SOMSC: Self-Organization-Map for High-Dimensional Single-Cell Data of Cellular States and Their Transitions

Tao Peng[a], Qing Nie[a,b,*]

[a]*Department of Mathematics, University of California Irvine, Irvine, 92697, USA*
[b]*Department of Developmental and Cell Biology and Department of Biomedical Engineering, University of California Irvine, Irvine, 92697, USA*

Abstract

Measurement of gene expression levels for multiple genes in single cells provides a powerful approach to study heterogeneity of cell populations and cellular plasticity. While the expression levels of multiple genes in each cell are available in such data, the potential connections among the cells (e.g. the cellular state transition relationship) are not directly evident from the measurement. Classifying the cellular states, identifying their transitions among those states, and extracting the pseudotime ordering of cells are challenging due to the noise in the data and the high-dimensionality in the number of genes in the data. In this paper we adapt the classical self-organizing-map (SOM) approach for single-cell gene expression data (SOMSC), such as those based on single cell qPCR and single cell RNA-seq. In SOMSC, a cellular state map (CSM) is derived and employed to identify cellular states inherited in the population of the measured single cells. Cells located in the same basin of the CSM are considered as in one cellular state while barriers among the basins in CSM provide information on transitions among the cellular states. A cellular state transitions path (e.g. differentiation) and a temporal ordering of the measured single cells are consequently obtained. In addition, SOMSC could estimate the cellular state replication probability and transition probabilities. Applied to a set of synthetic data, one single-cell qPCR data set on mouse early embryonic development and two single-cell RNA-seq data sets, SOMSC shows effectiveness in capturing cellular states and their transitions presented in the high-dimensional single-cell data. This approach will have broader applications to analyzing cellular fate specification and cell

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

1. Introduction

Heterogeneity of cell populations is considered functionally and clinically significant in normal and diseased tissues, and transitions among different subpopulations of cells play key roles in cell differentiation during development or disease recurrence (Tsioris *et al.*, 2014; Wilson *et al.*, 2014; Saadatpour *et al.*, 2015). In recent years, single-cell gene expression profiling technologies have emerged as an important tool in dissecting heterogeneity and plasticity of cell populations and in analysis of cell-to-cell variability on a genomic scale (Saliba *et al.*, 2014). For example, mammalian pre-implantation development has been analyzed from oocyte stage to morula stage in both human and mouse embryos using single-cell RNA sequencing (Xue *et al.*, 2013; Yan *et al.*, 2013) to identify stage-specific transcriptomic dynamics; In breast cancer, gene expression profiles of tumor subpopulations along a spectrum from low metastatic burden to high metastatic burden have been obtained using qPCR at the single-cell level (Lawson *et al.*, 2015); and multiple new phenotypes in healthy and leukemic blood cells have been identified using gene expression signatures through analysis of single-cell data (Levine *et al.*, 2015).

Distinguishing or clustering measured cells computationally through their transcriptomic data (e.g. gene expression) is challenging. The number of genes measured is usually significantly larger than the number of cells (Jiang *et al.*, 2004). Another challenge is that a group of cells collected at one temporal point from one sample may not be perfectly ordered in time compared to the cells collected at a slightly different temporal stage, due to cell-to-cell variability in sampling and the nature of unsynchronized cell divisions (Heppner, 1984; de Vargas Roditi and Claassen, 2015). As a result, the pseudotemporal ordering of single cells in a high-dimensional gene expression space was introduced (Trapnell *et al.*, 2014b). The difficulty in analyzing single-cell data becomes particularly evident for systems of differentiation in which new cell types emerge as time advances, such as identifying lineage specific markers of different cell subtypes during the development of murine lung (Treutlein *et al.*, 2014) and finding the differentiation trajectory of skeletal muscles (Trapnell *et al.*, 2014a).

Temporal ordering of single cells, grouping cells of similar transcriptomic profiles, finding transition points, and determining branches are the key steps

in analyzing the single-cell data. Clustering methods based on Principle Component Analysis (PCA) or Independent Components Analysis (ICA), such as MONOCLE algorithm (Trapnell *et al.*, 2014a), allow grouping cells according to the specific properties of interest. Several other clustering-based methods such as SPADE (Qiu *et al.*, 2011), t-SNE (Van der Maaten and Hinton, 2008), and viSNE (Amir *et al.*, 2013) were introduced to identify subpopulations within the measured cells without an explicit temporal ordering of cells. In the Wanderlust algorithm (Bendall *et al.*, 2014), a pseudo-temporal ordering technique incorporated the continuity concept in branching process, however, with an assumption that the cells consist of only one branch during differentiation. To study nonlinearity of the branching process in differentiation, a diffusion map technique was adapted to single-cell data by adjusting kernel width and inclusion of uncertainties, enabling a pseudo-temporal ordering of single cells in a high-dimensional gene expression space (Haghverdi *et al.*, 2015). SLICER is another method to capture highly nonlinear gene expression changes and select genes related to the process, and to detect multiple branches (Welch *et al.*, 2016). TASIC was developed to determine temporal trajectories, branching and cell assignments using a probabilistic graphical model (Rashid *et al.*, 2017). UNCURL incorporates prior knowledge to perform the cell state identification (Mukherjee *et al.*, 2017). With a focus on modeling the dynamic changes associated with cell differentiation, a bifurcation analysis method (SCUBA) was developed to extract lineage relationships (Marco *et al.*, 2014).

The Waddington landscape (Goldberg *et al.*, 2007) of gene expression provided a global and convenient view in describing stem cell dynamics and lineages. In this method, a forward stochastic model on a small gene network was first derived, then a landscape of cellular states was obtained by constructing an energy function that depends on each gene in the modeled regulatory network (Foster *et al.*, 2009; Zhang *et al.*, 2013; Zhou and Huang, 2011; Chen *et al.*, 2015). The prior knowledge of the gene regulatory network needs to be known in this method, and the landscape calculation did not require dimension reduction in the gene space. However, due to the computational cost associated with sampling solutions of stochastic differential equations or solving equations of probability density functions of the gene states, the network size in the landscape calculation can't be too large (Wang *et al.*, 2011).

Here, we propose a new method to analyze single-cell gene expression data by combining a machine learning method and a concept similar to the

landscape (Wang *et al.*, 2011; Kohonen, 1998) (Figure 1). In this approach, the high-dimensional data of single cells is first reduced to two dimensions through a classical unsupervised artificial neural network (ANN) method: a self-organizing map (SOM) (Kohonen, 1998) in which the topological properties of the input data are preserved through a neighborhood function. The cellular states are then identified by the watershed algorithm based on the U-matrix calculated by SOM (Vincent and Soille, 1991; Najman and Schmitt, 1996). By building transition paths among the cellular states, we obtain a cellular state map (CSM). In this map, the barriers separating different states provide information on transitions between cellular states. Moreover, a replication probability and transition probabilities are estimated in the cellular state transition paths. Next, the state-driven genes differentially expressed during a cellular state transition are determined by t-test, and then pathway enrichment analysis for each transition is performed based on the list of state-driven genes. In this approach, the transition path among the cellular states leads to a pseudo-temporal ordering of the cells. To study effectiveness and capability of the approach, we apply SOMSC to a set of simulated data and three real data sets based on qPCR or RNA-seq collected from cell at various stages of differentiation.

2. Methods

SOMSC has three major functions: identifying cellular states, reconstructing cellular state transition paths and building the pesudotime ordering of cells. The algorithm of SOMSC consists of six main steps as follows (Figure 1).

Step 1: Calculate a topographic chart of single cell data using a self-organizing map

SOMSC takes single-cell RNAseq and single-cell PCR data as the inputs $G = (g_1, g_2, \ldots, g_N)^T$, where g_i is the vector with the length n of the gene expression levels for the i-th sample and N is the number of the samples. Since two kinds of datasets are collected using different technology platforms, different preprocessing methods are necessary (See details in Supplementary file). A topographical chart of high dimensional expression data is calculated by a SOM. A SOM is an effective way of analyzing topology of high-dimensional data by projecting the data to low-dimensional surfaces through a rectangular, a cylindrical, or a toroidal map (Kohonen, 1998). In

the SOM, an ordered set of model vectors $x \in R^n$ is mapped onto the space of observation vectors $m_i \in R^n$ through the following iteration processes:

$$m_i(t+1) = m_i(t) + h_{c(x),i}(x(t) - m_i(t)) \tag{1}$$

where t is a regression index. This regression procedure is performed recursively for each sample $x(t)$. The scalar multiplier $h_{c(x),i}$ is a Gaussian neighborhood function, acting like a smoothing or blurring kernel over SOM computational grids:

$$h_{x(x),i} = \alpha(t) e^{-\frac{||r_i - r_c||^2}{2\sigma^2(t)}} \tag{2}$$

where $0 < \alpha(t) < 1$ is a learning-rate factor, which decreases monotonically in each regression step; $r_i \in R^2$ and $r_c \in R^2$ are the computational grid locations, and $\sigma(t)$ corresponds to the width of the neighborhood function that also decreases monotonically in each regression step. The subscript $c = c(x)$ is obtained when the following condition is achieved:

$$||x(t) - m_c(t)|| \leq ||x(t) - m_i(t)|| \tag{3}$$

Consequently, $m_c(t)$ is the "winner" that matches the best with $x(t)$. The comparison metric $||\bullet||$ is selected as the Euclidean metric for this study in Eq. 2, and Eq. 3. If multiple functions $c(t)$ satisfy Eq. 3 with discrete-valued variables, $c(t)$ is chosen randomly among those functions for the winner. In addition, a toroid map is used to reduce edge effects in the data on the overall mapping (Vesanto *et al.*, 1999). Applying the SOM to the single-cell gene expression data leads to a unified distance matrix (U-matrix) $U(x, y)$, representing distances between neighboring map units, where x and y denote the coordinates of a plane (Kohonen, 1998). Accordingly, $U(x, y)$ defines a 2-dimensional topographic chart.

Step 2: Identify basins of the topographic chart by the watershed algorithm

The watershed segmentation algorithm is employed to identify the basins of the topographic chart. By the SOM the high dimensional single cell data is projected onto $U(x, y)$ denoting the distance between adjacent grids. The algorithm identifies the boundaries of basins by constantly pouring water into the mountains of the chart (Najman and Schmitt, 1994). As water levels rise, the boundaries of basins are estimated (See more details of the watershed segmentation algorithm in Supplementary file). Let nb denote the number of basins in the topographic chart, and $C_1, C_2, \ldots, C_{nb}$ represent the

identified basins. The height of the barrier between two adjacent basins is the number of genes differentially expressed in those cells of the two adjacent basins, which are calculated by t-test with 5% significance. The adjacent matrix of basins is then calculated: $T = (t_{ij})_{n,n}$, where t_{ij} is the height of the barrier between C_i and C_j and $t_{ij} = 0$ means that C_i and C_j are not adjacent with each other.

Step 3: Classify the cellular states and construct their transition paths

The cellular states are classified based on the adjacent matrix calculated in Step 2. In order to avoid the over-segmentation occurrence when applying the watershed algorithm, we merge any two basins C_i and C_j when C_i and C_j form a cycle, which means the height of the barrier between these two basins is the smallest one between the basin C_i and all its adjacent basins as well as the smallest one between the basin C_j and all its adjacent basins. (C_i,C_j) is called a merging pair. After the merging, all cells in one basin labeled as S_i, (i = 1,...,m) are in the same cellular state. The adjacency matrix D =$(d_{ij})_{m,m}$ represents the distances between the cells in S_i and the ones in S_j (the default distance metric in our algorithm is the l_1 norm), and can be considered as an undirected network for cellular states in SOMSC, which is the cellular state network. All nodes in the network are labeled by the cellular states S_i (i = 1,2,...,m). The weight of the edge connecting the node S_i and the node S_j is d_{ij} in the adjacency matrix D and there is no edge between the node S_i and the node S_j if $d_{ij} = 0$. All edges in the cellular state network are denoted by F. We assume that the starting cellular state is S_s where $s \in \{1, 2, \ldots, m\}$. First, we find out the edge pair in the path (S_i, S_{i_p}) for each node S_i where $i \in \{1, 2, \ldots, m\}$. S_{i_p} is the closet adjacent cellular state of S_i. All edge pairs are denoted by E. Second, we make all nodes of the cellular states reachable. A node S_i is reachable when there exists a path from S_i to the node S_s of the starting cellular state, which consists of the edge pairs in E. Here we denote the set of all reachable nodes by R and the set of the rest nodes by Q. If the set Q is not empty, then we will find out the edges from $F \backslash E$ (the element in F but not in E) which connect the node in R and the one in Q and the one with smallest weight is selected to add to the set E and the sets of R and Q are updated correspondingly. We repeat these steps until Q is empty since each iteration results in more reachable nodes in R. Accordingly, all nodes become reachable with the edge pair set E, and the path P_i from S_i to S_s consists of a subset of the edge pair set E. Finally, all the paths P_i (i = 1,2,...,m) are combined to generate a transition path

tree of the cellular states. From $P_i = (S_{i_1}, S_{i_2}, \ldots, S_{i_{l_i}})$ ($i_k \in \{1,2,\ldots,m\}$, $k = 1, 2, \ldots, l_i$, $i_1 = i$, $i_{l_i} = s$) we can extract the transition pairs (S_{i_k},$S_{i_{k-1}}$) (k =1,2, ..., l_i), which mean that the cellular state S_{i_k} is the parent of $S_{i_{k-1}}$. All transition pairs determine the transition path tree (S_{p_1}, ..., S_{p_m}) together, where S_{p_i} is the parent cellular state of S_i in the path tree and S_i is the daughter cellular state of S_{p_i}.

Step 4: Construct the cellular state map for all cells

Each cell in the data is assigned a relative location in the cellular state map using the following method. Not all cells in the data show up in the topographic chart since only winners of the grids in the SOM stay, suggesting known cellular states of those cells. We then use k-nearest neighbors (KNN) algorithm to identify the states of the remaining cells (Altman, 1992).

To visualize the topologic structure of the data we define the cellular state map (CSM) as follows. First, we calculate the convolution function M by the following equations.

$$M(j,k) = \sum_p \sum_q \hat{U}(p,q) K(j-p+1, k-q+1) \tag{4}$$

where

$$\hat{U}(p,q) = \begin{cases} D(p,q) & (p,q) \text{ is on the boundaries} \\ 0 & \text{otherwise} \end{cases}$$

K is a kernel matrix whose elements define how to remove the high frequency components of the original data $\hat{U}$. The size of kernels may be different from the size of the $\hat{U}$. Small-sized kernels can smooth data containing only a few frequency components whereas larger size kernels can provide better precision for tuning frequency response, resulting in a smoother output. Accordingly, the CSM is the central part of the convolution M whose size is equal to the size of $\hat{U}$. Each cell is plotted as a three-dimensional sphere, whose center represents that cell's position in the CSM.

Step 5: Detect the state-driven genes and their enrichment analysis in each transition

We next identify the differentially expressed genes for each cellular state transition. We perform a t-test for the expression levels of each gene in all cells involved in the cellular state transition. A gene is taken to be state-driven if the p-value is within 1%. Accordingly, a list of all state-driven genes

for each cellular state transition can be obtained. The gene list is then used as an input into the Enrichr database (http://amp.pharm.mssm.edu/Enrichr/), which defines the members of each pathway as a gene list. Fisher's exact test is used to obtain a p-value for the number of pathway members present among the state-driven genes we obtained. If the p-value falls within our critical region (5%), we determine the pathway to be significantly enriched (Chen *et al.*, 2013).

Step 6: Estimate the cellular state replication probability and cellular state transition probabilities and determine the pseudo-time ordering of each cell during the cellular state transitions

One goal of SOMSC is to estimate the cellular state replication probability and the cellular state transition probability for each cellular state. First, we calculate the centroid for each cellular state and denote CT_i as the centroid of the cellular state S_i. Second, we define the cellular state replication axis PR_i and the transition axes DE_{d_1},DE_{d_2},...,DE_{d_k} for the corresponding k daughter cellular states S_{d_1},S_{d_2},...,S_{d_k}. They are calculated by $PR_i = CT_p - CT_i$ and $DE_{i_j} = CT_{i_j} - CT_i$ where j = 1,2,...,k; CT_p is the centroid of the parent cellular state of S_i; and C_{i_j} is the centroid of the jth daughter cellular state of S_i. Then we project the data $g_{i_1}-CT_i$,$g_{i_2}-CT_i$,...,$g_{i_{n_i}}-CT_i$, where n_i is the number of cells in the cellular state S_i, to the cellular state replication axis and the cellular state transition axises. Third, we obtain the vector (Pro_p,Pro_c, Pro_{i_1},Pro_{i_2},...,Pro_{i_k}) for the expression vector of each cell $g_{i_l}(l$ =1,2,...,$n_i)$, where Pro_p is the projection of the data of the l-th cell in the cellular state S_i to the cellular state replication axis PR_i, Pro_c is the distance between g_{i_l} and CT_i, and $Pro_{i_r}(r$ =1,2,...,$k)$ is the projection of g_{i_l} to the cellular state transition axis $DE_{i_r}(r$ =1,2,...,$k)$. We then determine the replication/transition state of the cells as follows: If the minimum positive element in (Pro_p, Pro_c,Pro_{i_1},Pro_{i_2},...,Pro_{i_k}) is the projection to one cellular state transition axis DE_{i_j}, then it means the l-th cell will transition to the daughter cellular state S_{i_j}. Otherwise, the l-th cell is in the replication cellular state. The number of cells in the replication and transition states are obtained and the normalized numbers are taken as the replication and transition probabilities, respectively. Fourth, the pseudotime ordering of each cell is determined by its projection to the replication axis or the transition axes or the distance from the cell to the centroid of the cellular state S_i. If the cell $g_{i_l}$is in the replication state and Pro_p is positive, then we get the ordering pair $(1, -Pro_p)$. If the cell g_{i_l} is in the replication

state and Pro_p is negative, then we get the ordering pair (2,Pro_c). If the cell g_{i_l} is in the transition state, then we get the ordering pair (3,$Pro_{\min}$), where $Pro_{\min}$ is the minimum positive element of $\{Pro_{i_1}, Pro_{i_2}, \ldots, Pro_{i_k}\}$. In this way, we get the ordering pair for each cell in the cellular state S_i. The pseudotime ordering of cells in S_i is based on the ordering pairs. We compute the pseudotime ordering of cells in different cellular states based on the cellular transition path. Finally, we construct the trajectories of the expression levels of each gene from all cells along the pseudotime ordering. Here we interpolate the average expression levels of the gene in disjoint small groups of cells along the pseudotime ordering.

Generate the simulation data

In order to effectively evaluate performance and choices of parameters of SOMSC, we next construct a toy system consisting of a small number of genes to mimic single-cell gene expression data. There are three stages in the system, and in each stage one type of cells makes a transition to two other types of cells (Figure 2A). Together, seven types of cells with three branches are present in the system. The cellular types are defined by the expression levels of six genes (Figure 2A). Specifically, in Type 1 cells Gene A and Gene B are activated and the other four genes are silenced; in Type 2 cells Gene A, Gene C, and Gene D are activated; in Type 3 cells Gene B, Gene E, and Gene F are activated; when one of Gene A and Gene B and one of Gene C, Gene D, Gene E and Gene F are activated, four other types of cells in the third stage are then defined as Type 4, Type 5, Type 6, and Type 7 cells, respectively.

The system of three-toggle modules consisting of six genes is modeled through a system of stochastic differential equations (Haghverdi *et al.*, 2015; Chen *et al.*, 2005; Ocone *et al.*, 2015). Starting with only Type 1 cells in the system (i.e. the initial state), the expression values of each gene are then collected at three different temporal stages for each stochastic simulation: the early, the middle, and the final stage, in order to mimic a typical set of temporal single-cell data (See Section II in the Supplementary file). Repeating the stochastic simulations using the same set of parameters and the same initial values of genes for 400 times produces a set of gene expression values, corresponding to 1200 sets of single-cell data.

3. RESULTS

3.1. SOMSC on the simulation data

We apply SOMSC to 346 cells, which are randomly selected from 1200 simulated cells. The SOM ($N = 16$) maps the high-dimensional gene-expression data to a 2-dimensional topographic chart, $U(x, y)$, and the watershed algorithm is employed to identify the basins of the topographic chart with the yellow boundaries between the adjacent basins (Figure 2B). There are seven basins identified. Since no cycles exist on the topographic chart, basins correspond to cellular states, which are labeled S_1, S_2,...,S_7 (Figure 2B). The red numbers on the yellow boundaries are the height of the boundaries between the adjacent cellular states. Based on the above information, we calculate the adjacency matrix of cellular states, $T = (t_{ij})_{7,7}$ and construct a cellular state network, in which the edge between the node S_i and the node S_j means that $t_{ij} \neq 0$ and the weight of the edge is t_{ij} (Figure 2C). Then we identify the edge pair for each cellular state (Table 1), highlighted in orange (Figure 2C). For example, the cellular state S_1 has five adjacent cellular states, S_2, S_3, S_5, S_6, and S_7 with edge weights, 2.6733, 2.1888, 1.2798, 1.1335, 5.0827 and S_6 is the one with the smallest weight, 1.1335. Therefore, (S_1, S_6) is the edge pair for the cellular state S_1.

Table 1: Edge pairs in the simulation data

Cellular state	Edge in the path	Cellular state	Edge in the path
S_1	(S_1, S_6)	S_2	(S_2, S_6)
S_3	(S_3, S_1)	S_4	(S_4, S_2)
S_5	(S_5, S_1)	S_6	(S_6, S_1)
S_7	(S_7, S_2)		

We track the paths from each cellular state to the starting cellular state S_6 using the highlighted edges. Since the edge pair for the cellular state S_1 is (S_1, S_6), the transition path P_1 is from the cellular state S_1 to the cellular state S_6, (S_1, S_6). Similarly, the transition path P_2 is from the cellular state S_2 to the cellular state S_6, (S_2, S_6). Because the edge pair for the cellular state S_3 is (S_3, S_1) and the path transition P_1 is (S_1, S_6), then the combination of (S_3, S_1) and P_1 results in the transition path P_3, (S_3, S_1, S_6). Similarly, We find the transition paths P_4, P_5, P_7 from these cellular

states S_4, S_5, S_7 to the starting cellular state S_6. Those transition paths are summarized as follows (Table 2).

Table 2: Path from one cellular state to the starting cellular state in the simulation data

State	Path	The path to the starting cellular state
S_1	P_1	(S_1, S_6)
S_2	P_2	(S_2, S_6)
S_3	P_3	(S_3, S_1, S_6)
S_4	P_4	(S_4, S_2, S_6)
S_5	P_5	(S_5, S_1, S_6)
S_6	P_6	(S_6, S_6)
S_7	P_7	(S_7, S_2, S_6)

The cellular state map is constructed to illustrate the relative locations of all cells and the relationship among different cellular states (Figure 2D) (See Methods). The SOM maps the winner of each computational grid to the topographical chart and the KNN algorithm places each non-winner cell at the mean position of its nearest neighbors. Cells are represented on the topographic chart as purple spheres (Figure 2D). Finally, we add the arrows from the parent cellular state to the daughter cellular state on the CSM (Figure 2E). The cellular state transition tree is obtained from all transition pairs, (S_1, S_3), (S_1, S_4), (S_3, S_5), (S_3, S_7), (S_4, S_2), and (S_4, S_6) (Figure 2F). To validate the calculated transition path we plot the cellular states of the cells in the topographic chart (Figure S4). All above results show that SOMSC can not only calculate the transition path but also identify the cellular states of over 90% cells.

3.2. SOMSC on qPCR data of mouse embryo development from zygote to blastocyst

Previously, the expression levels of 48 genes at seven time points were measured using qPCR for mouse early embryonic development from zygote to blastocyst (Guo *et al.*, 2010). Expression levels for each of the 439 single-cell libraries were normalized independently by the mean expression levels of two genes: Actb and Gapdh (Guo *et al.*, 2010).

Two different approaches might be applied to such data by either using the data at each temporal point individually or lumping the data of all seven

stages into one set. However, a series of single-state maps is unable to determine potential cellular state transition paths from the data because different basins or cellular states are obtained using different topographic charts (Figure S5). Therefore we choose to analyze all time points concurrently. Furthermore, prior knowledge about cell state is withheld during inference and subsequently used to validate the state-transition path calculated by SOMSC.

We generate a topographic chart from the expression-levels of all 439 cells and use the watershed algorithm to identify 20 basins. The basins are labeled by C_1, C_2, ..., C_{20} and the heights of the barriers between adjacent basins are the red numbers on the yellow barriers (Figure S6 and S7). The t-test is performed gene-by-gene between the cells of each basin-pair, C_i and C_j and then we calculate the number T_{ij} of the genes, which are not differentially expressed at 5% significance. Based on all calculated T_{ij}, we merge the following pairs: (C_{13}, C_{14}), (C_{17}, C_{18}), (C_{18}, C_{20}), (C_1, C_6), (C_2, C_3), (C_{15}, C_{16}), (C_4, C_7) (See Methods), yielding a new topographic chart with barriers highlighted in yellow (Figure 3B). Thirteen basins labeled by S_1, S_2, ..., S_{13} are identified in the new chart, corresponding to thirteen cellular states. The distance between adjacent states is labeled in red on the yellow boundaries between states (Figure 3B, S8 and S9).

Using the topographic chart, we construct a cellular state transition path tree in two steps. First, we construct the cellular state network (Figure 3C) and highlight the edges with the smallest weight values for each cellular state. The highlighted edges for the cellular states are summarized in Table 3.

Table 3: Edge pairs in the qPCR data of mouse embryo development from zygote to blastocyst

Cellular state	Edge in the path	Cellular state	Edge in the path
S_1	(S_1, S_9)	S_2	(S_2, S_9)
S_3	(S_3, S_8)	S_4	(S_4, S_{11})
S_5	(S_5, S_7)	S_6	(S_1, S_6)
S_7	(S_7, S_8)	S_8	(S_3, S_8)
S_9	(S_1, S_9)	S_{10}	(S_3, S_{10})
S_{11}	(S_{11}, S_{12})	S_{12}	(S_{11}, S_{12})
S_{13}	(S_6, S_{13})		

Second, we track the cellular state transition paths from each cellular

state to the starting cellular state S_5 using the highlighted edges. We find the paths from these cellular states S_3, S_4, S_7, S_8, S_{10}, S_{11}, S_{12} to the starting cellular state S_5 consisting of the edges from Table 3. They form one group R of reachable cellular states and the rest cellular states S_1, S_2, S_4, S_6, S_9, S_{11}, S_{12}, S_{13} form the group Q using the highlighted edges. Then the smallest weight value between R and Q is the one, 28.797, between the cellular state S_{11} and the cellular state S_{10}. Then $R \cup S_{11} \cup S_4 \cup S_{12}$ generates the new R and the new Q is the set consisting of S_1, S_2, S_6, S_9, and S_{13}. The edge with the smallest weight, 32.526, connecting R and Q is (S_2, S_{10}), which results all elements in Q are reachable. Accordingly, all these edges (S_1, S_9), (S_2, S_9), (S_3, S_8), (S_4, S_{11}), (S_5, S_7), (S_1, S_6), (S_7, S_8), (S_3, S_{10}), (S_{11}, S_{12}), (S_6, S_{13}), (S_2, S_{10}), and (S_{10}, S_{11}) result in that each cellular state can reach the starting cellular state S_5 by a transition path (Table 4).

Table 4: Path from one cellular state to the starting cellular state in the qPCR data of mouse embryo development from zygote to blastocyst

State	Path	The path to the starting cellular state
S_1	P_1	$(S_1, S_9, S_2, S_{10}, S_3, S_8, S_7, S_5)$
S_2	P_2	$(S_2, , S_{10}, S_3, S_8, S_7, S_5)$
S_3	P_3	(S_3, S_8, S_7, S_5)
S_4	P_4	$(S_4, S_{11}, S_{10}, S_3, S_8, S_7, S_5)$
S_5	P_5	(S_5, S_5)
S_6	P_6	$(S_6, S_1, S_9, S_2, , S_{10}, S_3, S_8, S_7, S_5)$
S_7	P_7	(S_7, S_5)
S_8	P_8	(S_8, S_7, S_5)
S_9	P_9	$(S_9, S_2, , S_{10}, S_3, S_8, S_7, S_5)$
S_{10}	P_{10}	$(S_{10}, S_3, S_8, S_7, S_5)$
S_{11}	P_{11}	(S_{11}, S_{12})
S_{12}	P_{12}	$(S_{12}, S_{11}, S_{10}, S_3, S_8, S_7, S_5)$
S_{13}	P_{13}	$(S_{13}, S_6, S_1, S_9, S_2, S_{10}, S_3, S_8, S_7, S_5)$

The combination of these paths results in the cellular state transition path tree, (S_9, S_{10}, S_8, S_{11}, 0, S_1, S_5, S_7, S_2, S_3, S_{10}, S_{11}, S_6). We know the state of the winner of each map grid and we use the KNN method to classify the remaining cells based on Euclidean distance to the nearest three winners. Then we smooth the data and produce the cellular state map for all cells (Figure 3E) (See details in Methods).

To study the dynamics of gene expression during the transitions, we calculate the pseudotime ordering of the cells, which is shown along the cellular state transition paths (Figure 4A) (See details in Methods). The colors of the cells represent the expression levels of CDX2. The variances of the expression levels of CDX2 in cellular states S_2, S_9, S_1, S_6, and S_{13} are smaller than the ones in other cellular states. The greater variances might be one of the reasons of resulting in the cellular state transition branches, $S_{10}\rightarrow S_2$, $S_{10}\rightarrow S_{11}$ and $S_{11}\rightarrow S_4$, $S_{11}\rightarrow S_{12}$ (Figure 4A). The cellular state transition trajectory of CDX2 shows that the expression levels of CDX2 in the trophectoderm (TE) is greater than in the inner cell mass (ICM) which includes the primitive endoderm (PE) and epiblast (EPI), consistent with the previous study (Jedrusik *et al.*, 2008). We define a gene as state-driven when its differential expression between the two adjacent states has a p-value of less than 5% as computed by t-test. Pathway enrichment is conducted using the Enrichr tool (See details in the Supplementary file). Briefly, Fisher's exact test is used to obtain a p-value for the number of pathway members present among the state-driven genes defined in our experiment. If the p-value falls within our critical region (5%), we determine the pathway to be significantly enriched. We present the pathway enrichment analysis for the cellular state transitions $S_{10}\rightarrow S_2$ (Figure 4C). They illustrate that PluriNetWork, ESC pluripotency pathway, regulation of actin cytoskeleton pathway and focal adhesion-PI3K-Akt-mTOR-signaling pathway involve significantly in the mouse early embryo development (Zhao and Guan, 2011; Reiske *et al.*, 1999). For the first transition branch $S_{10}\rightarrow S_2$, $S_{10}\rightarrow S_{11}$, we can see that the transition probability of $S_{10}\rightarrow S_{11}$ is 19% and the transition probability of $S_{10}\rightarrow S_{11}$ is 18%. Additionally the replication probability of S_2 is 67% and the replication probability of S_{11} is 54%. This provides a clear mechanism for the difference in population size between S_2 and S_{11}. Interestingly, we find the transition probability of $S_{11}\rightarrow S_4$, 28%, is greater than the one of $S_{11}\rightarrow S_{12}$, 18%. And the replication probability of S_4 is also greater than the one of S_{12}. However, the number of the cells in the cellular state S_4 is smaller than the one in the cellular state S_{12}. Perhaps the replication rate of cells in S_{12} is higher, despite a smaller replication probability for each individual cell, due to a much shorter cell cycle (Kelly *et al.*, 1978; Artus and Cohen-Tannoudji, 2008).

We validate the results calculated by SOMSC against our prior knowledge of the cells temporal order and differentiation trajectory. The temporal ordering of the cellular states identified by SOMSC is consistent with the stage

information of the cells in the data (Figure S10). The cells in the cellular state S_6, S_7, S_8, S_3, and S_{10} are collected at the 1-cell stage, the 2-cell stage, the 4-cell stage, the 8-cell stage, the 16-cell stage respectively. The cells harvested at the 32-cell stage are located in the areas of the cellular states S_2 and S_{11} in the cellular state map and the ones harvested at the 64-cell stage are at the regions of the cellular states S_4, S_{12}, S_9, S_1, S_6, and S_{13} in the map. Finally, SOMSC can track the two state transition branches during the mouse early embryo development (Zernicka-Goetz *et al.*, 2009; Pedersen *et al.*, 1986). The results calculated by SOMSC show that those two fate decisions occur at the 32-cell stage and the 64-cell stage.

3.3. SOMSC on scRNA-seq data of mouse haematopoietic stem cell differentiation

To analyze discrete genomic states and the transitional intermediates that span myelopoiesis, the previous study performed single-cell RNA sequencing (scRNA-seq) on 382 cells consisting of stem/multipotent progenitor cells, common myeloid progenitor (CMP) cells, granulocyte monocyte progenitor (GMP) cells, and LKCD34+ cells that includes granulocytic precursors (Olsson *et al.*, 2016). The quality control was performed in the original dataset and therefore the data of those 382 cells are used as the input to SOMSC (Olsson *et al.*, 2016). Out of 23955 genes measured in the original data, 1240 highly variable genes were selected.

We use the data from 382 cells to produces the topographic chart by SOM (Figure S11). We identify 14 basins by the watershed algorithm, labeled C_1, C_2, ..., C_{14} (Figure S11 and S12). The heights of the barriers between adjacent basins are the red numbers on the yellow barriers (Figure S12). The t-test is performed gene-by-gene between the cells of each basin-pair, C_i and C_j and then we calculate the number T_{ij} of the genes, which are not differentially expressed at 5% significance. Based on all calculated T_{ij}, we merge the following pairs to generate a new topographic chart: (C_1, C_5), (C_6, C_7), (C_7, C_9), (C_{10}, C_{11}), (C_1, C_{13}) (Figure S12). The basins in the topographic chart are labeled S_1, S_2, ..., and S_9 and each basin represents a cellular state. The distance between the adjacent cellular states in the chart is shown in red (Figure S13 and S14) and used to construct the cellular state network (Figure 5A).

The combination of these paths P_1, P_2, ..., and P_9 results in the cellular state transition path tree, (S_3, S_7, S_9, 0, S_7, S_7, S_1, S_1, S_4). The cellular state map including all 382 cells was generated as previously described

(Figure 5B). SOMSC identifies nine cellular states, which is consistent with previously reported classification of myelopoetic cells based on scRNA-seq (Olsson *et al.*, 2016). Importantly, SOMSC faithfully recovers the differentiation trajectories of the cells: cellular states S_4, S_9, S_3 are the LSK cells and S_1 consists of CMP and GMP cells. There is a branch from the cellular state S_1 to the cellular state S_7 and to the cellular state S_8. All cells in the cellular state S_8 are CMP cells and the ones in the cellular state S_7 are GMP cells. We find that Irf8 is expressed in the cellular state S_7 (Figure 5C). The cells in the cellular state S_2 are CMP, the ones in the cellular state S_5 are LSKs and the ones in the cellular state S_6 are GMP cells. From the violin plots in Figure 5C we can see that there is more than one mode in the cellular states S_4, S_9, S_3, S_1, and S_7, suggesting that the cell fate decision is made relatively early.

In order to determine the genes driving the cellular state transitions we identify state-driven genes by t-test as previously described. Based on the list of differentially expressed genes, the enriched pathways are obtained by gene set enrichment as previously described. We present the pathway enrichment analysis for the cellular state transitions $S_7 \rightarrow S_6$ (Figure 5E), indicating that TNFα-NFκB pathway is critical during the differentiation of GMP cells (See more results in the Supplementary file).

Next we estimate the replication probabilities and transition probabilities of the cellular states (Figure 5F). There are two branches in this cellular state transition path tree. The results show that the transition probability from S_1 to S_7 is 34% and the transition probability from S_1 to S_8 is 12%. The latter probability is smaller, corresponding to fewer cells classified in S_8 than in S_7. The results also show that the probability of the transition from S_7 to S_6, S_7 to S_2, and S_7 to S_5 are 22%, 16%, and 4% respectively. Correspondingly, S_6 represents the largest population among all three fates.

The cellular state transition path tree calculated by the SOMSC is consistent with the hematopoietic hierarchy. S_4 is the long-term HSC. S_9 is the short-term HSC. S_3 is the multipotent progenitor. S_1 is the common myeloid progenitor. S_7 is the granulocyte-macrophage progenitor. S_8 is the megakaryocyte-erythrocyte progenitor (Figure S15). Since Gfi1 is expressed only in the cells of S_6 among S_6, S_5, and S_2 (Figure S16), then S_6 is the granulocyte cell. Because Irf8 is highly expressed in the cells of S_2 (Figure 5C), then the cells in S_2 are the monocyte cells. We predict that the cells in S_5 are the dendritic cells. All of them need to be verified.

3.4. SOMSC on scRNA-seq data of adult mouse olfactory stem cell lineage trajectories

616 cells were collected to define a detailed map of the postnatal olfactory epithelium showing the cell fate potentials and branchpoints in olfactory stem cell lineage trajectories by whole transcriptome profiling scRNA-seq (Fletcher *et al.*, 2017). We select 824 highly variable genes out of 42127.

As before, we use an SOM to produce a topographic chart using the expression profiles of all 616 cells (Figure S17). We identify 22 basins by the watershed algorithm, labeled C_1, C_2, ..., C_{14} (Figure S18). Then 17 cellular states are determined by SOMSC (see Methods) and their corresponding cellular state network is constructed (Figure S19 and S20). Next, the CSM and the cellular state transition path are obtained (Figure 6). To validate our results, we utilize the prior categorization of these cells, which are available from (GEO accession GSE95601) (Fletcher *et al.*, 2017). They include the resting HBCs, immediate Neuronal Precursor 1 (INP1), Globose Basal Cells (GBCs), mature Sustentacular Cells (mSUS), transitional HBC 2, immature Sustentacular Cells (iSUS), transitional HBC 1, immature Olfacotry Sensory Neurons (iOSNs), Immediate Neuronal Precursor 3 (INP3), Microvillous Cells, type 1, mature Olfactory Sensory Neurons (mOSNs), Immediate Neuronal Precursor 2 (INP2), and Microvillous Cells (MV) (Fletcher *et al.*, 2017). The cellular states are denoted by the cluster labels used in the original data. The non-differentiating cell types are determined by the known marker genes (mature olfactory sensory neurons, mature sustentacular cells, and microvillus cells) (Fletcher *et al.*, 2017). All of these cellular types correspond to the cellular states, which are the endpoints in the cellular state transition paths, consistent with our prior knowledge (Figure S21). Notably, SOMSC identifies the sustentacular cluster as the endpoint in the cellular state transition path tree without using the prior knowledge (Figure 6CF) not possible using previous methods (Fletcher *et al.*, 2017). The SOMSC can not only identify those three primary cellular state transition paths but also detect more cellular state transition paths than what previous methods obtained (Figure 6CF). Interestingly, other lineage detection methods fail to recognize the three primary lineages in this dataset (Street *et al.*, 2017). We suggest here that GBC is categorized as an intermediate state, and a precursor cellular state of the immediate neuronal precursors. And that immature sustentacular cells may be the progenitor cells for mature olfactory sensory neurons. In addition, transitional horizontal basal cells are found in different branches. Finally, the microvillus cells are recognized as a separate branch

different from the neuronal lineage, which is consistent with the previous results (Fletcher *et al.*, 2017).

An interesting observation from the cellular trajectory of gene Trp63 shows that the expression levels of cells in S_{10} (HBC1) change from high to low and they become rather low before the cells transit into other cellular states, which was also observed in the previous study (Schnittke *et al.*, 2015) (Figure 6D). The pathway enrichment analysis indicates that p53 signaling pathway plays an important role during the cellular transition from HBCs to GBCs, which is consistent with the previous observations (Herrick *et al.*, 2017) (Figure 6E). It also shows that the pseudotime ordering of cells generated by SOMSC is consistent with previous study (Schnittke *et al.*, 2015). The estimated values of replication probability and the transition probabilities suggest that the length of cell cycle may play a critical roles at some cellular transitions. For example, the transition probability from S_{10} to S_4 (6%) is very close with the one from S_{10} to S_{13} (8%). The value of replication probability of S_{13} is smaller than the one of S_4 (Figure 6F). However, the number of cells in the cellular state S_{13} is much greater than the one in the cellular state S_4 indicating that the shorter cell cycle length of cells in the cellular states may be one of the key factors during the development of olfactory (Huard and Schwob, 1995).

4. Conclusion and Discussion

In this work we have presented a SOM based method for analyzing gene expression data of single cells that may contain multiple cellular states. Applications of SOMSC to a set of simulated data and three sets of experimental data have demonstrated the capabilities and effectiveness of SOMSC in identifying cellular states and their cellular state transition path trees.

The CSM based on the cellular states identified by SOM provides a global landscape view of the cellular states and their transitions. The location of each cell in the CSM may provide useful information on the cells viability, potential of transitions to different cellular states and the replication/transition probabilities using the unbiased selection of genes. These properties make our algorithm unique compared with many other methods for single-cell analysis.

The major computational cost of SOMSC comes from iteratively computing the U-matrix in the SOM, which has a complexity of $\mathcal{O}(N \cdot N_g \cdot D \cdot T)$ where D is the number of genes measured in the data, T is the number of iterations used in SOM, and N is the number of cells in the data set(Lee

and Verleysen, 2007). In practice, D is usually less than 10,000 (the number of genes significantly expressed), and both T and N are less than 1,000, implying an average complexity of $\mathcal{O}(10^{10})$.

The one parameter of SOMSC is the map size. A map with too many grids may produce too many small clusters while too few grids may lead to a map containing too few clusters (Figure S14A, S14B, S14C). In our analysis, we set the number of grids approximately equal to the number of cells in each dataset.

Single-cell data is often used to identify cellular states in heterogeneous cell populations (Kolodziejczyk *et al.*, 2015). SOMSC can capture the shape of data which lacks the convex or normal structure required by many other methods (Liebscher *et al.*, 2012). Another major feature of SOM is the identification of multiple minima since SOM searches the entire space of feasible solutions during early exploration, and divides the search space gradually until it finds an optimal solution (Liebscher *et al.*, 2012; Openshaw *et al.*, 1995). When SOM and k-means use the same initial guess near the optimal solution to approach the optimal, similar performance and similar results are obtained(Liebscher *et al.*, 2012; Openshaw *et al.*, 1995). This is consistent with the observation that SOMSC is rather stable in finding the basins of attraction and the transition paths from the CSM of the single-cell data.

Previous work has shown that confounding factors (e.g. batch errors, cell cycle effects) impede analysis of single-cell data (Buettner and Theis, 2012; Buettner *et al.*, 2015). PCA (Pickrell *et al.*, 2010), surrogate variable analyses (Leek and Storey, 2007), probabilistic estimation of expression residuals (Stegle *et al.*, 2010, 2012) and factor analysis (Risso *et al.*, 2014) have been explored to reduce the effects of confounders in gene expression studies on bulk cell populations (Stegle *et al.*, 2015). Most of these methods could be extended to the data of single cells, however, removal of cell cycle effects, an important source of variation in single-cell measurement, is still challenging. However, a linear mixed model has recently been utilized to remove cell-cycle dependent gene expression as a source of variation (Buettner *et al.*, 2015).

The CSM in many ways is similar to the landscape description although the typical landscape is a function of each gene. It would be interesting to make a comparison between a landscape computed by forward modeling based on a small size of network and the CSM generated by data. In general, SOMSC is a robust method, which is also convenient for visualization, to identify the cell states based on single-cell data. It facilitates the identification of pathway components, such as signature transcription factors, and

pseudo temporal ordering of cells involving complex differentiation trajectories.

Funding

This work is partially supported by National Institute of Health grants P50GM76516, R01GM107264, and R01ED023050, and National Science Foundation grants DMS1161621 and DMS1562176.

Acknowledgement

We are grateful to Matt Karikomi for manuscript proofreading.

Reference

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46**(3), 175–185.

Amir, E.-a. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, **31**(6), 545–552.

Artus, J. and Cohen-Tannoudji, M. (2008). Cell cycle regulation during early mouse embryogenesis. *Molecular and cellular endocrinology*, **282**(1), 78–86.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Peer, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, **157**(3), 714–725.

Buettner, F. and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, **28**(18), i626–i632.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, **33**(2), 155–160.

Chen, C., Zhang, K., Feng, H., Sasai, M., and Wang, J. (2015). Multiple coupled landscapes and non-adiabatic dynamics with applications to self-activating genes. *Physical Chemistry Chemical Physics*, **17**(43), 29036–29044.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Maayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, **14**(1), 128.

Chen, K.-C., Wang, T.-Y., Tseng, H.-H., Huang, C.-Y. F., and Kao, C.-Y. (2005). A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics*, **21**(12), 2883–2890.

de Vargas Roditi, L. and Claassen, M. (2015). Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. *Current opinion in biotechnology*, **34**, 9–15.

Fletcher, R. B., Das, D., Gadye, L., Street, K. N., Baudhuin, A., Wagner, A., Cole, M. B., Flores, Q., Choi, Y. G., Yosef, N., *et al.* (2017). Deconstructing olfactory stem cell trajectories at single-cell resolution. *Cell Stem Cell*, **20**(6), 817–830.

Foster, D. V., Foster, J. G., Huang, S., and Kauffman, S. A. (2009). A model of sequential branching in hierarchical cell fate determination. *Journal of theoretical biology*, **260**(4), 589–597.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, **128**(4), 635–638.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell*, **18**(4), 675–685.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, page btv325.

Heppner, G. H. (1984). Tumor heterogeneity. *Cancer research*, **44**(6), 2259–2265.

Herrick, D. B., Lin, B., Peterson, J., Schnittke, N., and Schwob, J. E. (2017). Notch1 maintains dormancy of olfactory horizontal basal cells, a reserve neural stem cell. *Proceedings of the National Academy of Sciences*, **114**(28), E5589–E5598.

Huard, J. M. and Schwob, J. E. (1995). Cell cycle of globose basal cells in rat olfactory epithelium. *Developmental dynamics*, **203**(1), 17–26.

Jedrusik, A., Parfitt, D.-E., Guo, G., Skamagki, M., Grabarek, J. B., Johnson, M. H., Robson, P., and Zernicka-Goetz, M. (2008). Role of cdx2 and cell polarity in cell allocation and specification of trophectoderm and inner cell mass in the mouse embryo. *Genes & development*, **22**(19), 2692–2706.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *Knowledge and Data Engineering, IEEE Transactions on*, **16**(11), 1370–1386.

Kelly, S., Mulnard, J., and Graham, C. (1978). Cell division and cell allocation in early mouse development. *Development*, **48**(1), 37–51.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, **21**(1), 1–6.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, **58**(4), 610–620.

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C.-Y., *et al.* (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**(9), e161.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., *et al.* (2015). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**(1), 184–197.

Liebscher, S., Kirschstein, T., and Becker, C. (2012). The flood algorithma multivariate, self-organizing-map-based, robust location and covariance estimator. *Statistics and Computing*, **22**(1), 325–336.

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, **111**(52), E5643–E5650.

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., *et al.* (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature cell biology*, **15**(4), 363–372.

Mukherjee, S., Zhang, Y., Kannan, S., and Seelig, G. (2017). Prior knowledge and sampling model informed learning with single cell rna-seq data. *bioRxiv*, page 142398.

Najman, L. and Schmitt, M. (1994). Watershed of a continuous function. *Signal Processing*, **38**(1), 99–112.

Najman, L. and Schmitt, M. (1996). Geodesic saliency of watershed contours and hierarchical segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, **18**(12), 1163–1173.

Ocone, A., Haghverdi, L., Mueller, N. S., and Theis, F. J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, **31**(12), i89–i96.

Olsson, A., Venkatasubramanian, M., Chaudhri, V. K., Aronow, B. J., Salomonis, N., Singh, H., and Grimes, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, **537**(7622), 698–702.

Openshaw, S., Blake, M., Wymer, C., *et al.* (1995). Using neurocomputing methods to classify britains residential areas. *Innovations in GIS*, **2**, 97–111.

Pedersen, R. A., Wu, K., and BaŁakier, H. (1986). Origin of the inner cell mass in mouse embryos: cell lineage analysis by microinjection. *Developmental biology*, **117**(2), 581–595.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, **464**(7289), 768–772.

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs Jr, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011).

Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, **29**(10), 886–891.

Rashid, S., Kotton, D. N., and Bar-Joseph, Z. (2017). Tasic: determining branching models from time series single cell data. *Bioinformatics*, page btx173.

Reiske, H. R., Kao, S.-C., Cary, L. A., Guan, J.-L., Lai, J.-F., and Chen, H.-C. (1999). Requirement of phosphatidylinositol 3-kinase in focal adhesion kinase-promoted cell migration. *Journal of Biological Chemistry*, **274**(18), 12361–12366.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, **32**(9), 896–902.

Saadatpour, A., Lai, S., Guo, G., and Yuan, G.-C. (2015). Single-cell analysis in cancer genomics. *Trends in Genetics*, **31**(10), 576–586.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, **42**(14), 8845–8860.

Schnittke, N., Herrick, D. B., Lin, B., Peterson, J., Coleman, J. H., Packard, A. I., Jang, W., and Schwob, J. E. (2015). Transcription factor p63 controls the reserve status but not the stemness of horizontal basal cells in the olfactory epithelium. *Proceedings of the National Academy of Sciences*, **112**(36), E5068–E5077.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, **6**(5), e1000770.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, **7**(3), 500–507.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, **16**(3), 133–145.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2017). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *bioRxiv*, page 128843.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014a). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, **32**(4), 381–386.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014b). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, **32**(4), 381.

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, **509**(7500), 371–375.

Tsioris, K., Torres, A. J., Douce, T. B., and Love, J. C. (2014). A new toolbox for assessing single cells. *Annual review of chemical and biomolecular engineering*, **5**, 455.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, **9**(2579-2605), 85.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., *et al.* (1999). Self-organizing map in matlab: the som toolbox. In *Proceedings of the Matlab DSP conference*, volume 99, pages 16–17.

Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 583–598.

Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, **108**(20), 8257–8262.

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016). Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data. *Genome biology*, **17**(1), 106.

Wilson, J. L., Suri, S., Singh, A., Rivet, C. A., Lu, H., and McDevitt, T. C. (2014). Single-cell analysis of embryoid body heterogeneity using microfluidic trapping array. *Biomedical microdevices*, **16**(1), 79–90.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., *et al.* (2013). Genetic programs in human and mouse early embryos revealed by single-cell rna [thinsp] sequencing. *Nature*, **500**(7464), 593–597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., *et al.* (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, **20**(9), 1131–1139.

Zernicka-Goetz, M., Morris, S. A., and Bruce, A. W. (2009). Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. *Nature Reviews Genetics*, **10**(7), 467–477.

Zhang, K., Sasai, M., and Wang, J. (2013). Eddy current and coupled landscapes for nonadiabatic and nonequilibrium complex system dynamics. *Proceedings of the National Academy of Sciences*, **110**(37), 14930–14935.

Zhao, X. and Guan, J.-L. (2011). Focal adhesion kinase and its signaling pathways in cell migration and angiogenesis. *Advanced drug delivery reviews*, **63**(8), 610–615.

Zhou, J. X. and Huang, S. (2011). Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends in Genetics*, **27**(2), 55–62.

Caption

Figure 1. A schematic diagram on steps of constructing cellular state maps and transition paths using the SOMSC method. (A) The gene expression data of single cells. (B) A topographic chart constructed by SOMSC using the data. The transitions among different basins are labeled by arrows: P1, P2, $\cdots$, and P5. (C) The cellular state lineage trees or differentiation processes are then summarized based on the transition paths. (D) The workflow of SOMSC.

Figure 2. SOMSC on the simulated model. (A) A three-stage lineage system. Stage 1 contains one type of cells in which the activated genes, A and B are highlighted by green; Stage 2 contains Type 2 cells and Type 3 cells in which the activated genes, A, C, and D are highlighted by orange in Type 2 cells while the activated genes, B, E, and F are highlighted by orange in Type 3 cells; Stage 4 contains four types of cells: Type 4 cells, Type 5 cells, Type 6 cells, and Type 7 cells. The activated genes, A and C, A and D, B and E, or B and F are highlighted in light green in Type 4, Type 5, Type 6, and Type 7 cells, respectively. (B) The topographic chart is constructed based on SOMSC algorithm with the map of 18×18 grids in the topographic chart computed for $N = 346$ single cells. The basins are labeled by S_1, S_2, S_3, S_4, S_5, S_6, S_7. Each basin represents a cluster of cells in one cellular state. The yellow areas are the boundaries between adjacent basins and the red numbers are the heights of the barriers of adjacent basins. (C) The cellular state network is built based on the topographic chart. The nodes are the cellular states. The weight of the edge is the height of the barrier of adjacent basins. The orange line is the edge with the smallest weight associated with each cellular state. (D) The cellular state map. The red dots are the cells. The basins correspond to the cellular states in Figure B. (E) The zoomed-in cellular state map. The white text is the label of the cellular state. The arrow is the direction of the cellular state transition. (F) The cellular state transition map. The percentage numbers present the probability of the cellular state transition replication and cellular state transitions.

Figure 3. SOMSC reconstructed the cellular state transition path using the qPCR data of mouse stem cells from zygote to blastocyst (Guo *et al.*, 2010). (A) The topographic chart is constructed based on the SOMSC algorithm with the map of 20×20 grids. (B) The topographic chart with barriers of

the adjacent cellular states. The yellow areas are the barriers of the adjacent cellular states. Each basin means one cellular state: S_1, S_2, ..., S_{13}. The red numbers are the heights of the barriers. (C) The cellular state network is built based on the topographic chart. The nodes are the cellular states. The weight of the edge is the height of the barrier of adjacent basins. The orange line is the edge with the smallest weight associated with each cellular state. (D) The cellular state map. The red dots are the cells. The basins correspond to the cellular states in Figure B. (E) The zoomed-in cellular state map. The white text is the label of the cellular state. The arrow is the direction of the cellular state transition. (F) The cellular state transition map. The percentage numbers present the probability of the cellular state transition replication and cellular state transitions.

Figure 4. The dynamics of gene expression levels during the cellular state transitions and pathway enrichment analysis of the cellular state transitions. (A). The pseudotime ordering of cells. The colors represent the expression levels of CDX2. The violin plots are the distributions of the expression levels of CDX2 in each cellular state. (B) The cellular state trajectory of CDX2. (C) The bubble charts of pathway enrichment for different cellular state transitions, $S_{10} \rightarrow S_2$. The x-axis is the pathway index and y-axis is $-log_{10}$(P-value). Each circle is one pathway. The pathways related with the data are highlighted. (D) The cellular state transition map. The percentage numbers present the probability of the cellular state transition replication and cellular state transitions.

Figure 5. SOMSC reconstructed the cellular state transition path using the scRNA-seq data of mouse haematopoietic stem cell differentiation (Moignard *et al.*, 2013). (A) The cellular state network is built based on the topographic chart. The nodes are the cellular states. The weight of the edge is the height of the barrier of adjacent basins. The orange line is the edge with the smallest weight associated with each cellular state. (B) The cellular state map. The red balls are the cells. The basins correspond to the cellular states. (C) The pseudotime ordering of cells. The colors represent the expression levels of Irf8. The violin plots are the distributions of the expression levels of Irf8 in each cellular state. (D) the cellular state trajectory of Irf8. (E) the bubble charts of pathway enrichment for different cellular state transitions, $S_7 \rightarrow S_6$. The x-axis is the pathway index and y-axis is $-log_{10}$(P-value). Each circle is one pathway. The pathways related with the data are highlighted. (F) The

cellular state transition path tree with the cellular state replication probability and the cellular state transition probabilities.

Figure 6. SOMSC reconstructed the cellular state transition path using the scRNA-seq data of adult mouse olfactory stem cell lineage trajectories (Fletcher *et al.*, 2017). (A) The cellular state network is built based on the topographic chart. The weight of the edge is the height of the barrier of adjacent basins. The orange line is the edge with the smallest weight associated with each cellular state. (B) The cellular state map. The red balls are the cells. The basins correspond to the cellular states. (C) The pseudotime ordering of cells. The colors are quantified by the expression levels of Trp63. The violin plots are the distributions of the expression levels of Trp63 in each cellular state. (D) the cellular state trajectory of Trp63. (E) the bubble charts of pathway enrichment for different cellular state transitions, $S_{10} \rightarrow S_{13}$. The x-axis is the pathway index and y-axis is $-log_{10}$(P-value). Each circle is one pathway. The pathways related with the data are highlighted. (F) The cellular state transition path tree with the cellular state replication probability and the cellular state transition probabilities.

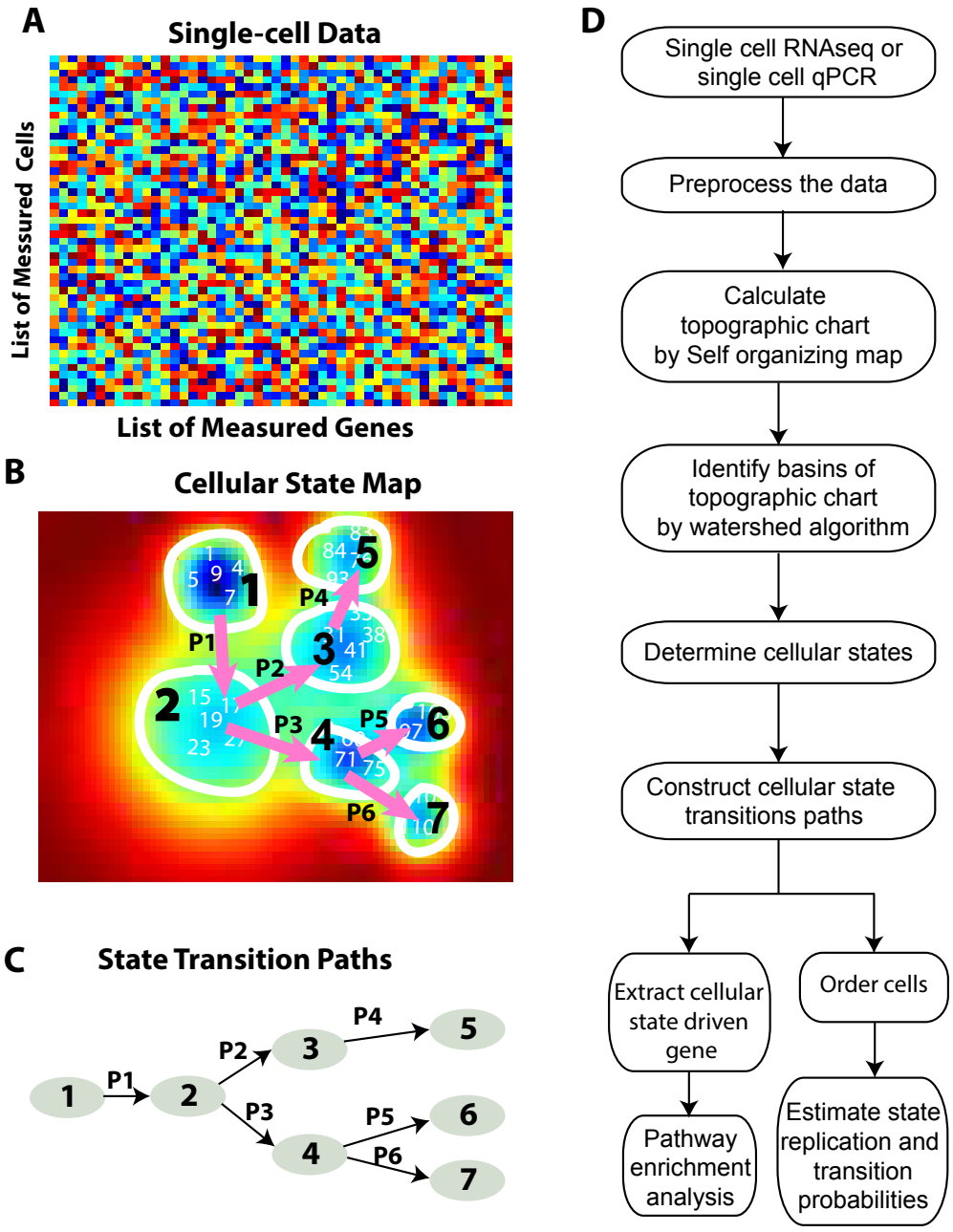

Figure 1:

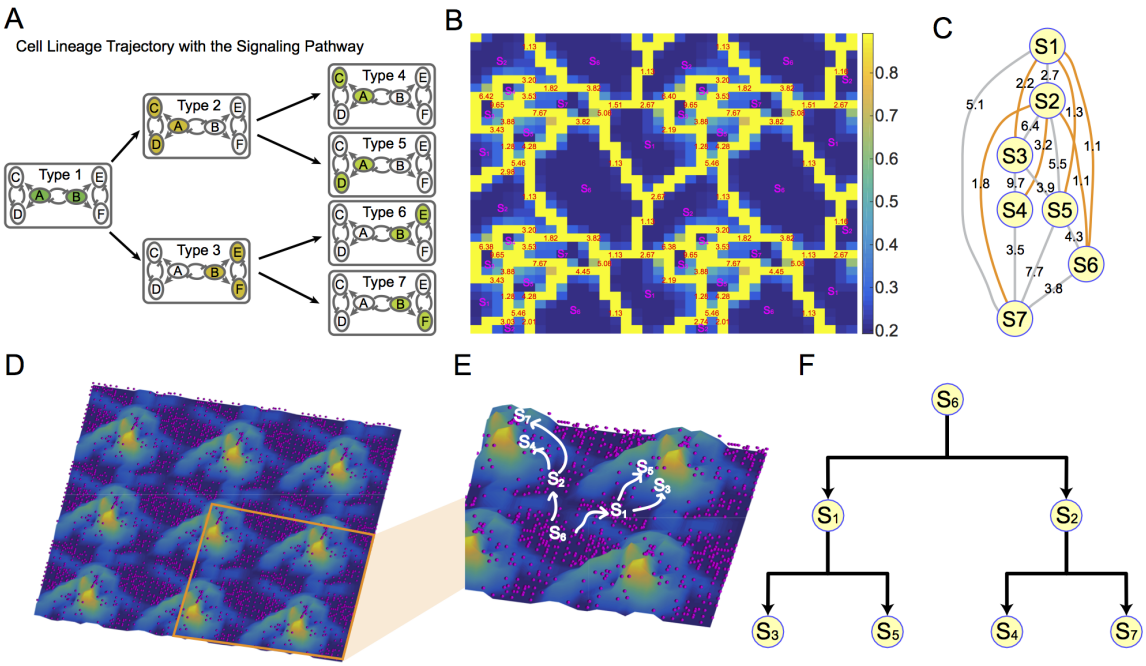

Figure 2:

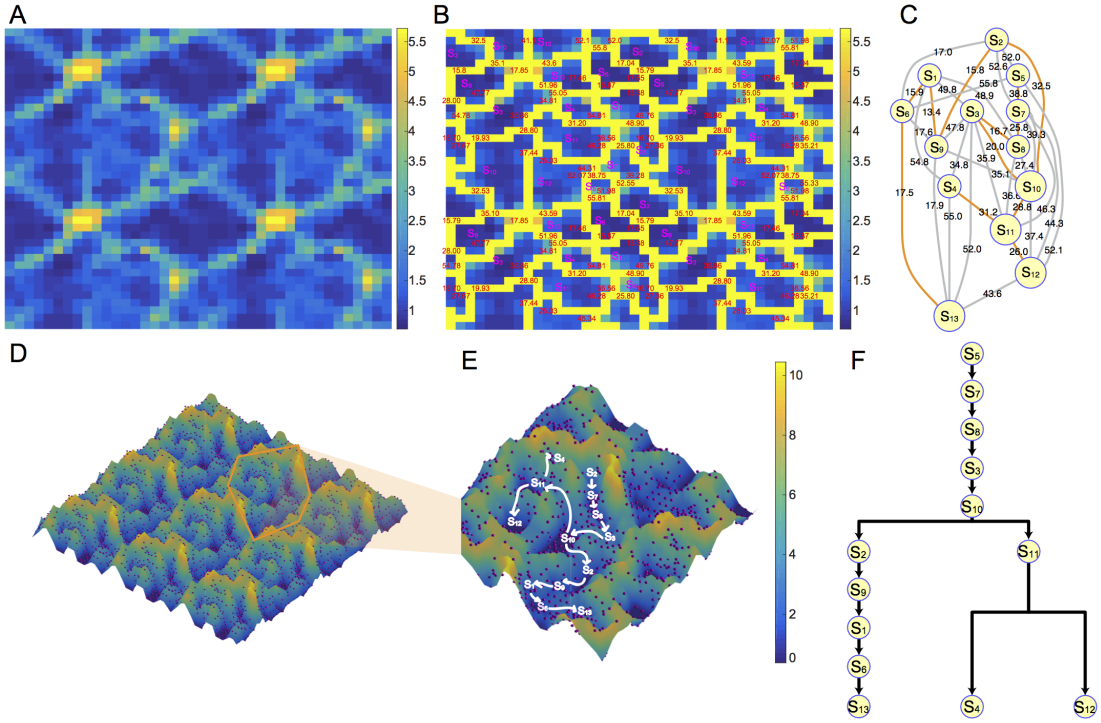

Figure 3:

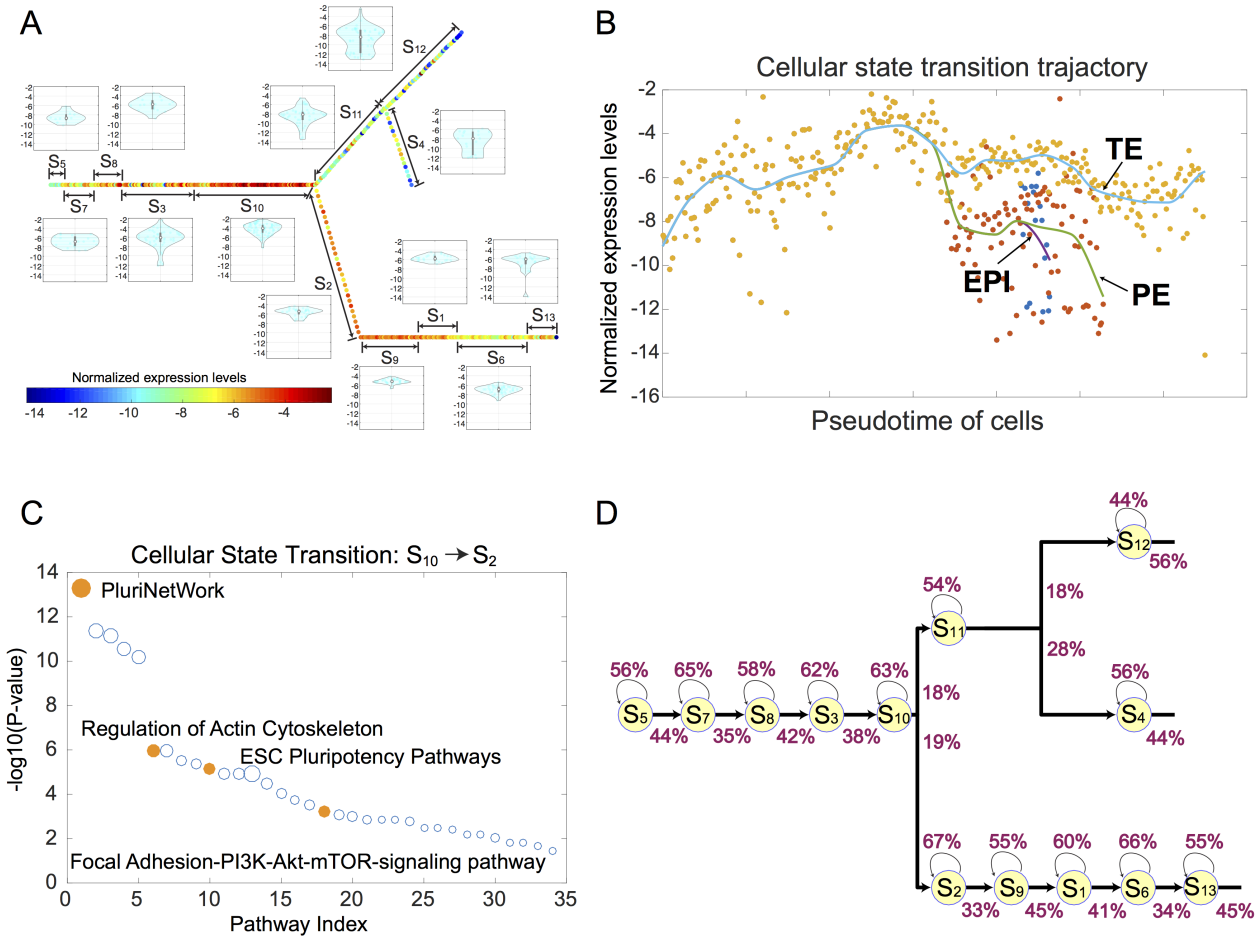

Figure 4:

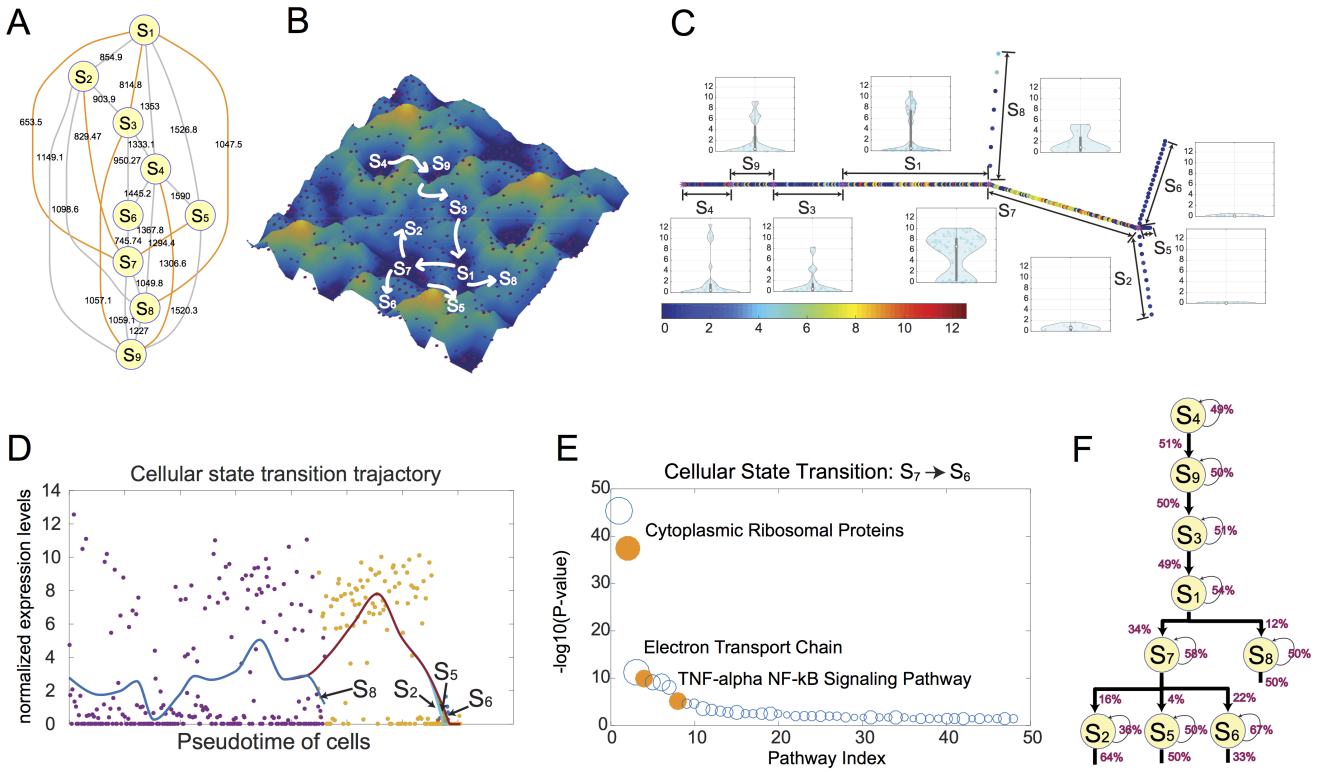

Figure 5:

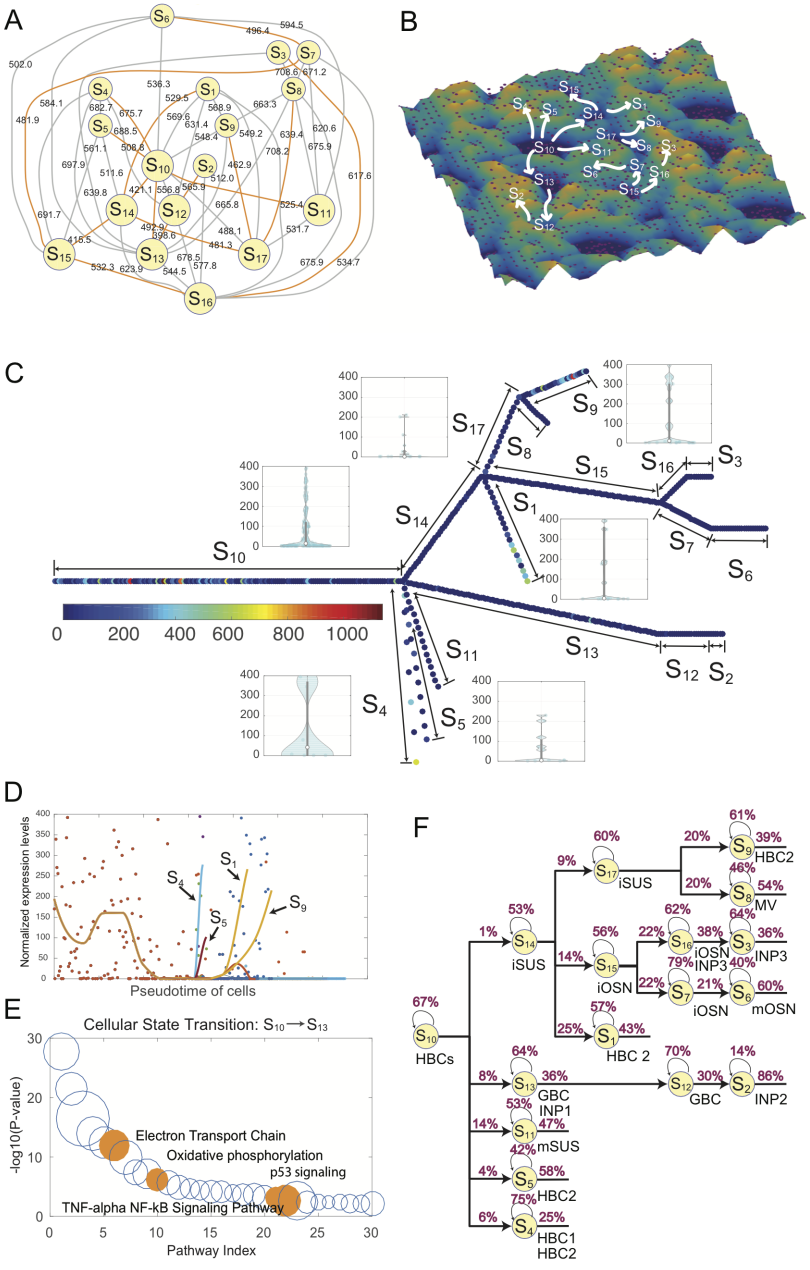

Figure 6: