Bioinformatics, 33(16), 2017, 2504–2512
doi: 10.1093/bioinformatics/btx173
Advance Access Publication Date: 4 April 2017
Original Paper

Gene expression

TASIC: determining branching models from time series single cell data

Sabrina Rashid[1], Darrell N. Kotton[2] and Ziv Bar-Joseph[1,3,*]

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]Department of Medicine, Department of Pathology and Laboratory Medicine, Center for Regenerative Medicine (CReM) of Boston University and Boston Medical Center, Boston, MA 02118, USA and [3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on August 3, 2016; revised on January 23, 2017; editorial decision on March 16, 2017; accepted on March 22, 2017

Abstract

Motivation: Single cell RNA-Seq analysis holds great promise for elucidating the networks and pathways controlling cellular differentiation and disease. However, the analysis of time series single cell RNA-Seq data raises several new computational challenges. Cells at each time point are often sampled from a mixture of cell types, each of which may be a progenitor of one, or several, specific fates making it hard to determine which cells should be used to reconstruct temporal trajectories. In addition, cells, even from the same time point, may be unsynchronized making it hard to rely on the measured time for determining these trajectories.

Results: We present TASIC a new method for determining temporal trajectories, branching and cell assignments in single cell time series experiments. Unlike prior approaches TASIC uses on a probabilistic graphical model to integrate expression and time information making it more robust to noise and stochastic variations. Applying TASIC to *in vitro* myoblast differentiation and *in-vivo* lung development data we show that it accurately reconstructs developmental trajectories from single cell experiments. The reconstructed models enabled us to identify key genes involved in cell fate determination and to obtain new insights about a specific type of lung cells and its role in development.

Availability and Implementation: The TASIC software package is posted in the supporting website. The datasets used in the paper are publicly available.

Contact: zivbj@cs.cmu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Most microarray and RNA-Seq studies to date have focused on profiling large populations of cells. While such approaches have led to many important results, they tend to overlook the heterogeneity of the population being profiled (Stegle *et al.*, 2015). This may be problematic in cases where the population contains a mixture of different cell types with different regulatory programs. Such cases arise in many different biological processes including cancer (Dalerba *et al.*, 2011), immune response (Shalek *et al.*, 2013) and development (Treutlein *et al.*, 2014). Similarly, in several differentiation protocols (including those utilizing Induced Pluripotent Stem (IPS) cells) only a small fraction of cells differentiate to the intended cell fate (Takahashi and Yamanaka, 2006).

While some work has relied on imaging technologies to analyze the expression of genes in single cells, these techniques are limited to a small number of proteins (Levsky *et al.*, 2002). More recently, new technologies based on RNA-Seq experiments were developed to profile global gene expression in single cells. By profiling different cells in the population the contribution of different cell types to changes in tissue level expression can be analyzed allowing researchers to address several of the problems mentioned above. However, the single cell based approaches have also raised new

computational challenges leading to new methods for the analysis of such data. These include issues related to sample quality (Stegle *et al.*, 2015), issues related to 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 the different components within a specific mixture/time point (Buettner *et al.*, 2015).

However, even with the early work on the analysis of this data, several issues remain. A key problem with single cell RNA-Seq data is the issue of time series analysis. While several methods have been developed for the analysis and modeling of temporal data in population based microarray and RNA-Seq experiments (Bar-Joseph *et al.*, 2012; Bonneau *et al.*, 2006; Patil and Nakai, 2014; Reid and Wernisch, 2015; 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 usually hold for the population as a whole, it clearly does not hold for all individual cells within the population. Consider for example a mixture of two different types of progenitor cells, A and B, each giving rise to a different cell fate CA and CB at the next time point. If we have a cell from A at time t, there is no reason to assume that this cell is correlated with the expression of CB cells at the next time point. Thus, a key issue when analyzing single cell RNA-Seq data is the ability to not only divide the different cells within a specific time point (for example, by clustering (Xu and Su, 2015)) but also to be able to link these cells over time by identifying the subsets of cells that belong to the same trajectory and using these to model the networks that are activated within this specific lineage.

A few recent methods have been developed to address the problem of connecting single cells along a temporal trajectory. However, these methods either completely ignore the time at which the cell was measured (Setty *et al.*, 2016; Trapnell *et al.*, 2014) or completely rely on the measurement time (Bendall *et al.*, 2014; Treutlein *et al.*, 2014) ignoring the fact that cells within a specific timepoint may still be in different developmental states. As we show in Results, methods that ignore the time in which the cell is profiled, while useful for cases where no branching is assumed, may face challenges when attempting to reconstruct the cell fate branching that controls fate decision at the different time points measured. Methods that only rely on time can fail to distinguish between differentiated and undifferentiated cells at a specific time points which, as we argue, leads to conclusions that may be inaccurate. Further, all previous methods utilize a deterministic model and distance function which is less appropriate for accommodating the noisy and stochastic nature of single cell expression data.

In this paper we present the Temporal Assignment of SIngle Cells (TASIC) method. TASIC uses a Hidden Markov Model (HMM) based on a probabilistic Kalman Filter approach to combine time and expression information for determining the branching process associated with time series single cell studies. We discuss the formulation of the model, how learning and inference is performed and how model structure is determined. Once a branching model is determined, each cell is associated with a state in the model and so the different expression trajectories for each cell fate can be reconstructed. As we show by applying our model to myoblast differentiation and lung development data, using the reconstructed cell fate trajectories we can identify key genes involved in the differentiation process for the different fates. In addition, the learned models can be used to infer functional assignment of cells and derive insights about the synchronization of the process being studied. The paper is organized as follows, below we first present the HMM model and then discuss learning, inference and down stream analysis. Then in the results section we discuss the performance of the algorithm on lung data and myoblast data and compare it with other existing methods to predict branching structures.

2 Methods

2.1 HMM model

We use HMMs to represent the branching process associated with dynamic transitions in cell states. HMMs are defined using a set of states, the transitions between these states and the emissions of each state (Fig. 1). In our case the HMM states correspond to potential cell states over time and the transition probabilities define the branching model. Each single cell in the dataset is assigned to one of these HMM states, such that the expression of that cell is assumed to be 'emitted' from the state the cell belongs to, allowing us to infer partial ordering for the measured cells. Given a branching model (for example the one presented in Fig. 1) we would like to learn the parameters for each of the states and, in cases where one state can transition to multiple other states the transition probabilities. This is achieved by using a Maximum Likelihood (MLE) framework which attempts to maximize $P(C|H)$ where C are the expression values and measurement times for the cells and H is the HMM model learned for this structure. If we know the parameters of the model, we can assign cells to states by finding, for each cell C_i, the state s to which it should be assigned. Such state can be computed by setting $P(C_i|H) = max_{s \in S}(C_i, s)$ where S is the set of states in the model and $P(C_i, s) = P(C_i|s)P(s)$. Thus the overall likelihood of the data given the model can be written as $P(C|H) = \prod_{i=1}^{n} P(C_i|H)$. As mentioned above, each state is associated with an expression profile and a time (g_s and t_s in Fig. 1). For a cell C_i with expression profile y_i collected at time v_i belonging to state s, we assume a Gaussian emission model for the gene expression (as commonly done to represent average expression profiles (Levsky *et al.*, 2002; Shapiro *et al.*, 2013; Wu *et al.*, 2014) and a Poisson distribution for time. These two components are independent of each other conditioned on the state. The model also contains transition probabilities between states. We set these probabilities to 0 for states that are not consecutive in time (for example, between state 1 and 3 in Fig. 1) and learn the probabilities for branching events (for example, from state 3 to state 5 in Fig. 1).

2.2 A Kalman filter approach for expression and temporal progression

To model our assumption that the time and expression of consecutive states are correlated we use a Kalman Filter approach that captures our assumption about the developmental trajectories. Specifically, we constrain the magnitude of change in expression and time between consecutive states (whether branching or not). In our model each state s has a single parent state P_s. Given the parent, we constrain the mean expression vector for state s by setting the transition equation as:

$$g_s = Ag_{P_s} + \eta_s, \eta_s \sim N(0, \Sigma_G)$$

The observed average gene expression value of the cells assigned at state s is estimated by the following emission equation:

$$c_s = Bg_s + \eta_s^c, \eta_s^c \sim N(0, \Sigma_C)$$

Σ_G and Σ_C represent the transmission and emission noise respectively. From this transition and emission equation, we perform forward backward technique to estimate the state parameters $\{g_s, \Sigma_G\}$ (Supplement).

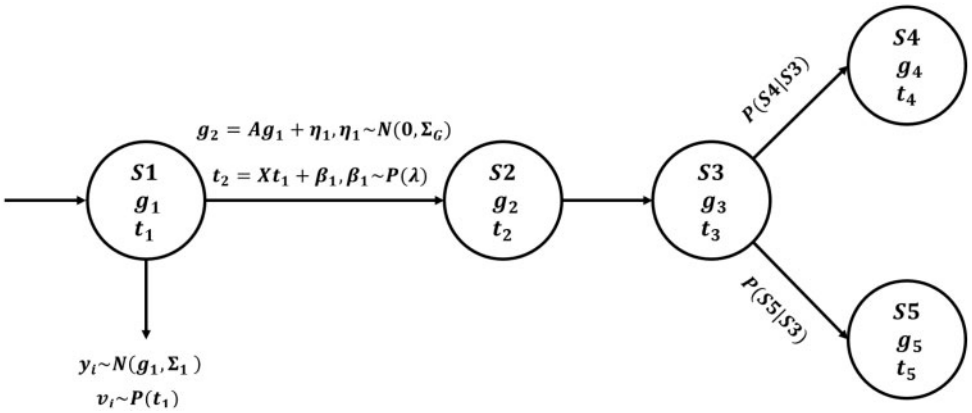

Fig. 1. Graphical representation of an example HMM structure with 5 states. g_s and t_s correspond to the gene expression profiles and average time associated with state s. y_i is the expression profile of cell C_i and v_i is the temporal value of cell C_i. $P(S5|S3)$ and $P(S4|S3)$ are the transition probabilities. See text for a detailed description of the parameters of the model

Similarly the transition equation for time at state s:

$$t_s = Xt_{p_s} + \beta_s, \beta_s \sim P(\lambda)$$

The observed average time value of the cells assigned at state s:

$$ct_s = Yt_s + \beta_s^c, \beta_s^c \sim P(\lambda^c)$$

Here A,B,X,Y are the state transition and emission matrices which we assume to be a unit diagonal and η_s and β_s represents the change in average expression and time between the two states. The full model is presented in Figure 1. We use the same forward backward technique here to estimate the Poisson time parameter for the state t_s (Supplement). After estimating the state parameters, the cells are assigned to the states by maximizing likelihood. For example, for a cell C_i with expression $y_i \in \mathbb{R}^M$ and time v_i in state s: $y_i \sim N(g_s, \Sigma_s)$ and $v_i \sim P(t_s)$. M is the total number of genes. and $P(.)$ denotes Poisson distribution.

2.2 A mixture model for single cell expression levels

Measured expression levels for single cell RNA-Seq data are more noisy than for bulk expression data. Unlike for bulk experiments, in single cell experiments the issue of *dropout* or complete absence of expression for lowly expressed genes is a major problem (Buettner *et al.*, 2015; Shapiro *et al.*, 2013). This issue needs to be taken into account when considering the emission model for our HMMs since cells assigned to specific states may contain a mixture of those with a detected expression for a specific gene and a dropout of that gene.

To address this issue, instead of using a Gaussian emission model for genes, as was used in the past (Ernst *et al.*, 2007) we use a gene specific mixture model. For each gene j the model includes two components, $g^j_{1,s}(x)$ is a Gaussian component whose parameters are learned from cells assigned to that state for which this gene is expressed and $g^j_{2,s}(x)$ which is used to represent the probability of the dropout event for cells assigned to this state:

$$g^j_s(x) = w^j_g * g^j_{1,s}(x) + (1 - w^j_g) g^j_{2,s}(x)$$

$$g^j_{1,s}(x) \sim N(\mu_s, \sigma_s)$$

$$g^j_{2,s}(x) \sim \begin{cases} k, x = 0 \\ 0, otherwise \end{cases}$$

here x is the gene expression measurement of gene j at state s. w^j_g is the mixture weight and k is a constant which represents the likelihood penalty for choosing the point model.

2.3 Learning model parameters

The above model contains several parameters that needs to learned (specifically the state emission parameters and the Kalman filter parameters as well the transition probabilities). In addition, inference is required to assign cells to states.

We use a modified Kalman filter estimation procedure to learn the parameters. For this, we employ an Expectation Maximization (EM) approach where cells are assigned to states in the E step and parameters are computed in the M step. Given initial cell assignments (which can be determined using the time cell were profiled or using clustering based on the expression), we can compute an estimate for the mean expression of each state (g_s) by using a modified forward-backward algorithm. In the forward step we use the values assigned to the parent of state s, p_s, to obtain an expected value for the error based on the Kalman filter transition parameters. In the backward step we include the errors assigned to the differences between state s and its child states as well as the emission of cells assigned to state s. The full derivations used in the EM algorithm are presented in the supplement.

2.4 Structure learning

The optimal HMM structure is learned via search of possible models. We start with the simplest structure (in this paper a three state chain, though any other structure can be used) and learn parameters for such model as discussed above. Next, we increase the complexity of the structure by adding branching. In this paper we constrain the maximum path length for each model to match the number of time points used in the study and the total number of splits to two at each state leading to $\sum_{t=0}^{T-1} 2^t$ possible models, where T is the total number of time points. Since T is usually very small (here we looked at studies with $T=3$ and $T=4$) this procedure is efficient (note that the complexity is only linear in the number of cells which are usually large while the number of time points is usually small). Although in our algorithm we restricted the splits as two way split, among our tested models we manually added one model with three way split (model 25) to test the Lung data which reportedly had five terminal states. The algorithm to find set of possible models is given the Supplement. Each time more branches are added to the structure, the likelihood increases (since the number of parameters increases). Thus, to find the optimal structure with use a penalized likelihood function (Bayesian Information Criteria, (BIC) citepschwarz) to compare the different models. The structure with lowest BIC score is selected as the optimal structure.

2.5 Gene selection

We use only genes that are expressed in a specific fraction of the cells (25% of the total cells in this paper). This reduces the total number of genes from 47k to 27k for myoblast dataset and from 27k to 13k for the lung development dataset. The genes are further filtered based on the normalized standard deviation of their expression values in all cells. In this paper, we have used the top 2500 filtered genes to select the optimal structure though as discussed in Results, robustness analyses indicates that using fewer genes leads to very similar results. Note that once the model is learned and cell are assigned to states we can examine the dynamics of all genes when looking for genes that are DE between different branches of the model regardless of whether these genes were used to learn the model or not.

2.6 Identifying branch and cell fate specific genes

The learned optimal structure often contains more than one terminal state. To determine genes that are uniquely activated or repressed in a specific branch versus other branches the average expression for genes in all cells assigned to a state is calculated. Using these expression values we compute an expression profile for each gene along each branch (from the initial state to the terminal state). Then a standard two-tailed two-sampled *t*-test is performed for each gene to find differentially expressed genes between the two paths. For scenarios with more than two terminal states, the *t*-test is performed once for each branch against the average gene expression of rest of the branches (Fig. 3). Cell fate specific genes are selected based on the *P*-values obtained from the one versus all *t*-test. Only genes with a corrected *P* values < 0.05 are used for further analysis (GO, literature search etc.). The dynamics of the top ranked genes by *P*-value are shown in Figures 3 and 6.

3 Results

To test the TASIC method, and to compare the results to prior approaches, we focused on two time series single-cell RNA seq datasets. The first is an *in vitro* study of the development of Human Skeletal Muscle Myoblast (Trapnell *et al.*, 2014). This dataset contains four time points: 0, 24 h, 48 h and 72 h following a switch from high-mitogen conditions (GM) to low-serum medium (DM) which induces differentiation. At each of the four time points between 49 and 77 cells were captured and mRNA-Seq was then sequenced resulting in a complete gene expression profile for each cell. The total number of cells in the dataset is 271. The second is an *in-vivo* study of mouse lung development (Treutlein *et al.*, 2014). This data consists of single cell RNA-Seq from mouse lung epithelial cells in three developmental and one adult time points (14.5d, 16.5d, 18.5d and adult). The dataset contains between 27 and 80 cells for each of the developmental time points. The cells at 18.5d were assigned in the original paper to one of 5 types: AT1, AT2, bipotential progenitor (BP), Ciliated and Clara. One of these cell types, BP, is new and was first identified by the single cell study.

3.1 Analysis of lung differentiation data

We first discuss the modeling and analysis of the *in vivo* lung data. For this data we used TASIC to select a branching model from several possible models based on BIC (Methods). As can be seen in Figure 2, the model with the highest penalized likelihood score (model 24) had two splits in each of the time points resulting in 4 paths (or, alternatively, 4 terminal states). This model scored higher than a model with 5 terminal states (model 25) indicating that for this data TASIC has only identified 4 different distinct cell types. Looking more closely at the assignment of cells with known markers (Table 1) indicates that while all 4 known cell types (AT1, AT2, Clara and Cilliated) are associated with a specific and distinct terminal state, BP is not associated with any such state. In fact, most BP cells are assigned by the model to earlier states despite their later time point (all marked cells were obtained in 18.5d). Thus, by modeling not just the snapshot expression at the last time point but also the progression of expression along the differentiation pathways, TASIC determines that BP cells are likely cells that have not fully differentiated despite their later developmental time. However, while BP cells from the 18.5d time point were assigned to earlier states in the model, the overall assignments of most other cells agreed with their terminal status and fate.

We have further analyzed the different paths in our model to identify genes that are differentially expressed and so likely unique to that path/fate. Specifically, we compared the top paths in the model (1,2,(4–6)) and lower paths (1,3,(5–7)) (Methods). We analyzed GO enrichment for the top 500 up and down regulated genes between these paths. We found a number of significant GO categories related to structural developmental processes. These include developmental process (GO: 0032502, corrected *P* value = 0.00141) and anatomical structure development (GO: 0048856, *P* value = 0.0479). We have also looked at differentially expressed genes in each of the four paths in the model, 1-2-4,1-2-6, 1-3-5, 1-3-7. Two of the top 10 ranked genes in each path are shown in Figure 3. We have further investigated the role of these predicted genes in lung development. Some of the genes are known to be important in lung development including JUND (Mariani *et al.*, 2002) and FOXA2 (Wan *et al.*, 2004). Other genes, such as Gabpb1 and Naa50 are known interaction partners of other known lung development genes (CIC and JMJD6, respectively (Lee *et al.*, 2011)) and our results suggest that they may be involved in this process as well. The complete list of DE genes for this model is available on the supporting website.

3.2 Robustness analysis and the importance of using time

While the results above indicate that TASIC is able to accurately reconstruct known aspects of the branching trajectory and identify known genes for this process, we wanted to determine single cell data noise and specific model parameters can impact these results. We have thus performed additional analysis to test the robustness of TASIC to random variations in the input data (applying TASIC to different subsets of the cells rather than to the entire dataset) and to variations in the number of DE genes used for learning the model (randomly using 1000 of the top 3000 DE genes). As can be seen in Supplementary Tables S2 and S3, in all cases, the majority of TASIC runs using random subsets of cells or genes identified the same model as the one identified using the full dataset. All other runs identified a similar model (model 23). Unlike model 24, model 23 has only 3 terminal states but is otherwise very similar and is able to successfully identify unique terminal cell fate states as well (Supplementary Table S4). The fact that TASIC was unable to identify model 24 in a small number of runs when using a subset of the cells may indicate that more data may be required to unambiguously separate the final terminal state.

We have also tested the importance of using the time information which, as mentioned above, is a unique aspect of TASIC. For this, we have performed the TASIC analysis without using the time

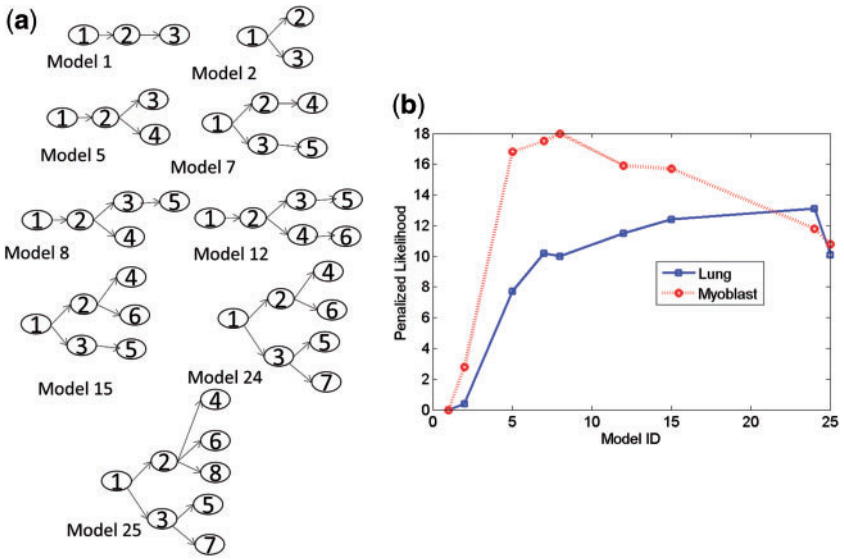

Fig. 2. Penalized likelihood scores for different branching models. (**a**) Structure of 9 of the models tested using our method, each containing at least 3 states. The 9 models include the highest scoring models for the two datasets and a number of other high scoring models out of the 25 models we tested (all models with at most 7 states), see Supplement for complete list of models and their penalized likelihood scores. (**b**) BIC scores obtained for these 9 models for the two datasets we studied in this paper. As can be seen, the optimal models greatly differed between the two datasets indicating the ability of the method to determine an accurate model for a specific set of trajectories

Table 1 Confusion matrix between cell assignments in our models and known cell types as determined using protein markers by (Treutlein et al., 2014) (Model 24)

	State 1	State 2	State 3	State 4	State 5	State 6	State 7
AT1	3	5	2	24	1	4	0
AT2	0	0	2	1	2	0	7
BP	8	1	2	0	0	2	0
Cilliated	0	0	0	1	0	2	0
Clara	1	1	1	0	6	0	2

As can be seen, while 4 of the cell types are associated with a terminal state, BP cells are mainly assigned to earlier states in the model indicating that they may not be completely differentiated despite the time point in which they were sampled.

information for each of the cells (removing transition and emission terms that depend on time). Without the time information, TASIC is unable to identify a model with the correct number of terminal states, identifying instead a model with only two terminal states (model 19). In addition, as can be seen in Supplementary Table S7, without the use of time the ability of TASIC to accurately assign cells to unique fate states is also severely reduced and the resulting model groups most cell fates into single (terminal) state. These results support the use of the time information when learning branching models for cell fate decisions.

3.3 Comparison with prior methods

It is not always easy to compare biological analysis methods since ground truth may not be known. We have thus relied on the lung data for comparisons between TASIC and prior methods for reconstructing expression trajectories from single cell data since for this data known markers were used to determine the cell state for cells at the 18.5d time point. As mentioned in the Introduction, those prior methods either do not use the time the cells were measured in and only rely on the expression data for their branching and ordering or completely rely on the measurement time for their assignments. In addition, none takes into account the stochastic nature of the data, using deterministic distance functions and projections instead.

Specifically, we compared our method to monocle (Trapnell *et al.*, 2014) SCUBA (Marco *et al.*, 2014), Wishbone (Setty *et al.*, 2016) and PCA (Treutlein *et al.*, 2014). Figure 4a presents the branching model obtained by monocle for this data. The Monocle analysis produces a trajectory with a single branch, with the unclassified E14.5 and E16.5 cells largely confined to two of the arms. The other arm includes cells from the E18.5 time point, most of which are AT1 cells while Clara and Cilliated cells appear to form a separate path in the middle of the AT1 path. We have also performed a separate analysis of Monocole without the Clara and Cilliated cells (Supplementary Fig. S2). In this analysis Monocole was able to separate early and adult cells, but unlike Tasic was unable to separate BP and AT1 cells. Wishbone is a recently published technique that also provides pseudo-temporal ordering of cells (Setty *et al.*, 2016). Similar to Monocle, Wishbone also does not utilize the time information and, currently, is only able to reconstruct models with a single branch (two terminal cell fates) which may not be appropriate for all types of developmental processes. While Wishbone identifies a bifurcation for the lung data, since the method is constrained to a single split it is unable to correctly determine the branching process and terminal states. Indeed, as can be seen in Supplementary Table S5, unlike TASIC, Wishbone groups cells from all fates in a single state in its model.

SCUBA (Marco *et al.*, 2014) provides a cellular hierarchy by iteratively clustering and mapping cells from different time points. While it works well for *in vitro* data, for the *in vivo* lung data SCUBA failed to identify any bifurcation events at both the time (Fig. 4c). For this data SCUBA outputted a single chain of three states with no branching. Further investigation of the results indicates that, since there was no bifurcation event detected by SCUBA for this data, the assignment of cells did not change much between

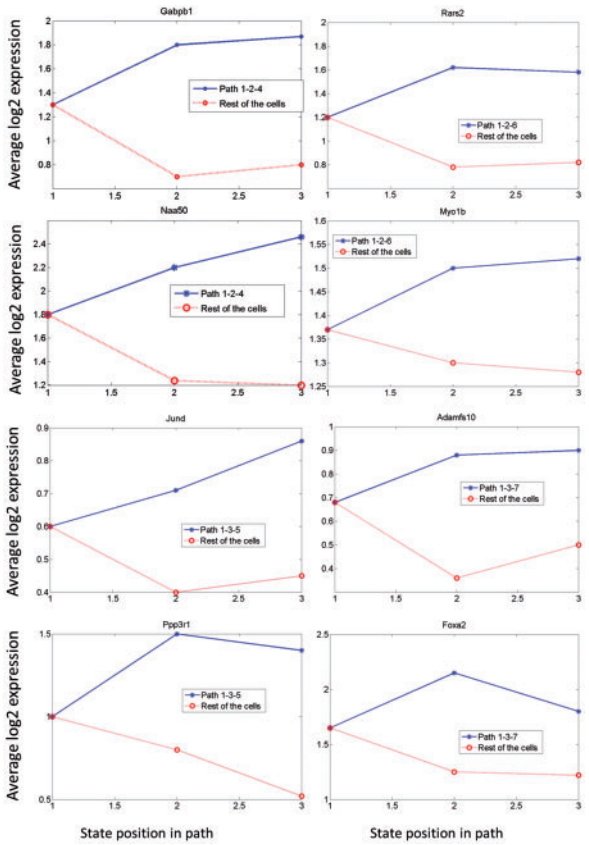

Fig. 3. Top DE genes identified in the lung models. Y axis is the average expression for the gene in cells assigned to the state along the path. X axis is the order of the state along the path. The top 2 up regulated genes for each of the 4 cell type specific paths in the optimal model for the lung data. Several of these genes are known to be involved in lung development. However, our ability to separate paths to different cell fate decisions allows the model to better match genes with the fates they likely regulate

iterations and so the distribution of cells in the final states are dominated by the initial clustering assignment.

In addition to methods specifically developed to address pseudo-ordering, several papers rely on PCA to analyze single cell expression data (Treutlein *et al.*, 2014). We have thus analyzed the top 1000 DE genes in the lung data using PCA. The results are presented in Figure 4c. As can be seen, while PCA is able, to a large extent, to separate the data according to the sampled time, it is unable to reconstruct branching trajectories. Moreover, the terminal cell fates were not well separated by PCA and some cells from the last time point, from mixed fates, are very separated from the others by PCA making it impossible to determine they relatedness from PCA alone.

3.4 Analysis of myoblast differentiation

'We next looked at the Myoblast differentiation data. Again, we used TASIC to compare several different models for this data based on the penalized likelihood score. For this data, model 8 had the highest penalized score (Fig. 2). The branching structure for model 8 agrees well with the results presented in the original paper (which also included a single split) and with prior biological knowledge about this differentiation process (Abmayr and Pavlath, 2012; Tapscott, 2005). While the overall model learned by TASIC agrees with the monocle branching model, the cell assignments differed for some of the cells (see Supplementary Table S8 for a confusion matrix). These differences can be attributed to the different ways in which expression data is used by the two methods. Whereas TASIC uses a probabilistic model, the monocle ordering is based on a deterministic model which uses the first two independent components of the overall gene expression matrix to construct the ordering.

Unlike the lung development data, where for some cells known markers were used to determine the specific cell state, the Myoblast data does not include similar 'ground truth' markers. Thus, to assess whether the model and cell assignments determined by our method are biologically relevant we analyzed genes that are differentially expressed (DE) between the two paths. We first looked at the enrichment of known Gene Ontology (GO) categories related to myoblast differentiation with the set of DE genes. As shown in Supplementary Figure S1 we see significant overlap between key categories and genes identified as DE between the top differential pathway (1-2-4) and the bottom non differentiating interstitial mesenchymal pathway (1-2-3-5) in the model identified by TASIC. These categories include 'myoblast differentiation' (corrected *P*-value of 0.0034), striated muscle differentiation (*P*-value 0.0), myotube differentiation (*P*-value 0.0071), skeletal muscle cell differentiation (*P*-value 0.00) and a number of other relevant categories (see Supplementary Fig. S1).

We investigated the dynamics of some known myoblast differentiation genes. We observe that the expression profiles reconstructed for these genes using cells assigned to the top and bottom path are very different (see Fig. 5). Specifically, genes that are known to drive myoblast differentiation, including MYOG, MSTN and MEF2C (Tapscott, 2005) are upregulated in the 1-2-4 path while they are much lower in the lower path (1-2-3-5). In contrast, ID3 and PBX1 which are known to have a transient down regulation effect on differentiating of cells (Trapnell *et al.*, 2014) are down regulated at the top path and higher in the bottom path. Given these trajectories we conclude that the 1-2-4 path most likely contains the myoblast cells whereas mesenchymal cells are assigned to the 1-2-3-5 path.

In addition to known genes (based on their GO annotations) associated with these paths, our model identifies several novel genes as DE between these two paths. Expression trajectories determined for these genes are presented in Figure 6. These genes include SDF2L1 which is determined to be upregulated in the myoblast path (1-2-4). SDF2L1 is found to be highly expressed in developed muscle cell (Sterrenburg *et al.*, 2004), which is consistent with our predicted upregulation in later states of myoblast differentiation. The study uses a different cell line of skeletal myoblast cells (GDS943 (Sterrenburg *et al.*, 2004)). The same study identified ARL2BP as transiently down regulated in the differentiation process, which is also consistent with our model prediction. A full list of the DE genes and their expression pattern (up or down regulated in the myoblast path) is available from the supporting website. We also find CACYBP up regulated in differentiation path 1-2-4. The effect of this gene on rat neonatal cardiomyocytes was studied in (Au *et al.*, 2006). Their conclusion was that the gene might play a key role in cardiomyogenic differentiation. Based on their prediction and our study, we can also predict that the gene might also be a key one in human myoblast differentiation. We also found that CBX1 gene is known interaction partner of MYOD, which actively takes part in myoblast differentiation (Philipot *et al.*, 2010).

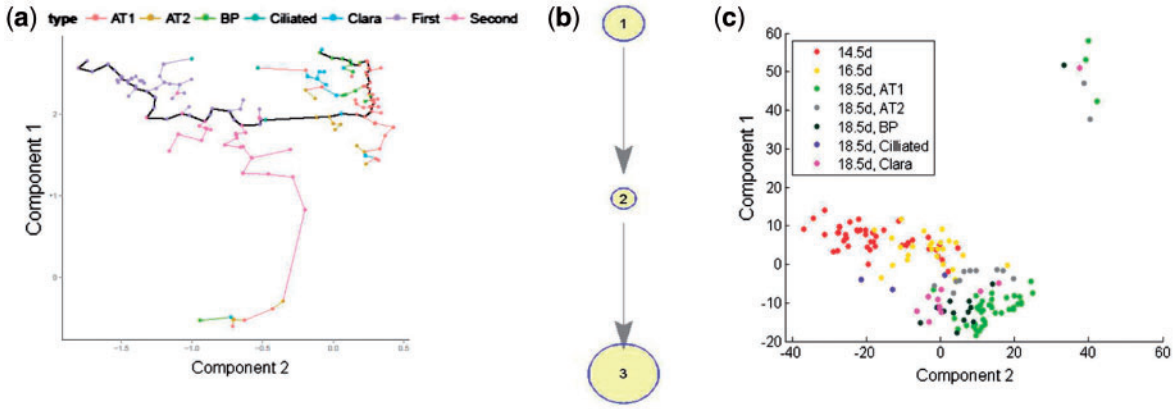

Fig. 4. Comparison with prior methods. (**a**) monocle: Results of Monocle analysis of the lung data. As can be seen, the reconstructed paths correctly separate the first two time points but provide little details about the groupings in the third time point. (**b**) SCUBA reconstruction for the same data. Diameter of nodes in the SCUBA model represents the number of cells assigned to these states by SCUBA. As can be seen, cells are mostly assigned by SCUBA to the first and last states though no branching is detected by the method and so all cells in the same time point were assigned to a single state. (**c**) PCA analysis. The Lung data was projected onto the first two principal components. Colors represent cells from different time points for the first two time points and cell fates for the last time point. As can be seen, while the PCA result provides some information on the trajectory of the expression pattern, no clear branching is observed and cells from different fates are mixed making it hard to infer the trajectory and progenitors for the different cell fates

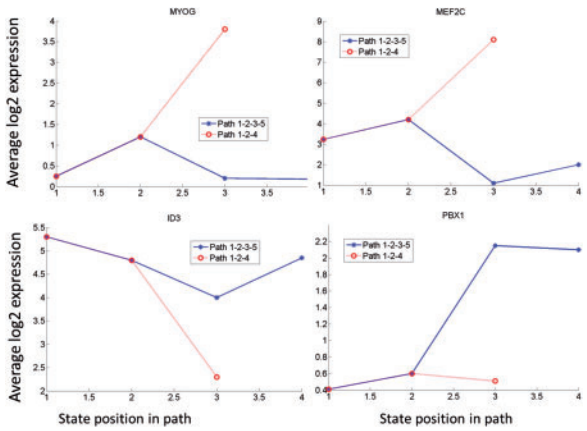

Fig. 5. Dynamics of known gene markers of Myoblast differentiation in our predicted model. The previously reported dynamics of these genes match well their expression along the 1-2-4 path making it the likely candidate for the differentiation pathway. The other path most likely correspond to interstitial mesenchymal cells

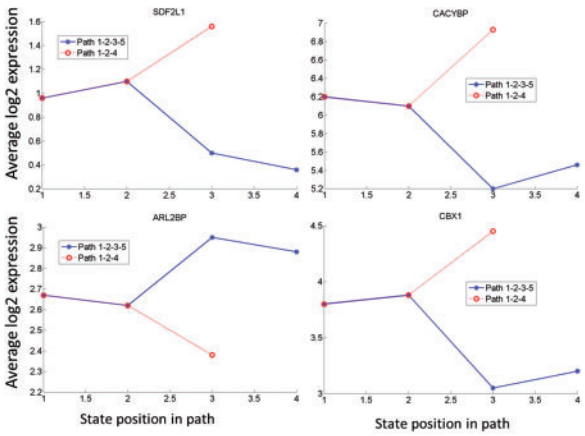

Fig. 6. Gene dynamics for the Myoblast model. Expression profiles for the top DE genes between the 1-2-4 and 1-2-3-5 paths in the model selected for the Myoblast differentiation data

4 Discussion

While the ability of profiling gene expression levels at the single cell level can lead to several new and important insights, it also raises new computational and analysis challenges. First, the expression data obtained in such experiment is much nosier and so models that analyze it should take that into account. Second, for time series data it is hard to determine the assignment of cells to developmental pathways, especially for cells measured at intermediate time points since in most cases no known markers exist to assign such cells. Finally, while the time in which the cell was sampled may be important for the correct assignment of the cell, there are several cases in cells, even in *in vivo* studies, are not fully synchronized leading to cells that are at different points along the differentiation trajectory to be sampled at the same time (Trapnell *et al.*, 2014).

TASIC uses a HMM model which represents the trajectories of cell fate decisions using a set of discrete states. While some prior methods have also utilized discrete states (for example, SCUBA) others have used a more continuous representation (monocle, Wishbone, etc.). There are pros and cons to both modeling strategies. On the one hand, continuous time assignment models are able to provide better resolution about the time and state assignments for each cell and can also help in determining gene trajectories since more points are provided along each branch. On the other hand, such models lack the ability to aggregate cells in order to learn accurate transitions from one state to the other (as we have shown with the lung data comparisons) and specific parameters (for example, expected expression for specific genes for a fate). Thus, both types of models are likely to yield different insights and should be applied when analyzing single cell time series data.

Using the reconstructed models researchers can determine both the number of terminal fates in their studies. As we showed, the ability to accurately assign cells to paths enabled us to determine that while so called bipotential progenitor (BP) cells indeed possess a unique expression profile when compared to other cells measured at 18.5d, these are likely cells that have not yet fully differentiated and are thus more closely related to cells in earlier time points. While these cells may be able to differentiate into both AT1 and AT2 cells,

as claimed by Treutlein *et al.*, the assignments by TASIC do not rule out an alternative explanations. First, it is possible that BP cells exist as a distinct subpopulation of cells at time points earlier than 18.5d, however, this possibility cannot be reconciled with the definition of BP cells proposed by Krasnow and co-workers (Desai *et al.*, 2014; Treutlein *et al.*, 2014) since their definition is based on the existence of this cell at 18.5d. Second, it is possible that BP cells do not represent a distinct functional population of cells at 18.5d, but rather comprise cells in transition between an AT2 and AT1 phenotype, and thus represent cells that have not fully differentiated. Indeed this second possibility is consistent with a long-standing model of alveolar differentiation (Adamson and Bowden, 1975; Kotton and Morrisey, 2014). In that model, distal lung bud epithelial cells differentiate first into early AT2 like cells, some of which are proliferative, and AT1 cells are then formed prior to birth by subsets of AT2-like cells that downregulate the AT2 genetic program while upregulating the AT1 program. This model suggests that some cells at 18.5d can still co-express portions of the transcriptomes characteristic of both cell types (AT2 and AT1) but would later associate preferentially with either AT2 or AT1 cells, as our model suggests. While the exact date of the common epithelial progenitor is still not fully determined (most put it earlier than E14.5 though others places it between E12.5 and E16.5) we have modeled a full hierarchy starting at E14.5. We do not claim that the E14.5 data includes such progenitor. Instead, we simply used the time points measured by Truetlein et al for our model.

Comparison of TASIC to other methods indicates that for some datasets it can provide a more detailed assignment that better groups cells to discrete states according to their function and developmental time. We note that part of the success of TASIC on these datasets may result from its reliance on the assumption of discrete states. In contrast, some of the other algorithms for single-cell trajectory inference including Monocle, Wishbone and DPT assume that the data in the experiments is collected from a smooth, continuous process. When these assumptions are violated (for example in the lung data where the cells are collected from developing mice rather than from cell lines) these prior methods may not be able to fully capture the developmental trajectories.

Another important aspect of the model is the ability to identify the set of genes that are uniquely associated with each terminal state and the path leading to it. Given the cells assigned to states along this path we can reconstruct the expression profiles of such genes and identify those that are DE between different paths in the model. As we showed, the set of DE genes agrees well with prior knowledge about key lung development and myoblast differentiation genes while at the same time identifying several additional genes associated with specific fates. Several of these genes may have been missed when studying population of cells because of their different trajectories in the pathways leading to these fates. Only by analyzing single cell data and reconstructing models with specific cell assignments can we accurately identify such genes.

TASIC is implemented in MATLAB and can be downloaded from the Supporting Website. While TASIC is appropriate for current datasets which usually profile only a small number of time points (resulting in a limited search space) an important future direction is to further improve TASIC so that it can learn both the structure (number of states and branching pattern) and parameters simultaneously. While this will likely require a heuristics search strategy (Schulz *et al.*, 2012) it can still enable the search a much larger set of possible models.

Given the ability of single cell time series studies to answer questions that could not be addressed by bulk studies we believe that methods that can efficiently and accurately analyze this data would prove to be essential for fully utilizing these type of experiments.

Funding

The work is supported in part by National Institute of Health [grants number 1R01HL128172 to D.K. and 1U01HL122626-01 to Z.B.J] and by the National Science Foundation [grant number DBI-1356505 to Z.B.J.].

Conflict of Interest: none declared.

References

Abmayr,S.M. and Pavlath,G.K. (2012) Myoblast fusion: lessons from flies and mice. *Development*, **139**, 641–656.

Adamson,I. and Bowden,D. (1975) Derivation of type 1 epithelium from type 2 cells in the developing rat lung. *Lab. Investig. J. Tech. Methods Pathol.*, **32**, 736–745.

Au,K.-W. *et al.* (2006) Calcyclin binding protein promotes DNA synthesis and differentiation in rat neonatal cardiomyocytes. *J. Cell. Biochem.*, **98**, 555–566.

Bar-Joseph,Z. *et al.* (2003) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 10146–10151.

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **13**, 552–564.

Bendall,S.C. *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, **157**, 714–725.

Bonneau,R. *et al.* (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.

Buettner,F. *et al.* (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. *et al.* (2011) Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.*, **29**, 1120–1127.

Desai,T.J. *et al.* (2014) Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature*, **507**, 190–194.

Ernst,J. *et al.* (2007) Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, **3**, 74.

Kotton,D.N. and Morrisey,E.E. (2014) Lung regeneration: mechanisms, applications and emerging stem cell populations. *Nat. Med.*, **20**, 822–832.

Lee,Y. *et al.* (2011) Atxn1 protein family and CIC regulate extracellular matrix remodeling and lung alveolarization. *Dev. Cell*, **21**, 746–757.

Levsky,J.M. *et al.* (2002) Single-cell gene expression profiling. *Science*, **297**, 836–840.

Marco,E. *et al.* (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A*, **111**, E5643–E5650.

Mariani,T.J. *et al.* (2002) Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *Am. J. Respir. Cell Mol. Biol.*, **26**, 541–548.

Patil,A. and Nakai,K. (2014) Timexnet: Identifying active gene sub-networks using time-course gene expression profiles. *BMC Syst. Biol.*, **8**, S2.

Philipot,O. *et al.* (2010) The core binding factor CBF negatively regulates skeletal muscle terminal differentiation. *PloS One*, **5**, e9425.

Reid,J.E. and Wernisch,L. (2015). Pseudotime estimation: deconfounding single cell time series. *bioRxiv*, 019588.

Schulz,M.H. *et al.* (2012) Drem 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst. Biol.*, **6**, 104.

Setty,M. *et al.* (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.*, **34**, 637–645.

Shalek,A.K. *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**, 236–240.

Shapiro,E. *et al.* (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618–630.

Stegle,O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133–145.

Sterrenburg,E. *et al*. (2004) Large-scale gene expression analysis of human skeletal myoblast differentiation. *Neuromusc. Disorders*, **14**, 507–518.

Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.

Tapscott,S.J. (2005) The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. *Development*, **132**, 2685–2695.

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

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

Wan,H. *et al*. (2004) Foxa2 is required for transition to air breathing at birth. *Proc. Natl. Acad. Sci. U. S. A*., **101**, 14449–14454.

Wu,A.R. *et al*. (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–46.

Xu,C. and Su,Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, btv088.

Young,W.C. *et al*. (2014) Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Syst. Biol*., **8**, 47.