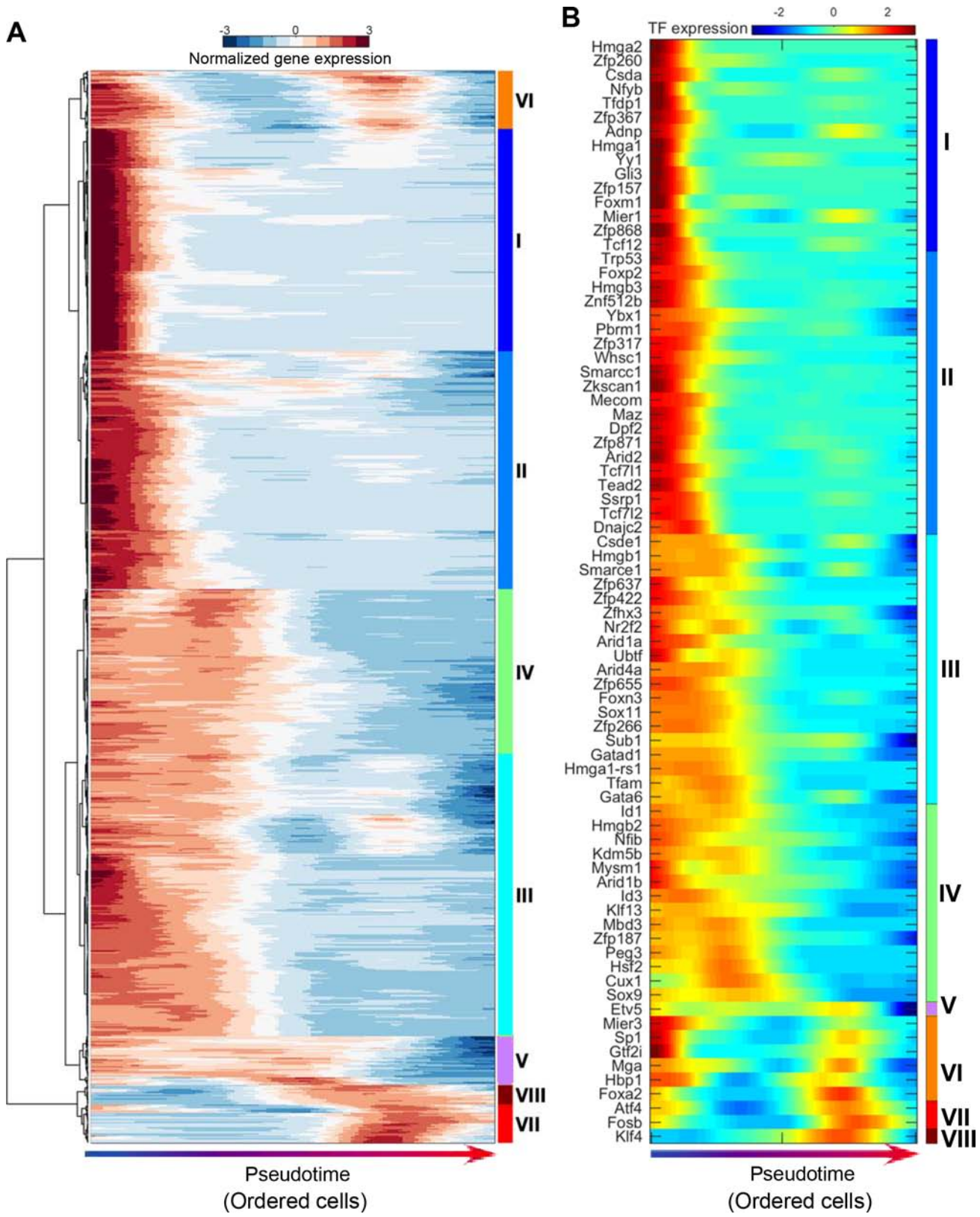

Figure S13. Heatmap of the expressions of pseudotime dependent genes and transcription factors along the reconstructed pseudotime during the progression of AT2 cell differentiation. (A) Heatmap of pseudotime dependent genes (n = 2068) organized according to a hierarchical clustering of their smoothed gene expression profiles using Pearson's correlation with Ward's method. Genes were clustered into eight groups (I–VIII). (B) "Rolling wave" showing the smoothed expression pattern of all the significantly differentiation-related transcription factors (TFs). TFs were ordered according to cluster membership (right) and peak expression.

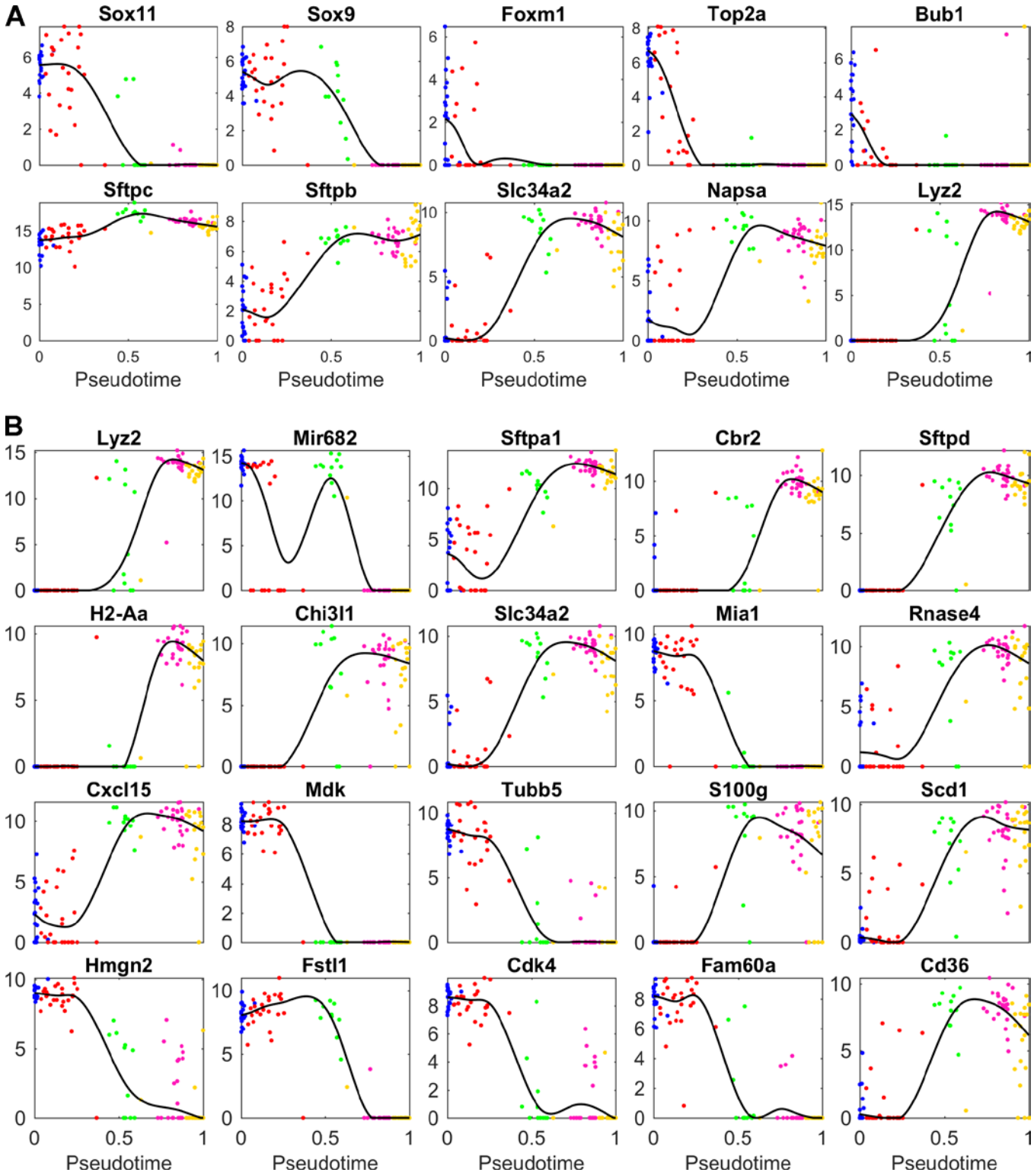

Figure S14. The expression patterns of genes along the reconstructed pseudotime during the progression of AT2 cell differentiation. (A) Validation of pseudotemporal ordering of AT2 cells using the known cell stage markers. Gene expression in AT2 cells plotted along pseudotime and fitted with a cubic smoothing spline (black line). Cells are colored according to cell clusters defined by scEpath. (B) The expression patterns of top 20 genes identified by scEpath show significant changes within pseudotime.

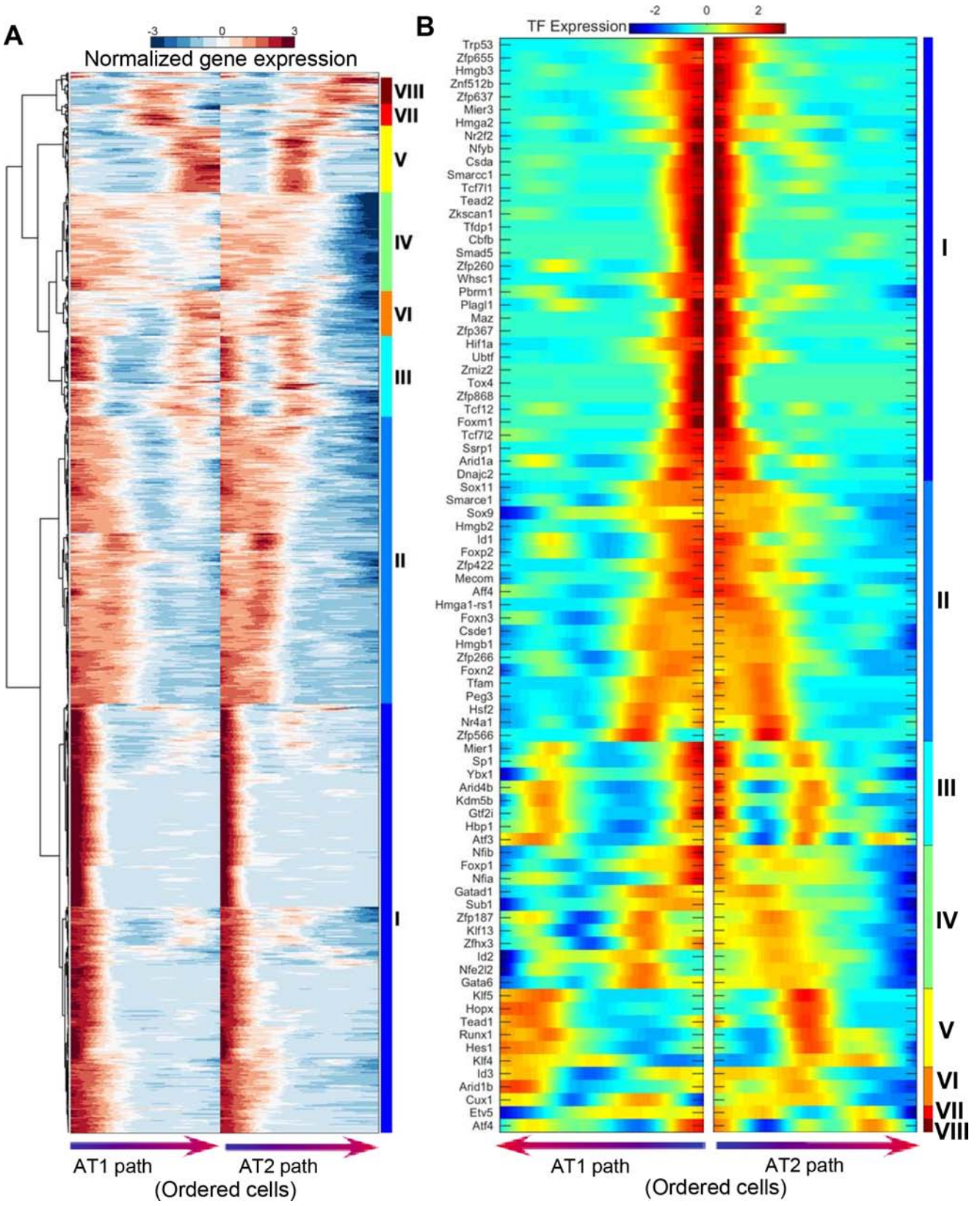

Figure S15. Heatmap of the expressions of pseudotime dependent genes and transcription factors along the reconstructed pseudotime during mouse lung epithelial specification. (A) Heatmap of pseudotime dependent genes (n = 1223) organized according to a hierarchical clustering of their smoothed gene expression profiles using Pearson's correlation with Ward's method during mouse lung epithelial specification. Genes were clustered into eight groups (I–VIII). (B) "Rolling wave" showing the smoothed expression pattern of all the significantly differentiation-related transcription factors (TFs). TFs were ordered according to cluster membership (right) and peak expression.

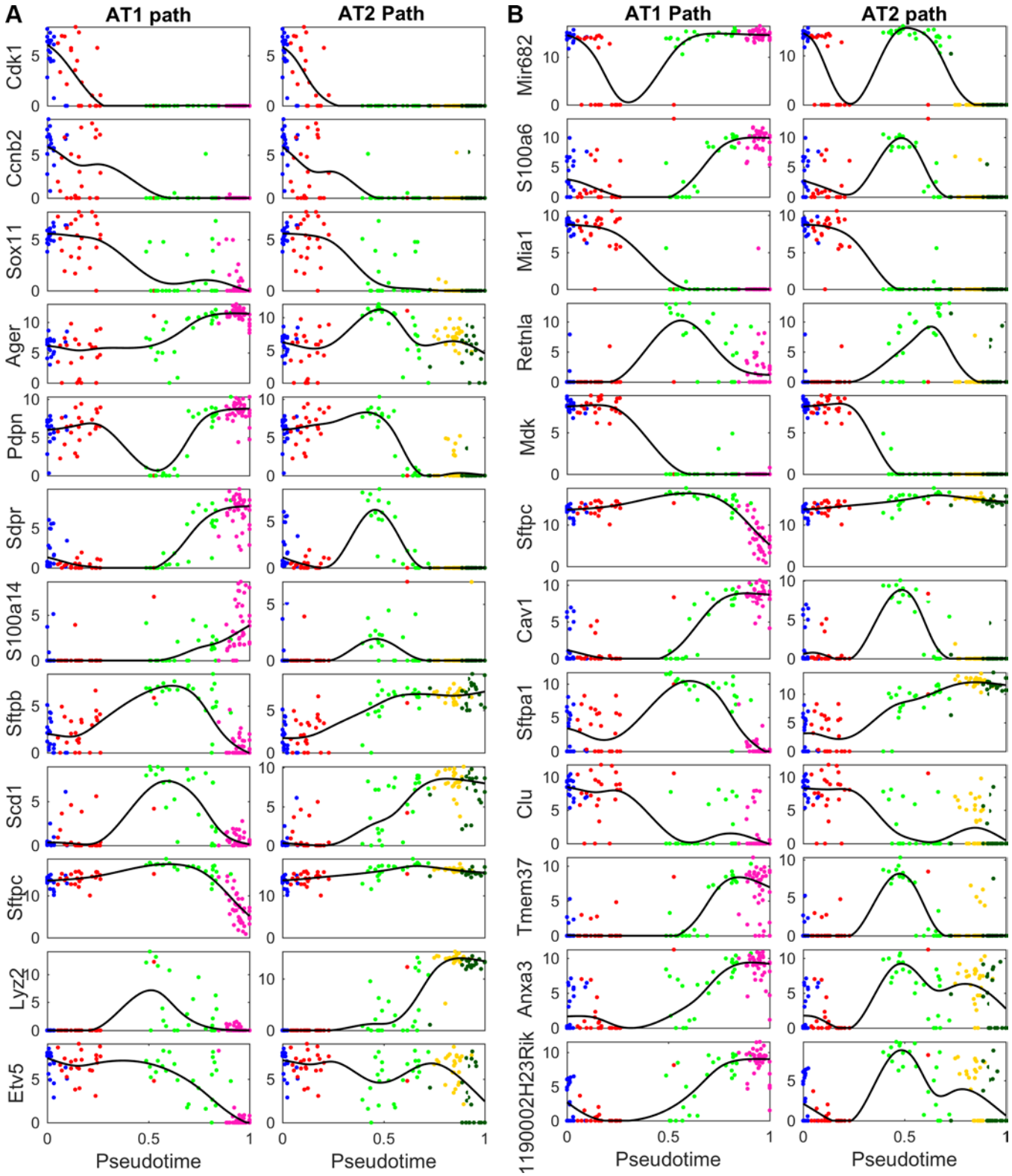

Figure S16. The expression patterns of genes along the reconstructed pseudotime during mouse lung epithelial specification. (A) Validation of pseudotemporal ordering using the known markers. Gene expression in cells plotted along pseudotime and fitted with a cubic smoothing spline (black line). Cells are colored according to cell clusters defined by scEpath. (B) The expression patterns of top 12 genes identified by scEpath show significant changes within pseudotime.

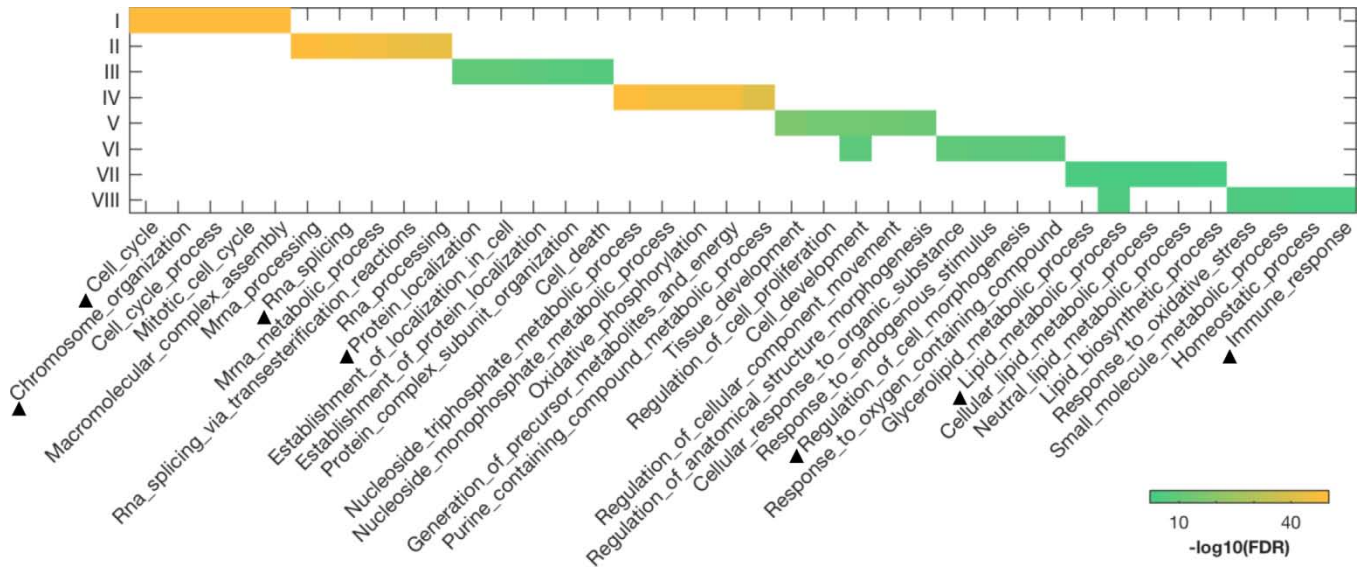

Figure S17. scEpath uncovered functional signatures during mouse lung epithelial specification. Functional map shows the top five enriched GO biological processes of pesudotime-dependent genes in each gene cluster defined by Figure 3H in main text. Go terms indicated by a triangle shows some representative functional signatures.

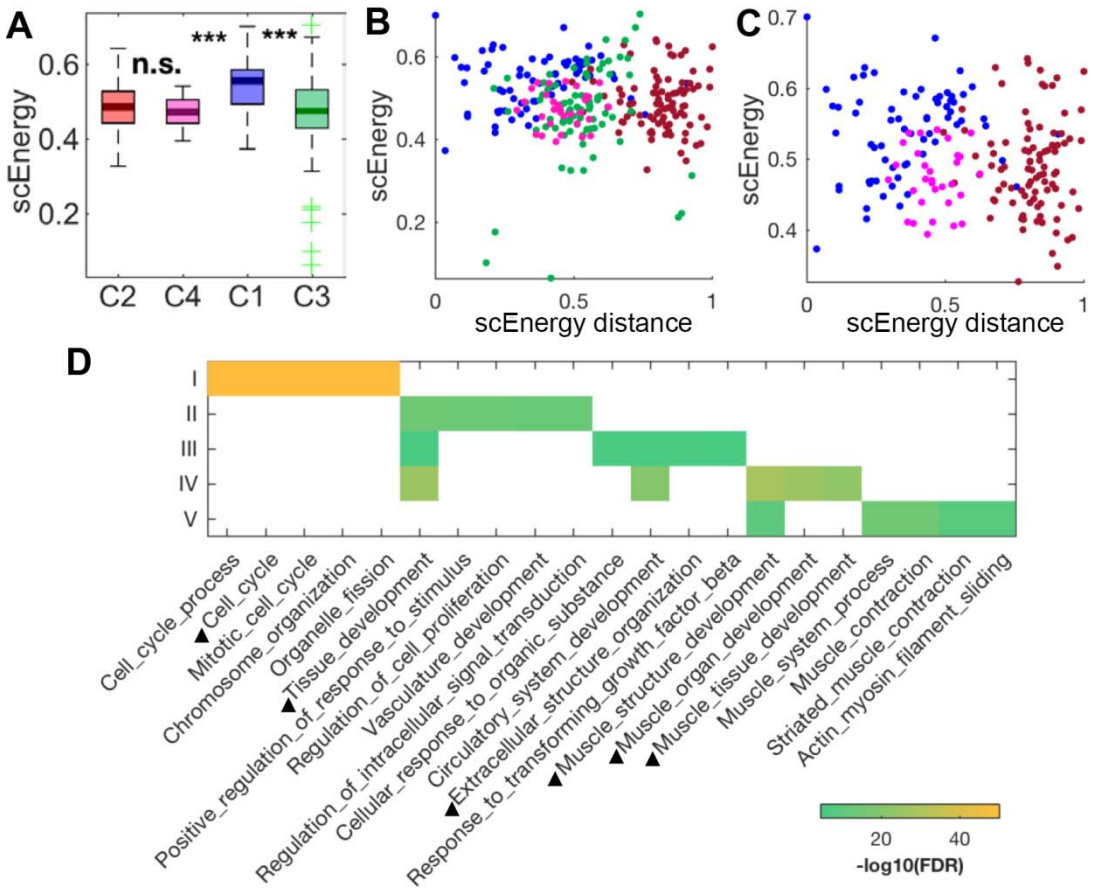

Figure S18. scEpath uncovered decreased scEnergies and functional signatures during myoblast differentiation. (A) Comparison of energy distribution among the identified cell clusters. "***" represents the p-value (a two-sided Wilcoxon rank-sum test) is less than 0.001, "n.s." represents the p-value is greater than 0.05. (B) All the cells visualized on the scEnergy distance--scEnergy space. (C) Cells involved in myoblast differentiation (i.e, cluster C3 containing mesenchymal cells was removed) were visualized on the scSimilarity--scEnergy space. The combination of scSimilarity and scEnergy well quantified the lineage relationship. (D) Functional map shows the top five enriched GO

biological processes of pesudotime-dependent genes in each gene cluster defined by Figure 4F in main text. Go terms indicated by a triangle shows some representative functional signatures.

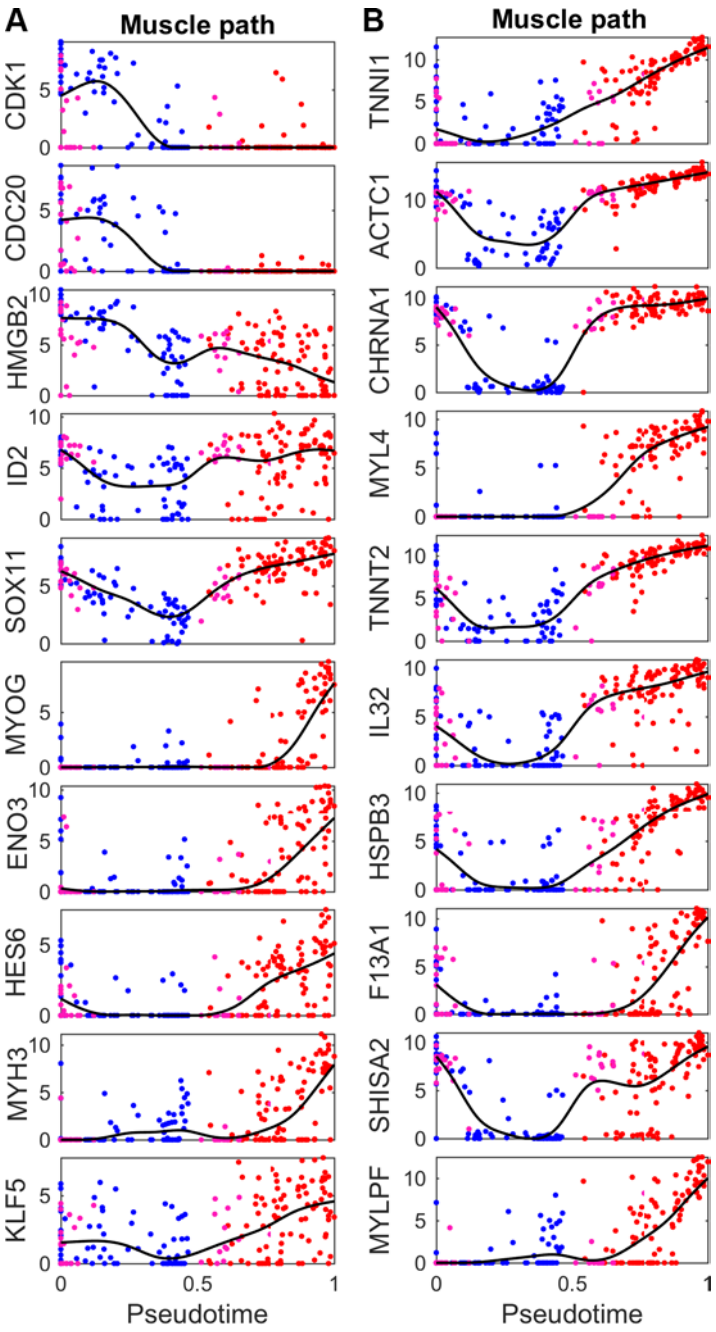

Figure S19. The expression patterns of genes along the reconstructed pseudotime during myoblast differentiation. (A) Validation of pseudotemporal ordering using the known markers. Gene expression in cells plotted along pseudotime and fitted with a cubic smoothing spline (black line). Cells are colored according to cell clusters defined by scEpath. (B) The expression patterns of top 10 genes identified by scEpath show significant changes within pseudotime.

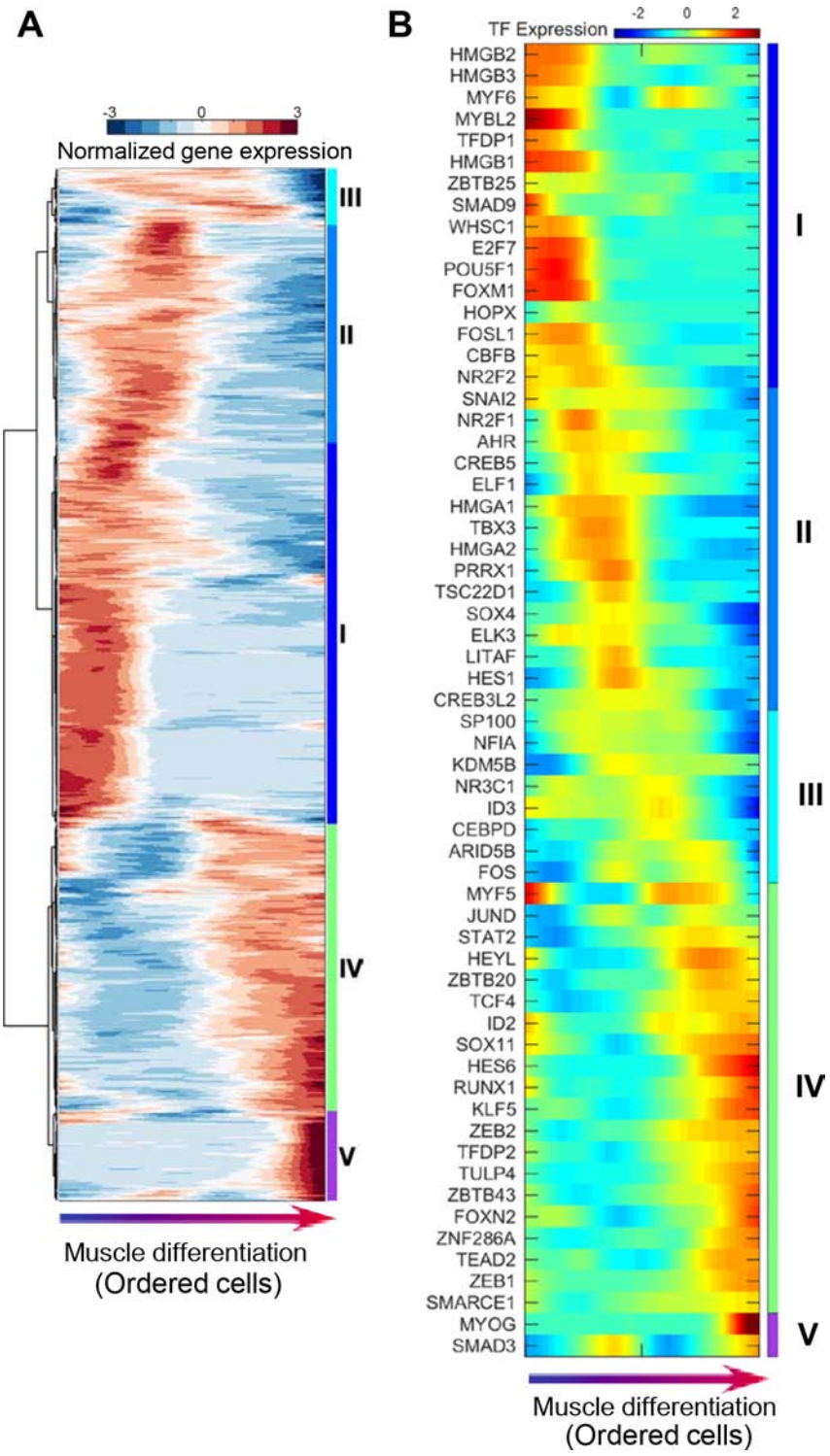

Figure S20. Heatmap of the expressions of pseudotime dependent genes and transcription factors along the reconstructed pseudotime during myoblast differentiation. (A) Heatmap of pseudotime dependent genes (n = 1116) organized according to a hierarchical clustering of their smoothed gene expression profiles using Pearson's correlation with Ward's method in HSMM data. Genes were clustered into five groups (I–V). (B) "Rolling wave" showing the smoothed expression pattern of all the significantly differentiation-related transcription factors (TFs). TFs were ordered according to cluster membership (right) and peak expression.

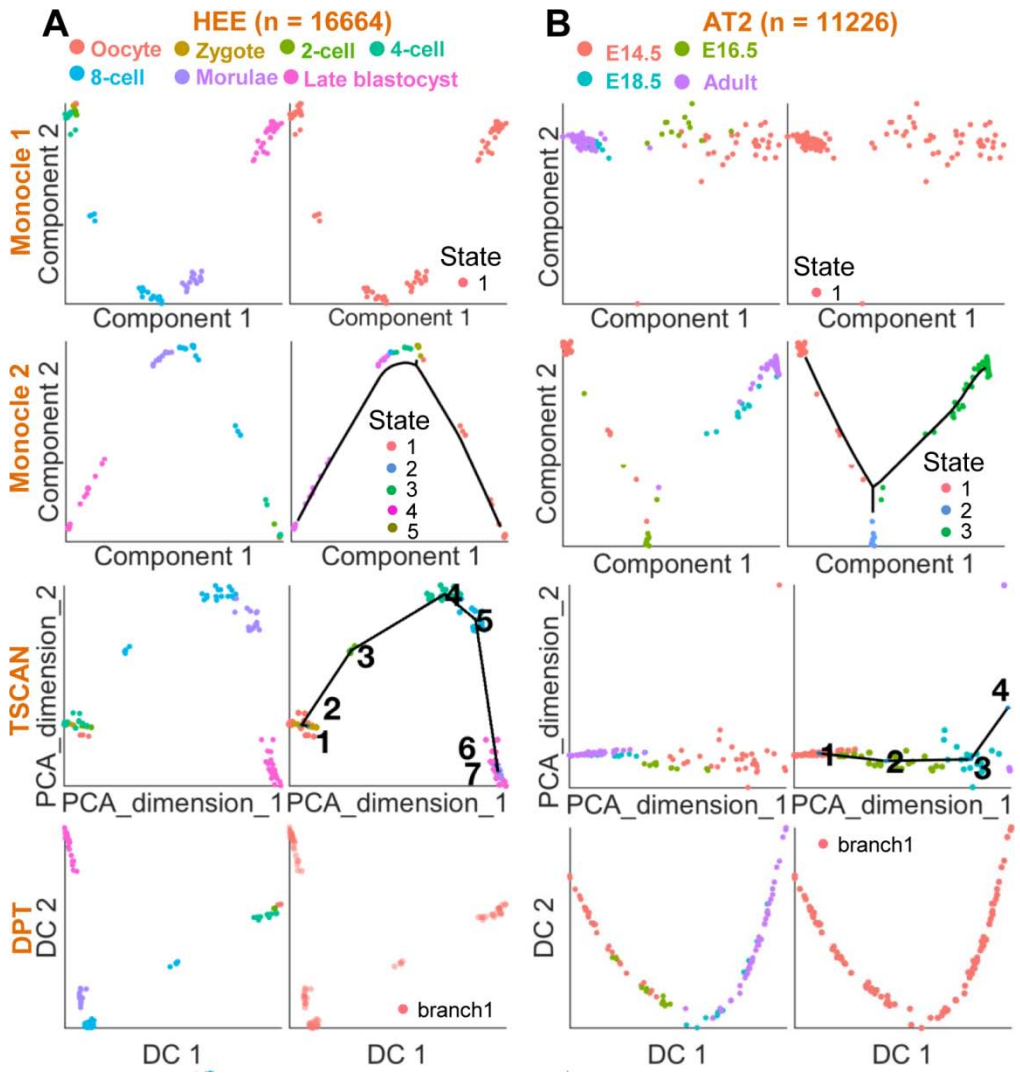

Figure S21. Monocle 1, Monocle 2, TSCAN and DPT were applied to the single-cell RNA-seq data from our first two examples. For each algorithm, the low-dimensional representation of cells and lineage trees was reported by the algorithms. (A) Application to HEE data. Left panel: Cells were colored based on the experimental stages ('Oocyte','Zygote','2-cell','4-cell','8-cell','Morulae' and 'Late blastocyst'). Right panel: Cells were colored based on the clusters (or states) identified by the algorithms. For DPT, cells were colored based on the branch assignment. (B) Application to AT2 data. Left panel: Cells were colored based on the experimental stages (E14.5, E16.5, E18.5 and Adult). Right panel: Cells were colored based on the clusters (or states) identified by the algorithms. Inferred lineage trees were also depicted.

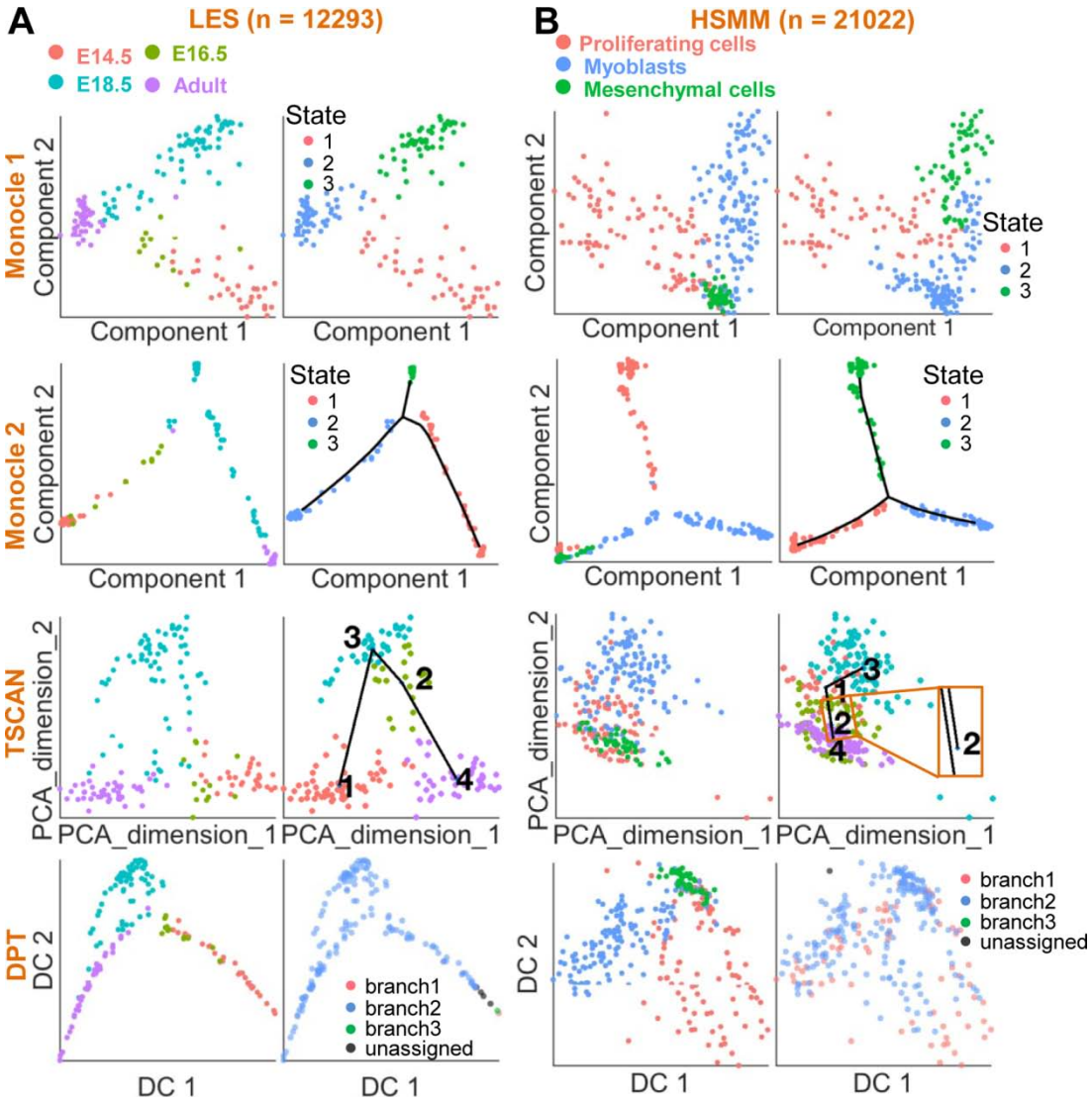

Figure S22. Monocle 1, Monocle 2, TSCAN and DPT were applied to the single-cell RNA-seq data from our last two examples. (A) Application to lung epithelial specification (LES) data. Left panel: Cells were colored based on the experimental stages (E14.5, E16.5, E18.5 and Adult). Right panel: Cells were colored based on the clusters (or states/branches) identified by the algorithms. (B) Application to human skeletal muscle myoblasts (HSMM) data. Left panel: Cells were colored based on the cell identity previously assigned by Monocle (Trapnell, et al., 2014) (proliferating cells, differentiating myoblasts cells and Interstitial mesenchymal cells). Right panel: Cells were colored based on the clusters (or states) identified by the algorithms. Inferred lineage trees were also depicted.

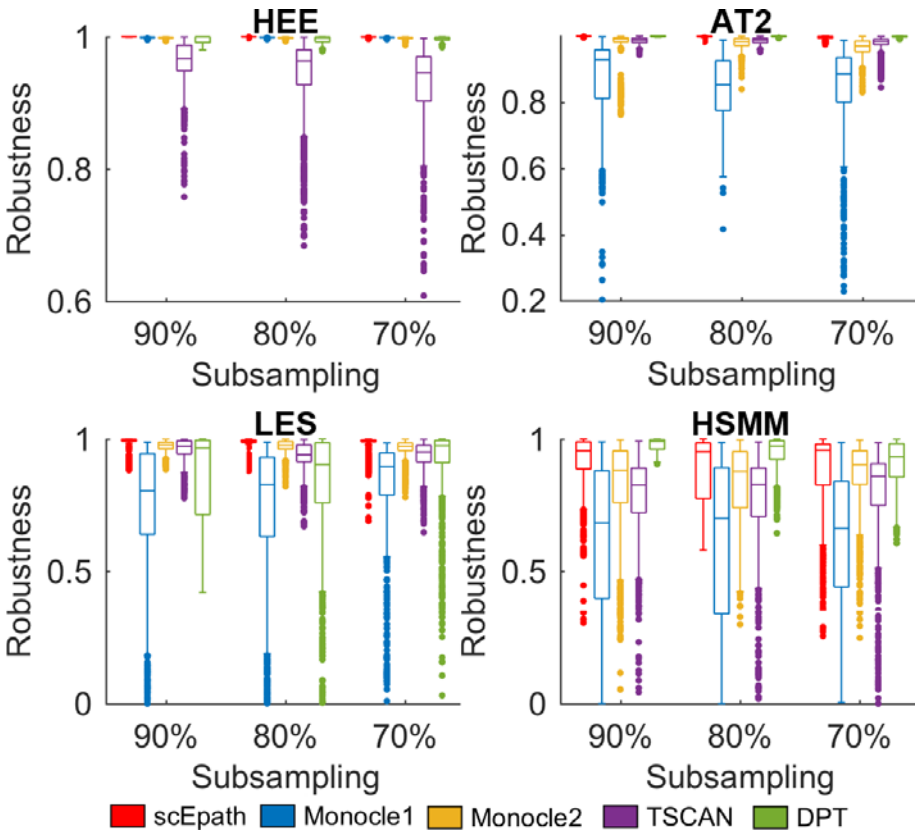

Figure 23. Comparison of robustness (by Pearson's correlation) of pseudotemporal ordering from each algorithm under repeated subsampling (independent 50 times) of the cells on each dataset. We used the Pearson's correlation to calculate the similarity of each pair of pseudotemporal ordering runs under repeated subsampling of 90, 80, or 70% of the total number of cells. In each scenario, the original data was independently subsampled 50 times, which resulted in a 50 by 50 (upper triangular) similarity matrix for each method.

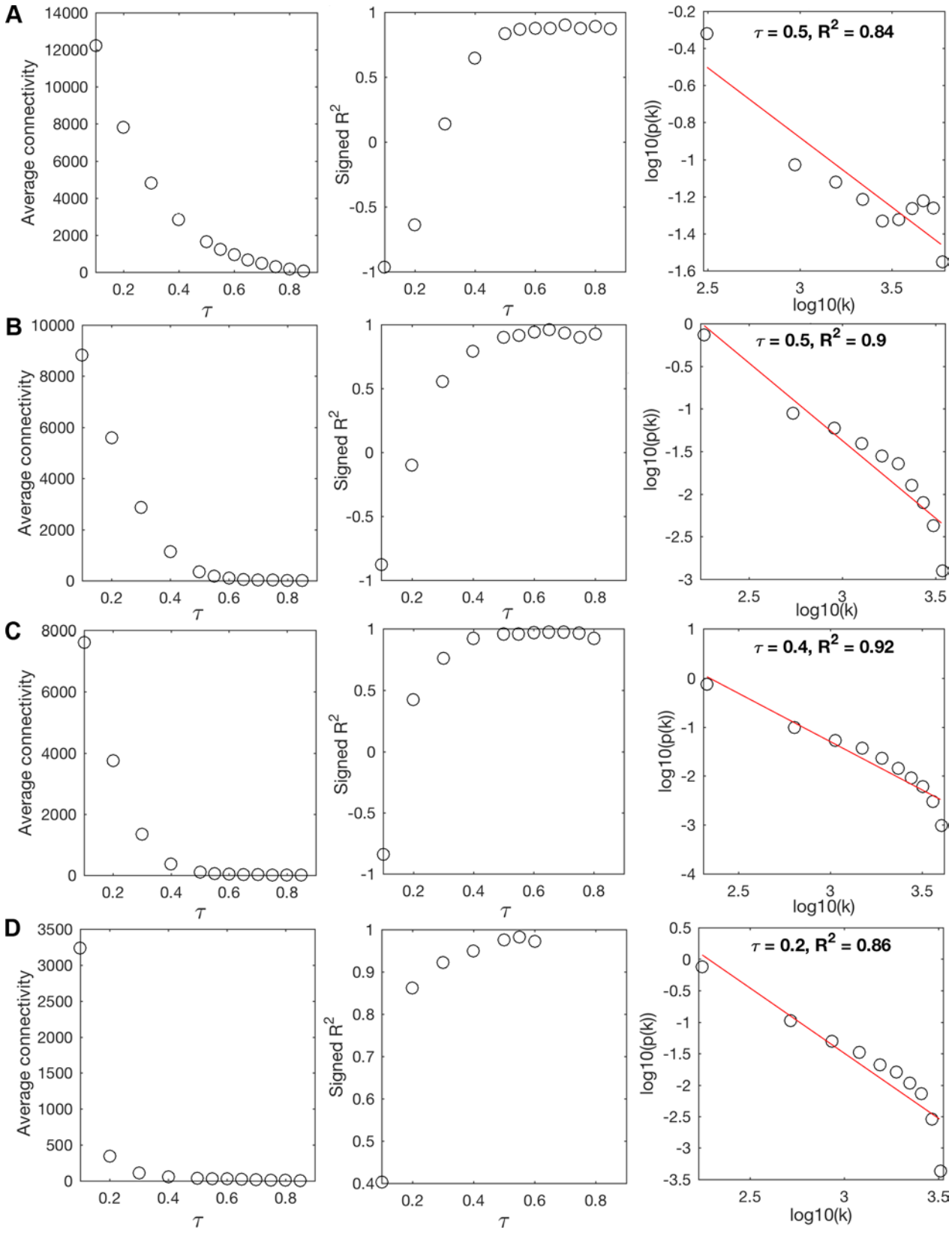

Figure S24. Analysis of topological properties of the constructed gene-gene interaction networks under different choices of the threshold τ. Left panel: The average connectivity of each constructed network under different threshold τ. Middle panel: Signed R^2 varies with the threshold τ. Right panel: Connectivity distribution ($p(k)$) under the selected threshold used in this study. $p(k)$ is defined by the fraction of nodes in the network with connectivity k. (A) HEE data (B) AT2 data (C) LES data (D) HSMM data.

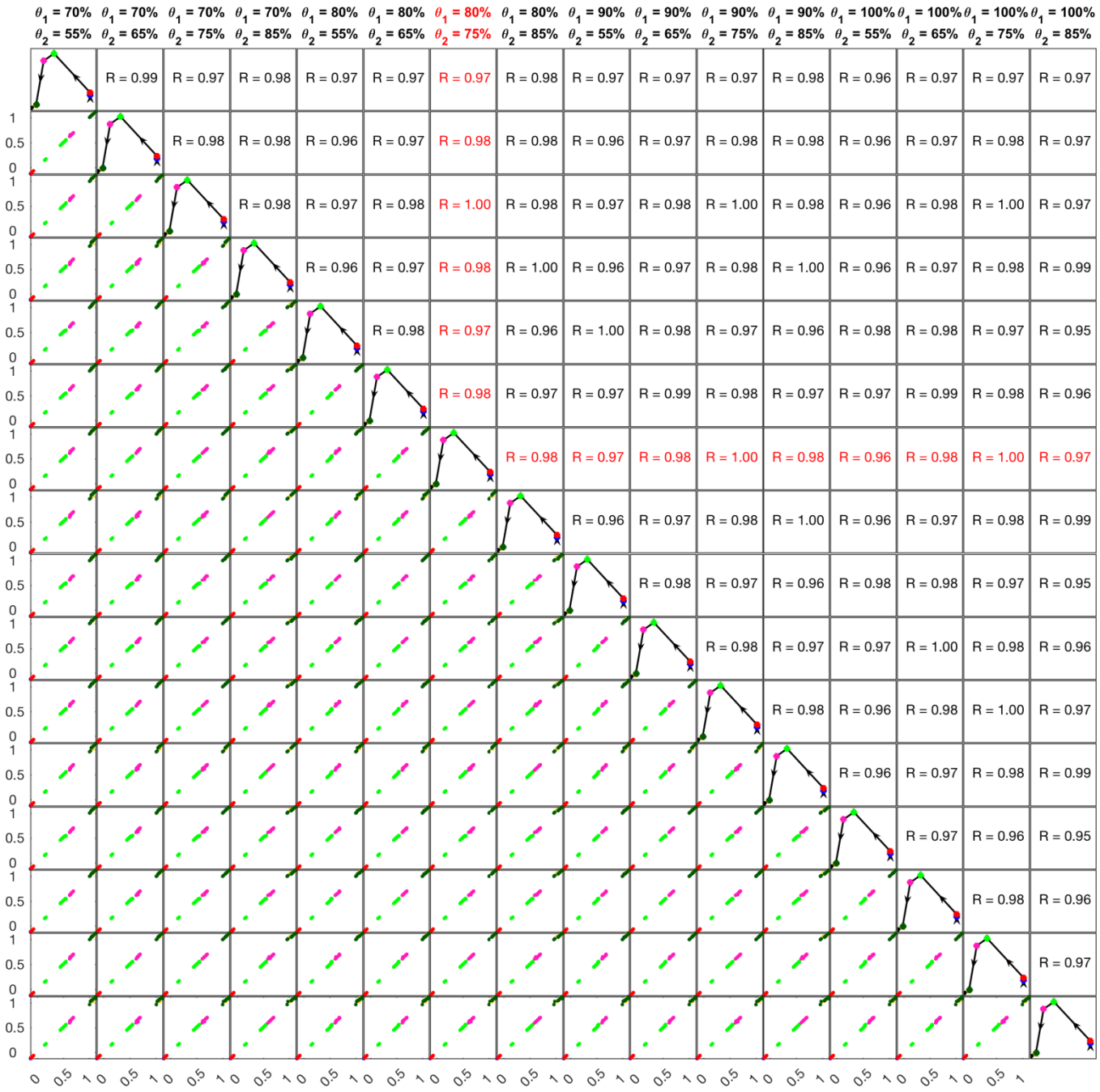

Figure S25. Robustness analysis of the inferred lineage and reconstructed pseudotime under different choices of θ_1 and θ_2 in scEpath using HEE data. The main diagonal displays the inferred lineage in each run (each dot is a cluster center). The lower part shows the scatter plot of the reconstructed pseudotime of single cells between pairwise runs (each dot is a cell). The upper part shows the Kendall rank correlation of the reconstructed pseudotime between pairwise runs. The default parameters in scEpath are indicated by a red font. The correlations between the default parameters and other choices of parameters are also indicated by a red font.

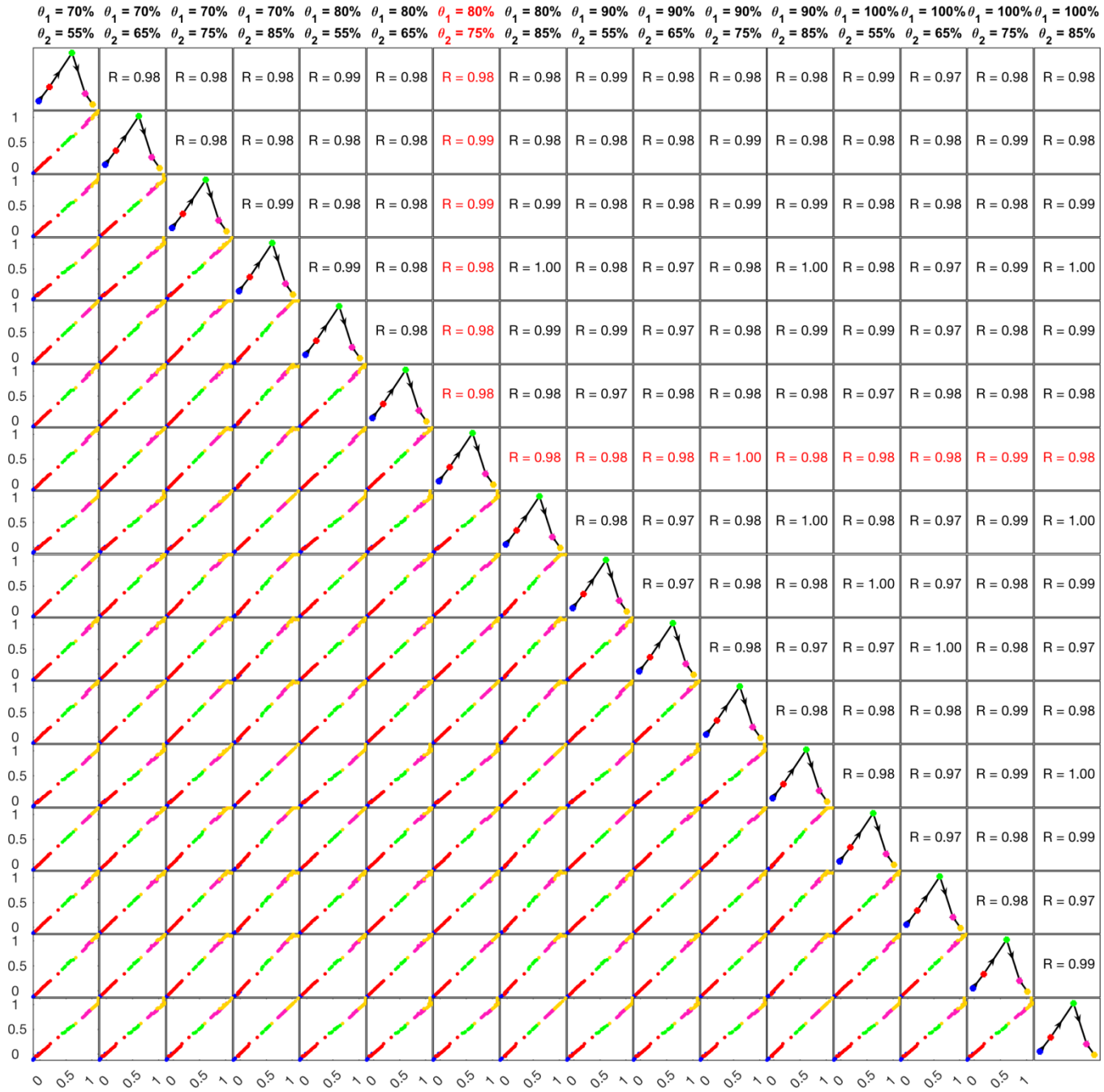

Figure S26. Robustness analysis of the inferred lineage and reconstructed pseudotime under different choices of θ_1 and θ_2 in scEpath using AT2 data. The main diagonal displays the inferred lineage in each run (each dot is a cluster center). The lower part shows the scatter plot of the reconstructed pseudotime of single cells between pairwise runs (each dot is a cell). The upper part shows the Kendall rank correlation of the reconstructed pseudotime between pairwise runs. The default parameters in scEpath are indicated by a red font. The correlations between the default parameters and other choices of parameters are also indicated by a red font.

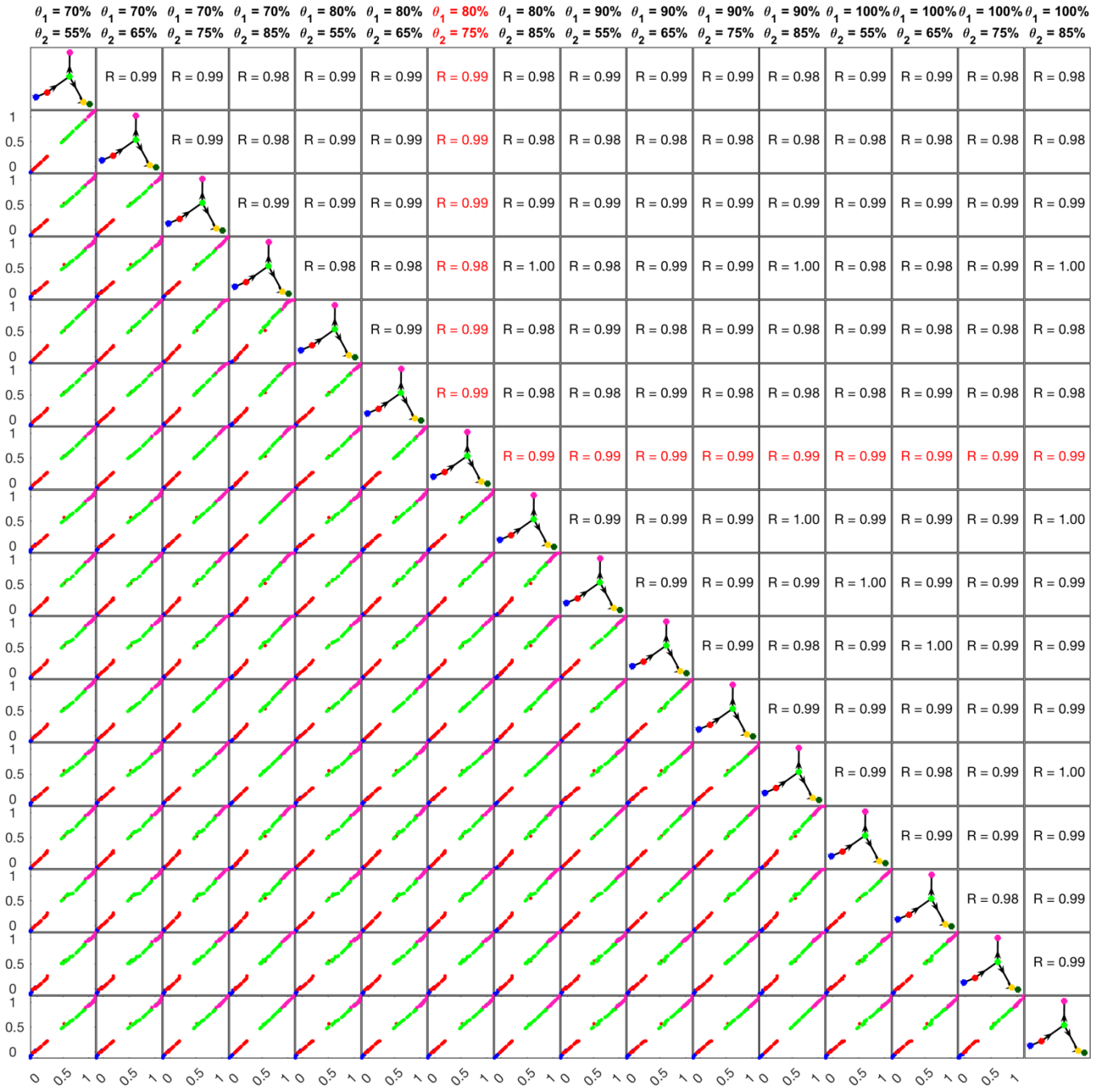

Figure S27. Robustness analysis of the inferred lineage and reconstructed pseudotime under different choices of θ_1 and θ_2 in AT1 branch using LES data. The main diagonal displays the inferred lineages in each run (each dot is a cluster center). The lower part shows the scatter plot of the reconstructed pseudotime of single cells between pairwise runs (each dot is a cell). The upper part shows the Kendall rank correlation of the reconstructed pseudotime between pairwise runs. The default parameters in scEpath are indicated by a red font. The correlations between the default parameters and other choices of parameters are also indicated by a red font.

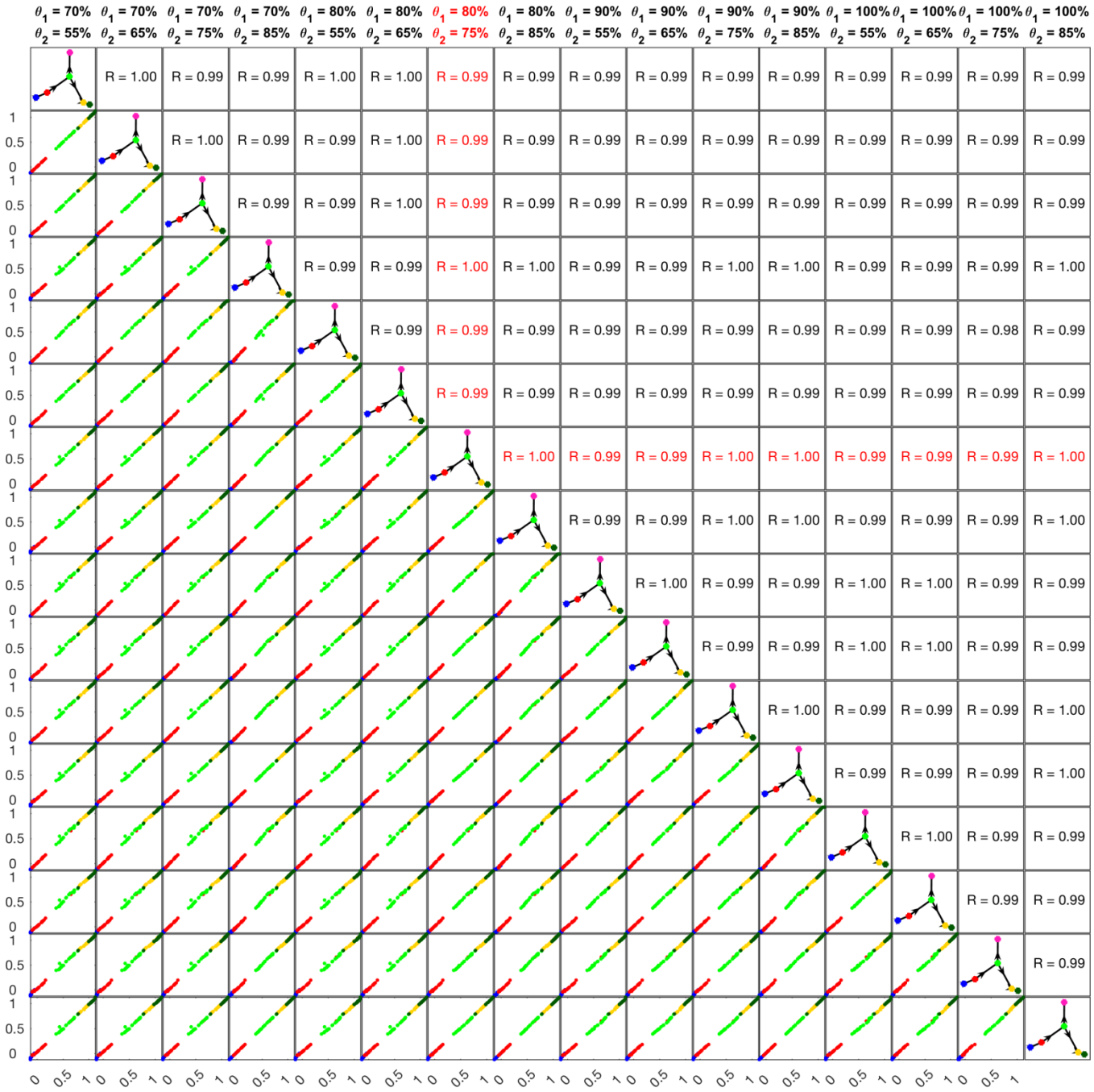

Figure S28. Robustness analysis of the inferred lineage and reconstructed pseudotime under different choices of θ_1 and θ_2 in AT2 branch using LES data. The main diagonal displays the inferred lineages in each run (each dot is a cluster center). The lower part shows the scatter plot of the reconstructed pseudotime of single cells between pairwise runs (each dot is a cell). The upper part shows the Kendall rank correlation of the reconstructed pseudotime between pairwise runs. The default parameters in scEpath are indicated by a red font. The correlations between the default parameters and other choices of parameters are also indicated by a red font.

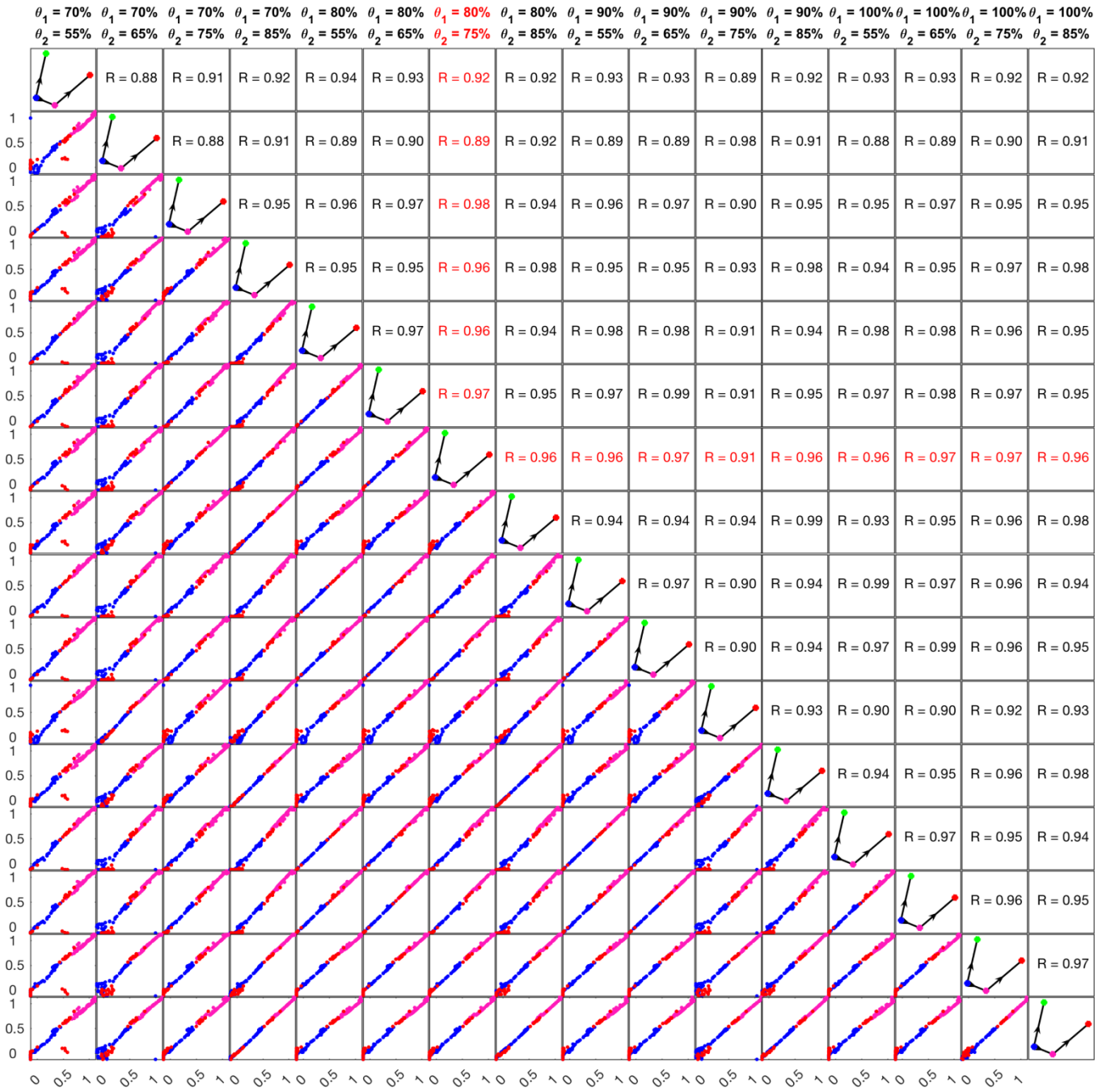

Figure S29. Robustness analysis of the inferred lineage and reconstructed pseudotime under different choices of θ_1 and θ_2 in muscle branch using HSMM data. The main diagonal displays the inferred lineages in each run (each dot is a cluster center). The lower part shows the scatter plot of the reconstructed pseudotime of single cells between pairwise runs (each dot is a cell). The upper part shows the Kendall rank correlation of the reconstructed pseudotime between pairwise runs. The default parameters in scEpath are indicated by a red font. The correlations between the default parameters and other choices of parameters are also indicated by a red font.

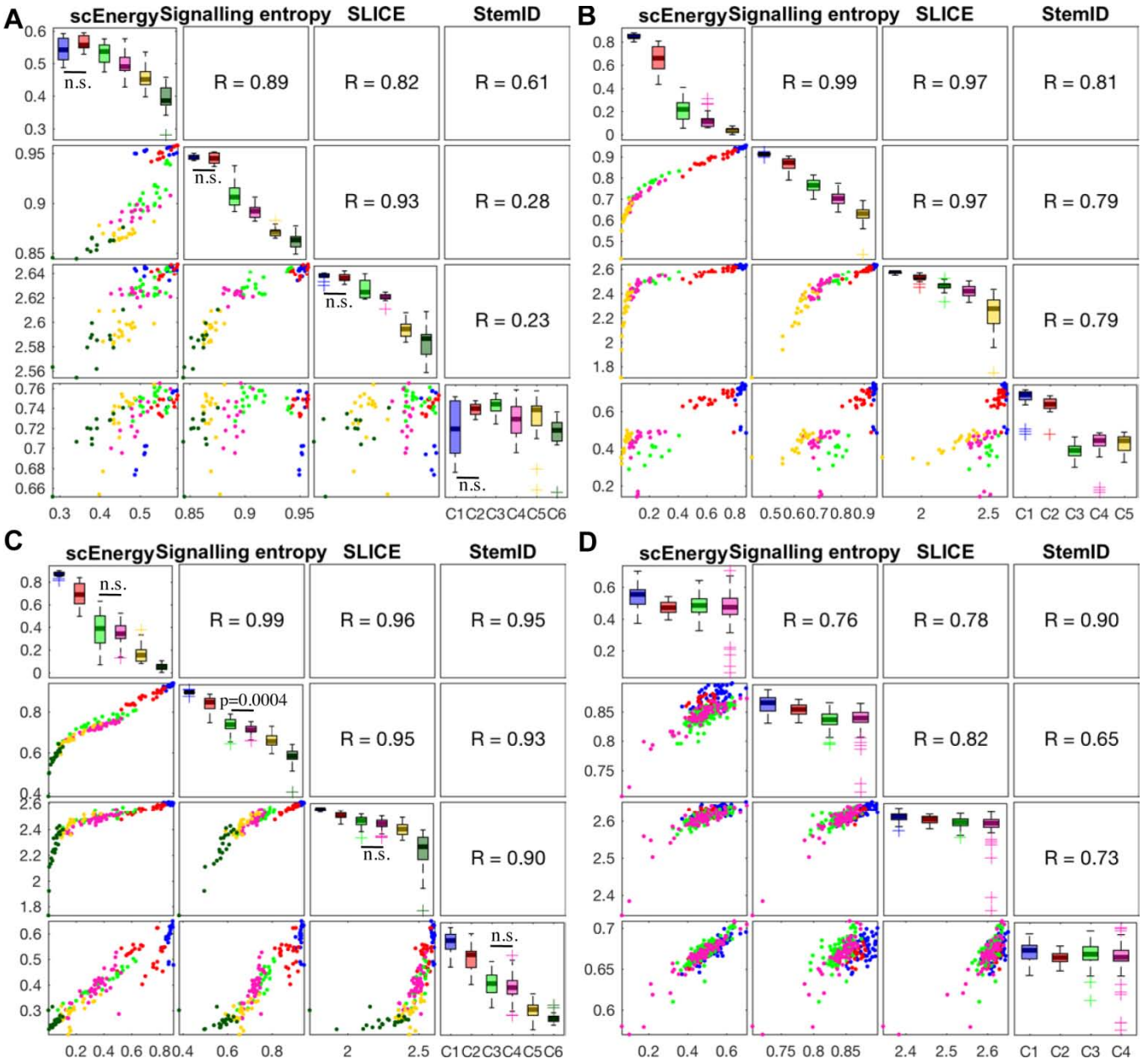

Figure S30. Comparative analysis of cell developmental states based on the energy calculated by scEnergy and the entropy calculated by signaling entropy (SCENT), SLICE and StemID using four experimental data. **(A)** Application to human early embryos (HEE) data. The main diagonal displays the distribution of energy/entropy among the cell clusters delineated by scEpath. The lower part shows the scatter plot of the energy/entropy between pairwise methods. The upper part shows the Spearman's correlation of the energy/entropy between pairwise methods. "n.s." means no significant difference (p-value > 0.1; a two sided Wilcoxon rank sum test). **(B)** Application to alveolar type 2 (AT2) data. **(C)** Application to lung epithelial specification (LES) data. "n.s." means no significant difference (p-value > 0.05; a two sided Wilcoxon rank sum test). **(D)** Application to human skeletal muscle myoblasts (HSMM) data. In sum, scEnergy showed high consistency with these existing measures. scEnergy, signaling entropy and SLICE successfully measured the developmental states. However, the transcriptome entropy proposed in StemID seems to be less closely correlated with the developmental states for HEE and AT2 data. Of note, StemID was originally proposed for unique molecular identifier (UMI) based scRNA-seq, while the four data were all read based. The results may be influenced by the compatibility of the data.

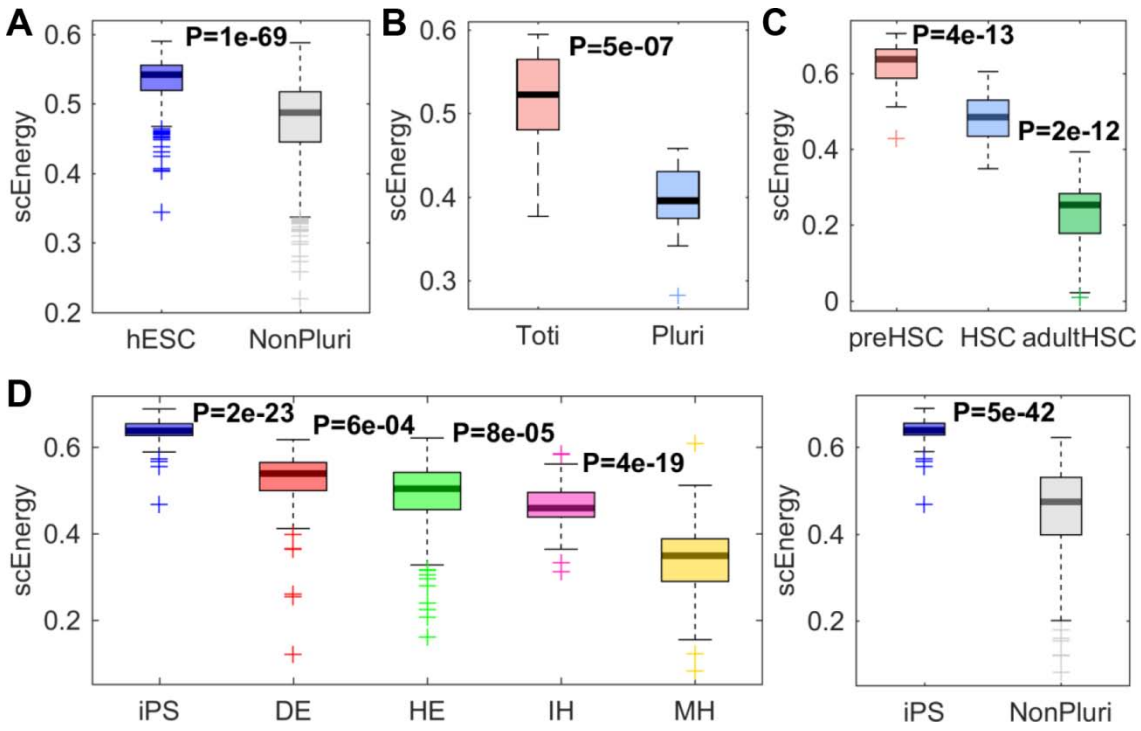

Figure S31. scEnergy correlates with developmental states of single cells. Box plot shows the comparison of scEnergy between the cells with different developmental states in four different scRNA-seq datasets. One tailed Wilcoxon rank-sum test P value is given. **(A)** The comparison is between the pluripotent human embryonic stem cells (hESC, n = 374) and non-pluripotent cells (NonPluri, n = 644, including 173 neural progenitor cells, 138 definite endoderm cells, 105 endothelial cells, 69 trophoblast-like cells and 159 human foreskin fibroblasts) from dataset GSE75748 (Chu, et al., 2016). **(B)** The comparison is between the totipotent cells (Toti, n = 44, including the cells from different stages: Oocyte, Zygote, 2-cell, 4-cell and 8-cell) and ICM cells (n = 12, from late blastocyst) from human embryonic development dataset GSE67120 (Yan, et al., 2013). **(C)** The comparison is between haematopietic stem cells (HSC, n = 37) and the precursors of HSC (pre-HSC, n = 50) or adult HSC (n = 30) capturing HSC formation during mouse embryogenesis from dataset GSE67120 (Zhou, et al., 2016). **(D)** scEnergy correlates with the hepatocyte-like lineage progression from the induced pluripotent stem (iPS) cells in the two-dimensional culture during human liver development from dataset GSE81252 (Camp, et al., 2017). **Left Panel:** Comparisons of scEnergy against different cell types: iPS (day 0, n = 80), DE (definitive endoderm, day 6, n = 70), HE (hepatic endoderm, day 8, n = 113), IH (immature hepatoblast-like, day 14, n =81) and MH (mature hepatocyte-like, day 21, n = 81). **Right panel:** The comparison is between iPS (n = 80) and all other (non-pluripotent, n = 345) cells.

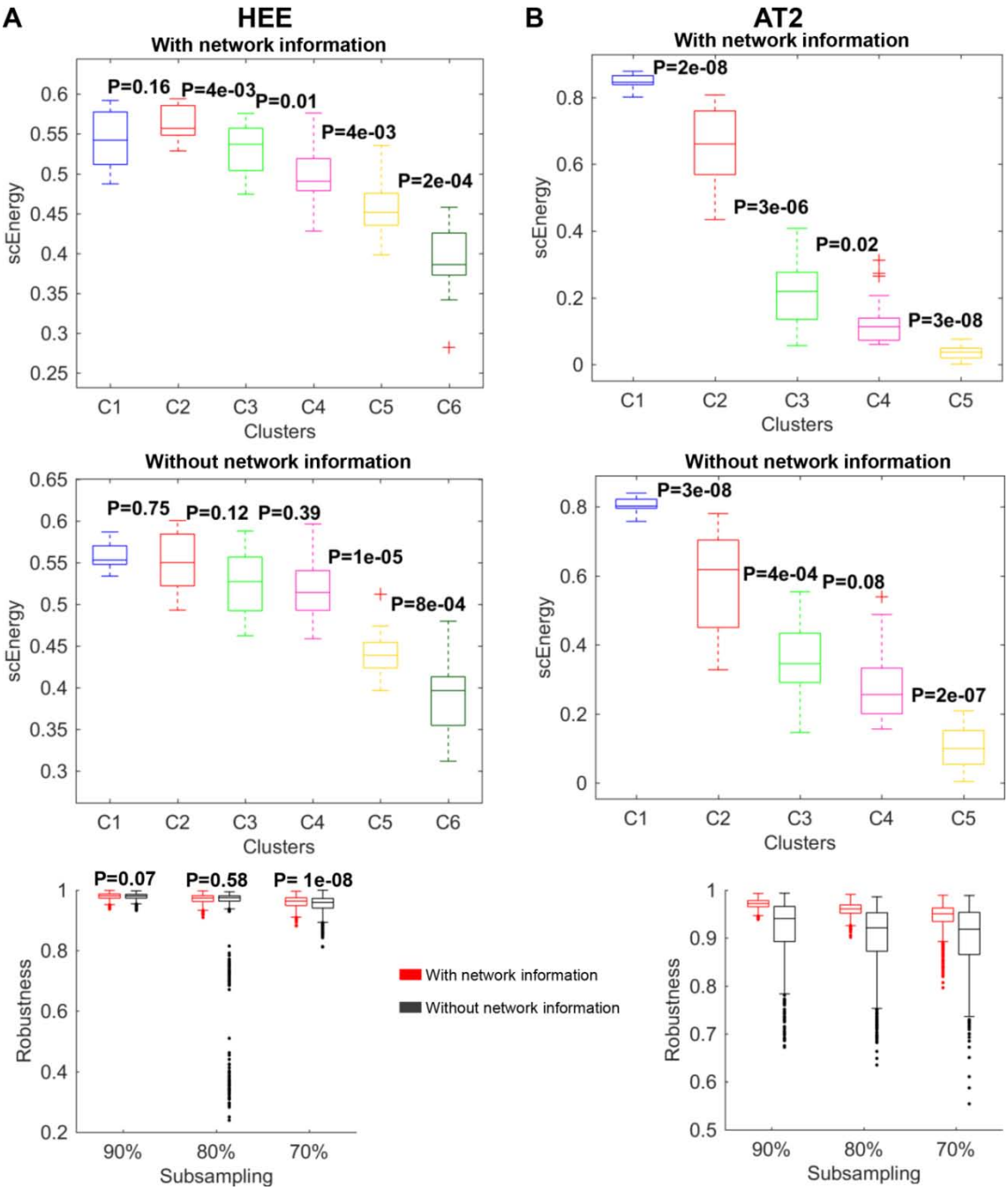

Figure S32. Comparison of the performance of scEpath with network information or not. **(A)** Application to HEE data. **Top:** Comparison of scEnergy distribution among the cell clusters delineated by scEpath with network information. Two tailed Wilcoxon rank-sum test P-value of two adjacent clusters is given. **Middle:** Comparison of scEnergy distribution calculated without network information (i.e., the simplified definition, see Supplementary Note 7). There is a significant difference between C2 (containing cells from 4-cell stage) and C3 (containing cells from 8-cell stage) with network information, while no significance was observed in the case without network information. Similar observation was shown between C3 and C4 (containing cells from Morulae). **Bottom:** Comparison of robustness (measured by PRS) of pseudotemporal ordering under repeated subsampling of the cells from HEE data. We used the Pseodotemporal Ordering Score (PRS) to calculate the similarity of each pair of pseudotemporal ordering runs under repeated subsampling of 90, 80, or 70% of the total number of cells. In each scenario, the original data was independently subsampled 50 times, which resulted in a 50 by 50 (upper triangular) similarity matrix for each method. One tailed Wilcoxon rank-sum test P-value (testing whether PRS is significant higher in the case with network information compared to the case without network information) is given. **(B)** Application to AT2 data. The significance level (P-value) of the case without network information increases compared to the case with network information. Particularly, there is a significant difference between C3 and C4 with network information, while no significance was observed in the case without network information.

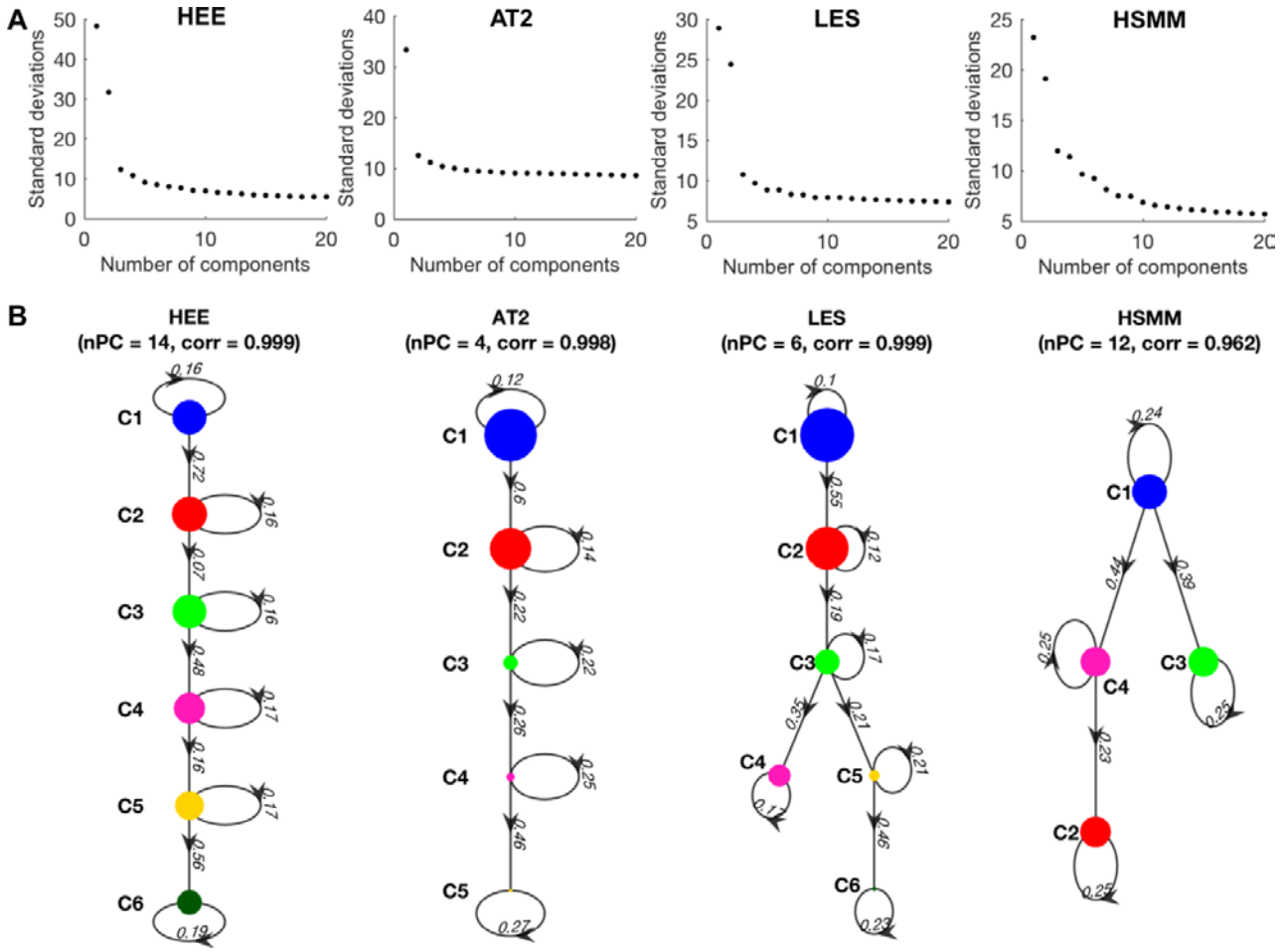

Figure S33. Robustness analysis of the number of components in the calculation of transition probability and cell lineages. **(A)** Standard deviations of the first 20 components in the four experimental datasets. Standard deviations of the first two PCs were significant larger than the remaining components. **(B)** Inferred lineages using the adaptively selected significant components based on the optimal hard threshold for singular values. In each example, the number of significant components (nPC) was given and the Pearson's correlation (corr) of transition probability matrix between the "optimal" one and the original one (using the first two components) was shown.

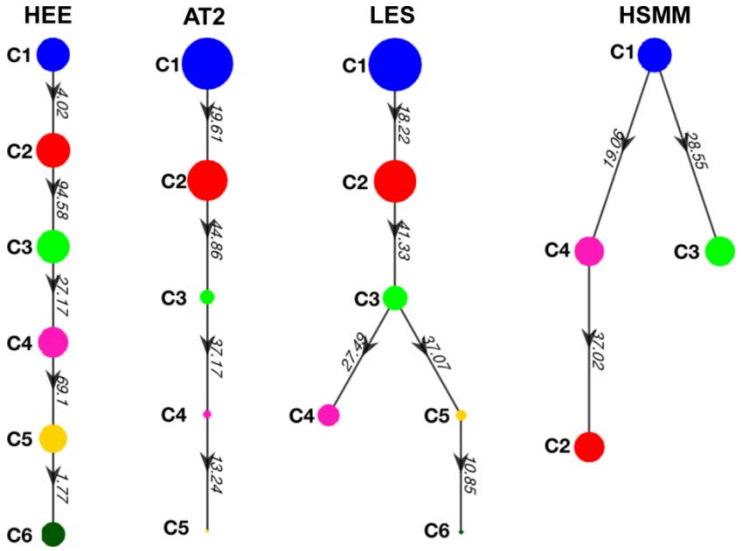

Figure S34. Inference of cell lineages by taking the Euclidean distance of pairs of metacell centroids as a weight of the edge. The same cell lineages were observed as our original ones in the four datasets.

REFERENCES

Albert, R. (2005) Scale-free networks in cell biology, *J. Cell Sci.*, **118**, 4947-4957.

Aldous, D. and Fill, J. (2002) Reversible Markov chains and random walks on graphs (https://www.stat.berkeley.edu/users/aldous/RWG/book.html). Berkeley.

Banerji, C.R.*, et al.* (2013) Cellular network entropy as the energy potential in Waddington's differentiation landscape, *Sci Rep*, **3**, 3039.

Barabasi, A.L., Gulbahce, N. and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease, *Nat. Rev. Genet.*, **12**, 56-68.

Camp, J.G.*, et al.* (2017) Multilineage communication regulates human liver bud development from pluripotency, *Nature*, **546**, 533-538.

Chu, L.F.*, et al.* (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm, *Genome Biol*, **17**, 173.

Gavish, M. and Donoho, D.L. (2014) The optimal hard threshold for singular values is 4/root 3, *IEEE. T. Inform. Theory*, **60**, 5040-5053.

Gibbons, A. (1985) *Algorithmic graph theory*. Cambridge university press.

Grun, D.*, et al.* (2016) De novo prediction of stem cell identity using single-cell transcriptome data, *Cell Stem Cell*, **19**, 266-277.

Guo, M.*, et al.* (2017) SLICE: determining cell differentiation and lineage based on single cell entropy, *Nucleic Acids Res*, **45**, e54.

Haghverdi, L.*, et al.* (2016) Diffusion pseudotime robustly reconstructs lineage branching, *Nat Methods*, **13**, 845-848.

Ji, Z. and Ji, H. (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, *Nucleic Acids Res*, **44**, e117.

Jin, S.*, et al.* (2014) Characterizing and controlling the inflammatory network during influenza A virus infection, *Sci. Rep.*, **4**, 3799.

Jin, S., Wu, F.-X. and Zou, X. (2017) Domain control of nonlinear networked systems and applications to complex disease networks, *Discrete Contin. Dyn. Syst. Ser. B*, **22**, 2169 - 2206.

Li, C., Hong, T. and Nie, Q. (2016) Quantifying the landscape and kinetic paths for epithelial-mesenchymal transition from a core circuit, *Phys. Chem. Chem. Phys.*, **18**, 17949-17956.

Logan, T.T., Rusnak, M. and Symes, A.J. (2015) Runx1 promotes proliferation and neuronal differentiation in adult mouse neurosphere cultures, *Stem Cell Res*, **15**, 554-564.

Moris, N., Pina, C. and Arias, A.M. (2016) Transition states and cell fate decisions in epigenetic landscapes, *Nat Rev Genet*, **17**, 693-703.

Qiu, X.*, et al.* (2017) Reversed graph embedding resolves complex single-cell trajectories, *Nat Methods*, **14**, 979-982.

Subramanian, A.*, et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

Teschendorff, A.E. and Enver, T. (2017) Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome, *Nat Commun*, **8**, 15599.

Trapnell, C.*, et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, *Nat Biotechnol*, **32**, 381-386.

Treutlein, B.*, et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq, *Nature*, **509**, 371-375.

Ustiyan, V.*, et al.* (2012) Foxm1 transcription factor is critical for proliferation and differentiation of Clara cells during development of conducting airways, *Dev. Biol.*, **370**, 198-212.

von Luxburg, U. (2007) A tutorial on spectral clustering, *Stat. Comput.*, **17**, 395-416.

Wagner, A., Regev, A. and Yosef, N. (2016) Revealing the vectors of cellular identity with single-cell genomics, *Nat Biotechnol*, **34**, 1145-1160.

Wang, B.*, et al.* (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning, *Nature Methods*, **14**, 414-416.

Yan, L.*, et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, *Nat Struct Mol Biol*, **20**, 1131-1139.

Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis, *Stat Appl Genet Mol Biol*, **4**, Article17.

Zhang, H.M.*, et al.* (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors, *Nucleic Acids Res*, **43**, D76-81.

Zhou, F.*, et al.* (2016) Tracing haematopoietic stem cell formation at single-cell resolution, *Nature*, **533**, 487-492.