Published online 20 December 2016

Nucleic Acids Research, 2017, Vol. 45, No. 7 ***e54***
doi: 10.1093/nar/gkw1278

SLICE: determining cell differentiation and lineage based on single cell entropy

Minzhe Guo[1,†], Erik L. Bao[2,†], Michael Wagner[3], Jeffrey A. Whitsett[1] and Yan Xu[1,3,*]

[1]The Perinatal Institute, Section of Neonatology, Perinatal and Pulmonary Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA, [2]Harvard Medical School, Boston, MA 02115, USA and [3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

Received November 22, 2016; Editorial Decision December 01, 2016; Accepted December 07, 2016

ABSTRACT

A complex organ contains a variety of cell types, each with its own distinct lineage and function. Understanding the lineage and differentiation state of each cell is fundamentally important for the ultimate delineation of organ formation and function. We developed SLICE, a novel algorithm that utilizes single-cell RNA-seq (scRNA-seq) to quantitatively measure cellular differentiation states based on single cell entropy and predict cell differentiation lineages via the construction of entropy directed cell trajectories. We validated our approach using three independent data sets with known lineage and developmental time information from both *Homo sapiens* and *Mus musculus*. SLICE successfully measured the differentiation states of single cells and reconstructed cell differentiation trajectories that have been previously experimentally validated. We then applied SLICE to scRNA-seq of embryonic mouse lung at E16.5 to identify lung mesenchymal cell lineage relationships that currently remain poorly defined. A two-branched differentiation pathway of five fibroblastic subtypes was predicted using SLICE. The present study demonstrated the general applicability and high predictive accuracy of SLICE in determining cellular differentiation states and reconstructing cell differentiation lineages in scRNA-seq analysis.

INTRODUCTION

Organogenesis is dependent on the progression of cells from less to more differentiated states (1). Cell fate decisions during differentiation are largely operative at the level of individual cells. Each cell has its own dynamically programed lineage trajectory path that is influenced by its developmental stages, epigenetic status, cell cycle, biological function and microenvironment (2,3). Determining the differentiation/stemness states of individual cells enables prediction of the dynamic genetic networks regulating cellular activities during organogenesis, repair and disease. Traditionally, the status of cell differentiation or stemness is determined by the expression of known differentiation markers (1) or by morphological features of cells (4,5) (e.g. size and shape). As such, these measurements rely on previous characterizations of known cell types, which may not be suitable for the study of novel or dynamically changing cell populations during organogenesis or in disease. Recent advances in single-cell RNA-seq (scRNA-seq) provide the feasibility of measuring the cellular heterogeneity and dynamic changes of individual cells during organ formation (6–8). Several algorithms, e.g. Wanderlust (1), Monocle (7), SCUBA (9) and Waterfall (10), have been developed to reconstruct lineage relationships of single cells from scRNA-seq data. However, these methods pseudotemporally classify cells based on transcriptome similarity rather than individual cell states. As such, they still require the use of external knowledge, such as time information, cell identity or marker gene expression, in order to determine the start and end points of dynamic processes and the directions of inferred pseudotemporal cell orderings.

We developed SLICE (Single Cell Lineage Inference Using Cell Expression Similarity and Entropy), a novel method for quantitatively measuring differentiation states of individual cells and reconstructing their lineages from scRNA-seq data. SLICE consists of two major functions: measuring cell differentiation states based on the calculation of single cell entropy (scEntropy) and predicting cell differentiation lineages by reconstructing cell trajectories directed by scEntropy-derived differentiation states (Figure 1). Entropy has been widely used in statistical mechanics, thermodynamics, and information theory as a measure of disorder or uncertainty in a system. Entropy is also a useful measure of cellular heterogeneity: cells with low entropy exhibit narrow, well-defined patterns of mRNA and protein expression (under strict regulatory constraints), whereas those with high entropy have broad, diverse patterns of expression (under weaker regulatory constraints and therefore

*To whom correspondence should be addressed. Tel: +1 513 636 8921; Fax: +1 513 636 7868; Email: yan.xu@cchmc.org
†These authors contributed equally to the work as first authors.

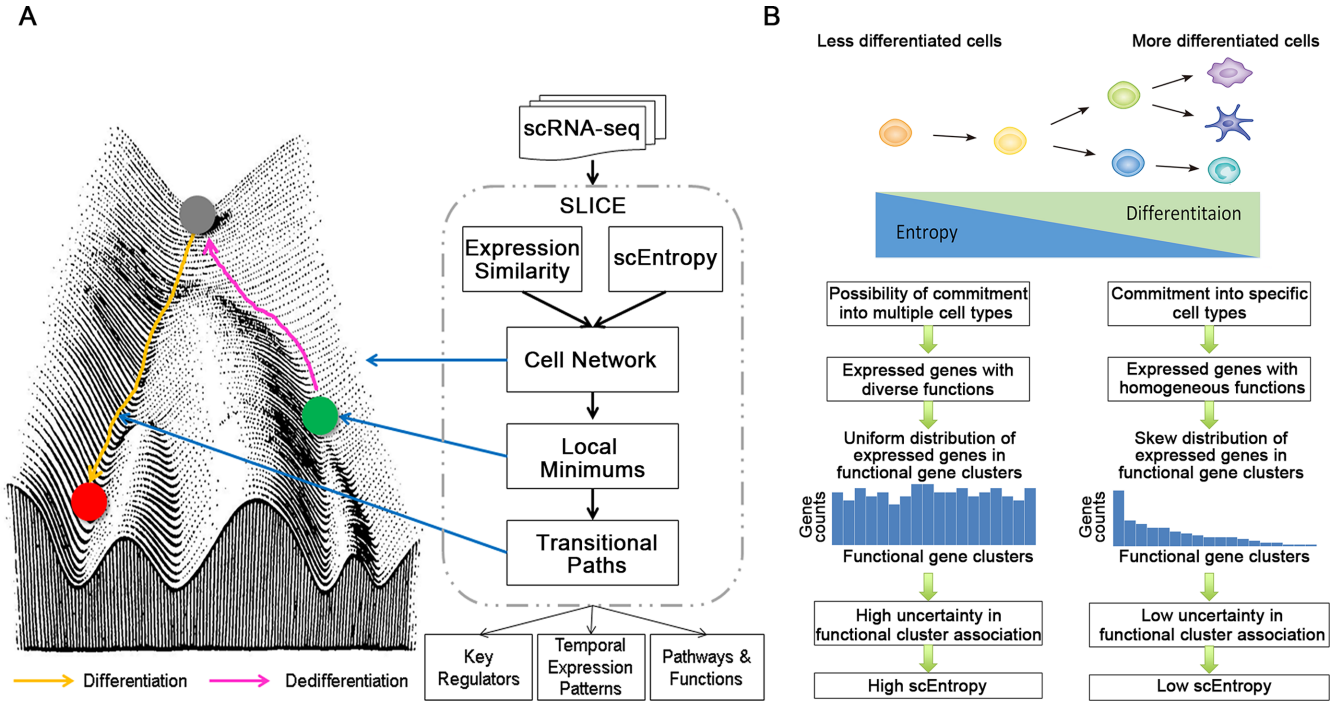

Figure 1. The design of the SLICE algorithm. (**A**) The schematic flow of the SLICE algorithm. (**B**) The schematic flow of scEntropy.

have multiple potentials) (11). Here, we propose that entropy inversely correlates with cell differentiation state: high entropy is associated with higher functional uncertainty and more differentiation potential (cell stemness), while low entropy is associated with more differentiated states with well-defined cell fates and functionalities (Figure 1B). Here, the term 'entropy' is not in reference to the noise or disorder in gene expression, but rather refers to the multiple potentials or uncertainty in a biological system. The algorithm assumes that cell-selective gene expression patterns during organogenesis are directly related to the diversity of active cellular functions. Undifferentiated cells express genes associated with diverse and heterogeneous functions in order to maintain multiple potentials of possible cell fate decisions. These cells thus maintain high scEntropy that derives from a more uniform distribution of functional activation, in which the activation probabilities of all functional classes are relatively equally distributed. In contrast, more fully differentiated cells express genes with restricted cellular functions and cell type commitment, thus exhibiting a skewed distribution of functional activation and a minimized scEntropy. Therefore, scEntropy quantifies the differentiation state of a given cell by measuring the uncertainty in the activation of its cellular functions. To calculate scEntropy, SLICE first computes pairwise gene functional similarities by applying Kappa statistics (12) to Gene Ontology annotations of the human or mouse genome, and then partitions genes into distinct functional clusters based on their functional similarity. The association between the functional clusters and gene expression in each single cell is determined using a Bayesian inference model with multinomial Dirichlet distribution as a prior to estimate the posterior probability distribution of the functional activation. The differentiation state of a cell represented by the degree of uncertainty in functional activation is then quantified by the entropy of its cell-specific posterior probability distribution. Using the differentiation states of individual cells measured by scEntropy, SLICE identifies relatively stable cell states, defined as the centroids of the cells with local minimum entropies, and then predicts transitional paths following entropy reduction between stable cell states to reconstruct cell lineages (Figure 1A). First, a cell–cell network is constructed with edges weighted by cellular expression profile similarity and nodes (cells) weighted by scEntropy. Neighboring cells in the network are grouped into cell clusters using a graph-based community detection algorithm, and local minimums within individual cell clusters are identified as relatively stable cell states in the network. Next, SLICE explicitly infers a lineage model from the data, representing the differentiation pathway among the stable cell states, and reconstructed the cell transitional path between two stable states by exploring a shortest path in the network, starting from the stable state with higher scEntropy, through a series of related cells, to the one with lower scEntropy.

To validate the accuracy and robustness of SLICE predictions, we applied SLICE to three independent scRNA-seq data sets (7,13,14) with known lineage and developmental time information. Results showed that scEntropy decreased with the progression of cellular differentiation stages; and the SLICE algorithm successfully reconstructed entropy-directed cell transitional paths that have been previously experimentally validated (7,13,14). We then applied SLICE to scRNA-seq data from fetal mouse lung to identify novel lung mesenchymal cell lineage relationships. Among five lung mesenchymal subsets identified in our previous

scRNA-seq study (15,16), we determined the relative order of cell differentiation status and identified a two-branched lineage model for lung fibroblast differentiation at the saccular phase of lung morphogenesis. SLICE offers unique features and significant improvements over existing methods of *in silico* lineage mapping, allowing the determination of cell differentiation states and differentiation lineage paths without the use of any external information. Results of this study demonstrated the general applicability and high predictive accuracy of SLICE in determining cellular differentiation states and reconstructing cell differentiation lineages in scRNA-seq analysis.

MATERIALS AND METHODS

SLICE consists of two major functions: quantitatively measuring cell differentiation state using single cell entropy and *in silico* reconstructing single cell lineages.

Single cell entropy (scEntropy) calculation

To calculate scEntropy, SLICE assumes that cellular gene expression patterns during organogenesis are directly related to the diversity of cellular functions. The scEntropy measurement derives from the level of uncertainty in the activation of cellular functions as reflected by the gene expression in individual cells, and it reflects the differentiation states of single cells. According to Boltzmann's entropy equation $H = k_B \log W$ (where k_B is the Boltzmann constant), the entropy (H) of a system positively correlates with the multiplicity (W) of the system, or the number of microstates corresponding to a given macrostate of the system. When the microstates of the system do not occur with equal probabilities, Boltzmann's equation can be rewritten to express entropy in terms of a probability distribution of microstates, i.e., $H = -k_B \sum_{i=1}^{W} p_i \log p_i$, in which p_i is the occurring probability of microstate i. A system with a flatter probability distribution has higher entropy, and a system with a more skewed probability distribution has lower entropy. SLICE applies this principle to calculate the entropy of single cells based on their probability distributions of functional activation. Less differentiated cells express genes with more diverse and heterogeneous functions to maintain the potential for multiple possible cell fate decisions, resulting in flatter probability distributions of functional activation and higher entropy. In contrast, more fully differentiated cells primarily express genes associated with a specific cell type commitment, therefore exhibiting a skewed probability distribution of functional activation, and establishing a minimized scEntropy. In this way, the scEntropy of a cell corresponds to the differentiation state of the cell.

Let S be the set of single cells from an scRNA-seq experiment, G be the set of all annotated genes detectably measured in the experiment, and $R_{ij} \geq 0$ be the RNA abundance (e.g. measured by FPKM, RPKM, or TPM values) of gene $i \in G$ in cell $j \in S$. Given an abundance threshold $\theta > 0$, we consider that a gene $i \in G$ is expressed in a cell $j \in S$ if $R_{ij} \geq \theta$. Thus, $G_j^\theta = \{i | i \in G, R_{ij} \geq \theta\}$ constitutes the set of genes expressed in cell j and $G^\theta = \cup_{j \in S} G_j^\theta$ is the set of genes expressed in at least one cell in S.

The scEntropy for each cell $j \in S$, denoted by H_j, is computed as an expected value (E) according to

$$H_j = E\left[-\sum_{k=1}^{m} p_j^k(B, F) \log\left(p_j^k(B, F)\right)\right]$$

in which B is a bootstrap sample of G^θ, $F = \{F_1, \ldots, F_m\}$ is a partition of B into m distinct functional groups, and $p_j^k(B, F)$ denotes the activation probability of functional group $F_k \in F$ based on the expression pattern of B in cell $j \in S$. The more genes from F_k expressed in cell j, the higher the probability that F_k is activated by the gene expression in j. Assuming that the activation probabilities $p_j^1, \ldots, p_j^m$ follow a multinomial distribution with a Dirichlet prior parameterized by $\alpha^1, \ldots, \alpha^m$, the Bayesian estimate (17) of the posterior probabilities is given by

$$p_j^k(B, F) = \frac{c_j^k(B, F) + \alpha^k}{C_j(B, F) + A},\ k \in \{1, \ldots, m\}$$

in which $c_j^k(B, F)$ denotes the number of genes from functional group F^k expressed in cell j that were included in bootstrap sample B, $C_j(B, F) = \sum_{k=1}^{m} c_j^k(B, F)$, and $A = \sum_{k=1}^{m} \alpha^k$. $C_j(B, F)$ denotes the total number of genes from all functional groups expressed in cell j that were included in the bootstrap sample B, so $C_j = |G_j^\theta \cap B|$. The parameters of the prior $\alpha^1, \ldots, \alpha^m$ are set to be proportional to the number of genes in each functional group in order to reduce the impact of the size imbalance of the functional groups on the entropy calculation.

$$\alpha^k = |F_k| \Big/ \sum_{t=1}^{m} |F_t|,\ k \in \{1, \ldots, m\}$$

To obtain F from B, SLICE first retrieves the functional annotations associated with each gene in B. In the present work, we used the DAVID Functional Annotation(12) (subset: GOTERM_BP_FAT) dataset for calculations. The GO_FAT database, developed as part of the Annotation Tool of the DAVID suite of bioinformatics resources terms (18), is derived from GO slim with additional filters to filter out the broadest terms so that they will not overshadow the more specific terms. These annotations are then used to compute $\kappa(x, y) \in [0, 1]$, $\forall x, y \in B$, the functional similarity between two genes x and y, using Kappa statistics (12). The functional similarity measurement using Kappa statistics was proposed by Huang *et al.* (18), and implemented in DAVID Bioinformatics Resource (https://david.ncifcrf.gov/) to measure gene pair functional similarity and identify functional clusters of genes. Based on the genome-wide gene-to-gene functional similarity matrix, SLICE then uses a K-means clustering algorithm to partition B into m distinct functional groups F with $d(x, y) = 1 - \kappa(x, y)$ as the distance measure.

In our analyses of all the four datasets, scEntropies were calculated using the following parameterization: $\theta = 1$, $|B| = 1000$, $m = \sqrt{|B|/2}$, and 100 bootstrap samples. Ribosomal genes were excluded from the scEntropy calculation.

Single cell lineage reconstruction

Cell differentiation is likely to transition through a sequence of intermediate states on the way to becoming fully mature. Single cells isolated at any particular developmental time may yield a mixture of cells at different stages in an unsynchronized manner: some cells are in more stable states while others may be in a transitional phase from one stable state to another. Multiple stable states may co-exist in a given scRNA-seq dataset. Using the differentiation states of individual cells measured by scEntropy, SLICE can unbiasedly determine the stable states in a given scRNA-seq dataset and reconstruct cell differentiation lineages by discovering entropy directed cell trajectories among the stable states. This is achieved through the following steps: (i) stable state identification, (ii) lineage model inference and (iii) cell trajectory reconstruction. A detailed schematic flow of using SLICE for lineage reconstruction can be found in Supplementary Figure S1.

Stable state identification. To identify stable states in a given scRNA-seq dataset, SLICE first divides cells into distinct clusters, representing distinct cell states or cell types in the dataset, and then identifies a closely-located core cell set with local minimum scEntropies within each cluster to define the stable state for the cluster.

We implemented two independent approaches for cell cluster identification. The first one is a graph based approach, in which we first construct a cell–cell network with edges weighted by cellular expression profile dissimilarity and nodes (cells) weighted by scEntropy, and then use a network community detection algorithm to partition the nodes in the network into distinct cell communities (clusters). We consider the set of single cells S as points in a reduced expression space obtained from a dimension reduction analysis (e.g. principal component analysis) of the full expression space defined by all genes detectably measured in scRNA-seq experiment. From this space, SLICE first constructs a complete weighted graph, where vertices represent cells, and edges are weighted by the Euclidean distance between cells in the expression space. Next, SLICE finds $T(S)$, the minimum spanning tree on the complete graph. If S contains a sufficient number of cells sampled without error from the cellular differentiation process underlying S, $T(S)$ is an accurate polygonal reconstruction of the differentiation process (7). Nevertheless, most current scRNA-seq datasets were under-sampled. For example, the C1 Fluidigm only captures 96 cells per run. Due to this sampling issue in scRNA-seq, some transitional states between cells may be missing. Therefore, SLICE performs a local wiring procedure to extend the tree $T(S)$ into a cell-cell network $N(S)$ by connecting cells in $T(S)$ that have distances smaller than a wiring threshold φ, thus rescuing cells representing potentially missed transitions due to the aforementioned cell sampling issue. By assigning the differentiation states of cells (estimated using scEntropy) as the node weights in $N(S)$, the cell-cell network $N(S)$ approximates the differentiation landscape underlying S. After constructing the cell-cell network $N(S)$, the Louvain algorithm (19) was utilized to partition the cells in $N(S)$ into multiple non-overlapping cell communities (clusters).

Alternatively, SLICE implemented a clustering based approach to partition cells into distinct groups. In this approach, we also consider the set of single cells S as points in a reduced expression space, and apply the Partitioning Around Medoids (PAM) algorithm (20) to divide cells into distinct clusters. The number of clusters can be determined by Gap statistic (20,21). We utilized the implementation of PAM and Gap statistic in the R 'cluster' package (https://cran.r-project.org/web/packages/cluster). This alternative approach is independent of the local wiring procedure and graph construction, providing users with more options to better fit individual datasets.

After identifying cell clusters, SLICE finds a closely-located core cell set with local minimum scEntropies within each cluster and identifies the centroid of the core cell set as a stable state. The scEntropy of a constructed stable state is the mean scEntropy of the core cells that define the stable state. Similarly, the expression of a gene in a stable state is the mean expression of the gene in the core cells that define the stable state. For the graph based approach, SLICE uses the Linear Prize-Collecting Steiner Tree (LPCST) problem (22) to identify the core cells in a cell cluster. Given an undirected graph with vertices associated with non-negative profits (node-prizes) and edges associated with non-negative costs, the LPCST problem finds a connected subgraph that maximizes a profit metric defined as the sum of all node-prizes taken into the solution minus the costs of the edges needed to establish the network. Here, to detect the core cell set in a cell cluster, SLICE considers the sub-graph of the cell cluster in $N(S)$ as the undirected graph in LPCST. The profits associated with the cluster cells (vertices) are set to be inversely proportional to their scEntropies, and the profit of a cell is set to 0 if the scEntropy of the cell is higher than the η quantile of the scEntropies of all cells in the cluster ($\eta = 0.25$ in our current analyses). SLICE then utilizes a fast heuristic approach implemented in the Bioconductor 'BioNet' package (23) to find a subnetwork that represents an approximated solution to the LPCST problem. The core cell set of the cluster is thus comprised of the cells in the subnetwork. For the clustering based method, SLICE uses the cells with top 25% of low entropy in each cluster as the core cell set.

Lineage model inference. Once we identified all the stable states, SLICE explicitly infers a lineage model for the data by constructing a directed minimum spanning tree among the stable states. This step facilitates the uncovering of heterogeneous branched lineage relationships among individual cells in a scRNA-seq dataset. Let $\Psi = \{\psi_1, \ldots, \psi_m\}$ be the set of identified stable states. SLICE constructs a complete weighted graph, where vertices represent stable states, and edges are weighted by the Euclidean distance between stable states in the reduced expression space. The directions of the edges are determined by scEntropy. According to Waddington's differentiation landscape, stable states with higher scEntropies correspond to shallower valleys in the landscape and represent less differentiated cell types. In contrast, stable states with lower scEntropy are the deeper valleys in the differentiation landscape and represent more differentiated cell types. Since we focus on inferring differentiation lineage model in this work, we keep the edges from

the stable states with higher scEntropy to the ones with lower scEntropy. SLICE then infers the differentiation lineage model $L(\Psi)$ by finding the minimum spanning tree on the complete directed graph of Ψ. One can also use SLICE to infer a dedifferentiation lineage model by setting the parameter to reverse the directions of edges.

Lineage dependent cell transitional path reconstruction. SLICE implemented two approaches to reconstruct the cell transitional paths: a shortest-path based approach and a principal curve based approach. Let $\Psi = \{\psi_1, \ldots, \psi_m\}$ be the set of stable states, $L(\Psi)$ be the inferred lineage model, and ψ_x and ψ_y be any two stable states where the scEntropy of ψ_x is higher than the one of ψ_y, SLICE defines the differentiation lineage from ψ_x to ψ_y as the shortest path from ψ_x to ψ_y in $L(\Psi)$, denoted by $P_{xy} = (\psi_x, \psi_{(1)}, \psi_{(2)}, \ldots, \psi_{(k)}, \psi_y)$, where $\psi_{(1)}, \psi_{(2)}, \ldots, \psi_{(k)}$ are the intermediate stable states in the path.

Next, we used a shortest-path approach to reconstruct the cell transitional path underlying the differentiation lineage from ψ_x to ψ_y. We first identify the cell transitional path between each pair of two successive stable states $\psi_{(q)}$ to $\psi_{(q+1)}$ in P_{xy} as the shortest path from $\psi_{(q)}$ to $\psi_{(q+1)}$ in $T(S)$ or $N(S)$ that we constructed in the 'stable state identification' step, and then we concatenate these individual pairwise transitional paths to form the full cell transitional path from ψ_x to ψ_y. The dynamic expression profile of a given gene following a specific cell transitional path is constructed by merging its expression with the neighboring cells along the path, addressing the high variance of gene expression in scRNA-seq data. Assume that the cell sequence $(\psi_x, v_1, v_2, \ldots, v_q, \ldots, v_n, \psi_y)$ represents the cells in the transitional path underlying the differentiation lineage from stable state ψ_x to stable state ψ_y, $\{v_1, v_2, \ldots, v_q, \ldots, v_n\} \subseteq S$, and $R_{i,q}$ represents the expression of gene i in cell v_q in the sequence. SLICE first identifies $D(q)$, the nearest neighbors (distance smaller than a threshold δ) of cell v_q in the path, considers cells in $D(q)$ as replicates of v_q, and then measures the gene expression along the path as $(\bar{R}_{i,D(x)}, \bar{R}_{i,D(1)}, \ldots, \bar{R}_{i,D(q)} \ldots \bar{R}_{i,D(n)}, \bar{R}_{i,D(y)})$ with variance $(\mathrm{var}(R_{i,D(x)}), \mathrm{var}(R_{i,D(1)}), \ldots, \mathrm{var}(R_{i,D(q)}), \ldots \mathrm{var}(R_{i,D(n)}), \mathrm{var}(R_{i,D(y)}))$, where $\bar{R}_{i,D(q)}$ and $\mathrm{var}(R_{i,D(q)})$are the mean and variance of $\{R_{i,t} | t \in D(q)\}$, respectively. In our analyses, δ was set to 0.8 quantile of the edge weights of $T(S)$ or $N(S)$; expression values were increased by 1 and log2 normalized.

Alternatively, SLICE implemented a principal curve based approach to reconstruct the cell transitional paths. In this approach, SLICE also defines the differentiation lineage from ψ_x to ψ_y as the shortest path from ψ_x to ψ_y in $L(\Psi)$, denoted by $P_{xy} = (\psi_x, \psi_{(1)}, \psi_{(2)}, \ldots, \psi_{(k)}, \psi_y)$, where $\psi_{(1)}, \psi_{(2)}, \ldots, \psi_{(k)}$ are the intermediates stable states in the path. Then SLICE identifies $S_{xy} \subseteq S$, the set of cells in the cell clusters whose stable states are in P_{xy}, fit a principal curve, f_{xy}, through S_{xy} in the reduced expression space, and project each cell in S_{xy} onto the principal curve in order to assign each cell to a projection index. The cell transitional path underlying P_{xy} is reconstructed by ordering cells in S_{xy} according to their principal curve projection indexes. The dynamic expression profile of a given gene following a specific cell transitional path is constituted by the expression of the gene in the cells in the transitional path. The fitted principal curve f_{xy} is a smooth curve that passes through the middle of S_{xy} in the reduced expression space in an orthogonal sense, and the projection index for each cell is the value λ where $f_{xy}(\lambda)$ is closest to the cell (24). We used the R 'princurve' package (https://cran.r-project.org/web/packages/princurve) for principal curve fitting and cell projection.

Identification of lineage dependent differentially expressed genes. To identify lineage dependent differentially expressed genes, SLICE first smoothens the expression profiles of each gene using cubic regression splines and fits to two models (model M1: gene expression changes in the path; and model M0: gene expression does not change in the path), assigns a *P*-value to each gene using the likelihood ratio test to compare the goodness of fit of the two models, and then identifies differentially expressed genes as the ones with high variance and low *P*-value. The generalized additive models with integrated smoothness estimation method (25) is utilized in SLICE for smoothening and model fitting. In our analyses, lineage dependent differentially expressed genes were identified using the following criteria: expressed ($>= 1$) in at least two cells in the lineage, variance of the fitted expression values greater than or equal to 0.5, and false-discovery rate adjusted *P*-value <0.1.

Identification of lineage dependent temporal patterns. Once we identified lineage dependent differentially expressed genes, temporal gene expression patterns in a lineage can be discovered by using a clustering algorithm to divide the differentially expressed genes into clusters (patterns) based on their model M1 predicted expression values.

Data sets, mRNA abundance estimates, and cell type/state assignments

We demonstrated the utility of SLICE using four independent scRNA-seq data sets from both *Homo sapiens* and *Mus musculus*. Dataset 1 consists of 101 mouse lung alveolar type 2 (AT2) cells collected at four developmental time points from the embryonic mouse lung, from E14.5 to adulthood (13). Dataset 2 consisted of 266 differentiating human skeletal muscle myoblasts (HSMM) (7). Dataset 3 consisted of 88 cells from seven stages in human early embryos (HEE) (14) including oocytes, zygotes, 2-cell embryos, 4-cell embryos, 8-cell embryos, morula and blastocysts. Dataset 4 contained single cells ($n = 79$) of five predicted fibroblastic subtypes from mouse lung at E16.5, including 'proliferative mesenchymal progenitor' cells, 'intermediate fibroblast 1' cells, 'intermediate fibroblast 2' cells, 'myofibroblast/smooth muscle-like' cells, and 'matrix fibroblast' cells (15,16). The scRNA-seq and cell type information of AT2, HSMM, HEE cells were obtained from Treutlein *et al.* (13), Trapnell *et al.* (7), Yan *et al.* (14), respectively. The scRNA-seq and cell type information of single cells from embryonic mouse lung at E16.5 were obtained from Du *et al.* (15); Guo *et al.* (16). The original quality control, alignment, quantification and cell type assignment of scRNA-seq data of each dataset were described in the corresponding manuscripts.

RESULTS

SLICE predicted differentiation states and lineages of mouse lung alveolar type 2 cells

We first used the AT2 data set to validate the SLICE algorithm. The scRNA-seq and cell type assignment of AT2 cells from four time points (E14.5, E16.5, E18.5 and adult) of mouse lung were obtained from Treutlein *et al.* (13). Cells ($n = 101$) with cell type assignments of 'AT2', '*Sftpc*+' or '*Sftpc*+*Scgb3a2*+' were selected for SLICE analysis. ERCC spike-ins were excluded from the analysis. The scEntropies of AT2 cells were calculated using the method and criteria as described in the Materials and Methods section. To construct the cell–cell network, SLICE first utilized the following criteria to select potentially informative genes, i.e. genes ($n = 2973$) expressed (FPKM > 1) in at least 30% of the cells and with a non-zero variance (greater than 0.5 variance in $\log_2$-transformed FPKM). Next, a principal component analysis was performed using the $\log_2$-transformed expression values of the selected genes, and the first two principal components were used to measure the Euclidean distance between cells for the network construction. The local wiring threshold φ was set to 0.5 fraction of the maximum weight of the edges in the minimum spanning tree.

We found that the scEntropy of differentiating AT2 cells consistently decreased along the four developmental time points from embryonic day E14.5 to adulthood (Figure 2A). When AT2 cells were ordered by their scEntropy in descending order, lung epithelial progenitor cell markers, e.g. *Sox11* and *Sox9,* decreased, whereas mature AT2 markers, e.g. *Sftpb* and *Sftpc,* increased (Figure 2B). Differentiation stages of AT2 cells determined by scEntropy aligned well with the known developmental times and the expression profiles of the associated RNA markers. To predict AT2 cell lineage transitions, SLICE first divided neighbouring cells into four clusters, each mainly containing cells from different time points; and the stable state in each cluster was also identified (Figure 3B). Then, SLICE inferred the lineage model that contains a single lineage (Figure 3C) and reconstructed its corresponding cell transitional path (Figure 3D) from the cluster C1, the stable state of which had the highest scEntropy, through cluster C2 and C3, to cluster C4, the stable state of which had the lowest scEntropy. Along the predicted path, the expression of AT2 cell differentiation markers, *Sftpc*, *Sftpb*, *Napsa* and *Slc34a2*, was increased (Figure 3E), while the expression of progenitor cell markers, e.g. *Sox11* and *Sox9*, and cell cycle genes, *Foxm1*, *Bub1* and *Top2a*, was decreased (Figure 3E), indicating that the reconstructed transitional path represents the progression of AT2 cell differentiation.

SLICE predicted differentiation states and lineages of single cells from differentiating human skeletal muscle myoblasts

We then validated the utility of SLICE using human skeletal muscle myoblasts (HSMM). The scRNA-seq of HSMM cells ($n = 271$) was obtained from Trapnell *et al.* (7). After evaluating the number of expressed genes (FPKM > 1) in each cell, we detected five cells ('T0_CT_A11', 'T0_CT_E10', 'T0_CT_C09', 'T48_CT_C02', 'T48_CT_H04") as outliers with a relatively low number of expressed genes. Outliers were defined as values lower than the 0.75 fraction of the 0.1 quantile of all values. The remaining 266 HSMM cells were used in the SLICE analysis. The scEntropies of HSMM cells were calculated using the method and criteria as described in the Methods section. To construct the cell-cell network, SLICE first performed a principal component analysis using the $\log_2$-transformed expression values of a set of marker genes known to be important to the myogenesis process (Supplementary Table S1), and then used the first two principal components to define the Euclidean distance between cells for the network construction. The local wiring threshold φ was set to the maximum weight of the edges in the minimum spanning tree.

The differentiation states and lineages of HSMM cells predicted by SLICE were consistent with previous analyses based on Monocle (7). The 'Proliferating cells' ($n = 96$) assigned by Monocle (7) were associated with higher scEntropy, while the previously defined 'differentiating myoblasts' ($n = 127$) were associated with lower scEntropy (Figure 2C). Four clusters of HSMM cells and the stable state in each cluster were identified (Figure 4B). Cluster C1 consisted of 'proliferating cells', cluster C2 contained both 'proliferating cells' and 'differentiating myoblasts', cluster C3 was comprised of 'differentiating myoblasts', and cluster C4 mainly consisted of the 'interstitial mesenchymal cells' previously defined by Monocle (7). SLICE inferred a two-branched lineage model (Figure 4C). Since the cluster C4 mainly contained contaminated cells that were excluded after pseudotime inference in the original analysis (7), we also only analyzed the lineage from cluster C1, through cluster C2, to cluster C3 cells. The cell transitional path following this lineage was reconstructed (Figure 4D). Expression patterns of myogenic differentiation markers along the SLICE-predicted transitional path (left, Figure 4E) largely reproduced the patterns according to pseudotime using Monocle (7) (right, Figure 4E), suggesting that the path obtained by entropy-reduction predicted the differentiation process of HSMM cells from proliferating to more differentiated myoblast cells.

SLICE inferred differentiation states and lineages of single cells from human early embryo development

Next we tested SLICE using single cells from very early stages of human embryo development. The scRNA-seq of single cells ($n = 90$) from seven stages of human early embryonic development, including oocyte ($n = 3$), zygote ($n = 3$), 2-cell embryo ($n = 6$), 4-cell embryo ($n = 12$), 8-cell embryo ($n = 20$), morula ($n = 16$) and late blastocyst ($n = 30$), were obtained from Yan *et al.* (14). Two cells from the morula stage, 'Morula #1 – Cell #3', 'Morula #1 – Cell #8', were excluded as outliers from our analysis since they were consistently clustered with cells from the late blastocyst stage in both the original analysis (14) and our previous analysis (16). The scEntropies of HEE cells were calculated using the method and criteria as described in the Methods section. To construct the cell–cell network for the HEE cells, SLICE first utilized the following criteria to select potentially informative genes, i.e. genes ($n = 10\,601$) expressed (RPKM>1) in at least 30% of the cells and with a non-zero variance (>0.5 variance in $\log_2$-transformed RPKM).

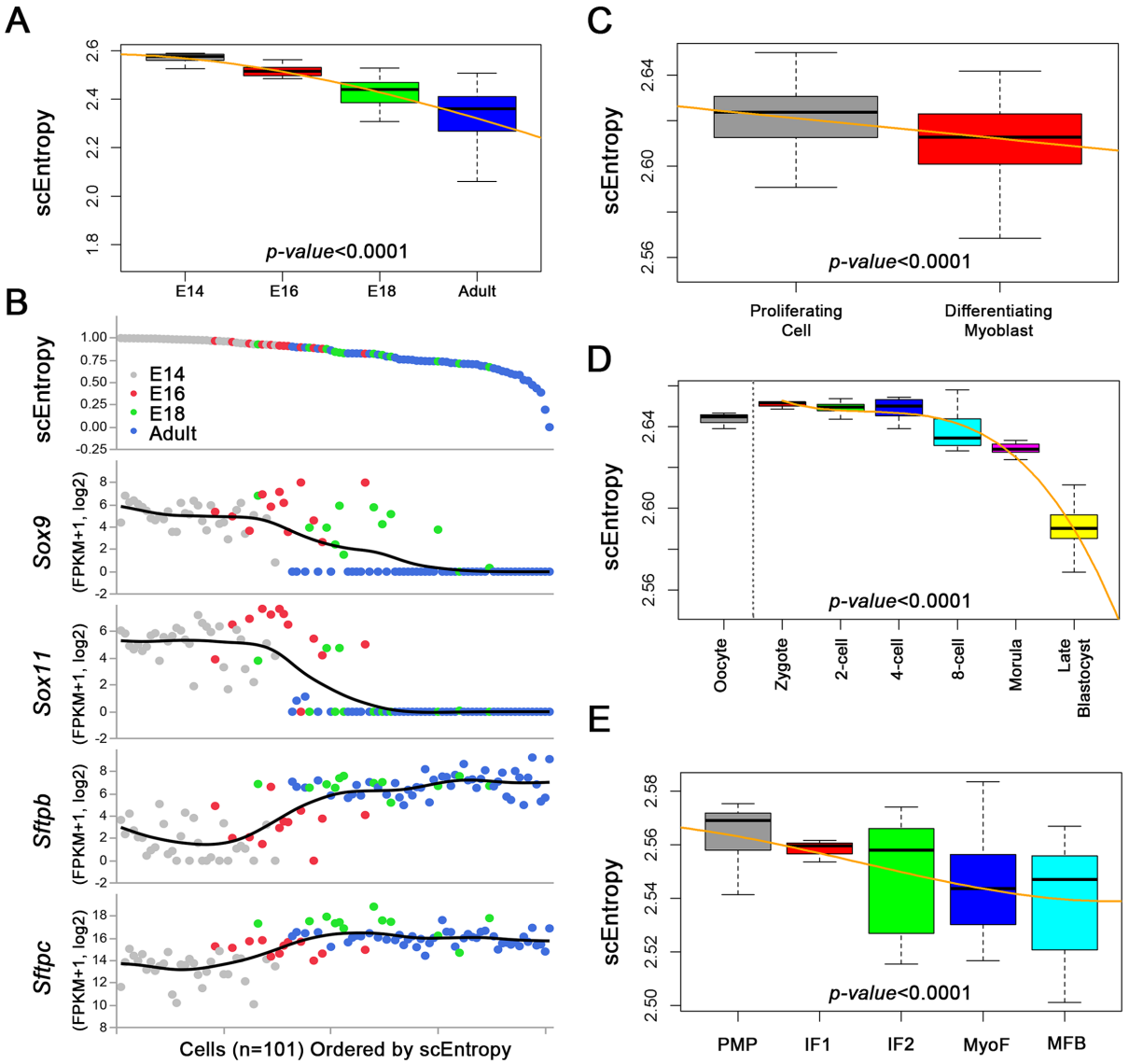

Figure 2. Single cell entropy (scEntropy) measures differentiation states of individual cells. (**A**) The scEntropies of lung alveolar type 2 (AT2, $n = 101$) cells significantly decreased along with mouse lung development. The cells were isolated from E14.5, E16.5, E18.5 and adult mouse lung (13). (**B**) The decrease of scEntropies of AT2 cells correlates with the progression of AT2 differentiation. The expression patterns of early progenitor markers (*Sox9* and *Sox11*) and mature AT2 markers (*Sftpb* and *Sftpc*) were used to validate the order of scEntropies. (**C**) The scEntropies of human skeletal muscle myoblast (HSMM, $n = 223$) cells significantly decreased during myoblast differentiation. The assignments of cells to proliferating cells ($n = 96$) and differentiating myoblast cells ($n = 127$) were pre-defined by Trapnell *et al.* (7). (**D**) The scEntropies of human early embryonic (HEE, $n = 88$) cells significantly decreased during human early embryo development. The cell developmental stages were defined by Yan *et al.* (14). **E**) The scEntropy measurement of the predicted fibroblast subtypes from E16.5 mouse lung (FB, $n = 79$). Fibroblast subtypes were defined by Guo *et al.* (16); PMP: proliferative mesenchymal progenitor, IF1: intermediate fibroblast 1, IF2: intermediate fibroblast 2, MyoF: myofibroblast/smooth muscle-like, MFB: matrix fibroblast. In (B), (C), (D) and (E), orange lines represent the polynomial fits of degree 3 of scEntropies using least squares regression and *P*-values were obtained using the Analysis of Variance (ANOVA), with *P*-value <0.0001 as the threshold for significance.

Next, a principal component analysis was performed using the $\log_2$-transformed expression values of the selected genes, and the first two principal components were used to measure the Euclidean distance between cells for the network construction. The local wiring threshold φ was set to 0.2 fraction of the maximum weight of the edges in the minimum spanning tree.

Starting from zygote stage, the scEntropy of HEE cells decreased throughout the progression of human embryo development (Figure 2D). SLICE identified four distinct cell clusters and reconstructed an entropy-reduction cell transition path that represented the progression of the human early embryo development from zygote to late blastocyst (Figure 5A–D), a prediction that was validated by the expression of marker genes associated with the inferred trajectory (Figure 5E).

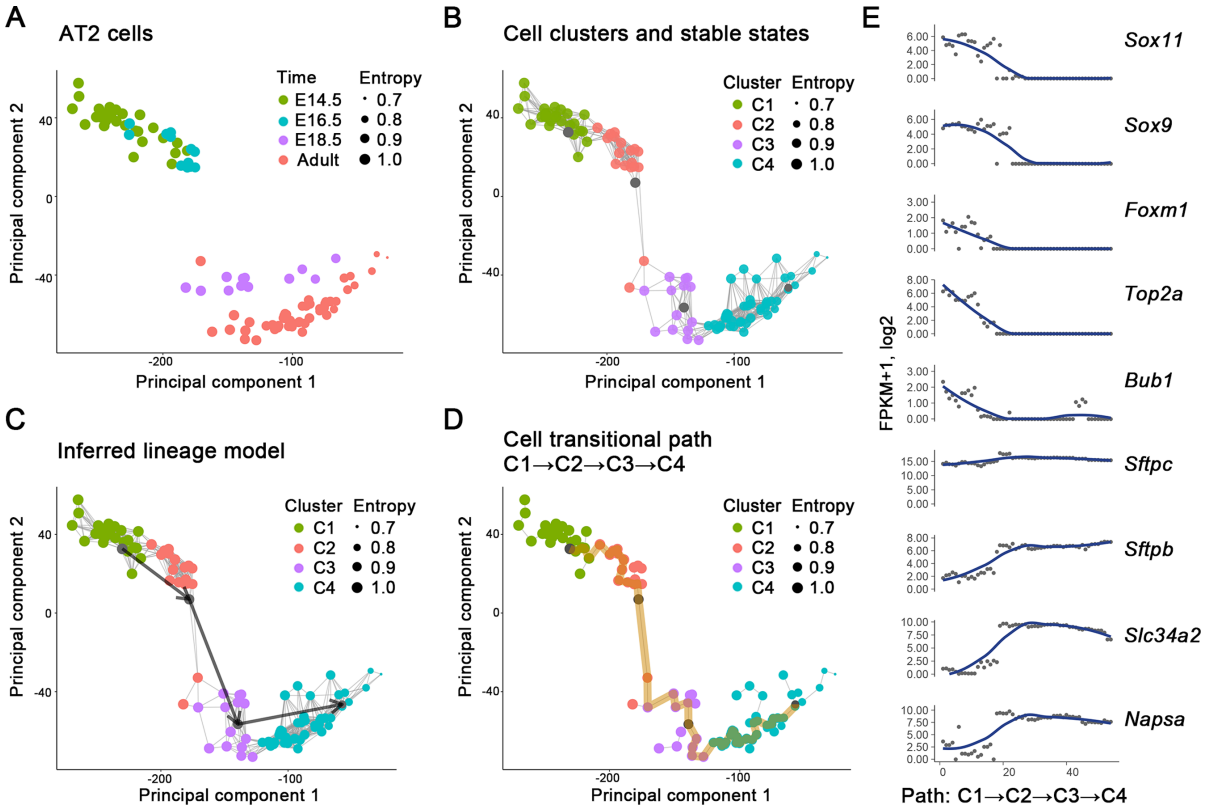

Figure 3. SLICE reconstructed the differentiation lineage of mouse lung alveolar type 2 (AT2) cells. (**A**) Principal component analysis (PCA) of AT2 cells ($n = 101$) isolated from E14.5, E16.5, E18.5 and adult mouse lung. Genes ($n = 2973$) expressed in at least 30% of the cells and with a non-zero variance in $\log_2$-transformed expression were used in the PCA analysis. (**B**) Four cell clusters and their stable states were identified by SLICE. (**C**) The lineage model of AT2 cells during lung development inferred by SLICE. (**D**) The reconstruction of transitional path from C1→C2 →C3→C4 of AT2 cells. The path was reconstructed by concatenating pairwise shortest-paths between successive stable states (C1, C2, C3 and C4) in the minimum spanning tree of the cells. (**E**) The expression patterns of known markers indicated that the reconstructed trajectory represented the progression of AT2 cell differentiation. In (A)–(D), sizes of cells were proportional to the max-normalized scEntropies of the cells, and dark-gray nodes represent the detected stable states.

SLICE predicted a two-branched differentiation pathway of lung fibroblasts at E16.5 mouse

While lineage relationships among pulmonary epithelial cells are increasingly understood in the developing mouse (13,26,27), identities and lineage relationships among mesenchymal cell populations remain poorly defined. Hierarchical relationships, dynamic gene expression patterns and associated functions of the diverse cell types comprising the fetal lung mesenchyme remain largely unknown. Here, we applied SLICE to scRNA-seq of embryonic mouse lung at E16.5 (15,16) to identify the lung mesenchymal cell lineage relationships at a developmental stage during which lung growth and differentiation associated with tissue morphogenesis are highly active. Unlike the AT2, HSMM and HEE datasets, in which cases cell type and developmental time stages are known, this is a cross-sectional dataset obtained from single cells isolated from whole lung at a single time point (E16.5 mouse lung). Since each single cell is a unique entity and the differentiation states among closely related individual cells are not synchronous, we believe that single cells analyzed at a given developmental time, such as the E16.5 time point in this dataset, will reveal cells representing distinct differentiation states.

Our previous analyses of scRNA-seq data from E16.5 mouse lung identified three distinct fibroblast subtypes and two intermediate fibroblasts from a total of 148 lung single cells, which we termed 'Proliferative Mesenchymal Progenitor' (PMP, $n = 25$), 'Myofibroblast/smooth muscle-like' (MyoF, $n = 16$), 'Matrix Fibroblast' (MFB, $n = 23$), 'Intermediate Fibroblast 1' (IF1, $n = 3$), and 'Intermediate Fibroblast 2' (IF2, $n = 12$) (15,16). We termed cells largely based on the marker expression; the real hierarchical relationship among these cells is unknown. Signature genes for each cell type were predicted (15,16). The scEntropies of these FB cells were calculated using the method and criteria as described in the Materials and Methods section. To construct the cell-cell network for the E16.5 fibroblast cells, SLICE performed a principal component analysis using the $\log_2$-transformed expression values of the signature genes predicted by the original analysis (15,16) for the five fibroblast subtypes and with a high variance (greater than 8 variance in $\log_2$-transformed FPKM) ($n = 252$, Supplementary Table S1), and then used the first two principal components to measure the Euclidean distances between cells for the network construction. The local wiring threshold φ was set to the 1.2 times of the maximum weight of the edges in the minimum spanning tree.

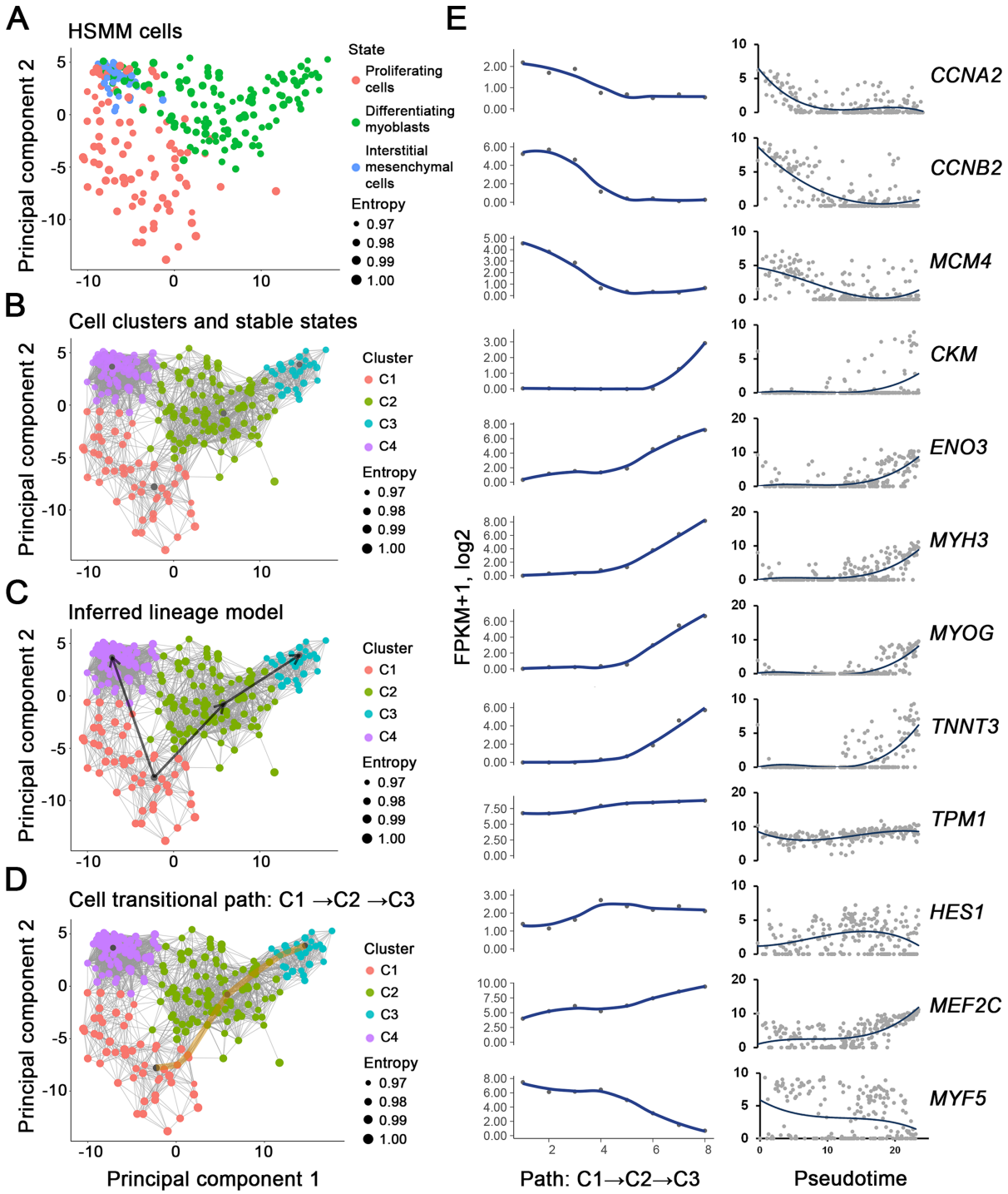

Figure 4. SLICE reconstructed the differentiation lineage of single cells from human skeletal muscle myoblasts. (**A**) Human skeletal muscle myoblast (HSMM) cells ($n = 266$) in the two-dimensional space derived from a principal component analysis using a set of marker genes (Supplementary Table S1). Cell states (indicated by three distinct colours) were defined by the original analysis (7). (**B**) Four cell clusters and their stable states identified by SLICE. (**C**) The lineage model of HSMM cells inferred by SLICE. (**D**) The reconstruction of transitional path from cluster C1, through cluster C2, to cluster C3 of HSMM cells. The path was reconstructed by connecting pairwise shortest-paths between successive stable states (C1, C2 and C3) in the cell-cell network. (**E**) The expression patterns of myogenic differentiation markers along the reconstructed transitional path (left) were consistent with the patterns based on pseudotime (right) using Monocle (7). In (A)–(D), sizes of cells were proportional to the max-normalized scEntropies of the cells, and dark-grey nodes represent the detected stable states.

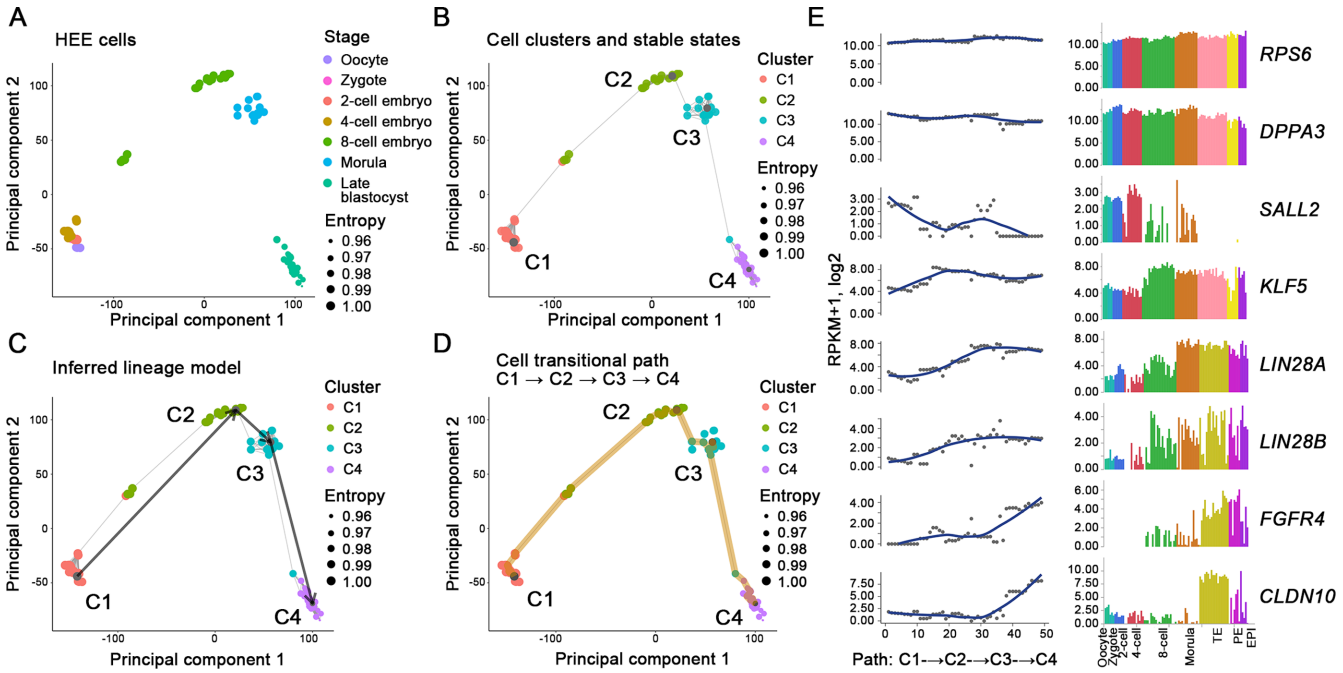

Figure 5. SLICE reconstructed the differentiation lineage of single cells from human early embryo (HEE) development. (**A**) HEE cells ($n = 88$) with development stage information in the two-dimensional space derived from a principal component analysis using genes ($n = 10\,601$) expressed in at least 30% of the cells and with a non-zero variance in $\log_2$-transformed expression. (**B**) Four cell clusters and their stable states identified by SLICE. (**C**) The inferred lineage model of HEE cells during human early embryo development. (**D**) The reconstruction of transitional path from cluster C1, through C2 and C3, to C4 of HEE cells. The path was reconstructed by concatenating pairwise shortest-paths between successive stable states (C1, C2, C3 and C4) in the minimum spanning tree of cells. (**E**) The expression patterns of markers along the inferred transitional path (left) were consistent with the stages defined in Yan *et al.* (14). TE, trophectoderm; PE, primitive endoderm; EPI, epiblast. In (A)–(D), sizes of cells were proportional to the max-normalized scEntropies of the cells, and dark-grey nodes represent the detected stable states.

We calculated the scEntropy of lung fibroblastic cells (Figure 2E). Highest entropy was found in PMP cells; intermediate fibroblast subtypes (IF1 and IF2) had intermediate levels of entropy; MFB and MyoF had relative low entropy levels, support their more differentiated states. We project the predicted cell states at E16.5 to the dynamic RNA expression patterns obtained from whole lung tissue during perinatal mouse lung development (E15 to PN0) (28) to understand the ontogenic changes in gene expression in each fibroblast subtype (Supplementary Figure S2). The expression of PMP signature genes decreased during lung development, in a pattern similar to that of cell cycle genes, supporting their role as proliferative progenitors. In contrast, the expression of MyoF and MFB signature genes increased during lung development. Thus, differentiation states of mesenchymal cell subtypes measured by scEntropy were cross validated via the temporal changes in the expression of cell specific signature genes during the perinatal period of lung maturation using data from whole lung microarray (24).

A two-branched transitional trajectory consisting of five fibroblastic subtypes was identified using SLICE (Figure 6A–F), predicting a branched differentiation pathway of mesenchymal cells. In both branches, the fibroblast cell marker *Pdgfra* and the mesenchymal glucocorticoid receptor, *Nr3c1*, were expressed, while the expression of cell cycle genes, including *Foxm1* and *Top2a,* was decreased (Figure 6G), supporting the concept that the trajectory represent the differentiation pathway of lung fibroblasts. Along the transitional path ' C1→C3→C4' (Figure 6E), the expression of matrix fibroblast cell markers, *Fn1*, *Tcf21*, and *Vcam1*, was increased, while the expression of myofibroblast/smooth-muscle markers, *Myocd*, *Myh11* and *Actg2*, was decreased (left, Figure 6G). This pattern of gene expression was reversed along the other cell transitional path from the cluster C1 to C2 cells (Figure 6D and right, Figure 6G). Expression patterns of cell-specific marker genes were closely associated with the two branched lineage predictions. To further validate the prediction, we identified lineage dependent differentially expressed genes using the method and criteria as described in the Methods section and assessed their enriched functional annotations using Toppgene suite (29). 1839 genes were identified as differentially expressed in the transitional path from C1 to C4 and these genes were clustered into three temporal patterns (top, Figure 6H): genes ($n = 493$) in 'Pattern 3' were mostly induced along the path, and their enriched functional annotations including 'extracelluar matrix', 'collagen and elastic fibre formation' were closely associated with the functions of matrix fibroblasts (top, Figure 6I), while genes ($n = 979$) in 'Pattern 1' were mostly down-regulated along the path, and they enriched functional annotations that were closely related to 'cell cycle' or 'proliferation' (Supplementary Table S2), therefore the cell transitional path from C1 to C2 represented the differentiation process of matrix fibroblast cells. For the transitional path from C1 to C2, we identified 2,871 differentially expressed genes and clustered them into three temporal patterns (bottom, Figure 6H): genes in 'Pattern 3' ($n = 622$) were mostly up-regulated along the

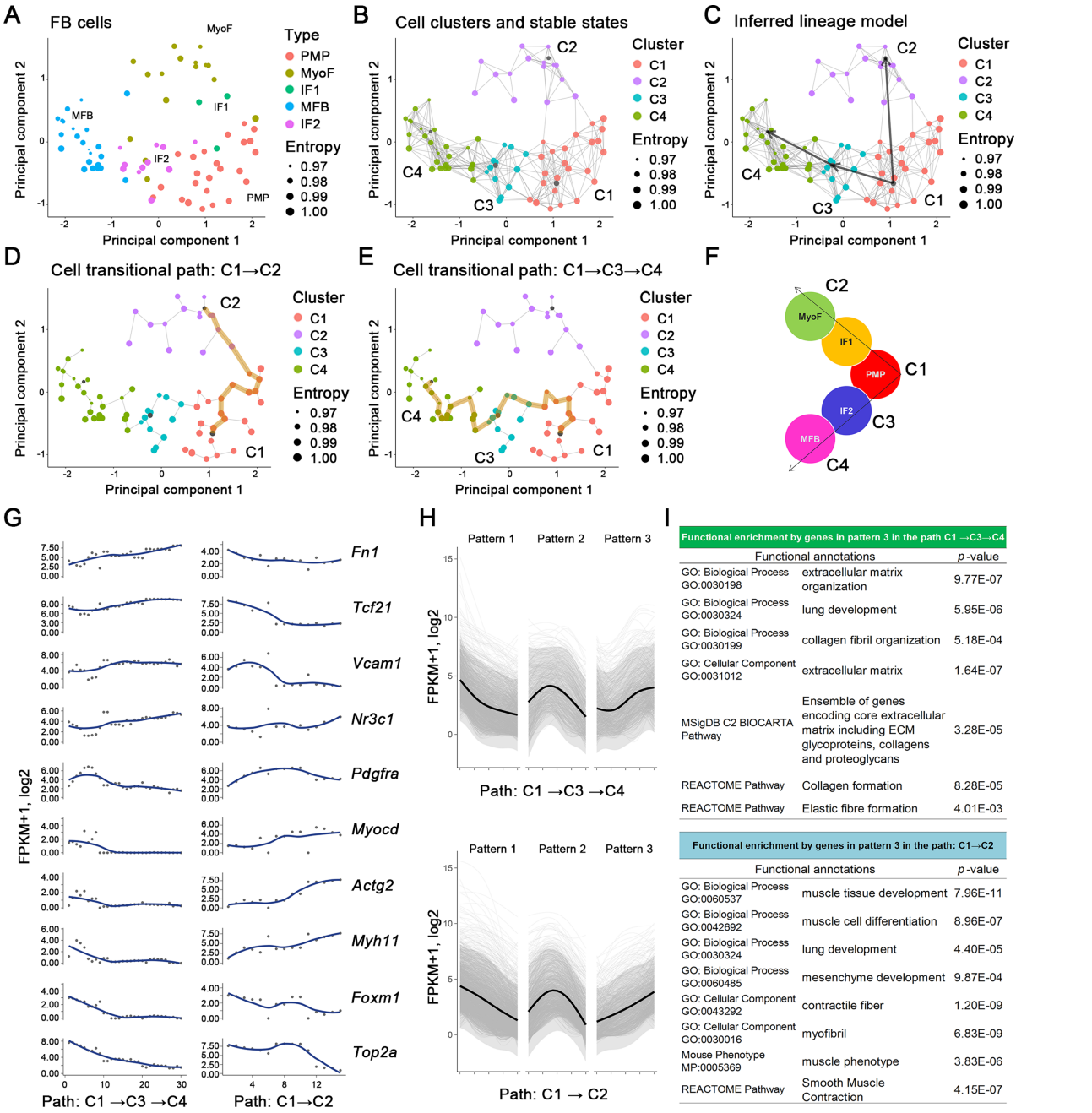

Figure 6. SLICE predicted a two-branched fibroblast cell differentiation pathway at E16.5 mouse lung. (**A**) Principal component analysis of five fibroblast (FB) subtype cells ($n = 79$) using a set of highly variable genes ($n = 252$, Supplementary Table S1). Our previous analyses of single cell RNA-seq data from E16.5 mouse lung identified five distinct fibroblast subtypes, termed as PMP, MyoF, MFB and two unknown fibroblast (IF1 and IF2). Their hierarchical relationships are unknown. (**B**) Four cell clusters and their stable states identified by SLICE. (**C**) A two-branched lineage model of FB cells at E16.5 mouse lung inferred by SLICE. (**D**) and (**E**) showed the reconstructed cell transitional path for each branch, respectively. Each path was reconstructed by concatenating pairwise shortest-paths between successive stable states in the minimum spanning tree of cells. (**F**) A schematic presentation of the inferred two-branched differentiation pathway among the five predicted FB subtypes. (**G**) The expression patterns of markers along the two transitional paths 'C1→C3→C4' (left) and 'C1→C2' (right) suggested a two-branched cell differentiation pathway of lung fibroblasts at E16.5 (i.e. PMP→MFB and PMP→MyoF). (**H**) Temporal patterns of differentially expressed genes in transitional path 'C1→C3→C4' (top) and in transitional path 'C1→C2' (bottom). In (**H**), temporal pattern 1 mainly consisted of differentially down-expressed genes while temporal pattern 3 is primarily comprised of differentially up-expressed genes. (**I**) Enriched functions of genes ($n = 493$) induced in transitional path 'C1→C3→C4' (top) including 'extracelluar matrix' and 'collagen and elastic fiber formation' which were closely associated with the functions of matrix fibroblasts. Enriched functions of genes ($n = 622$) induced in transitional path 'C1→C2' (bottom) including 'muscle development and differentiation' and 'smooth muscle contraction'. In (A)–(E), sizes of cells were proportional to the max-normalized scEntropies of the cells, and dark-grey nodes represent the detected stable states.

path, their enriched functional annotations include 'muscle development and differentiation' and 'smooth muscle contraction' (bottom, Figure 6I). Genes ($n = 1445$) in 'Pattern 1' were mostly down-regulated along C1 to C2 and functionally enriched in 'cell cycle' and 'proliferation' (Supplementary Table S2). These analyses validated the functionalities of the two branched lineage prediction by SLICE, supporting the concept that PMP represents a common lung fibroblast progenitor from which two major lineages, MFB and MyoF, are derived (Figure 6F).

Methodologies comparison and evaluation

SLICE performs two major functions: (i) measuring the differentiation states of single cells and (ii) reconstructing cell differentiation trajectories. Recently, several methods (1,7,9,30–32) have been developed for determining cellular differentiation state or inferring lineage reconstruction from single cell data. We performed a comparative evaluation of SLICE using multiple scRNA-seq data sets produced by different techniques from a variety of contexts in human and mouse. Supplementary Table S3 lists these related methods and their functions.

Among these, signaling entropy (32) and StemID (30) utilized the concept of entropy to measure cell stemness or differentiation states. Although the hypotheses and principles underlying the entropy calculations are different from those of SLICE, the application concept is similar. We compared the cell differentiation states measured by StemID, SLICE and signaling entropy (32). Both SLICE and signaling entropy successfully measured the known cell states in all dataset, while the transcriptome entropy proposed in StemID failed to correctly measure the known differentiation states (Supplementary Figure S3). Of note, StemID was originally proposed to measure cell state using unique molecular identifier (UMI) based scRNA-seq data, while the data sets we collected for the present study were all read-based; the results may be influenced by the compatibility of the data. While the differentiation stages measured by scEntropy and signaling entropy showed high consistency (Supplementary Figure S3), signaling entropy and SLICE are fundamentally distinct methods and complementary. Signaling entropy uses gene expression patterns to measure the amount of uncertainty in how information (signaling) is passed on in the molecular interaction network and relies on a fine-tuned protein–protein interaction (PPI) network for the calculation, which limits the genes that can be analysed using this method to the genes within the PPI network. Our approach enables an entropy calculation using the expression patterns of a larger portion of the transcriptome (i.e. genes with GO annotations), reducing the information loss in the calculation. Another advantage of using SLICE over signaling entropy is that it will automatically lead to the next step of performing single cell lineage reconstruction once the cell differentiation states are determined.

For the other methods that we listed in Supplementary Table S3, including Monocle, SCUBA, Wanderlust and Waterfall (1,7,9,31), none of them can quantitatively measure the cell states and order them accordingly. All utilized transcriptome similarity to reconstruct the temporal ordering of single cells and relied on the known time information or known markers expression to determine the direction of the ordering. Here, we chose Monocle (version 1.6.1 from Bioconductor) as the representative algorithm of this class to compare with the second part of SLICE function (i.e. reconstruct cell differentiation trajectories) using known lineage or previously experimentally validated developmental time information as common reference. As shown in Figure 4, SLICE reproduced the human skeletal muscle myoblasts lineage originally constructed by Monocle using the HSMM dataset. In addition, we compared the performance of SLICE and Monocle using three independent datasets, including AT2 ($n = 101$) (13), FB ($n = 79$) (15,16), HEE ($n = 88$) (14) and E18.5 mouse lung *Epcam*+ epithelial cells (EPI, $n = 80$) (13). Among the four datasets, AT2 and HEE contain relatively homogeneous cell populations, while FB and EPI data contain more heterogeneous cell populations. For the comparison compatibility, we selected the same set of genes for Monocle and SLICE to perform dimension reduction and lineage reconstruction.

Both SLICE and Monocle correctly reconstructed the differentiation pathway in the HEE data (Figure 5 and Supplementary Figure S4). For the AT2 data, SLICE correctly inferred the differentiation trajectory of AT2 cells from E14.5→E16.5→E18.5→adult mouse lung (Figure 3), while Monocle reconstructed the trajectory from E16.5→E14.5→E18.5→adult mouse lung, mis-placed the order of E14.5 and E16.5 cells (Supplementary Figure S5). For the E18.5 EPI data set, Treutlein *et al.* identified five distinct cell types, including alveolar type 2 cells, alveolar type 1 cells, ciliated cells, Clara cells, and alveolar bipotential progenitor (BP) cells and proposed a differentiation pathway of BPs into alveolar type 1 and type 2 cell lineages. Here, we selected the alveolar type 1, type 2 and BP cells ($n = 66$) from the EPI data and aimed to reconstruct the branched differentiation pathway. SLICE correctly clustered cells into three corresponding populations, with highest scEntropy in BP cells to indicate the progenitor states, and then unbiasedly reconstructed the branched differentiation pathway from BP cells into alveolar type 1 and type 2 lineages (Supplementary Figure S6). On the other hand, Monocle only partially uncovered the three cell populations and ordered all the cells into a single lineage (Supplementary Figure S6). Similarly, for the E16.5 lung FB data, SLICE predicted a two-branched differentiation pathway of five fibroblastic subtypes, from PMP →IF2→MFB and PMP→IF1→MyoF (Figure 6); while Monocle was able to identify three major cell states (i.e., PMP, MyoF and MFB), it failed to infer the hierarchical relationships among the three cell states to reconstruct a branched tree lineage (Supplementary Figure S7).

The comparative data analysis showing that the performance of SLICE is comparable to Monocle when reconstructing cell trajectories from scRNA-seq data sets following a homogenous biological process. More importantly, SLICE can correctly infer branched lineage models from cross-sectional scRNA-seq from heterogeneous populations and reconstruct cellular transitional path along each branch, while Monocle did not resolve the complexity of the branched models.

DISCUSSION

Single-cell RNA-seq offers an unprecedented opportunity to measure the molecular states of individual cells and elucidate their lineage relationships, which is fundamentally important for understanding the formation and functions of complex organs. SLICE is a novel algorithm designed for quantitatively measuring cellular transitional stages and predicting cell differentiation lineages from scRNA-seq data independent of external knowledge, such as cell identity, marker gene expression, or time information. Using SLICE, we reproduced previously validated experimental findings of three independent scRNA-seq datasets, which supports the general applicability and high predictive accuracy of SLICE in determining cellular differentiation states and reconstructing cell differentiation lineages. In addition, we applied SLICE to scRNA-seq of embryonic mouse lung at E16.5 to identify lung mesenchymal cell lineage relationships which are largely unknown. SLICE predicted a two-branched transitional trajectory from the common progenitor cell (PMP) to lung matrix fibroblast and myofibroblast. The prediction may serve as an important lung mesenchymal lineage model and merit further experimental validation.

The main tuning parameters in the scEntropy calculation include: m - the number of functional groups, θ is the abundancy threshold for determining the expressed genes and $|B|$ is the size of each bootstrap sample B. In the present work, m was determined by the rule: $m = \sqrt{|\Omega|/2}$, where Ω is the set of genes for clustering. Alternatively, an optimal m can be determined by a clustering validation criterion, such as Bayesian information criterion (BIC), Akaike information criterion (AIC), or Gap statistic (20,21). This optimization may be computationally intensive due to the potentially large size of Ω. In our analyses of all four datasets, we set θ to 1 and $|B|$ to 1000, and calculated the expected values using 100 bootstrap samples. We compared scEntropies calculated using different choices of θ and $|B|$ and showed that the results were robust against perturbations of θ and $|B|$ (Supplementary Figures S8–S11). We also performed a deterministic calculation of scEntropy by setting B as the union set of the top expressed genes ($n = 500$) from each cell, and found that the deterministic calculations of scEntropy were consistent with the bootstrap estimates of scEntropy in all four scRNA-seq datasets (Supplementary Figure S12). We chose to use Gene ontology (GO) terms to represent gene function because compared to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway or Medical Subject Headings (MeSH), GO terms by far is more complete and relatively unbiased for genome-wide gene functional characterization. Nevertheless, we were fully aware that GO terms are incomplete and have their own limitations. In the future, we will test using combined annotation resources and fuzzy clustering algorithms, such as fuzzy K-means (allow overlapping genes with different degree of memberships in different functional clusters) (33,34) to test whether these methods improve the gene functional classification of individual cells. Identifying cell stable states and reconstructing cell trajectory are two of the essential steps of SLICE. We introduced two alternative approaches for each step: network based and clustering based approaches for the cell states identification, and shortest path based and principal curve based approaches for cell trajectory reconstruction. These approaches reached consensus in our analysis and lineage prediction (Supplementary Figures S13–S16) and are independent and complementary to each other; one can choose different approaches for better data fitting, or use them in combination to reach a consensus to improve performance.

Recently, several algorithms, e.g. Wanderlust (1), Monocle (7), SCUBA (9) and Waterfall (31), were developed to infer temporal orderings of single cells from scRNA-seq data, aiming to map individual cells onto specific points in the progression of biological processes by connecting cells based on gene expression similarity. Nevertheless, they do not measure cell states, and therefore, require use of external knowledge, such as time information, cell identity, or marker gene expression, to determine the start and end points of dynamic processes, and the directions of the inferred pseudotemporal orderings. In contrast, SLICE directly and quantitatively measures the entropy of individual cells and uses entropy to predict cell states and lineages independently of experimental knowledge.

We performed a comparison of SLICE and Monocle using multiple datasets from both human and mouse (see Methodologies comparison and evaluation). The performance of SLICE is comparable to Monocle when the data are from a homogeneous cell population. More importantly, this comparison demonstrated the unique advantages of applying SLICE in two situations: (i) when dealing with cross-sectional dataset with no sequential order or time information, SLICE unbiasedly determined the end points and the directions of the transitions; and (ii) when dealing with dataset composed of heterogeneous cell population, SLICE correctly discover the branched lineage models and reconstruct cellular transitional paths along each branch. In disease, injury or other specific stimulations, different cells may have different dynamic responses to the stimulation and not much is known about disease stages or stage dependent activation markers. Therefore, SLICE has unique advantages to quantitatively measure the entropy changes of cells of interest independent of external knowledge or data structure, which will provide new insights into the signaling mechanisms controlling cell fate and functions in relation to disease pathogenesis and potential new therapeutics.

In summary, SLICE is generally applicable to single cell transcriptomic data, provides comprehensive functionalities and options for different data models, and presents unique features and improvements over existing methods for *in silico* lineage mapping, allowing the determination of cell differentiation states and entropy-based lineage paths without external information. While we demonstrated the utility of SLICE using several developmental and differentiation data sets, the approach can also be readily generalized to other biological processes. In the future, we plan to apply SLICE to conditions other than developmental cues including cancer, injury and other disease situations.

AVAILABILITY

SLICE was implemented in R. The source code and demonstrations are available for download at http://research.cchmc.org/pbge/slice.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author's contributions: M.G. and Y.X. conceived and designed the experiments. M.G., E.L.B., M.W. and Y.X. designed the mathematical models and the algorithm. M.G. and E.L.B. implemented the algorithm. M.G., E.L.B., M.W., J.A.W. and Y.X. contributed to the data interpretation, troubleshooting, manuscript writing and editing. All of the authors read and approved the final manuscript. We thank Dr J. Meller for insightful discussions. We also thank Y. Du for assistance with RNA-Seq data processing.

FUNDING

National Heart, Lung and Blood Institute of National Institutes of Health [LungMAP Grant U01HL122642 to J.A.W.,Y.X. LRRC Grant U01HL110967 to J.A.W,Y.X. LRRC Young Investigator Pilot Project Grant to M.G.]. Funding for open access charge: National Institutes of Health.
Conflict of interest statement. None declared.

REFERENCES

1. Bendall,S.C., Davis,K.L., Amir el,A.D., Tadmor,M.D., Simonds,E.F., Chen,T.J., Shenfeld,D.K., Nolan,G.P. and Pe'er,D. (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 714–725.
2. Li,L. and Clevers,H. (2010) Coexistence of quiescent and active adult stem cells in mammals. *Science*, **327**, 542–545.
3. Pujadas,E. and Feinberg,A.P. (2012) Regulated noise in the epigenetic landscape of development and disease. *Cell*, **148**, 1123–1131.
4. Chalut,K.J., Ekpenyong,A.E., Clegg,W.L., Melhuish,I.C. and Guck,J. (2012) Quantifying cellular differentiation by physical phenotype using digital holographic microscopy. *Integr. Biol. (Camb.)*, **4**, 280–284.
5. Weber,P., Wagner,M., Kioschis,P., Kessler,W. and Schneckenburger,H. (2012) Tumor cell differentiation by label-free fluorescence microscopy. *J. Biomed. Opt.*, **17**, 101508.
6. Grun,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251–255.
7. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
8. Patel,A.P., Tirosh,I., Trombetta,J.J., Shalek,A.K., Gillespie,S.M., Wakimoto,H., Cahill,D.P., Nahed,B.V., Curry,W.T., Martuza,R.L. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
9. Marco,E., Karp,R.L., Guo,G., Robson,P., Hart,A.H., Trippa,L. and Yuan,G.C. (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5643–E5650.
10. Shin,J., Berg,D.A., Zhu,Y., Shin,J.Y., Song,J., Bonaguidi,M.A., Enikolopov,G., Nauen,D.W., Christian,K.M., Ming,G.L. *et al.* (2015) Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*, **17**, 360–372.
11. MacArthur,B.D. and Lemischka,I.R. (2013) Statistical mechanics of pluripotency. *Cell*, **154**, 484–489.
12. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
13. Treutlein,B., Brownfield,D.G., Wu,A.R., Neff,N.F., Mantalas,G.L., Espinoza,F.H., Desai,T.J., Krasnow,M.A. and Quake,S.R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.
14. Yan,L., Yang,M., Guo,H., Yang,L., Wu,J., Li,R., Liu,P., Lian,Y., Zheng,X., Yan,J. *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.
15. Du,Y., Guo,M., Whitsett,J.A. and Xu,Y. (2015) 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax*, **70**, 1092–1094.
16. Guo,M., Wang,H., Potter,S.S., Whitsett,J.A. and Xu,Y. (2015) SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput. Biol.*, **11**, e1004575.
17. Hausser,J. and Strimmer,K. (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, **10**, 1469–1484.
18. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
19. Blondel,V.D., Guillaume,J.-L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, P10008.
20. Reynolds,A.P., Richards,G., de la Iglesia,B. and Rayward-Smith,V.J. (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model Algorithms*, **5**, 475–504.
21. Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **63**, 411–423.
22. Ljubić,I., Weiskircher,R., Pferschy,U., Klau,W.G., Mutzel,P. and Fischetti,M. (2006) An algorithmic framework for the exact solution of the prize-collecting steiner tree problem. *Math. Prog.*, **105**, 427–449.
23. Beisser,D., Klau,G.W., Dandekar,T., Muller,T. and Dittrich,M.T. (2010) BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, **26**, 1129–1130.
24. Hastie,T. and Stuetzle,W. (1989) Principal curves. *J. Am. Stat. Assoc.*, **84**, 502–516.
25. Wood,S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **73**, 3–36.
26. Rock,J.R. and Hogan,B.L. (2011) Epithelial progenitor cells in lung development, maintenance, repair, and disease. *Annu. Rev. Cell Dev. Biol.*, **27**, 493–512.
27. Whitsett,J.A. and Weaver,T.E. (2015) Alveolar development and disease. *Am. J. Respir. Cell Mol. Biol.*, **53**, 1–7.
28. Xu,Y., Wang,Y., Besnard,V., Ikegami,M., Wert,S.E., Heffner,C., Murray,S.A., Donahue,L.R. and Whitsett,J.A. (2012) Transcriptional programs controlling perinatal lung maturation. *PLoS One*, **7**, e37046.
29. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.
30. Grun,D., Muraro,M.J., Boisset,J.C., Wiebrands,K., Lyubimova,A., Dharmadhikari,G., van den Born,M., van Es,J., Jansen,E., Clevers,H. *et al.* (2016) De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, **19**, 266–277.
31. Shin,J., Berg,D.A., Zhu,Y., Shin,J.Y., Song,J., Bonaguidi,M.A., Enikolopov,G., Nauen,D.W., Christian,K.M., Ming,G.L. *et al.* (2015) Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*.
32. Banerji,C.R., Miranda-Saavedra,D., Severini,S., Widschwendter,M., Enver,T., Zhou,J.X. and Teschendorff,A.E. (2013) Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Sci. Rep.*, **3**, 3039.
33. Bezdek,J.C. (1980) A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2**, 1–8.
34. Valafar,F. (2002) Pattern recognition techniques in microarray data analysis: a survey. *Ann. N. Y. Acad. Sci.*, **980**, 41–64.