Transiently increased intercommunity regulation characterizes concerted cell phenotypic transition

Weikang Wang[1*], Dante Poe[1,2], Ke Ni[1,2], Jianhua Xing[1,3,4,*]

[1] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15232, USA.

[2]Joint CMU-Pitt Ph.D. Program in Computational Biology, University of Pittsburgh, Pittsburgh, PA, USA.

[3] Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15232, USA.

[4] UPMC-Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA.

• To whom correspondence should be addressed. Email: xing1@pitt.edu, weikang@pitt.edu

ABSTRACT

Phenotype transition takes place in many biological processes such as differentiation and reprogramming. A fundamental question is how cells coordinate switching of expressions of clusters of genes. Through analyzing single cell RNA sequencing data in the framework of transition path theory, we studied how such a genome-wide expression program switching proceeds in three different cell transition processes. For each process we reconstructed a reaction coordinate describing the transition progression, and inferred the gene regulation network (GRN) along the reaction coordinate. In all three processes we observed common pattern that the effective number and strength of regulation between different communities increase first and then decrease. The change accompanies with similar change of the GRN frustration, defined as overall confliction between the regulation received by genes and their expression states, and GRN heterogeneity. While studies suggest that biological networks are modularized to contain perturbation effects locally, our analyses reveal a general principle that during a cell phenotypic transition intercommunity interactions increase to concertedly coordinate global gene expression reprogramming, and canalize to specific cell phenotype as Waddington visioned.

INTRODUCTION

A lasting topic in science and engineering is how a dynamical system transits from one stable

attractor to a new one in a corresponding state space (Hanggi *et al*, 1990). For example, a substitution reaction in organic chemistry can proceed either through first breaking an existing chemical bond to form an intermediate planar structure followed by forming a new bond, termed as a SN1 mechanism, or through forming a trigonal bipyramidal intermediate complex where breaking of the old bond and formation of the new bond take place concertedly, termed as a SN2 mechanism (Fig. 1A top) (Morrison & Boyd, 2010). Which mechanism dominates a process is determined by both the relative thermodynamic stability of the two intermediate structures, and the kinetics of forming them.

Another example of transitions that attracts increased interest recently is transitions between different cell phenotypes, partly due to available genome-wide characterization of the cell gene expression state throughout a transition process aided with advances of single cell genomics techniques. A cell is a nonlinear dynamical system governed by a complex regulatory network. The latter is formed by a large number of interacting genes, and can have multiple stable attractors corresponding to different cell phenotypes. Typically a large number of phenotype-specific genes maintain a specific phenotype through mutual activation while suppressing expression of genes corresponding to other exclusive phenotypes. In some sense it resembles a spin system segregating into upward and downward domains. When a cell phenotypic transition (CPT) takes place, the genes need to switch their expression status, analogous to flipping some upward and downward spin domains.

A question arises as how a CPT proceeds. The transition may be sequential with gene silence first to form an intermediate with the initial cell phenotype destabilized without commitment to a new phenotype, followed by activation of other genes to instruct the cell into one specific final stable phenotype (similar to the SN1 mechanism). Alternatively gene activation and silence may happen concurrently as in the SN2 mechanism, with hybrid intermediate states co-expressing genes corresponding to the two phenotypes. One can vision two qualitatively different characteristics of the two mechanisms.

Testing the two mechanisms requires examining how a genome-wide gene regulatory network (GRN) changes during a CPT, for which we exploited a recently developed RNA velocity formalism (La Manno *et al*, 2018). While scRNA-seq data only provide snapshots of cell transcriptomic states, RNA velocity analysis makes it possible to extract some dynamical information RNA velocity is a high-dimensional vector that can be inferred from the quantity of spliced and un-spliced RNA, and predicts the future state of individual cells on a timescale of hours. Using the data we further developed a dynamical model, and analyzed the transition dynamics in the framework of reaction rate theories and network science theories. A concerted mechanism is supported by characterizations of three CPT processes with a number of statistical quantities, notably a conserved pattern of peaked intercommunity interactions at an intermediate stage of each transition.

RESULTS

Dynamical model reconstructed from scRNA-seq data of epithelial-mesenchymal transition

We first analyzed a scRNA-seq dataset of epithelial-mesenchymal transition (EMT) of human A549 treated with TGF-β (Cook & Vanderhyden, 2020), a total of $N = 3003$ single cell samples measured at several time points (Fig. 1B). During EMT cells change from epithelial to mesenchymal phenotypes with increased EMT hallmark gene set score. We selected $M = 583$ genes whose variations correlate with the transition direction during EMT to form an M-dimensional state space (see Materials and Methods for details). Following Qiu et al. (Qiu *et al*, 2021), from the RNA velocities we constructed a Markov transition model with an $N \times N$ transition matrix T specifying the transition probabilities among the N measured cells using a Fokker-Planck kernel (Fig. 1B).

For rate theory analyses we estimated the kernel density of the day 0 samples using Scikit-learn (Pedregosa *et al*, 2011), and defined the cells whose local densities ρ_{0d} are in top 100 as within the initial state A. Similarly we estimated the kernel density of the day 3 samples and selected those cells whose ρ_{3d} are in top 100 as within the final state B. We randomly selected pairs of cells in state A and the state B, and obtained an ensemble of Dijkstra shortest paths between them based on the transition matrix (Fig. 1C). We applied a modified finite temperature string method (E *et al*, 2005; Vanden-Eijnden & Venturoli, 2009; Wang & Xing, 2020) to the ensembles of transition paths (see Materials and Methods for details), and obtained an array of reaction coordinate (RC, denoted by $\{r\}$) points from the simulated shortest paths (Fig.1d). The RC is a central concept in rate theories (Hanggi *et al.*, 1990) that reflects progression of the EMT process. The RC points divide the M-dimensional state space into Voronoi cells, so the value of RC of each cell was assigned by the Voronoi cell that it locates in.

Next, to study how the regulatory network reconfigures along the RC, we need the governing dynamical equations of the EMT process. Lamanno et al. showed that from scRNA-seq data one can obtain both the single cell expression vectors of the spliced mRNAs $\{\mathbf{x}^\alpha\}$, and estimate the instant RNA velocity vectors $\{\mathbf{v}^\alpha = (d\mathbf{x}/dt)^\alpha\}$ from reads of spliced and unspliced mRNAs, with α representing the α-th cell. Qiu et al. (Qiu *et al.*, 2021) further developed a procedure of reconstructing the generally nonlinear equations from the data. Here we adopted a simpler linear model by assuming the governing equation as $\mathbf{v} = \mathbf{F}\mathbf{x} + \varepsilon$, with ε being random white noises, and F_{ij} quantifying the regulation of gene j on gene i so the node strength and direction in the intracellular gene regulatory network (GRN). The regulation can be direct such as gene j acting as a transcription factor on gene i, or indirect mediated through molecular species not resolved by the scRNA-seq measurements. We inferred the matrix $\mathbf{F}$, which is in general asymmetric, from the data $\{\mathbf{x}^\alpha, \mathbf{v}^\alpha\}$ with the partial least square regression (PLSR) together with local false discovery rate (LFDR) methods to ensure $\mathbf{F}$ is sparse (see Materials and Methods) (Pihur *et al*, 2008).

With $\mathbf{F}$ being the same for all cells, the expression states of genes within the GRN may differ between different cells. Notice that a prerequisite for gene j acting on i is that gene j is expressed in the cell, otherwise the $j \rightarrow i$ edge is treated as non-existent in this specific cell.

Reconstructed gene regulatory network reveals increased intercommunity interactions at intermediate EMT stage.

Similar to the SN1 v.s. SN2 mechanisms, the concerted but not the sequential mechanism predicts significantly increased gene-gene interactions. Therefore, to evaluate the two possible mechanisms, we calculated the number of effective edges, i.e., edges with nonzero F_{ij} and gene j expressed (s_j = 1) in the cell, which indeed increases first and then decreases during EMT (Fig. EV1A).

To examine the nature of the increased interactions, we divided the inferred GRN into four communities using the Louvain method (Blondel *et al*, 2008; Traag *et al*, 2019). Each community contains both E and M genes, and the number of effective intra-community edges correlates with the number of active genes. We did not observe a universal pattern among the four communities on how intracommunity interactions change along the reaction coordinate. Actually the number of intracommunity edges for community 0 and 1 have different trends of variation,, while the community 2 and 3 have peaked changes (Fig. EV1B). In contrast, the number of effective inter-community edges increases first and then decreases (Fig. 1E). For visual inspection, we examined intercommunity interaction strengths, defined as the total number of effective edges between different communities, at several points along the RC, which clearly show strongest intercommunity interactions at r = 10 (Fig. 1F). This variation of intercommunity effective edges is not related with the number of genes that are active (Fig. EV1C).

Network analyses identify the increase of frustration and network heterogeneity during EMT

In a SN2 mechanism, the increased bond number is due to coexistence of the bonds that are to form and break. Analogously, for a concerted mechanism one expects co-expression of genes that normally only express in one stable phenotype, leading to confliction on the expression state of a gene and the regulation acting on it. To quantify existence of such conflicting interactions, with the gene expression binarized as 0 for silence, and 1 for active expression, we defined a cell-specific effective matrix, $\bar{F}_{ij} = (2s_i - 1)F_{ij}$. Then we defined a frustration value for the interaction between a pair of genes (i, j) as $fs_{ij} = s_j \,\mathrm{sgn}\left(\bar{F}_{ij}\right)$, assuming a value 1 (not frustrated), 0 (no regulation), and -1 (frustrated), and sgn(x) is the usual sign function,

$$\mathrm{sgn}\left(x\right) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

Furthermore we defined the overall frustration score of a cell-specific GRN as the fraction of frustrated edges out of all edges in the whole network of the cell (Tripathi *et al*, 2020). For EMT, the average frustration score along RC increases first and reaches a peak when cells were treated with TGF-β for about one day, then decreases (Fig. 1G), consistent with the concerted but not the sequential mechanism. We also calculated $H = -\sum_{i,j} fs_{ij}$, analogous to the pseudo-Hamiltonian of a cell defined by Font-Clos et al (Font-Clos *et al*, 2018) but with directed regulation, which also shows similar peaked profile along the RC (Fig. EV1D).

Gao et al defined three additional quantities to characterize network topological structures (Gao *et al*, 2016). Define vectors of outgoing and ingoing weighted degrees of gene connectivity in the network as, $d^{out} = \mathbf{1}^T \bar{\mathbf{F}}'$, and $d^{in} = \bar{\mathbf{F}}'\mathbf{1}$, respectively, with $\mathbf{1}$ the unit vector $\mathbf{1} = (1, \ldots, 1)^T$, and $\bar{\mathbf{F}}'$ the $\bar{\mathbf{F}}$ matrix excluding the diagonal terms. Network heterogeneity measures how homogenously the weighted connections are distributed among the genes, $\mathcal{H} = \sigma^{in}\sigma^{out}/\langle d \rangle$, with σ^{in} and σ^{out} the square roots of variances of the elements of d^{in} and d^{out}, respectively. For the EMT data the network heterogeneity shows a similar pattern of first increase upon leaving the epithelial region, reaching a maximum, then decrease when approaching the mesenchymal region (Fig. 1H).

To rule out the possibility that the observed properties are specific for the Markov model we used, we repeated the above analyses with a transition matrix reconstructed from a correlation kernel of Bergen et al. (Bergen *et al*, 2020) The analyses gave similar RC, and the frustration score as well as the number of effective edges (Fig. EV2A-C) change along the RC similarly as observed with the model obtained using that of Qiu et al.

In total, several lines of evidence support that EMT proceeds through a concerted mechanism. Indeed, both in vivo and invitro studies have identified intermediate states of EMT that have co-expressed epithelial and mesenchymal genes (Pastushenko *et al*, 2018; Zhang *et al*, 2014). A schematic summary of the concerted mechanism is shown in Fig. 1I. Co-expression of conflicting genes leads to increased intercommunity edges, and frustrated edges. Some of the genes transiently act as hub genes, which lead to increased network heterogeneity.

Reconstructed gene regulatory network reveals similar concerted mechanism in two additional developmental systems.

To investigate whether the concerted mechanism is general for CPTs, we performed the same analyses on two additional CPT scRNA-seq datasets. One is on development of pancreatic endocrine cells (Bastidas-Ponce *et al*, 2019). During embryonic development Ngn3-low progenitors first transform into Ngn3-high precursors then Fev-high cells. The latter further develop into endocrine cells, specifically glucagon producing α-cells that we focus on here (Fig. 2A). A calculated RC characterizes this transition process (Fig. EV3A). The number of effective inter-community edges increases first and then decreases while low Ngn3 expression cells transit into α-cells (Fig. 2B &2C). The number of effective edges shows similar trend (Fig. EV3B).

The cells in state A and B were selected from Ngn3-low progenitors and glucagon producing α-cells, respectively. Along the RC, the peak of the average frustration score locates at the cell population with high Ngn3 expression (Fig. 2D). The network heterogeneity and the pseudo-Hamiltonian value show a trend similar to the frustration score (Fig. 2E and Fig. EV3C).

Another system is on development of the granule cell lineage in dentate gyrus, where radial glia-like cells differentiate through nIPCs, Neuroblast 1 and 2, immature granule cells, and eventually into mature granule cells (Fig. 2F & Fig. EV4A) (Hochgerner *et al*, 2018). The number of effective inter-community edges increases first and then decreases when low radial glia like cells transit into mature granule cells (Fig.2G &2H). The number of effective edges shows a similar trend (Fig. EV4B).

Along the calculated RC, the neuroblast has the highest frustration score (Fig. 2I). The network heterogeneity and the pseudo-Hamiltonian value again exhibits dynamics similar to the frustration score (Fig. 2J and Fig. EV4C). Therefore the dynamical properties of the two processes are consistent with the concerted mechanism.

DISCUSSIONS

The idea of relating CPTs and chemical reactions has been discussed in the literature (Moris *et al*, 2016). Here we presented a procedure of reconstructing the RC of a CPT process from scRNA-seq data. A related concept is the transition state. In chemical reactions it typically refers to short-lived intermediates or a state of maximal potential energy along the RC. It is tempting to identify the intermediate state with highest frustration as the "transition state", while it is unclear whether it is indeed a dynamical bottleneck of the associated CPT process.

The observed common pattern of transiently peaked intercommunity interactions provides a new angle to examine the structure-function relation of a biological network. Previous theoretical and experimental studies have shown that a biological network is generally modularized with dense intracommunity interactions and sparse intercommunity interactions, which helps insulating perturbations in one community from propagating globally and increases functional robustness of each module (Gardner & Ashby, 1970; Gilarranz Luis *et al*, 2017; May, 1972). The observed pattern supports that intercommunity interactions of a GRN are indeed minimized at stable phenotypes. During a CPT, a cell needs to escape a stable phenotype, and the increased intercommunity interactions help on coordinating gene expression profile change among communities. The decreased modularity is consistent with a critical state transition mechanism (Mojtahedi *et al*, 2016) that individual components become more connected and correlated, as what observed near the critical point of a phase transition.

In the two developmental processes the frustration score of initial state A is higher than that of final state B. Notice that the initial states are stem-like cells, Ngn3-low progenitors and radial glia-like cells, respectively. The EMT process seems to be different with the final mesenchymal state being less frustrated than the initial epithelial state. However, EMT is generally regarded as a dedifferentiation process. Therefore the results of all three datasets suggest that frustration decreases with differentiation. Gulati et al. identified single-cell transcriptional diversity as a hallmark of developmental potential (Gulati *et al*, 2020). It remains to examine whether frustration provides an alternative and complementary measure on developmental potential.

In summary, in this work through analyzing scRNA-seq data of CPTs in the context of dynamical systems theory we identify that many CPTs may share a common concerted mechanism. This conclusion is also supported by an increasing number of studies on various CPT processes reporting existence of intermediate hybrid phenotypes that have co-expression of marker genes of both the initial and final phenotypes such as the partial EMT state (Zhang *et al.*, 2014). Notice that a cell typically has multiple target phenotypes to choose, functionally the concerted mechanism may allow canalized transition for directing the cells to transit to a specific target phenotype, as visioned by the developmentalist C. H. Waddington (Waddington, 1942).

MATERIALS AND METHODS

1. Data sets

The scRNA-seq data of EMT, development of pancreatic endocrine cells, development of granule cell lineage were obtained from the GEO website with GEO number GSE121861 (Cook & Vanderhyden, 2020), GSE132188 (Bastidas-Ponce *et al.*, 2019), and GSE95753 (Hochgerner *et al.*, 2018), respectively.

2 Gene selection

We focused on genes showing switch-like behavior during the phenotype transition. For the EMT dataset, first we selected high-expression genes across the whole dataset. To avoid selection bias, we selected high-expression genes in 0 d, 8 h, 1 d and 3 d, separately. The filtering criterion is based on minimum number of counts and minimum number of cells, two parameters in scVelo. We set the minimum number of counts to be 20, and minimum number of cells to be 5% of the number of cells of the corresponding cell group. The gene set is the union set of these selected genes, excluding genes unrelated to the transition process (see below for the filtering procedure). A total of the top 2000 high-expression genes of the whole dataset was selected for subsequent analyses.

For the pancreatic endocrinogenesis dataset, cells have been grouped into four types, and we first selected genes for each type separately, then combined them to select the top 2000 high-expression genes of the whole dataset, excluding genes not related to the transition processes. We set the minimum number of counts to be 20 and minimum number of cells to be 10% of the number of cells of the corresponding cell type.

For the Dentate gyrus neurogenesis dataset, cells have been grouped into seven types. We used the same parameters as for the pancreatic endocrinogenesis dataset.

To filter out genes that are not related to a transition process under study, we used a number of regression methods including f-regression and mutual information regression (Pedregosa *et al.*, 2011). The regression targets were set as the sample time (for the EMT dataset) or the cell types (for the other two sets) along the transition direction of the CPT processes. For the EMT process, the sample time of 0 day, 8 hour, 1 day and 3 day were assigned values of 0, 1, 2, and 3, respectively. For the development of pancreatic endocrine cells, the regression target values of Ngn3-low progenitors, Ngn3-high precursors, Fev-high cells and glucagon producing α-cells were set as 0, 1, 2, and 3, respectively. For development of granule cell lineage in dentate gyrus, the regression target values of radial glia-like cells, nIPCs, Neuroblast 1, Neuroblast 2, immature granule cells 1, immature granule cells 2, and mature granule cells are set as 0, 1, 2, 3, 4, 5 and 6, respectively. The values of f-regression and mutual information were normalized by the maximum value of each gene. If a gene's f-regression score is larger than the threshold h_f or mutual information is larger than the threshold h_m, it was chosen for later analysis. For EMT

process, we set h_f = 0.1, and h_m = 0.15. For the other two processes, we set h_f = 0.1, and h_m = 0.5.

To binarize the expression state of a selected gene within one cell, we use the Kmeans method to determine its ON and OFF state by grouping cells into two clusters based on the expression value of the gene.

3 Path analysis from single cell RNA velocity analysis

With scRNA-seq velocity analysis, we reconstructed the velocity graph of the whole cell population, which is a transition matrix between all pairs of the cells. Each cell is treated as a node in this network. The distance between different cells is $-log\ P_{ij}$, where P_{ij} is the transition probability between cells. Here we added one constrain in the velocity graph that only the transition between the cells that their sample time points are successive or they have the same sample time. We also took the cell density at each sample time point into consideration. The density of each single cell ρ_t (normalized by the maimum) at its sample time was evaluated by kernel density estimation. The transition from cell i to cell j is penalized by the relative cell density of cell j . The distance from cell i to cell j with penalty is $-log\ P_{ij} \times (1 + 5 * \exp(-(\rho_t - r)))$. For EMT, we used r = 0.5, while for the other two datasets we used r = 0.8. A single cell trajectories is more likely to pass the high-density region of each sample time. A total of 100 cell pairs (c_f, c_l) were randomly selected from high density regions in the first and last sample populations. The shortest paths between the cell pairs were calculated with Dijkstra's algorithm. These shortest paths are the probable single cell trajectories in the phenotype transition.

4 Procedure for determining a RC

We follow a procedure adapted from what used in the finite temperature string method for numerical searching of RC and non-equilibrium umbrella sampling (Dickson *et al*, 2009; Vanden-Eijnden & Venturoli, 2009).

a) Identify the starting and ending points of the reaction path as the means of data points in the state A and state B, respectively. The two points are fixed in the remaining iterations.
b) Construct an initial guess of the reaction path that connects the two ending points in the feature space through linear interpolation. Discretize the path with N (= 15) points (called images, and the k_{th} image denoted as r_k with corresponding coordinate $\mathbf{X}(r_k)$) uniformly spaced in arc length.
c) For a given trial RC, divide the multi-dimensional state space by a set of Voronoi polyhedra containing individual images, and calculate the score function, $F = \sum_k \sum_u \sum_t \left\| s_k - X_{u,t} \mid X_{u,t} \in s_k \right\|^2 + w \sum_k \sum_u d_{k,u}$, where X_{u,t_α} stands for the points on simulated trajectory u at step t_α that reside within the k_{th} polyhedron (containing image point s_k); $d_{k,u}$ is the distance between image s_k and trajectory u, defined as the distance between each image on the path to the closest point on the trajectory, $d_{k,u}^2 = \arg\min \left\| r_k - X_{u,\arg\min \|r_k - X_{u,t}\|^2} \right\|^2$; w is a parameter

that specifies the relative weights between the two terms in the right hand of the expression, here we use 2.

d) Carry out the minimization procedure through an iterative process. For a given trial path defined by the set of image points, we calculate a set of average points using the following equations, $\overline{X}(r_k) = \frac{\sum_u \sum_t \{X_{u,t} \mid X_{u,t} \in r_k\} + w \sum_u X_{u,\arg\min\|r_k - X_{u,t}\|^2}}{1+w}$. Next we update the continuous reaction path through cubic spline interpolation of the average positions (Jones *et al*, 2001), and generated a new set of *N* images $\{X(r_k)\}$ that are uniformly distributed along the new reaction path. We set a smooth factor, *i.e.,* the upper limit of $\sum_{k=1}^{N}\left(\overline{X}(r_k) - X(r_k)\right)$, as 1 for calculating the RC.

e) Iterate the whole process in step 3 until there was no further change of Voronoi polyhedron assignments of the data points.

5 Network inference

We adopted a partial linear square regression (PLSR) method to infer the gene regulation networks. We used the velocity vector of each single cell $d\boldsymbol{x}/dt$ and the level of spliced mRNA $\boldsymbol{x}$ for inferring the GRN with the PLSR method, $d\boldsymbol{x}/dt = \boldsymbol{F}\boldsymbol{x} + error$, where $\boldsymbol{F}$ is a constant and generally asymmetric matrix describing gene regulation strength. Regression methods are widely used in network inference, and among which the PLSR has several advantages. First, it can be used when the number of features is larger than the number of samples. In the scRNA dataset, the number of genes is often comparable to or larger than the number of cells. Second, it can avoid over-fitting because it uses major components for regression. However, the regulation relation *F* obtained from PLSR is typically a dense matrix, while most GRNs are sparse. To generate a sparse network, we further adopt the method of local false discover rate (LFDR) to select those regulation relations that are statistically significant. This procedure ensures the GRN is sparse (Pihur *et al.*, 2008). Since cells within each Voronoi cell can scatter dispersively in the orthogonal space, we select only cells close to the RC for infering the $\boldsymbol{F}$ matrix. That is, among cells within each Voronoi cell, we selected the k-nearest-neighboring (KNN) cells of the corresponding RC point. Such cells from all Voronoi cells collectively form the set for **F** matrix inference. We performed the inference using scikit-learn (Pedregosa *et al.*, 2011) by maximizing the covariance between $\boldsymbol{x}$ and $d\boldsymbol{x}/dt$ in the PLSR method. The value of components was set to be 2 and data were standardize . In LFDR, the null hypothesis H_0 assumes that $F_{i,j}$, which is regulation from gene *j* to gene *i*, is 0. An interaction is identified as nonzero when $fdr(F_{i,j}) < q$, where $fdr(F_{i,j})$ is the false discover rate and q is the threshold (Efron, 2007). The following R package was used for calculation (*https://rdrr.io/cran/locfdr/*), with a central matching estimation method. The degrees of freedom was set as 10 in all calculations, and q was set as 0.1.

Code availability: A python notebook is included in the package *dynamo*.

Acknowledgements: We thank Yan Zhang for help discussions on using the *dynamo* package. This work was partially supported by National Cancer Institute (R37 CA232209), and National Institute of Diabetes and Digestive and Kidney Diseases (R01DK119232) to JX.

Competing interests: There is no competing interest to declare.

Figure captions

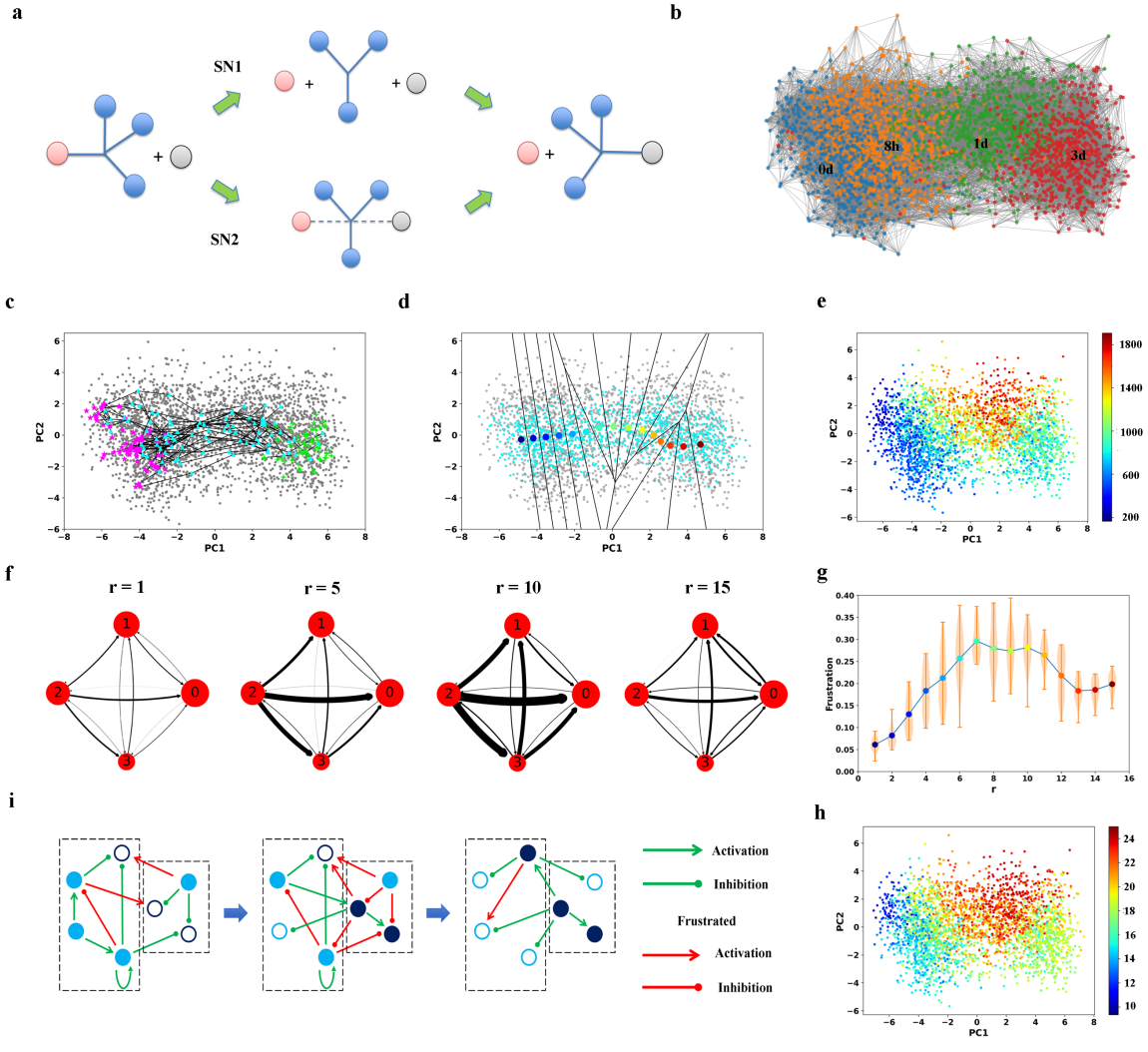

Figure 1 Dynamical systems theory analyses of scRNA-seq data of A549 cells undergoing TGF-β-induced epithelial-to-mesenchymal transition reveal a concerted transition mechanism.

(A) Competing mechanisms for substitution reactions. SN1 and SN2 mechanisms for chemical reactions. The nodes and edges represent chemical groups and chemical bonds, respectively.

(B) scRNA-seq data and RNA velocity-based transition graph shown in the cell expression state space (shown in the 2D leading PCA space). Each dot represents a cell, and each edge between two dots indicates a transition between cell states corresponding to the two cells.

(C)Dijkstra shortest path sampled on the transition graph, illustrated in the 2D leading PCA space. Each path is a single cell trajectory (labeled with cyan triangle) that starts from the epithelial state (labeled with magenta star) and transit into the mesenchymal state (labeled with lime cross).

(D) 1-D Reaction coordinate (RC) reconstructed from the Dijkstra shortest paths using a revised finite temperature string method. The colored dots represent the RC points (start from blue and ends in red). The cyan dots are cells close to the RC to form a reaction tube, i.e., cells within each Voronoi cell that are k-nearest-neighbors of the corresponding RC point. These cells were used to infer the ***F*** *matrix in the gene space.*

(E) Cell-specific network structure characterization with the number of effective intercommunity edges in the GRN of EMT. Each dot represents a cell.

(F) Evolution of the number of effective intercommunity edges along the RCs during EMT. Each node represents a community (with the size of the node standing for the number of genes in the community). The width of an edge represents the number of effective edges. Arrow represents direction of regulation.

(G) Frustration score along the RC of EMT. The mean and variance at each RC point were calculated using all cells within the reaction tube (cyan dots) and the corresponding Voronoi cell.

(H) Cell-specific network structure characterization of GRN heterogeneity.

(I) Schematic of the concerted mechanism for a cell phenotypic transition. Filled circles represent active genes. Empty circles represent silent genes. Colors represent marker genes of different cell states. The dash-line boxes represent communities.

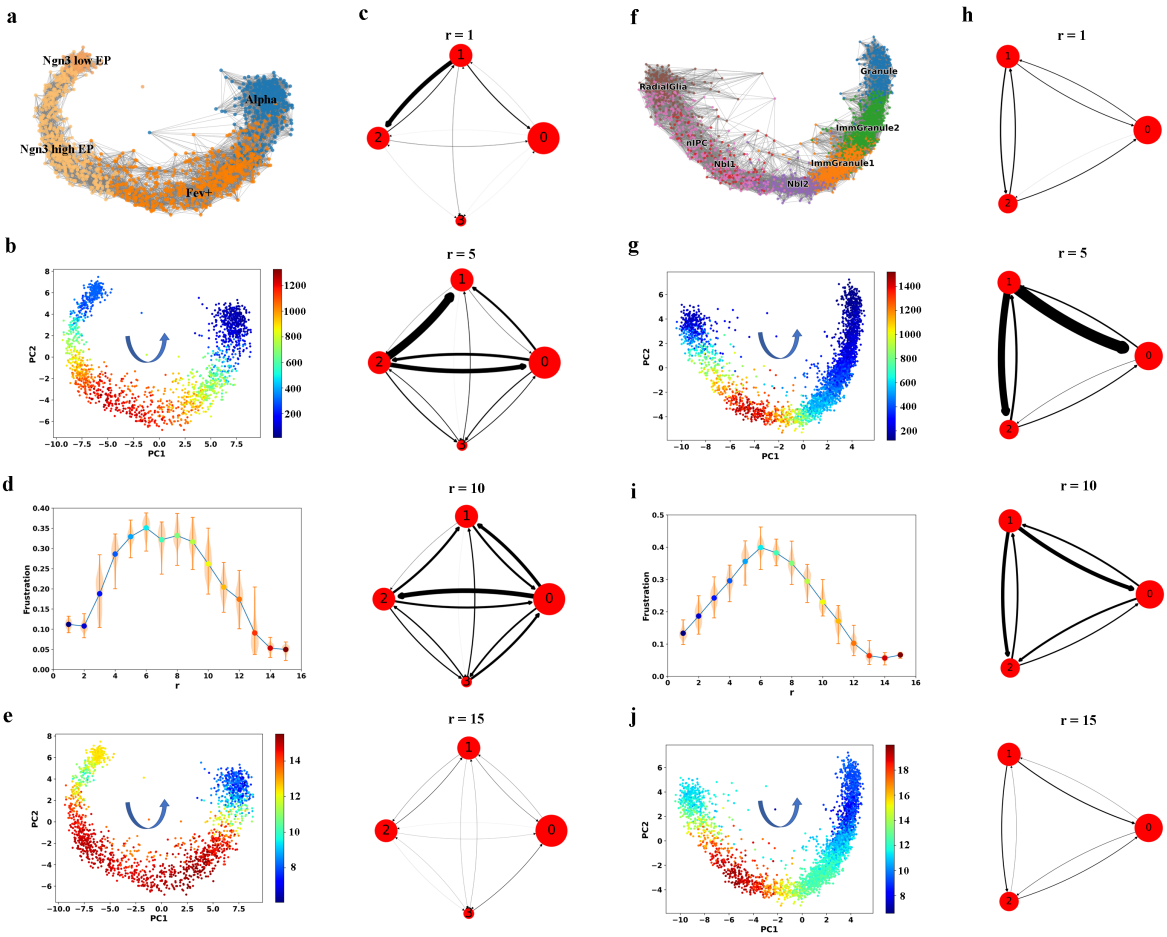

Figure 2 Analyses on pancreatic endocrinogenesis and dentate gyrus neurogenesis

(A) Transition graph of pancreatic endocrinogenesis based on RNA velocity.

(B) Cell-specific variation of effective intercommunity regulation in endocrine cell development. Colors represent the number of effective intercommunity edges within each cell in the GRN. Arrow represents the direction of development.

(C) Evolution of the number of effective intercommunity edges along the RC during pancreatic endocrinogenesis. Each node represents a community (with the size of the node standing for the number of genes in the community). The width of an edge represents the number of effective edges between two communities.

(D) Frustration score along the RC in endocrine cell development.

(E) Cell-specific variation of heterogeneity in endocrine cell development.

(F)-(J) Same as in panel (a)-(e), respectively, except for the granule cell lineage development in dentate gyrus neurogenesis dataset.

Expanded View figure captions

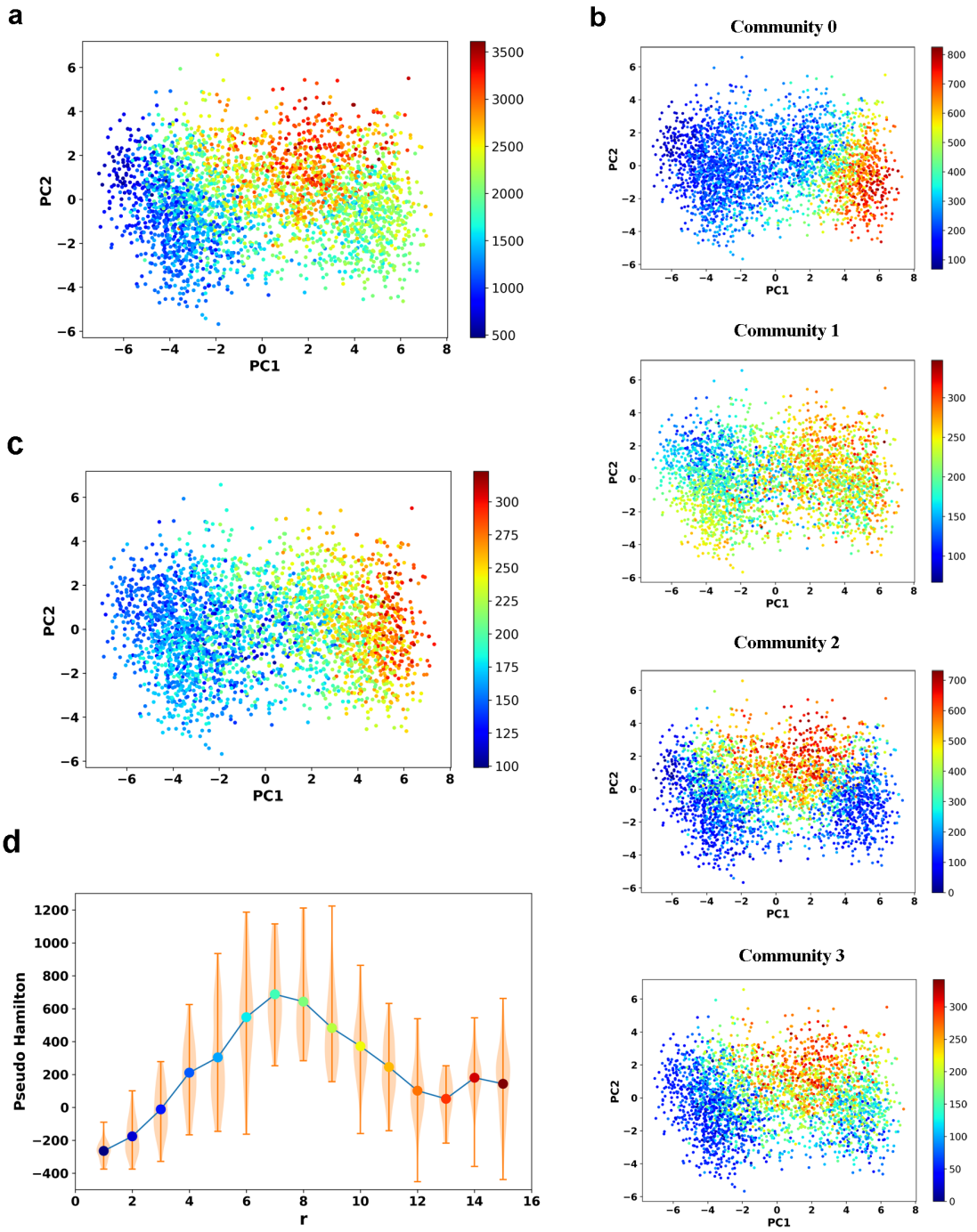

Figure EV1 Additional results on transition path analyses of the EMT dataset.

(A) Cell-specific network structure characterization with the number effective regulation edges by color. Each dot represents a cell.

(B) Cell-specific network structure characterization with the number effective regulation edges inside each community by color. Each dot represents a cell.

(C) Cell-specific characterization with the total number of effective genes by color. Each dot represents a cell.

(D) Pseudo Hamilton values along the RC.

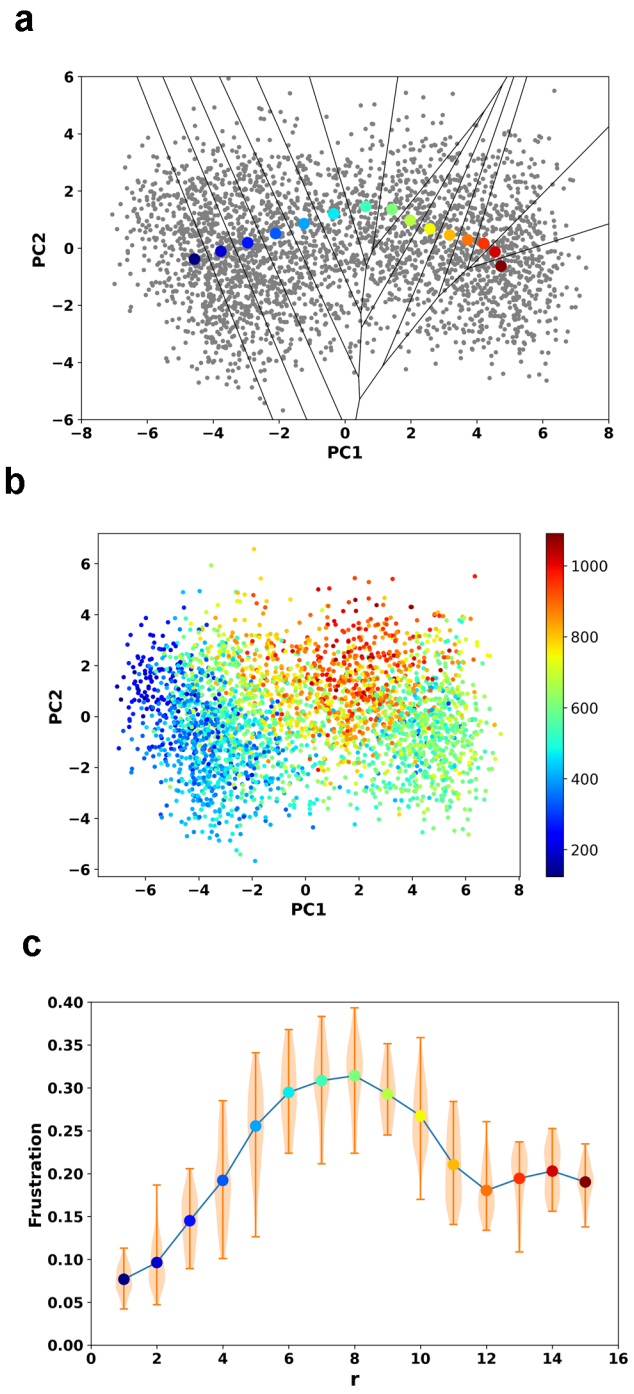

Figure EV2 Analysis of the EMT dataset with the package scvelo.

(A) The RC calculated from the Dijkstra shortest paths. The colored dots represent the RC points (start from blue and ends in red). Each grey dot represents a cell.

(B) Variation of effective inter-community regulation edges in the GRN in the processes of EMT. Colors represent the number of effective inter-community regulation edges in individual cells.

(C) Frustration score along the RC of EMT.

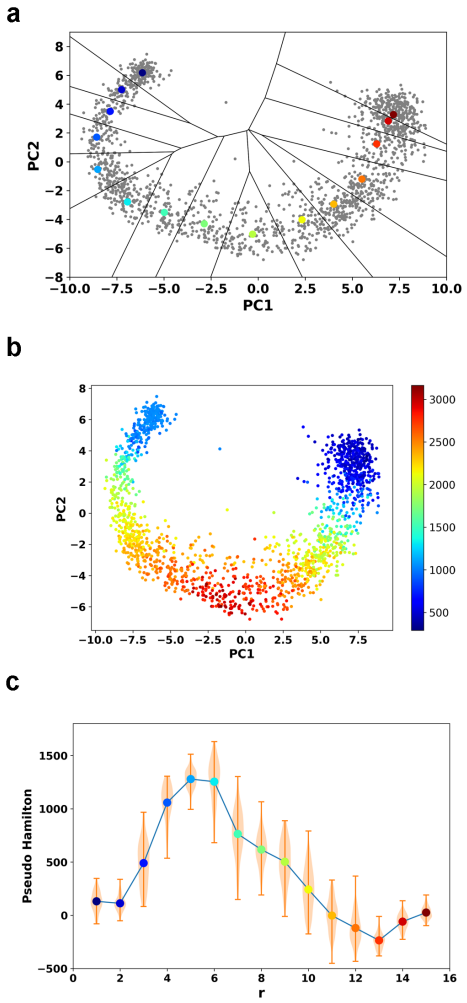

Figure EV3 Transition path analyses of the pancreatic development dataset.

(A) Voronoi cells defined by an array of points equally distributed along the RC divide the expression space of pancreatic endocrinogenesis into different regions. The colored dots represent the RC points (start from blue and ends in red). Each grey dot represents a cell.

(B) Cell-specific network structure characterization with the number of effective regulation edges by color. Each dot represents a cell.

(C) Pesudo-Hamiltonian values along the RC.

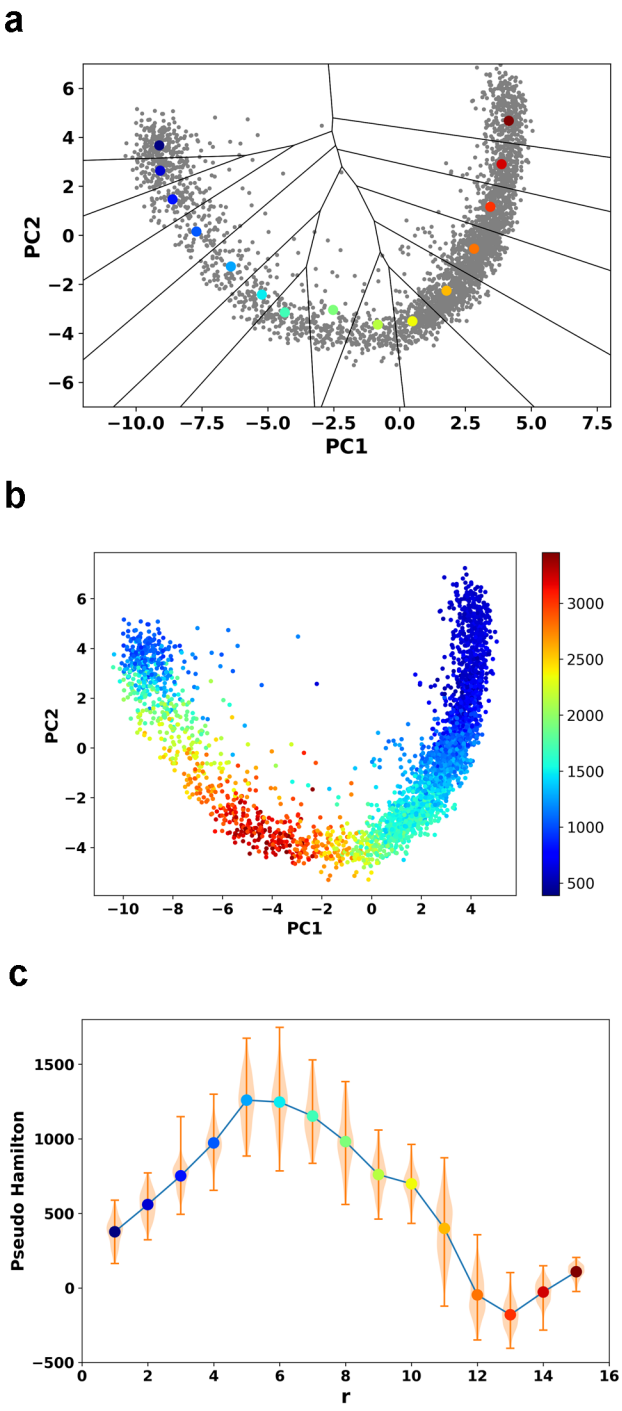

Figure EV4 Transition path analyses of the dentate gyrus neurogenesis dataset.

(A) Voronoi cells defined by an array of points equally distributed along the RC divide the expression space of dentate gyrus neurogenesis into different regions. The colored dots represent the RC points (start from blue and ends in red). Each grey dot represents a cell.

(B) Cell-specific network structure characterization with the number of effective regulation edges by color. Each dot represents a cell.

(C) Pesudo-Hamiltonian values along the RC.

References

Bastidas-Ponce A, Tritschler S, Dony L, Scheibner K, Tarquis-Medina M, Salinno C, Schirge S, Burtscher I, Böttcher A, Theis FJ *et al* (2019) Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development (Cambridge, England)* 146

Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008

Cook DP, Vanderhyden BC (2020) Context specificity of the EMT transcriptional response. *Nature Communications* 11: 2142

Dickson A, Warmflash A, Dinner AR (2009) Nonequilibrium umbrella sampling in spaces of many order parameters. *The Journal of chemical physics* 130: 02B605

E W, Ren W, Vanden-Eijnden E (2005) Finite Temperature String Method for the Study of Rare Events. *J Phys Chem B* 109: 6688-6693

Efron B (2007) Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association* 102: 93-103

Font-Clos F, Zapperi S, La Porta CAM (2018) Topography of epithelial–mesenchymal plasticity. *Proceedings of the National Academy of Sciences* 115: 5902-5907

Gao J, Barzel B, Barabási A-L (2016) Universal resilience patterns in complex networks. *Nature* 530: 307-312

Gardner MR, Ashby WR (1970) Connectance of Large Dynamic (Cybernetic) Systems: Critical Values for Stability. *Nature* 228: 784-784

Gilarranz Luis J, Rayfield B, Liñán-Cembrano G, Bascompte J, Gonzalez A (2017) Effects of network modularity on the spread of perturbation impact in experimental metapopulations. *Science* 357: 199-201

Gulati GS, Sikandar SS, Wesche DJ, Manjunath A, Bharadwaj A, Berger MJ, Ilagan F, Kuo Angera H, Hsieh RW, Cai S *et al* (2020) Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* 367: 405

Hanggi P, Talkner P, Borkovec M (1990) Reaction-rate theory: 50 years after Kramers. *Rev Mod Phys* 62: 254-341

Hochgerner H, Zeisel A, Lönnerberg P, Linnarsson S (2018) Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nature Neuroscience* 21: 290-299

Jones E, Oliphant T, Peterson P (2001) SciPy: Open source scientific tools for Python.

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A *et al* (2018) RNA velocity of single cells. *Nature*

May RM (1972) Will a Large Complex System be Stable? *Nature* 238: 413-414

Mojtahedi M, Skupin A, Zhou J, Castaño IG, Leong-Quong RYY, Chang H, Trachana K, Giuliani A, Huang S (2016) Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biology* 14: e2000640

Moris N, Pina C, Arias AM (2016) Transition states and cell fate decisions in epigenetic landscapes. *Nature reviews Genetics* 17: 693-703

Morrison RT, Boyd RN (2010) *Organic Chemistry*. Pearson education

Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, Van Keymeulen A, Brown D, Moers V, Lemaire S *et al* (2018) Identification of the tumour transition states occurring during EMT. *Nature* 556: 463-468

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et*

al **(2011) Scikit-learn: Machine Learning in Python.** ***J Mach Learn Res*** **12: 2825–2830**

Pihur V, Datta S, Datta S (2008) Reconstruction of genetic association networks from microarray data: a partial least squares approach. ***Bioinformatics (Oxford, England)*** **24: 561-568**

Qiu X, Zhang Y, Hosseinzadeh S, Yang D, Pogson AN, Wang L, Shurtleff M, Yuan R, Xu S, Ma Y ***et al*** **(2021) Mapping Transcriptomic Vector Fields of Single Cells.** ***bioRxiv*****: 696724**

Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. ***Scientific Reports*** **9: 5233**

Tripathi S, Kessler DA, Levine H (2020) Biological Networks Regulating Cell Fate Choice Are Minimally Frustrated. ***Phys Rev Lett*** **125: 088101**

Vanden-Eijnden E, Venturoli M (2009) Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. ***The Journal of chemical physics*** **130: 05B605**

Waddington CH (1942) Canalization of development and the inheritance of acquired characters. ***Nature*** **150: 563-565**

Wang W, Xing J (2020) Analyses of multi-dimensional single cell trajectories quantify transition paths between nonequilibrium steady states. ***bioRxiv*****: 2020.2001.2027.920371**

Zhang J, Tian XJ, Zhang H, Teng Y, Li R, Bai F, Elankumaran S, Xing J (2014) TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. ***Science signaling*** **7: ra91**