Reconstructing differentiation networks and their regulation from time series single-cell expression data

Jun Ding,[1] Bruce J. Aronow,[2] Naftali Kaminski,[3] Joseph Kitzmiller,[4] Jeffrey A. Whitsett,[4] and Ziv Bar-Joseph[1]

[1]*Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA;* [2]*Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio 45229, USA;* [3]*Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, Connecticut 06520, USA;* [4]*Section of Neonatology, Perinatal and Pulmonary Biology, Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio 45229, USA*

Generating detailed and accurate organogenesis models using single-cell RNA-seq data remains a major challenge. Current methods have relied primarily on the assumption that descendant cells are similar to their parents in terms of gene expression levels. These assumptions do not always hold for in vivo studies, which often include infrequently sampled, unsynchronized, and diverse cell populations. Thus, additional information may be needed to determine the correct ordering and branching of progenitor cells and the set of transcription factors (TFs) that are active during advancing stages of organogenesis. To enable such modeling, we have developed a method that learns a probabilistic model that integrates expression similarity with regulatory information to reconstruct the dynamic developmental cell trajectories. When applied to mouse lung developmental data, the method accurately distinguished different cell types and lineages. Existing and new experimental data validated the ability of the method to identify key regulators of cell fate.

[Supplemental material is available for this article.]

Most methods for reconstructing regulatory networks using high-throughput data relied on microarray and RNA-seq studies profiling large populations of cells (Liao et al. 2003; Margolin et al. 2006; Ernst et al. 2007; Schulz et al. 2012). While such approaches have led to many important results, they tend to overlook the heterogeneity of the population being profiled. This may be problematic where a mixture of different cell types, with different regulatory programs, is being profiled, for example, in cancer (Dalerba et al. 2011), immune response (Shalek et al. 2013), and development (Treutlein et al. 2014).

Single-cell RNA-seq data addresses this problem by profiling the contribution of different cell types to changes in tissue level expression, allowing for much more detailed and accurate models. However, such data has also raised new computational challenges, some of which were recently addressed, including issues related to sample quality (Stegle et al. 2015), normalization of single-cell data (which is more challenging, especially for lowly expressed genes) (Shapiro et al. 2013; Wu et al. 2014), and the development of clustering methods to identify distinct components within a specific mixture/time point (Buettner et al. 2015; Guo et al. 2017).

Another challenge with single-cell RNA-seq data is the analysis of time series. While several methods have been developed for the analysis and modeling of temporal data in population-based microarray and RNA-seq experiments (Bonneau et al. 2006; Bar-Joseph et al. 2012; Patil and Nakai 2014; Young et al. 2014), they all relied on one key assumption: that consecutive time points measure a continuously evolving process. In other words, the assumption is that measurements at time point $t+1$ are correlated with measurements at the previous time point t (either the $t+1$ expression levels continuously evolve from the expression of the same genes at time point t [Bar-Joseph et al. 2003] or they are regulated by genes expressed at the previous time point [Bar-Joseph et al. 2012]). While these assumptions may hold for the population as a whole, it clearly does not hold for all individual cells whose functions, proliferation, and differentiation vary dynamically within the population. Thus, a key issue when analyzing single-cell RNA-seq data is the ability to not only identify different cells within a specific time point (e.g., by clustering) (Xu and Su 2015) but also link these cells over time by identifying the subsets of cells that belong to the same trajectory. A further challenge is to derive the regulatory networks that control different cell fates or states that are profiled in the study.

A few recent methods have been developed to address the problem of connecting single cells along a temporal trajectory. Some of these methods are limited and can only reconstruct models with no branching (a single trajectory) (Bendall et al. 2014) or with a single branch point (Setty et al. 2016). While these may be useful for in vitro data, they are less appropriate for in vivo studies in which multiple types of cells are studied (Treutlein et al. 2014). Other methods either completely ignore the time at which the cell was measured (Trapnell et al. 2014) or rely on the measurement time (Marco et al. 2014; Treutlein et al. 2014), ignoring the fact that cells may be in different developmental states at a single time point. Indeed, both types of methods cannot accurately reconstruct complex developmental trajectories (Rashid et al. 2017) and fail to distinguish between differentiated and undifferentiated cells at a specific time point. While these methods differ

Corresponding author: zivbj@cs.cmu.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.225979.117.

in the computational models they use, they generally rely on the same underlying assumption that consecutive cells (or states) in the ordering should be very similar in terms of expression levels of their genes. While this assumption makes sense when sampling rates are very high, they do not always hold for in vivo studies (e.g., the lung developmental data discussed in this paper which is sampled every 2 d). In such cases, additional information can be used to determine the ordering and branching in the model. One such source of information is the set of transcription factors (TFs) that are active at each developmental stage. If these can be inferred, then states in which the factors are active could be linked to downstream states in which their targets are activated or repressed even if the overall correlation between the two states is not very high. An advantage of such an approach is that in addition to the ordering or branching model, we also obtain a network model that describes which regulatory events lead to the different cell fates, the TFs controlling these events, and their time of activation.

In the present work, we have utilized single-cell RNA expression data during the critical period of lung morphogenesis as the embryo prepares for air-breathing at birth. At this time the lungs consist of many distinct, mesenchymal and endodermally derived epithelial cells that are rapidly dividing and differentiating to form a functional organ. During late gestation, dramatic changes in organ structure and epithelial cell differentiation and function create a functional gas exchange unit in the alveolar regions of the lung via a process that is highly active, but still not fully understood at the molecular level (Whitsett and Weaver 2015).

To model the process of lung epithelial growth and differentiation, we present a model that integrates time series single-cell RNA-seq data with general protein–DNA interaction data. We applied our model to reconstruct differentiation networks and their regulation based on single-cell lung development data. The model accurately distinguishes between cell types and trajectory of branching during cell differentiation. The model predicts several TFs as important regulators at various stages of development. While many of these were known, others are novel. We used existing and new data to validate some of these TFs and their activation times.

Results

We developed a new method that learns a probabilistic model for constructing regulatory networks from single-cell time series expression data. An overview of the method is presented in Figure 1 (for an illustration using real data, see also Supplemental Fig. S1). We initially start by clustering cells within each time point. We use several clustering evaluation metrics to determine the number of clusters at each time point (Methods). While the measured time provides useful information about the state of the different cells, previous studies demonstrated that cells from the same time point can be unsynchronized. To address this issue, we allow for cells to be moved to states representing other time points than the ones they were measured in (see below), and we further test each of the clusters to determine whether certain clusters in the same time point are actually composed of cells at different differentiation stages (Methods). Following this analysis, we arrive at an initial model in which we link each cluster to the cluster at the preceding time point that is most similar to it in expression space (Fig. 1C). We next iterate between two steps: reassigning cells to states in the model and determining the set of states and their connectivity (parent–child relationship) until the likelihood does not increase. As part of the model learning, we determine the set of TFs predicted to regulate genes in each of the states. We use this information to improve our model by requiring that factors regulating descendant states be expressed at parental states and by ensuring that genes expressed or repressed in cells assigned to descendant cells that are regulated by the identified TF follow their predicted trajectory (up- or down-regulation). Both requirements further impact cell assignment and model learning. If, after reassignment, states become empty (no assigned cells), they are

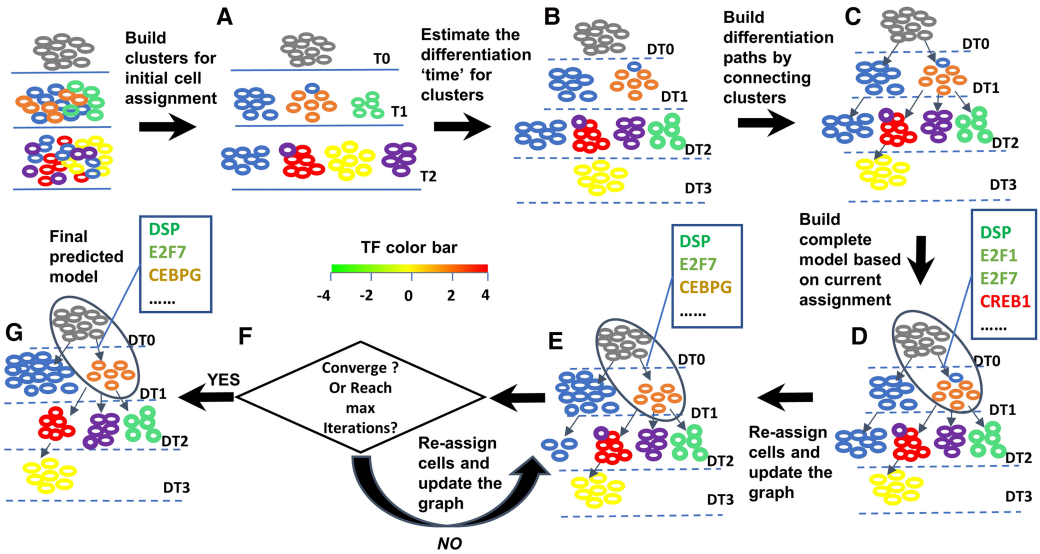

Figure 1. Learning differentiation models from single-cell RNA-seq data. (*A*) Initial clusters are determined using spectral clustering. T0, T1, and T2 represent the measurement time. (*B*) Initial "differentiation time" is estimated for clusters based on difference with clusters for the first time point. DT0, DT1, DT2, and DT3 denote the estimated differentiation time. (*C*) Differentiation paths are constructed by connecting clusters at lower levels to their most similar parent at the level above them. (*D*) Regulating TFs are determined for each edge. TFs are colored based on their expression change along the edge. (Red) Increased expression; (green) decreased expression; (blue) stable expression. Shades represent the extent of the expression change. (*E*) Initial model. (*F*) Iterating between cells and state reassignments and parameter learning until convergence. (*G*) Final model.

removed from the model. Thus, unlike all prior methods for determining trajectories in single-cell time series data, our model does not only rely on expression similarity but also takes into account potential regulation that may be important for in vivo studies in which sampling is not frequent.

Application of the model to lung developmental data

To test our method, we first applied it to study cell fate trajectories in mouse lung development. We used a time series data set with 152 cells from Treutlein et al. (2014). Of these, 45 cells were profiled at day E14.5, 27 cells at E16.5, and 80 cells at E18.5. Known cell fate markers were used to determine the cell type for the E18.5 cells (earlier cells are progenitors and may not express these markers). We applied our method to these data and observed that it converged after five iterations. While convergence is obviously data dependent, we note that we observed similarly fast convergence when analyzing other single-cell data sets (e.g., six and eight iterations for different sets of mouse embryonic fibroblasts (MEFs) reprogramming data sets), indicating that the method can likely be widely applied.

As can be seen in Figure 2, the model correctly separated all four terminal cell types (AT1, AT2, Club, and ciliated) into different terminal states. It identified an additional terminal state (E2_18_0; state numbers are arbitrary) that contains a mixture of AT1 cells and bipotential progenitors (BP) cells. These latter cells were first identified by Treutlein et al. (2014) and predicted to be nonterminal fates that can serve as progenitors to both AT1 and AT2 cells. Treutlein et al. (2014) have also observed that intermediate states are present at E18.5 such as early AT1 (*Pdpn*, *Ager* positive; *Sftpc* low) and early AT2 cells (*Sftpc* positive; *Pdpn*, *Ager* low). This is captured in our model. There are two AT1 states in our model, the first (E2_18_0) contained BP cells, and the second (E3_18_4) is more homogeneous. Average expression of *Sftpc* (an AT2 marker) is 7.81 in the AT1 cells from first state and only 4.72 in the AT1 cells from the second (*P*-value difference of 0.0107 based on rank test), suggesting that the first AT1 group represents an early AT1 group while the second is a more differentiated state. The model also correctly reconstructs parts of the known branching process, indicating that ciliated cells are derived from a different set of progenitors at E16.5 (Rawlins et al. 2007).

In addition to the assignment of cells to states, the model also highlights several TFs as playing an important role in cell differentiation. Several of these factors are known to be involved in this process, including NKX2-1 (Herriges and Morrisey 2014), SOX9 (Rockich et al. 2013), FOXA1/FOXA2 (Wan et al. 2004), and GATA6 (Yang et al. 2002).

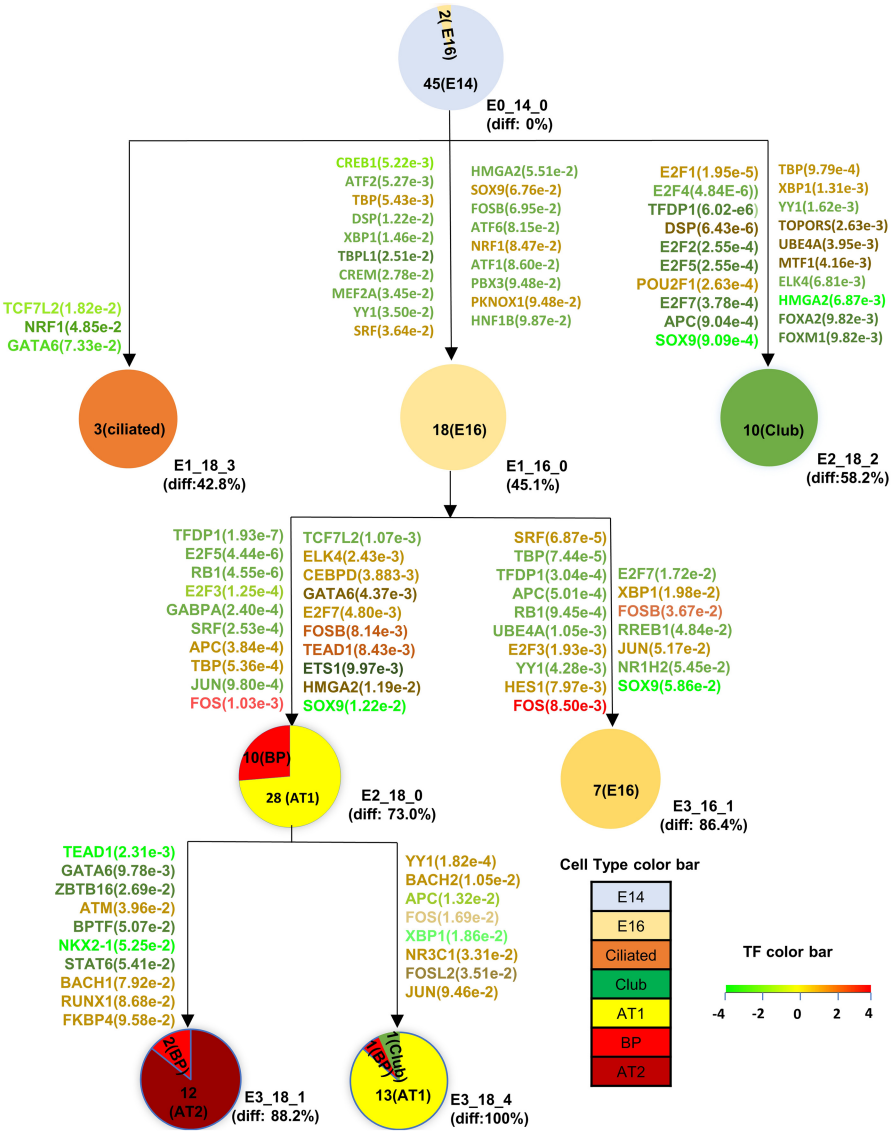

Figure 2. Differentiation model on Treutlein et al. (2014) lung single-cell data set. (diff) Differentiation stage 0% (most undifferentiated)–100% (most differentiated) from the first state. Number of cells assigned to each state was given inside the pie chart. Cell types for each state (node) are based on the assignments of Treutlein et al. (2014) and so are only available for cells in the last time point (E18.5); cells from earlier time points are labeled using the time point (E14, E16). TFs are associated with paths they are predicted to regulate and color-coded based on their expression change (fold change) along the regulated path. TF *P*-values are based on the set of targets associated with the states they are predicted to regulate.

Increasing the number of E16.5 cells in our model

While the model in Figure 2 agrees with many known aspects of lung cell differentiation, it does not provide enough information about the less studied parts of this process, specifically the role TF plays in driving cells to different fates. Such understanding is a key point since it can provide information about why certain types of cells are absent from diseased lungs and may even provide directions for treatments in such cases. One of the problems is the

fact that relatively few cells were profiled for the intermediate time point (only 27 cells at E16.5). Thus, to improve the model, we replaced the E16.5 cells with epithelial cells from a larger study that only focused on E16.5 single cells in lung development (Du et al. 2015). Forty-nine of these cells are epithelial progenitors and were further associated with various potential fates based on marker expression profiles (note that these assignments were not used in model learning that is unsupervised but will be discussed below when analyzing the results) (Guo et al. 2015). Since the data now come from two different groups, we performed an analysis, using housekeeping genes, to determine if further normalization was needed (Supplemental Methods; Supplemental Fig. S2).

A detailed view of epithelial cell differentiation in lung development

Figure 3 presents the revised model using the Du et al. (2015) E16.5 data. As can be seen, while the assignment to terminal states in this model is similar to the one in Figure 2, we see differences in the overall structure with a more detailed view of the differentiation

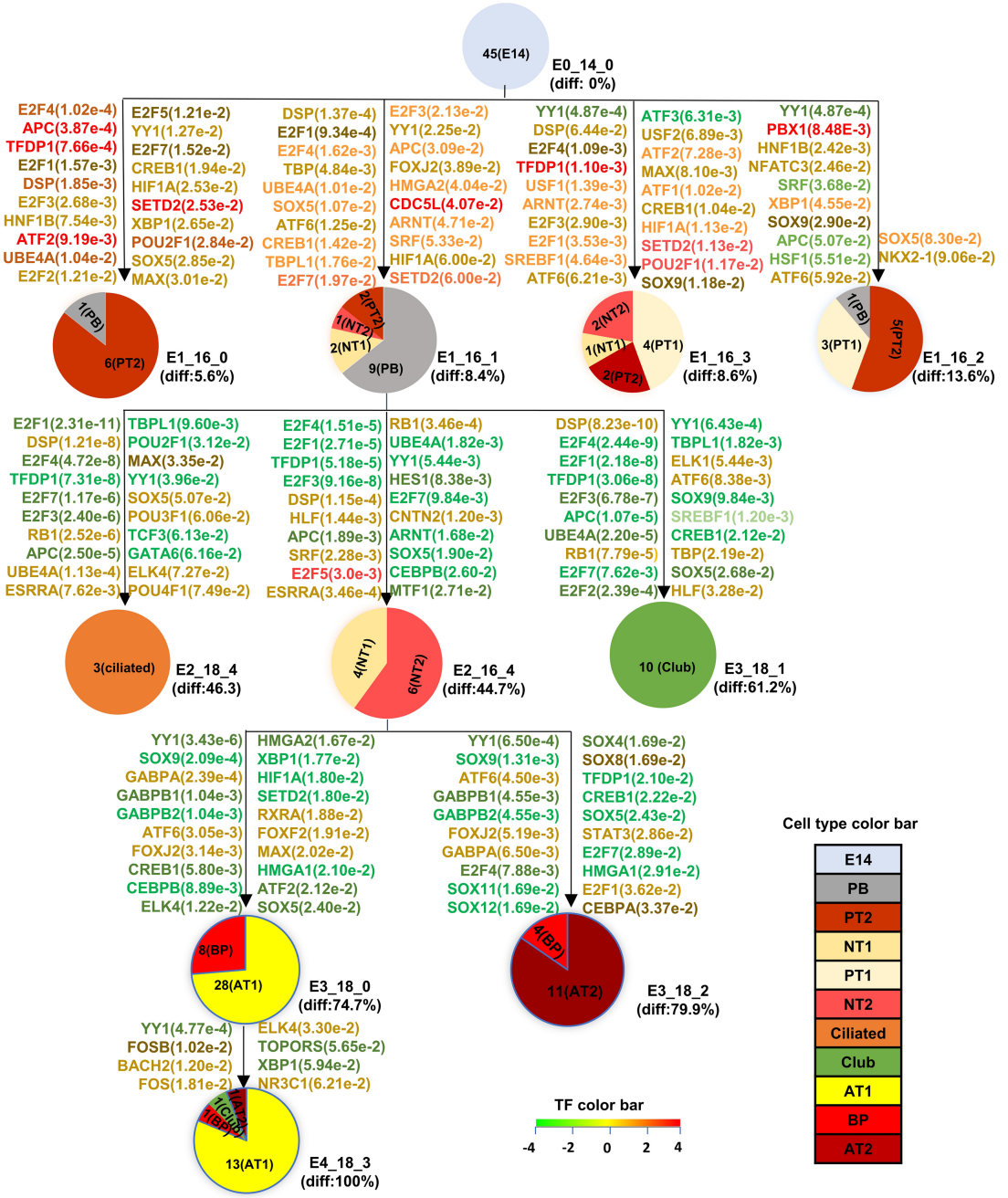

Figure 3. Differentiation model using data from both Treutlein et al. (2014) and Du et al. (2015). Cell types taken from both Treutlein et al. (2014) and Du et al. (2015). (PT2) Proliferative AT2 early precursor; (PT1) proliferative AT1 early precursor; (PB) proliferative bipotential precursor; (NT1) noncycling AT1 precursor; (NT2) noncycling AT2 precursor. Differentiation scores, TFs color, and *P*-value have the same meaning as in Figure 2.

process. For example, in this model, we see an earlier separation of ciliated and Club cells, as has previously been observed (Rawlins et al. 2007). In contrast, the separation of AT1 and AT2 cells is through a set of progenitors, and their fate is determined later in the process. Once again, we see BP cells mainly clustered with AT1 cells (though this time also as progenitors to these cells, e.g., the link from the third to the fourth level in the model). The cell identities used to evaluate the cell assignment were obtained from the original studies (Treutlein et al. 2014; Guo et al. 2015).

Given the better agreement with prior knowledge along with the more elaborate view, we have studied this model in more detail. To validate some of these predicted TFs, we performed Gene Ontology (GO) analysis using PANTHER version 11.0 (Mi et al. 2013). We found that TFs predicted by the model were significantly enriched for GO terms associated with lung epithelial cell differentiation (1.75×10^{-5}) and regulation of epithelial cell proliferation (2.0×10^{-6}) (Supplemental Table S1). While many of the predicted TFs are novel (see also below), several are supported by prior studies. For example, E2F4 is required for the development of ciliated and Club cells (Danielian et al. 2007). We found E2F4 to be ranked as one of the top TFs (first and second) for the ciliated and Club states. SREBF1 (also known as SREBP1) is required for the development of alveolar epithelial cells (AECs) (Mason 2006), consistent with model predictions (E2_16_0→E3_18_0 (AT1,BP mixture)). Similarly, the model identifies CEBPA/CEBPB as regulating the development of alveolar and airway epithelial cells, an observation that is supported by prior work (Martis et al. 2006; Roos et al. 2012). For a complete list of known TFs identified by our model, see Supplemental Table S2.

Simulation and robustness analysis

We performed simulation studies to test the ability of our method to handle noise and dropouts, which are often prevalent in single-cell data (Kharchenko et al. 2014). For each cell, we simulated different dropout rates by setting the expression of randomly chosen 5%–80% genes to zero. Results indicate that our method is robust against such noise. For 5% and 10% simulated dropouts, the predicted differentiation structures are the same as the one presented above. Cell assignments are only slightly worse but generally agree with the ones obtained using the original data without simulated dropouts (Supplemental Table S3). When the dropout rate increases to 20%, the predicted differentiation structure changes slightly (Supplemental Fig. S3), ciliated cells and Club cells were assigned to the same cluster, and AT1 cells were assigned to three different clusters, while the overall cell assignment is still in good agreement with the predictions on the original data and also with the known labels. Beyond 40% dropout, we see more changes though scdiff is still able to separate proliferative and noncycling AT1/AT2 precursors and AT1 and AT2 cells. For more details, see Supplemental Figures S3 through S7.

To simulate the expression noise, we also added varying levels of random Gaussian noise to the expression of all genes (Supplemental Results). Again, we observe that for low noise levels, the predicted differentiation structures is the same as the one of the original data, while for higher levels, the structure is only slightly different (Supplemental Table S4; Supplemental Fig. S8). We also performed a bootstrap analysis in which we used a subset of the cells to learn the model (randomly sampling 80%, 82.5%, 85%, 87.5%, and 90% of all cells). We compared the resulting models (Supplemental Figs. S9–S13) and observed that both the models and cell assignments are similar to the ones obtained when using the full set of cells (~90% agreement for cell assignment) (Supplemental Table S5). We also tested the impact of some of the parameters used by the model and observed that within a reasonable range the changes did not have a large impact on the resulting model (Supplemental Figs. S14, S15).

Comparisons to prior methods

While pseudo-time-ordering methods differ from scdiff in several aspects (including the use of the profiled time for the initial assignment and the ability to infer continuous vs. discrete ordering), both types of methods attempt to infer models for the progression of cell states in developmental studies. We have thus compared scdiff to pseudo-time-ordering methods. Past work has shown that some of these methods, including Monocle (Trapnell et al. 2014), SCUBA (Marco et al. 2014), and principal component analysis (PCA), fail to accurately model cell assignment and trajectories for the lung development data discussed above (Rashid et al. 2017). Here we further analyze the performance of another method, diffusion pseudotime (DPT) (Haghverdi et al. 2016) on the lung (Treutlein et al. 2014) and on MEFs reprogramming data (Treutlein et al. 2016). As can be seen in Figure 4, while DPT finds some structure in both data sets, it fails to correctly separate cell types and identify branching for the lung data and does not accurately order the reprogramming data. In contrast, our method is able to both correctly assign cells to states and identify the progression of time from embryonic cells to developed neurons. For comparison using bone marrow data, see also Supplemental Figure S16 (Olsson et al. 2016).

In addition to direct comparisons that focus on the ordering and cell assignments (which is the focus of all prior methods including TASIC) (Rashid et al. 2017), these prior methods do not use protein–DNA interaction data and so cannot directly identify the set of TFs that regulate each branching point. Thus, a major advantage of our method is the ability to rely on such data to infer not just cell assignment but also the regulatory events that drive this process. To assess the impact of this novel aspect of our method, we have applied the method without using TF information (i.e., similar to prior methods which only use expression similarity). Results are presented in Supplemental Figure S17. As can be seen, cell assignments and enriched TFs (based on their targets) were different when not using the TF–gene interaction information to construct the model. Six terminal cells (7.5%) are assigned to an incorrect state in this case. We also see differences in the set of significant TFs (calculated as a post-processing step when not using them for the learning). Specifically, several TFs that are known to be involved in epithelial lung development are missing from the non-TF model. These include FOXA1, which regulates the lung epithelial differentiation (Besnard et al. 2005); GATA6, which regulates the differentiation of distal lung epithelium (Yang et al. 2002); RFX3, which affects the airway epithelium development (Didon et al. 2013); and others.

Staining experiments agree with predicted TF activity time

To test model predictions for the activity of TFs, we used staining experiments in developing mouse lungs. We selected a number of factors that were either novel predictions or for which the prediction of their regulatory timing was novel. These include overexpression (OE) of the hypoxia-inducible factors (HIF1A), which has been identified in a variety of developmental, physiologic, and pathogenic processes within the lung (Shimoda and Semenza 2011; Tibboel et al. 2015); SOX9, which has multiple roles in the

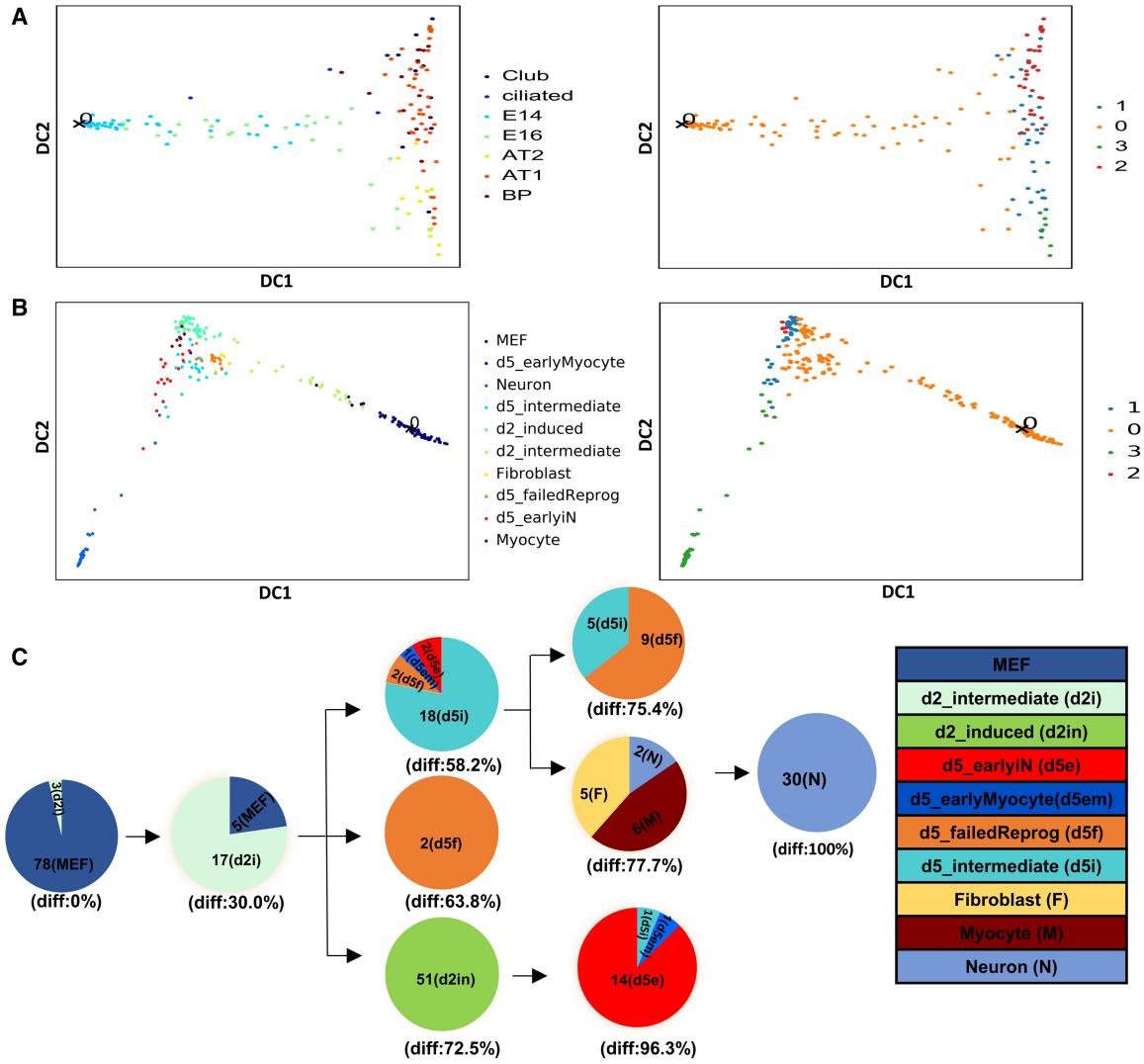

Figure 4. Performance comparison with diffusion pseudotime (DPT). (*A*) DPT analysis of the Treutlein et al. (2014) mouse lung single-cell data (the data used in Fig. 2). (*Left*) Cells colored by their type as determined by Treutlein et al. (2014). (*Right*) Cell assignment by DPT using default parameters. While DPT finds some structure in the data, it is unable to separate the AT1 and Club cells and does not show any major branching prior to E18.5. (*B*) DPT analysis of mouse embryonic fibroblasts (MEFs) reprogramming data (Treutlein et al. 2016) setting 2. While the DPT model finds a branch leading from the MEF cells to the neurons, it does not order correctly the intermediate day 2 and day 5 cells (note that day 5 cells are mostly on the other branch and only day 2 cells are close to neurons). (*C*) In contrast, a model based on our method for the same data correctly places most day 2 cells in the second level with day 5 cells closer to the neurons. See also text for discussion.

lung epithelium, including the regulation the extracellular matrix (Rockich et al. 2013); and known epithelial cell markers.

HIF1A is predicted to be regulating AT1 and AT2 states, and its targets are predicted to be down-regulated in these states, indicating that OE of *Hif1a* at E18.5 may affect AT1/AT2 cell differentiation or function. A modest sacculation defect, increased dilation, and regional difference in the SFTPC (AT2 marker) and HOPX (AT1 marker) distribution were observed in the staining results (Fig. 5A,B) at E18.5, consistent with model predictions.

SOX9 (Supplemental Fig. S18) was highly expressed in peripheral regions of proliferating AEC progenitor cells at E16.5. SOX9 staining decreased dramatically by E18.5 matching our prediction. In our model, SOX9 is predicted to have a relatively high activity in edges from states at E16.5 and also the predicted expression of *Sox9* is highest in early proliferative progenitor states, consistent with the loss of SOX9 staining at E18.5.

Perturbation experiments support model predictions

While the overall model structure provides some insights about the process of epithelial cell differentiation, an important advantage of TF-based assignment is the ability of the model to make specific predictions about possible perturbation experiments and their outcome. Specifically, if a TF is predicted to regulate a specific path in the model (e.g., the edge from E1_16_0 to E2_16_0) but not the other fates that descend from the same parent state, then a possible prediction is that the knockout (KO) or OE of that TF (depending on the impact the TF has on its downstream genes) would impact the specific path it regulates and the cell fates associated with it but much less so for other fates. We have thus collected available KO and ChIP-chip data for three TFs identified in our model and compared the results to the model for the specific TFs analyzed. For each such expression experiment, we compare the

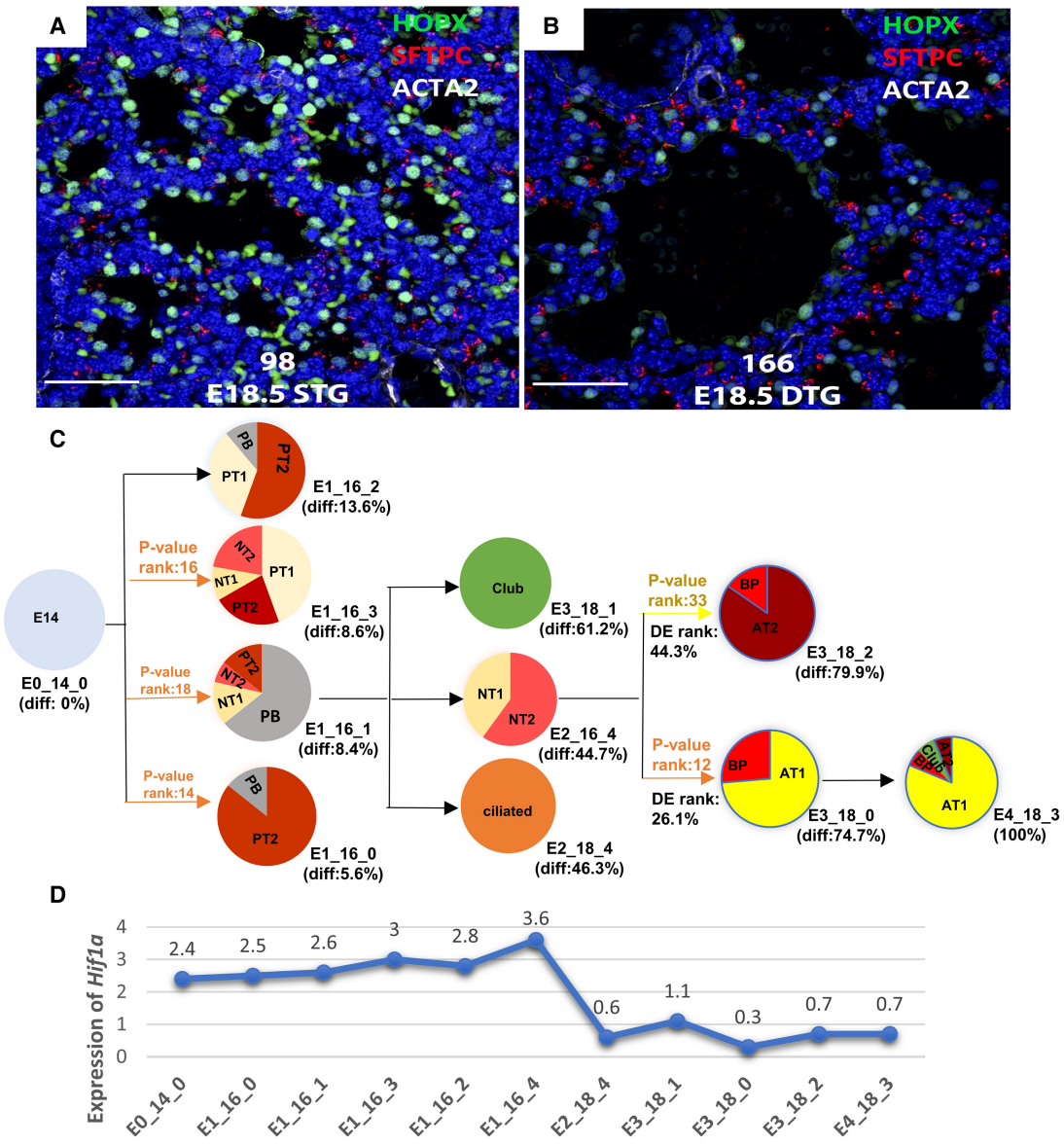

Figure 5. Increased HIF1A activity disrupts sacculation and influences AT1/AT2 cell distribution. (*A*,*B*) Experimental results for HIF1A staining. HIF1A (three-point mutant) was expressed under conditional control of SFTPC-rtTA, $(otet)_7$-HIF1A-TPM. Doxycycline was provided to the dam from E12.5 to E18.5. Single transgene (STG) controls, lacking HIF1A-TPM expression ($n = 3$), were compared with double transgenic (DTG) mice expressing HIF1A-TPM under doxycycline control in airway epithelial cells ($n = 4$). Staining of lung tissue for ACTA2 (smooth muscle actin), HOPX (an AT1 cell marker), and SFTPC (proSP-C, and AT2 cell marker) are shown. (C) Model prediction for HIF1A. HIF1A is identified as a top regulator of AT1 cells (ranked as the 12th TF) with a lower, though still significant, impact on AT2 cells (ranking as the 33rd TF). (*D*) mRNA expression of *Hif1a* in the different states reconstructed by the model. As predicted by the model, OE of *Hif1a* influences AT1/AT2 cell distribution with a larger impact on AT1 cells compared with AT2 cells.

correlation of the wild-type (WT) and KO differentially expressed (DE) genes to the average expression profile for those genes in each of states. For each state in our model, the KO correlation can either be similar to the WT correlation, in which case we cannot infer a large impact of the TF, or be different than the WT correlation, in which case we can infer that the TF is impacting the expression of genes in the state/cell subtype.

The first TF we looked at was CREB1, which was predicted to regulate both the AT1 (E2_16_4→E3_18_0 (AT1 28, BP 8)) and AT2 (E2_16_4→E3_18_2 (AT2 11, BP 4)) edges. It was also found to be regulating ciliated (E1_16_1→E2_18_4) and Club (E1_16_1→E3_18_1) cells. We used WT and KO data for *Creb1* from an experiment profiling lung epithelial cells at E17.5 from (Bird et al. 2011). We identified 273 DE genes between the *Creb1* KO and WT samples and then calculated the correlation between the expression of these DE genes in the WT/KO *Creb1* study and the expression of the predicted states in our model. As predicted by the model, and as can be seen in the first three columns of Table 1, *Creb1* KO had the strongest impact on AT1 and AT2 gene expression. Specifically, while WT *Creb1* data exhibited a highly significant correlation with AT1 and AT2 cells (correlation coefficient = 0.401), a KO of *Creb1* led to much lower correlation (correlation coefficient = 0.124). This suggests that CREB1 may indeed be required for the differentiation of AT1 and AT2 cells. These results

Table 1. Spearman correlation of DE genes between *Creb1* KO data and predicted clusters

	E3_18_0 AT1(28),BP(8)	E3_18_2 AT2(11),BP(4)	E4_18_3 AT1(13),BP(1),Club(1),AT2(1)	E2_18_4 ciliated(3)	E3_18_1 Club(10)
Creb1 WT	(0.449, 1.1×10^{-12})	(0.371, 2.6×10^{-9})	(0.383, 2.32×10^{-9})	(0.209, 0.0015)	(0.352, 4.83×10^{-8})
Creb1 KO	(0.196, 0.00306)	(0.038, 0.569)	(0.137, 0.0390)	(0.0516, 0.439)	(0.0637, 0.339)

were supported by Besnard et al. (2011) and Antony et al. (2016), demonstrating a severe lack of AECs in *Creb1*-deleted mice. Similarly, WT *Creb1* data show a strong correlation with Club cells (correlation coefficient = 0.352), while the correlation with the KO *Creb1* experiment is much lower. We have observed much weaker correlation between ciliated cells and *Creb1* WT data and no correlation with the *Creb1* KO data.

HMGA2 was identified as a TF required for proper cell differentiation in our model. While it is known to be involved in lung cancer (Di Cello et al., 2008), its role in lung development is much less clear. HMGA2 was predicted to regulate the proliferative bipotential precursor edge (E0_14_0→E1_16_1 [proliferative bipotential precursor 9]) and its descendent AT1 edge (E2_16_4→E3_18_0 (AT1 28, BP 8)) and AT2 (E2_16_4→E3_18_2 (AT2 11, BP 4)). We looked at *Hmga2* KO experiments performed at E18.5 from Singh et al. (2015). We identified a set of 298 DE genes. WT *Hmga2* expression levels are highly correlated with AT1 and AT2 states (correlation coefficient 0.509 for AT1 and 0.440 for AT2) (Table 2). This correlation disappears for the KO *Hmga2* experiment, which supports the model predictions. The model also predicts HMGA2 as a regulator of Club cell differentiation (E1_16_1, the direct parent of the Club cells state). We did not observe an impact of *Hmga2* gene deletion on the correlation with ciliated cells.

NKX2-1 is a critical factor regulating lung epithelial differentiation (Minoo et al. 1995). Our model predicts NKX2-1 (TTF1) to be active at the early stage of lung epithelial cell differentiation. It was predicted as the regulating factor for edge E0_14_0 (common ancestor)→E1_16_2 (proliferative AT2 early precursor) and edge E1_16_1 (proliferative bipotential precursor)→E2_16_4 (noncycling AT1 precursor and noncycling AT2 precursor). In order to validate this prediction, we downloaded ChIP-chip experiment for NKX2-1 performed in lung epithelial cells from Tagne et al. (2012). We compared DE genes in each state (defined as genes whose expression in the descendant state is different from their expression in the parent state), the observed targets of NKX2-1 from the ChIP-chip experiment using hypergeometric test (Supplemental Table S6). The experimental results match the model well. Edges predicted to encode active NKX2-1 TF are much more enriched for targets of NKX2-1 (e.g., *P*-value of 0.007 for the edge from E0_14_0 to the E1_16_2 node based on hypergeometric distribution). In contrast, several of the other edges, which were not predicted to be regulated by NKX2-1, do not overlap with targets, indicating that the model can discriminate between active factors for specific fates.

Staining and OE experiments further support model predictions

We performed staining experiments in developing mouse lungs to see if factors identified by the model based on expression and regulation are indeed active at the protein level at time predicted. For this, we looked at SOX9, which has multiple roles in the lung epithelium, including the regulation the extracellular matrix (Rockich et al. 2013), and at the hypoxia-inducible factor (HIF1A), which has been identified in a variety of developmental, physiologic, and pathogenic processes within the lung (Shimoda and Semenza 2011; Tibboel et al. 2015).

SOX9 (Supplemental Fig. S18) was highly expressed in peripheral regions of proliferating AEC progenitor cells at E16.5. SOX9 staining decreased dramatically by E18.5 similar to its assignment in the model and its expression in these points. In our model, SOX9 is predicted to have a relatively high activity in edges from states at E16.5, consistent with the loss of SOX9 staining at E18.5.

HIF1A is predicted to be regulating AT1 and AT2 states (Fig. 5C) and its targets are predicted to be down-regulated in these states, indicating that OE of *Hif1a* at E18.5 may affect AT1/AT2 cell differentiation or function. HIF1A and its targets are down-regulated at later stages of the model (Fig. 5D). A modest sacculation defect, increased dilation, and regional differences in SFTPC (AT2 marker) and HOPX (AT1 marker) distribution were observed in the staining results (Fig. 5A,B). As predicted, increased activity (OE) of *Hif1a* disrupted sacculation and impaired epithelial cell differentiation, consistent with model predictions.

Given the staining results obtained for HIF1A, we performed additional experiments to test the impact of its expression on downstream genes. As mentioned above, HIF1A was predicted as a regulator for both AT1 edge (E2_16_4→E3_18_0 (AT1 28, BP 8)) and AT2 edge (E2_16_4→E3_18_2 (AT2 11, BP 4)). However, unlike the other TFs mentioned above, we observed a decline in the expression levels of *Hif1a* at target states, indicating that the activity of this TF activator needs to be reduced during lung development (Fig. 5). Following (Bridges et al. 2012), we used a cDNA construct that constitutively activates *Hif1a* in normoxic conditions. We compared two versions of OE *Hif1a*: single transgenic samples (STGs) and double transgenic samples (DTSs). We identified 223 DE genes between STG and DTG and used this to examine the correlation between states in our model and *Hif1a* OE. The results are presented in Table 3.

Our results support the role of HIF1A as a regulator of (repressed) lung development. Although the OE results for *Hif1a*

Table 2. Spearman correlation of DE genes between *Hmga2* KO data and the predicted states

	E3_18_0 AT1(28),BP(8)	E3_18_2 AT2(11),BP(4)	E4_18_3 AT1(13),BP(1),Club(1),AT2(1)	E2_18_4 ciliated(3)	E3_18_1 Club(10)
Hmga2 WT	(0.509, 4.78×10^{-21})	(0.440, 1.6×10^{-15})	(0.397, 1.1×10^{-12})	(0.176, 0.00231)	(0.441, 1.351×10^{-15})
Hmga2 KO	(0.0889, 0.126)	(−0.0433, 0.457)	(0.0582, 0.317)	(0.167, 0.00378)	(0.0154, 0.792)

Table 3. Spearman correlation of DE genes between *Hif1a* OE experiment and predicted clusters

	E3_18_0 AT1(28),BP(8)	E3_18_2 AT2(11),BP(4)	E4_18_3 AT1(13),BP(1),Club(1),AT2(1)	E2_18_4 ciliated(3)	E3_18_1 Club(10)
Hif1a STG	(0.134, 0.0495)	(0.197, 0.00383)	(0.167, 0.0144)	(0.137, 0.0455)	(0.194, 0.00442)
Hif1a DTG	(0.04447, 0.516)	(0.0133, 0.847)	(−0.046, 0.503)	(−0.05112, 0.456)	(−0.0486, 0.479)

(Table 3) did not show significant correlation difference between STG and DTG mice samples, the staining experiments (Fig. 5) as well as the direction of correlation coefficient change in Table 3, support the model prediction.

Analyzing additional data sets

To further test if our method can be generally applied to analyze progression pathways from single-cell RNA-seq data, we have also used it to analyze time series single-cell data sets from MEFs reprogramming (Treutlein et al. 2016) and from mouse bone marrow (Olsson et al. 2016). The MEFs reprogramming data set focused on two settings: The first studied cells treated by ASCL1, and the second looked at cells treated with a combination of ASCL1, POU3F2 (previously known as BRN2), and MYT1L. Our model identified ASCL1 as a key regulator for both conditions even though the information about the specific gene perturbation experiment was not used in the learning process. Several other known factors were identified. The model accurately assigned cells to states (Supplemental Figs. S4, S19) and provided a map for the differentiation of MEFs to multiple cell fates (for interactive model, see Supplemental website). Similarly, for the mouse bone marrow data (Olsson et al. 2016), our predicted model correctly determines that HSCP cells differentiate to Mono and Gran cells through a series of intermediate states. For more details about performance comparison on additional data sets, see the Supplemental Results and Supplemental Figures S20 and S21.

Discussion

We developed and tested a computational method for reconstructing dynamic regulatory networks from single-cell time series data. Unlike prior methods for pseudotemporal ordering of such data, our method uses static information about targets of TFs both to improve the learning of a branching model and to identify TFs that regulate various stages in the process. Applying our method to single-cell lung development data from multiple laboratories allowed us to reconstruct developmental pathways for a number of different types of lung epithelial cells. As we show, the reconstructed models both capture known biology (in terms of cell groupings and temporal assignment of events) and raise new hypotheses about the roles that certain TFs play in the development of specific cell types. We validated these predictions using both immunofluorescence staining and expression experiments identifying new roles for a number TFs in regulating lung development.

One of the predicted TFs, HIF1A, is known to decrease in expression with advancing gestation in the fetal mouse lung (Bridges et al. 2012). To assess the effects of HIF1A on epithelial cell differentiation, an oxygen-stable form of HIF1A was conditionally expressed in respiratory epithelial cells. As predicted, OE of *Hif1a* inhibited maturation of AT1 cell precursors, indicated by decreased intensity and numbers of HOPX-stained AT1 cells, and increased proportion of proSP-C–stained AT2 cells. Additional support to the model was obtained using immunofluorescence confocal microscopy analysis of fetal mouse lung from the canalicular (E16.5) to saccular stage (E18.5) of lung morphogenesis. At E16.5, lung mesenchyme was prominent and epithelial cells lining peripheral regions of acinar buds stained for both NKX2-1 and SOX9, as predicted by the model (Supplemental Fig. S18). At E18.5, the peripheral acinar buds had dilated, mesenchyme had thinned, the levels SOX9 were markedly decreased, and HOPX had increased, consistent with differentiation of AT1 and AT2 cell progenitors. Phosphohistone H3, a marker of cell proliferation, expressed in the SOX9-positive epithelial progenitors and associated mesenchyme at E16.5, was markedly decreased at E18.5, consistent with the decreased proliferation that occurs with advancing gestation in the mouse lung.

The limited knowledge of TF–DNA interaction is one bottleneck of our method. For example, we did not have the targets for HOPX in our database and thus were unable to predict it as an active regulator. In order to overcome this problem, our method provides the ability to predict top DE genes for each edge. By combining predicted regulators and top DE genes information, our model is able to identify potential regulators even without accurate target information. For example, the aforementioned missing regulator *Hopx* was predicted as top DE genes (top up-regulated DE genes in AT1 paths and top down-regulated DE gene in AT2 paths), which is consistent with the fact that *Hopx* is an AT1 marker.

To further test if our method can be generally applied to analyze progression pathways from single-cell RNA-seq data, we have also used it to analyze single-cell RNA-seq from MEFs reprogramming data (Treutlein et al. 2016) and time series mouse bone marrow (Olsson et al. 2016). The reconstructed models in both cases agreed with known biology while highlighting several novel TFs as potential regulators. These results highlight the global applicability of the method, which we hope can be used to study a wide range of developmental and differentiation processes.

Methods

Single-cell RNA-seq data sets

We downloaded time series lung single-cell data from Treutlein et al. (2014) and MEFs reprogramming single-cell data from Treutlein et al. (2016). We also used lung single-cell E16.5 data from Du et al. (2015) and Guo et al. (2015). We preprocessed these data sets as was done in the original paper (Treutlein et al. 2014). Specifically, (1) if the FPKM of gene expression is smaller than one, the gene will be regarded as not expressed; (2) genes with zero variance across cells are removed (we also tried a more stringent criterion, please refer to the Supplemental Results and Supplemental Figure S22 for details); and (3) transform to Log FKPM.

Initial clustering of single cells

We start by clustering the single cells at each individual time point to get an initial cell assignment. For this, we use a correlation-based method that was shown to be more suited than Euclidean distance

when dealing with noisy (and sometimes partial) data (Zimek et al. 2012). We use Spearman correlation to compute a similarity matrix across cells. Next, spectral clustering (Ng et al. 2001) is used to cluster single cells based on the similarity matrix. For larger data sets with thousands of cells, the time complexity of the spectral clustering ($O(n^3)$, where n is the number of cells) may be prohibitive. For such data sets, we have also implemented an alternative initial clustering strategy: PCA+k-Means, which is much faster and does not significantly impact results. For the complete details, see Supplemental Methods, Supplemental Results, and Supplemental Figures S23 and S24. To determine the number of clusters for each time point (or states in our initial model), we used several quality assessment scores. We combined these scores using an ensemble strategy similar to random forest to determine the optimal number of Clusters k for each time point. For the discussion of methods used and how they were combined, see Supplemental Methods.

Reassigning clusters and initial model construction

The initial clustering was based on the time point associated with each cell in the time series experiment. However, several recent studies indicate that cells may be unsynchronized with respect to their state even if they are collected at the same time point (Goranov et al. 2009; Trapnell et al. 2014). Thus, some of the clusters at a specific time point may represent states that are either earlier or later than other clusters in the same time. In this work, we developed and used the 'Similarity To Ancestor-STA' (STA) strategy to infer an initial cluster assignment to various levels in the model. STA computes the Spearman correlation between the expression of all cells (where the expression of cell is defined as the expression vector of all genes in the cell) within every cluster and the expression of the cluster(s) at the first time point. STA of a cluster represents a vector of Spearman correlation values between the expression of each cell within the cluster and the expression of cluster(s) at the first time point. Clusters (except for the ones belonging to the first time points) are sorted based on the average STA of the cluster. Next, we compute the significance of the difference in correlation between consecutive clusters in the ordering using ranksum test $pv = ranksums(STA_X, STA_Y)$ for a pair of clusters X and Y. If we find a point in the ordering where the difference is significant (P-value <0.05), we assign the clusters that follow that break to a new level. This process is continued for all levels until reaching the last cluster (see also Supplemental Methods).

Once we determined the set of levels in the model and the clusters associated with each level, we next connect clusters in each level to the most similar cluster (in terms of correlation) at the level right above it. By connecting all clusters to their parents, we get a graph (clusters as Nodes, parent–child relationship as Edges) that structurally represents the differentiation model.

Predicting TFs regulating differentiation pathways

An important aspect of our method is the ability to both reconstruct and analyze the differentiation pathways based on the set of TFs that regulate various state transitions. TFs whose targets are active in later stages of the process are likely active at earlier stages (in order to activate or repress their targets), and so expression levels of TF at a specific time point can be used to determine cell assignment and state connections at the next time point. We discuss below how we use TFs to impact these aspects. Here we discuss how we identify a set of TFs that are used to seed the model and the transition and emission probabilities.

We used the TF–gene interaction data from Ernst et al. (2007) and Schulz et al. (2012). "TF–gene interaction data" refers to the information about potential targets for TFs. The data are in a form of matrix with each entry denoting a TF–gene pair. Values are either binary (yes/no evidence for the interaction) or probabilities (between zero and one) depending on the source used to infer the interaction. For complete information on how the data are collected and processed, see Schulz et al. (2012). Following the initial model construction, we first identify a set of DE genes (from parent cluster to current cluster) for each cluster (state) in our model. By using this set, we identify TFs that are enriched for DE targets in each state using the hypergeometric distribution. TFs with P-value <0.1 are kept as the candidate regulators. Next, we check which of the candidate TFs are expressed in the parent node of the state (expressed in at least 20% cells of the cluster). TFs that are both significantly enriched and expressed are used in the expression progression Kalman filter model as discussed below. A ranking was also provided beside the P-value for each regulating TF to demonstrate the relative regulating power at each specific edge.

To select a subset of TFs for each of the edges in the model, we use a Lasso regression method (Tibshirani 1996), which uses the TF–gene interaction data to predict the expression values for target genes in the downstream state. We first classify genes in that state as up-regulated ↑, down-regulated ↓, or not-changing ≈ comparing the parent state (Supplemental Methods). Next, a logistic regression classifier that uses the interactions between selected TFs and the genes as input is trained with the target of maximizing the ability to predict the level of the target gene expression based on the interaction data alone. The idea behind this is that TFs that are active would be selected by the Lasso method since they provide useful information about their targets, whereas those that are inactive or less significant would have very small coefficients and be removed from the model. For complete details, see Supplemental Methods.

A Kalman filter model for differentiation progression

To model expression changes and regulation during single-cell differentiation, we use a Kalman Filter model. Similar to hidden Markov models (HMMs), when using a Kalman filter we need to estimate transition and emission models, though unlike the unconstrained version in HMMs, these take a specific, linear form. Our Kalman Filter model assumes that gene expression at cluster s is related to the expression of its parent cluster P_s based on the following transition model:

$$X_s = A_s X_{P_s} + B_s + w_s, \tag{1}$$

$$w_s \sim N(0, Q), \tag{2}$$

where X_s denotes the gene expression vector at cluster s, X_{Ps} denotes gene expression of the parent cluster of s, and w_s is the process noise, which is assumed to be drawn from a zero mean Gaussian noise. A is the linear transition matrix, and B is the offset matrix. We set A to be the identity matrix to denote the fact that genes in descendant states are expected to be similar to genes in their parent states, as was done in prior methods for modeling temporal progression in single-cell studies (Bendall et al. 2014; Marco et al. 2014; Trapnell et al. 2014; Juliá et al. 2015; Shin et al. 2015). However, unlike these prior methods, our model allows for a divergence in gene expression between parent and child states for genes that are regulated by TFs that are predicted to be active in the parent state. This is the goal of the B matrix. To encode this, we use the logistic regression model discussed above. Once the model is learned, we have a set of active TFs for each state. We then use these TFs and the parameters learned for them to assign a label to each gene in the descendant state. Following these assignments, each gene in state s is either up-regulated ↑, down-regulated ↓, or not-regulated ≈. Note that these labels are a function of the parent,

and so if we reassign a state to another parent in the model (see below), they may change, allowing the model to refine assignments in cases where cell memberships change.

Next, we label gene expression changes in the descendant states as follows. If gene g is determined to be "up-regulated," then its expected expression value in cluster s will be the expression of its parent cluster P_s multiplied by an up-regulation scaling factor U ($1/U$ for down-regulation). If g is predicted to be not regulated, then its expected expression is the same as the one at P_s.

To handle dropouts, we use a variant of the zero-inflated negative binomial model (Kharchenko et al. 2014), which we adjust to handle continuous values (in our case, Gaussian emission distribution). Specifically, we use, similar to zero inflated models, a mixture model for the expression emission probability. This enables us to account for dropouts (zeros in the expression matrix) without fully penalizing the cells when computing their likelihood of being emitted from the state. Specifically, we set the emission probability to

$$P(g|s) = w_g p_1(g|s) + (1 - w_g)p_2(g|s), \tag{3}$$

$$p_1(g \mid s) \sim N(X_s^g, \sigma_s), \tag{4}$$

$$p_2(g|s) = \begin{cases} k, & \text{if } g = 0. \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where X_s^g is the mean of expression of gene g in cluster s as discussed above and σ_s denotes the variance of gene g; $(1 - w_g)$ is the fraction of dropped out genes for that cluster obtained by maximum likelihood estimation (MLE) as the ratio of cells with nonzero values for that gene in the cluster. k is an arbitrary probability value, which is the same for all dropped genes in all clusters. The Kalman filter model for STA was defined similarly as the expression model described above. Please refer Supplemental Methods for complete details.

Learning parameters for the Kalman filter model

We used the initial assignments discussed above to fit gene expression transition and emission models. Given current assignments, we can compute the MLE of the transition and emission noise variance. As mentioned above, we also use the initial assignments and structure to determine parameters for the logistic regression model, which in turn determine the set of values used in the B transition offset matrix for each parent–child relationship. For complete details, see Supplemental Methods.

Model refinement and cell reassignments

Once we learned the initial transition and emission parameters, we can determine the global likelihood based on the assignment of cells to different states in the model.

$$\log(\text{Likelihood}(c_1, c_2, \ldots, c_n, A|M)) = \sum_{i=1}^{n} \log P(c_i, s_i|M) \tag{6}$$

$$= \sum_{i=1}^{n} [\log(P(s_i)P(c_i|s_i))] \tag{7}$$

$$= \sum_{i=1}^{n} [\log(P(s_i)) + \log(P(STA_{c_i}|s_i)) + \log(P(G_i|s_i)] \tag{8}$$

$$= \sum_{i=1}^{n} \{\log(P(s_i) + \log(P(\text{STA}_{c_i}|s_i)) + \sum_{g_k \in g^i} \log(P(g_k|s_i)\}, \tag{9}$$

$$\log(P(s_i)) = \log(\prod_{q \in Q_i} p(q|q_p)) \tag{10}$$

$$= \sum_{q \in Q_i: \text{path to } s_i} \log(q|q_p). \tag{11}$$

Here n is the number of all cells, A represents the current assignments of cells to states, g^i is the set of all genes for cell i (c_i), and s_i is the state to which cell i is assigned. $P(G_i|s_i)$ is the expression probability of c_i, which indicates the agreement of gene expression between c_i and the state s_i; $P(g_k|s_i)$ is the expression probability of gene g_k, which represents the probability that g_k is emitted by state s_i. $P(\text{STA}_{ci}|s_i)$ is the time probability of c_i, which indicates the agreement of STA values between c_i and state s_i. Q_i is the path from the root node to state (node) s_i, including the root node: $P(\text{root}|\text{root}_{\text{parent}}) = P(\text{root}|\text{None}) = P(\text{root})$. $\log(q|q_p)$ modeled the transition relations and was estimated based on the current assignment: $P(q|q_p) = |C_q|/|CP_{p_q}|$. $|C_q|$ is number of cells at state q. CP_{q_p} denotes the number of cells, which are from all children states of q_p (parent state of q).

We next attempt to improve the likelihood of the model by refining the model structure (i.e., changing parent–descendant assignments) and reassigning cells to states in the model. To reassign cells, we compute the maximal probability for cell c_i, $P(c_i|s)$ for all states s in the model. Specifically we find

$$\text{Assign}(c_i) = \underset{s}{\text{argmax}}\, P(c_i, A|M) \tag{12}$$

$$= \underset{s}{\text{argmax}}\, P(c_i, s) \tag{13}$$

$$= \underset{s}{\text{argmax}}\, P(s)P(STA_{c_i}|s)P(G_i|s) \tag{14}$$

$$= \underset{s}{\text{argmax}}\, P(s)P(\text{STA}_i|s) \prod_{g_k \in g_i} P(g_k|s) \tag{15}$$

$$= \underset{s}{\text{argmax}} \log\,(P(s)) + \log\,(P(\text{STA}_i|s)) + \sum_{g_k \in g^i} \log\,(P(g_k|s)). \tag{16}$$

Note that assignment can lead to states becoming empty. If this happens, these states are removed from the model. After reassigning cells to states, we further refine the model by updating nodes (states) and edges (parent relationship). We remove states that become empty and recompute the edges (fromNode, toNode, regulating TFs) by updating the parent for each remaining state. For this, we use the (re)assigned cells to recompute a set of DE genes for that state, test which potential parent state in the preceding level maximizes the transition function for that state (based on the logistic regression model computed for the parent), and select the parent with the highest likelihood. Once a parent is assigned, we recompute the set of TFs for the edge by using the new set of DE genes for each state (if reassignment of cell changed the set).

Identifying epithelial cells in a large cohort of E16.5 lung cells

In this work, we integrated data from multiple prior lung development single-cell studies. One of the data sets we used was the LunGENS (Du et al. 2015; Guo et al. 2015), which profiled 49 single cells from fetal mouse lung at E16.5 using RNA sequencing of cells separated-using the Fluidigm C1.

Perturbation and imaging experiments

All mouse experiments were performed under AAALAC-approved protocols reviewed at Cincinnati Children's Hospital Medical Center (CCHMC). For immunofluorescence confocal microscopy, lung tissue from embryos (E16.5 and E18.5) was fixed in 4% PFA (PBS). Tissue was sectioned at 5 microns for paraffin and 7 microns for frozen samples. Slides were incubated with antisera versus NKX2-1 (catalog number: RB1231; rabbit, Seven Hills Bioreagents), SOX9 (catalog number: AB5335; rabbit, Millipore), SFTPC (catalog number: SC-7706; goat, Santa Cruz), HOPX (catalog number: SC-30216; rabbit, Santa Cruz), and ACTA2 (catalog number: A5228; mouse, Sigma-Aldrich) or phosphohistone H3 (catalog number:SC-12927; goat, Santa Cruz). Detailed

methodologies are provided in the Lung Image website accessible at https://research.cchmc.org/lungimage. Sections were imaged on a Nikon A1Rsi confocal microscope. For studies of HIF1A, tissue was obtained from fetuses (E18.5) of transgenic mice engineered to express a HIF1A mutant protein under control of the human SFTPC-rtTA promoter construct by expressing (tetO)7/CMV/HIF1A/ODD/N803, a normoxia stable form of HIF1A. Administration of doxycycline activates expression of the transgene in fetal respiratory epithelial cells. Dams were treated with doxycycline from E16.5 until E18.5, the time of sacrifice. Doxycycline-treated single transgenic and double transgenic fetuses were identified by genotyping. Imaris and Nikon Elements software was used to export images, and Adobe Photoshop used to adjust levels of fluorescence for data display.

Mouse studies

An activated form of HIF1A, HIF1A(TPM), was expressed under conditional control of the SFTPC-rtTA, $(otet)_7$-HIF1A TPM. Double and single transgenic littermates were compared from dams treated with doxycycline chow from E12.5 until sacrifice, as previously reported (Bridges et al. 2012). Confocal immunofluorescence microscopy was performed for ACTA2, HOPX, and SFTPC.

For fetal mouse studies, C57BL/6 mice were time mated to obtain litters at E14.5, E16.5, and E18.5 for immunofluorescence staining for SFTPC (proSP-C), HOPX, ACTA2 (αSMA), SOX9, phosphohistone H3 (pHisH3), and NKX2-1 (TTF1) as described in the LungMAP data repository at www.lungmap.net.

Software availability

scdiff is primarily written in Python, available as an open source tool at GitHub (https://github.com/phoenixding/scdiff). This GitHub repository includes detailed instructions on how to use the method. The scdiff source code is also available as the Supplemental code. All the data and results on the Supplemental website (http://www.cs.cmu.edu/~jund/scdiff/) are provided as the Supplemental Materials (Supplemental website).

Acknowledgments

This work is supported in part by the National Institutes of Health (grant numbers 1R01GM122096 and U01HL122626-01 to Z.B.-J. and U01HL122642 to J.A.W.) and by the National Science Foundation (grant number DBI-1356505 to Z.B.-J.). We thank Yina Du for RNA data support and Dr. James Bridges for HIF1A (TDM) tissue. We also thank Easwaran Ramamurthy for testing our software.

References

Antony N, McDougall A, Mantamadiotis T, Cole T, Bird A. 2016. Creb1 regulates late stage mammalian lung development via respiratory epithelial and mesenchymal-independent mechanisms. *Sci Rep* **6:** 25569.

Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS. 2003. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc Natl Acad Sci* **100:** 10146–10151.

Bar-Joseph Z, Gitter A, Simon I. 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet* **13:** 552–564.

Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Peer D. 2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157:** 714–725.

Besnard V, Wert S, Kaestner K, Whitsett J. 2005. Stage-specific regulation of respiratory epithelial cell differentiation by Foxa1. *Am J Physiol Lung Cell Mol Physiol* **289:** L750–L759.

Besnard V, Wert SE, Ikegami M, Xu Y, Heffner C, Murray SA, Donahue LR, Whitsett JA. 2011. Maternal synchronization of gestational length and lung maturation. *PLoS One* **6:** e26682.

Bird AD, Flecknoe SJ, Tan KH, Olsson PF, Antony N, Mantamadiotis T, Hooper SB, Cole TJ. 2011. cAMP response element binding protein is required for differentiation of respiratory epithelium during murine development. *PLoS One* **6:** e17843.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol* **7:** R36.

Bridges JP, Lin S, Ikegami M, Shannon JM. 2012. Conditional hypoxia inducible factor-1α induction in embryonic pulmonary epithelium impairs maturation and augments lymphangiogenesis. *Dev Biol* **362:** 24–41.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33:** 155–160.

Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, et al. 2011. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* **29:** 1120–1127.

Danielian PS, Kim CFB, Caron AM, Vasile E, Bronson RT, Lees JA. 2007. *E2f4* is required for normal development of the airway epithelium. *Dev Biol* **305:** 564–576.

Di Cello F, Hillion J, Hristov A, Wood LJ, Mukherjee M, Schuldenfrei A, Kowalski J, Bhattacharya R, Ashfaq R, Resar LM. 2008. HMGA2 participates in transformation in human lung cancer. *Mol Cancer Res* **6:** 743–750.

Didon L, Zwick RK, Chao IW, Walters MS, Wang R, Hackett NR, Crystal RG. 2013. RFX33 modulation of FOXJ1 regulation of cilia genes in the human airway epithelium. *Respir Res* **14:** 70.

Du Y, Guo M, Whitsett JA, Xu Y. 2015. 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax* **70:** 1092–1094.

Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. 2007. Reconstructing dynamic regulatory maps. *Mol Syst Biol* **3:** 74.

Goranov AI, Cook M, Ricicova M, Ben-Ari G, Gonzalez C, Hansen C, Tyers M, Amon A. 2009. The rate of cell growth is governed by cell cycle stage. *Genes Dev* **23:** 1408–1422.

Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. 2015. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol* **11:** e1004575.

Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y. 2017. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* **45:** e54.

Haghverdi L, Buettner M, Wolf FA, Buettner F, Theis FJ. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13:** 845–848.

Herriges M, Morrisey EE. 2014. Lung development: orchestrating the generation and regeneration of a complex organ. *Development* **141:** 502–513.

Juliá M, Telenti A, Rausell A. 2015. *Sincell*: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. *Bioinformatics* **31:** 3380–3382.

Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11:** 740–742.

Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci* **100:** 15522–15527.

Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. 2014. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci* **111:** E5643–E5650.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7:** S7.

Martis PC, Whitsett JA, Xu Y, Perl AKT, Wan H, Ikegami M. 2006. C/EBPα is required for lung maturation at birth. *Development* **133:** 1155–1164.

Mason RJ. 2006. Biology of alveolar type II cells. *Respirology* **11:** S12–S15.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the panther classification system. *Nat Protoc* **8:** 1551–1566.

Minoo P, Hamdan H, Bu D, Warburton D, Stepanik P, deLemos R. 1995. TTF-1 regulates lung epithelial morphogenesis. *Dev Biol* **172:** 694–698.

Ng AY, Jordan MI, Weiss Y. 2001. On spectral clustering: analysis and an algorithm. *NIPS* **14:** 849–856.

Olsson A, Venkatasubramanian M, Chaudhri VK, Aronow BJ, Salomonis N, Singh H, Grimes HL. 2016. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537:** 698–702.

Patil A, Nakai K. 2014. TimeXNet: identifying active gene sub-networks using time-course gene expression profiles. *BMC Syst Biol* **8:** S2.

Rashid S, Kotton D, Bar-Joseph Z. 2017. TASIC: determining branching models from time series single cell data. *Bioinformatics* **33:** 2504–2512.
Rawlins EL, Ostrowski LE, Randell SH, Hogan BL. 2007. Lung development and repair: contribution of the ciliated lineage. *Proc Natl Acad Sci* **104:** 410–417.
Rockich BE, Hrycaj SM, Shih HP, Nagy MS, Ferguson MA, Kopp JL, Sander M, Wellik DM, Spence JR. 2013. Sox9 plays multiple roles in the lung epithelium during branching morphogenesis. *Proc Natl Acad Sci* **110:** E4456–E4464.
Roos AB, Berg T, Barton JL, Didon L, Nord M. 2012. Airway epithelial cell differentiation during lung organogenesis requires C/EBPα and C/EBPβ. *Dev Dyn* **241:** 911–923.
Schulz MH, Devanny WE, Gitter A, Zhong S, Ernst J, Bar-Joseph Z. 2012. DREM 2.0: improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst Biol* **6:** 104.
Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34:** 637–645.
Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498:** 236–240.
Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14:** 618–630.
Shimoda LA, Semenza GL. 2011. HIF and the lung: role of hypoxia-inducible factors in pulmonary development and disease. *Am J Respir Crit Care Med* **183:** 152–156.
Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, Enikolopov G, Nauen DW, Christian KM, Ming GL, et al. 2015. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17:** 360–372.
Singh I, Ozturk N, Cordero J, Mehta A, Hasan D, Cosentino C, Sebastian C, Krüger M, Looso M, Carraro G, et al. 2015. High mobility group protein-mediated transcription requires DNA damage marker γ-H2AX. *Cell Res* **25:** 837–850.
Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16:** 133–145.
Tagne JB, Gupta S, Gower AC, Shen SS, Varma S, Lakshminarayanan M, Cao Y, Spira A, Volkert TL, Ramirez MI. 2012. Genome-wide analyses of Nkx2-1 binding to transcriptional target genes uncover novel regulatory patterns conserved in lung development and tumors. *PLoS One* **7:** e29907.
Tibboel J, Groenman FA, Selvaratnam J, Wang J, Tseu I, Huang Z, Caniggia I, Luo D, van Tuyl M, Ackerley C, et al. 2015. Hypoxia-inducible factor-1 stimulates postnatal lung development but does not prevent O_2-induced alveolar injury. *Am J Respir Cell Mol Biol* **52:** 448–458.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol* **58:** 267–288.
Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32:** 381–386.
Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509:** 371–375.
Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SAM, Sim S, Neff NF, Skotheim JM, Wernig M, et al. 2016. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534:** 391–395.
Wan H, Kaestner KH, Ang SL, Ikegami M, Finkelman FD, Stahlman MT, Fulkerson PC, Rothenberg ME, Whitsett JA. 2004. Foxa2 regulates alveolarization and goblet cell hyperplasia. *Development* **131:** 953–964.
Whitsett JA, Weaver TE. 2015. Alveolar development and disease. *Am J Respir Cell Mol Biol* **53:** 1–7.
Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, et al. 2014. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11:** 41–46.
Xu C, Su Z. 2015. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31:** 1974–1980.
Yang H, Lu MM, Zhang L, Whitsett JA, Morrisey EE. 2002. GATA6 regulates differentiation of distal lung epithelium. *Development* **129:** 2233–2246.
Young WC, Raftery AE, Yeung KY. 2014. Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Syst Biol* **8:** 47.
Zimek A, Schubert E, Kriegel HP. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat Anal Data Min* **5:** 363–387.

Received June 4, 2017; accepted in revised form December 21, 2017.