Revealing cell-fate bifurcations from transcriptomic trajectories of hematopoiesis

Simon L. Freedman[a,b], **Bingxian Xu**[c], **Sidhartha Goyal**[d,e,*], **and Madhav Mani**[a,b,c,*]

[a]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208; [b]NSF-Simons Center for Quantitative Biology, Northwestern University, Evanston, IL 60208; [c]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208; [d]Department of Physics, University of Toronto, Toronto, ON M5R 2M8; [e]Institute of Biomedical Engineering, University of Toronto, ON M5R 2M8; [*]To whom correspondence should be addressed

Inspired by Waddington's illustration of an epigenetic landscape, cell-fate transitions have been envisioned as bifurcating dynamical systems, wherein the dynamics of an exogenous signal couples to a cell's enormously complex signaling and transcriptional machinery, eliciting a qualitative transition in the collective state of a cell – its fate. Single-cell RNA sequencing (scRNA-seq) measures the distributions of possible transcriptional states in large populations of differentiating cells, making it possible to interrogate cell fate-transitions at whole-genome scales with molecular-scale precision. However, it remains unclear how to bridge the disparate scales of the dynamics of whole transcriptomes to the molecules that define the collective fate-transitions in Waddington's geometric vision. We bridge these scales by showing that bifurcations in transcriptional states can be analytically pinpointed and their genetic bases revealed, directly from data. We demonstrate the power of our conceptual framework and analytical scheme in the context of a recent scRNA-seq based investigation of a classic case-study of sequential fate decisions – the transition of hematopoietic stem cells to neutrophils. Our work provides a rigorous and model-independent mathematical framework for detecting and categorizing transitions in cell-fate directly from sequencing data, aiding in gene network inference and determination of the salient properties of cellular differentiation, such as when cell fate transitions are reversible and how these transitions generate a diversity of cell types.

1. Introduction

During development and tissue regeneration, cells progress through multiple transitions to ultimately adopt a distinguishable function. While each transition en-route to a terminal fate involves the coordination of myriads of molecules and complex gene regulatory networks interacting with external factors, there is a common view that cellular specification depends on significantly fewer underlying variables and control parameters. This view was notably explicated by Conrad Waddington, in an illustration of an epigenetic space as a tilted, bifurcating landscape, where a vast number of nodes (genes) provide the scaffold for the smooth hills and valleys (cell state) along which a pebble (cell) can reliably roll down until it finds a resting position (terminal fate) (Fig. 1A) (1).

Many of the characteristics of Waddington's landscape can be codified into the language of dynamical systems, including that cell fates resemble valleys (attractors) in gene expression, or transciptomic, space (2–5), that a small amount of stable states can emerge from large interconnected simple boolean networks (6), and that known genetic interactions can yield multiple cell fates (bistability) (7, 8). Waddington's illustration has motivated analysis of the wealth of data captured in single-cell RNA-sequencing (scRNA-seq) as well, in which the transcriptome of populations of cells are measured, often at multiple time-points as they differentiate. For example a mathematical model of a pitchfork bifurcation can be fit to scRNA-seq data and yield predictions for developmental perturbations (9), large transcriptomic matrices can be dimensionally reduced to enhance the resolution of bifurcations to precisely determine the genes enabling a cell fate decision (10, 11), and well characterized cell-lineage relationships can be used to extract predictive models of gene regulation (12, 13). More

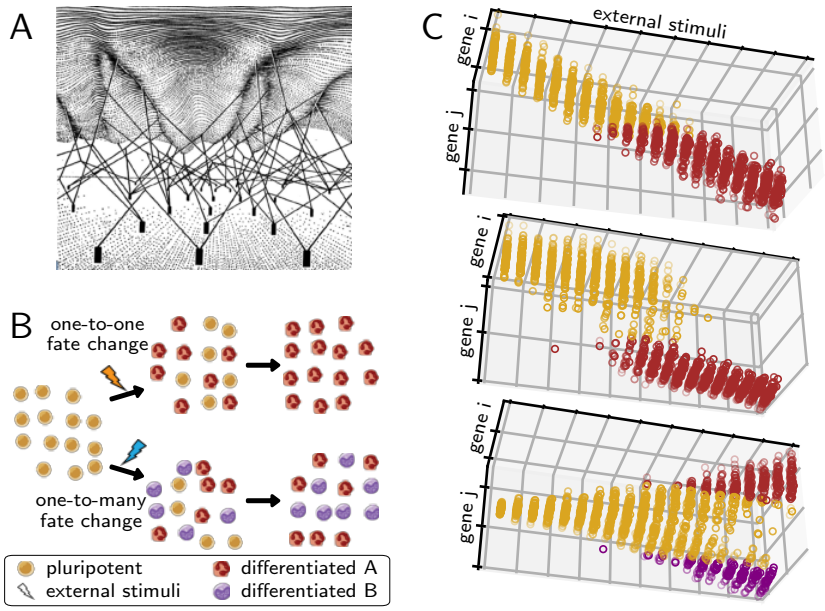

Fig. 1. Cell type differentiation as a dynamical process. (A) Waddington's landscape of cell fate commitment in which cell fates are represented as valleys, committment barriers as hills, and gene activity as pegs underneath that control the heights of hills and valleys. Adapted from Refs. (1, 15). (B) Schematic cell-population snapshots of maturation (top), in which one cell fate transitions to a different one, and a cell fate decision (bottom), in which a pluripotent cell differentiates to either of two lineages. Cell type images by Mikael Häggström, and A. Rad, used with permission and recolored. (C) Gene expression trajectories for cells (dots) at varying levels of a differentiating stimuli for cases where the differentiation landscape does not bifurcate (top), undergoes a saddle-node bifurcation (middle), or undergoes a pitchfork bifurcation (bottom).

recently, it has been shown that the proliferation in the number of accessible cell fates during development evident in Waddington's landscape is a statistical feature of time-course scRNA-seq data (14).

While it is now common place to employ statistical dimensionality reduction methods to analyze scRNA-seq datasets, spanning multiple cell fates and decisions, it is often not apparent where, in the low dimensional representation, cell fates change (15–18). This has motivated the formulation of pseudotime algorithms, parametric curve fitting tools that use the transcriptome similarity of cells to determine their relative ordering in developmental time, and yield transcriptomic trajectories that span developmental decisions (10, 19, 20). It remains unclear, however, how to bridge the measurements of transcriptomic trajectories to dynamical systems theory and Waddington's perspective: that development comprises distinct, discontinuous changes of cell-fate. Additionally, it is unclear if and how transcriptomic trajectories can be used to unbiasedly learn more about the genetic mechanisms driving cell fate changes.

Our main insight in this paper is that developmental pseudotime can be viewed as a control, or bifurcation, parameter for studying changes in transcriptomic state. As an example of how control parameters relate to the transcriptional state within a cell, consider the classical relationship between thio-methylgalacatoside (TMG) (a molecule similar to lactose) and the cellular process of Lactose metabolism (21). Lactose metabolism is maintained via the *lac* operon, a DNA functional unit comprised of three genes: *lacZ*, which encodes enzymes that metabolize Lactose, *lacY* which encodes proteins that facilitate the uptake of TMG and

similar molecules from outside the cell, and *lacA* which encodes enzymes for sugar metabolism. TMG activates transcription of the operon by inhibiting LacI from repressing the operon. A positive feedback loop via *lacY* that encodes for extracellular TMG uptake ensures that high concentrations of extracellular TMG enable the cell to adopt a lactose-metabolic state, while low concentrations yield a non-lactose-metabolic state. Most interestingly, intermediate extracellular TMG concentrations yield bistability between the two states, a signature of saddle-node bifurcations (21). Thus, dynamical molecular processes involving transport and the regulated expressions of genes, ultimately yield a cell's adoption of a specific transcriptional state (e.g., high LacZ and LacY) and corresponding function (lactose-metabolic), but a functional change driven by a bifurcation can only be triggered if the control parameter (the amount of extracellular TMG) is varied past a threshold. While recent work has used pseudotime as a proxy for developmental time, we take this further and show that it can be viewed as an axis along which a control parameter(s) steadily varies, permitting a dynamical systems style investigation of possible bifurcations in the collective transcriptomic state. This dynamical systems view is further suggested by the observation that the dynamic molecular processes that lead to changes in the transcriptomic state, such as signal transduction and transcription, generically occur on the order of seconds and minutes (22), while cell fates change over the course of hours or days (4), suggesting that developmental control parameters vary at significantly slower rates than their downstream molecular processes.

Here, we layout and demonstrate a statistical formalism for detecting and interrogating bifurcations in developmental fate transitions directly from transcriptomic pseudotime trajectories. Unlike previous studies (7–10), we do not assume any specific mathematical form for the underlying genetic interactions, but instead focus on the observation that at some point along a bifurcative developmental fate change, there must be a window of developmental time during which the observed cells exhibit multistability between the transitioning fates (Fig. 1B) (23, 24). We show that our data-driven approach enables us to distinguish between three different types of transcriptomic variation directly from systems-level data: a non-bifurcative cell fate change that is due to continuous changes in gene expression (Fig. 1C, top) without multistability; a cell fate change that is due to a one-to-one state transition (Fig. 1C, middle), such as those that may occur during terminal-fate maturation (25); and a cell fate change that is due to a one-to many-state transition, resembling those depicted in Waddington's landscape that occur when pluripotent cells decide between multiple cell lineages (Fig. 1C, bottom). We apply our framework to a class of *in-silico* gene networks, to demonstrate its ability to recover the salient features of a bifurcating dynamical system, and examine the effects of high dimensionality and noise. We demonstrate the utility of our framework in the context of a recently published scRNA-seq exploration of hematopoiesis, and show that cell-fate bifurcations can be pinpointed and analyzed in scRNA-seq data even without detailed knowledge of the underlying system's dynamics and controls. Finally, we demonstrate that our framework can be used to extract genetic relationships that are pivotal to the dynamics underlying a bifurcative cell-fate change, in order to recover genes of known importance in hematopoiesis and generate new testable hypotheses for genetic networks that underly a cell fate decision.

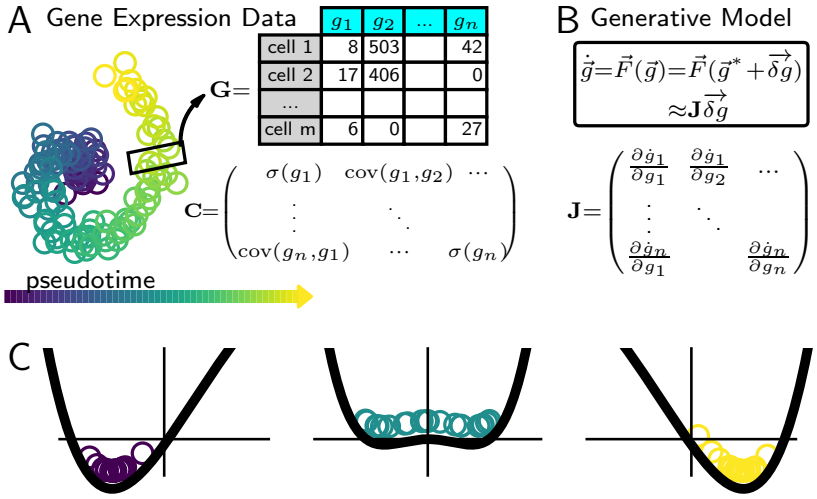

Fig. 2. Relating the Jacobian to the Covariance at a bifurcation. (A) Schematic of a single-cell RNA-seq dataset arranged by each cell's developmental (pseudo-) time. (Left) Visualization of dataset in two collective gene-expression dimensions. (Right) Gene expression matrix at the psuedotime indicated by the black rectangle, and its corresponding covariance matrix. (B) Schematic of a generative model ($\vec{F}$) that could yield the gene expression matrix in (A), and its connection to the Jacobian ($\mathbf{J}$). In this model, $\overrightarrow{\delta g}$ is the deviation of the gene expression vector, $\vec{g}$ from the fixed point, $\vec{g}^*$. (C) Snapshots of a collection of particles at steady state following the dynamical process defined by $\dot{x} = -2x^3 + x/2 + a$ and uniformly sampled noise for $a = -1$ (left), $a = 0$ (center) and $a = 1$ (right). Black line shows the underlying landscape, $\int \dot{x} dx$.

2. Theory: the relationship between the Jacobian and the Covariance

In this section we outline the theoretical foundations that enables gleaning key factors driving cellular differentiation from scRNA-seq data. Summarizing our key insights, we layout a conceptual framework tailored to the analysis of sequencing data that, despite the absence of a generative mechanistic model for the dynamics of this high-dimensional and complex system, can reveal and investigate the system's most salient dynamical features, its bifurcations, from the transcriptomic trajectories alone (24, 26).

A scRNA-seq measurement yields a transcriptomic matrix, where each row is a different cell and each column is a different gene (Fig. 2A). Various dimensionality reduction and visualization tools, such as t-SNE (27), and k-nearest neighbor maps (28), have been used to visualize this data and determine pairwise distances between cells based on their transcriptomes (for example, the methods described in Secn. S3). More recently, parametric curve fitting tools have been developed to determine how cells in transcriptomic space vary with respect to a state variable, such as developmental time (pseudotime) (10, 19, 20) (Secn. S3). Where a pseudotime trajectory contains a cellular state change, then one could perhaps determine which genes are responsible, by binning the cells along the trajectory to statistically compare $\mathbf{G}(\tau)$, the transcriptomic matrix at pseudotime τ, with a neighboring matrix at a time $\Delta\tau$ in the future, $\mathbf{G}(\tau + \Delta\tau)$ (29). While such an analysis may reveal which genes exhibit dynamic signals, it will not reveal if the underlying cell-fate landscape has bifurcated, and may be sensitive to the discretization (13). Here, we present an alternative method, fundamentally rooted in dynamical systems theory and aimed at pinpointing cell-fate bifurcations, that only relies on being able to calculate $\mathbf{C}$, the gene-gene covariance at a particular pseudotime (Fig. 2A).

Our analytical framework hinges on a fundamental mathematical relationship between the covariance matrix $\mathbf{C}$ estimated from a data matrix and the most salient properties of the underlying, unknown, generative model for the dynamics (illustrated in Fig. 2B). Regarding

the underlying biochemical processes codified in a mechanistic model, we assume that (1) they are stochastic and Markovian (30–33), and (2) occur at significantly faster timescales (seconds to hours) than the timescales over which transitions in cellular fates are observed (hours to days) (4, 22). A consequence of these assumptions (details in Secn. A) is that the local time evolution of a cell's transcriptomic profile is controlled by a single matrix, the Jacobian ($\mathbf{J}$), where $J_{ij} = \partial \dot{g}_i / \partial g_j$ is the effect of the amount of gene j on the dynamics of gene i (Fig. 2B). Thus, the Jacobian matrix encodes the update rules that determine the future state of the transcriptome based on its current state and the exogenous signals presented to the cell. Note that the Jacobian matrix, in general, changes with pseudotime.

Under these assumptions, $\mathbf{J}$ relates to $\mathbf{C}$ through the continuous-time Lyapunov equation (34),

$$\mathbf{JC} + \mathbf{CJ}^T + \mathbf{D} = 0 \qquad [1]$$

where $\mathbf{D}$ is the expected noise amplitude for individual genes and their their interactions (derivation in Secn. 5A) (26). Eq. (1) is a remarkable relationship between the Jacobian and Covariance matrices of a system, implying that some mechanistic details of a system are encoded in, and recoverable from, the fluctuations of the state variables alone! While leveraging this relationship doesn't make it possible to determine every element of $\mathbf{J}$, since $\mathbf{D}$ is generically unknown and $\mathbf{J}$ (unlike $\mathbf{C}$) is generically not a symmetric matrix, its most salient properties corresponding to its eigen-decomposition are inferrable from the covariance matrix near bifurcations.

Before we discuss the quantitative ramifications of this relationship in the context of a realistic high-dimensional dynamical system we first investigate a simple, but easy to visualize, one-dimensional toy-model, ($\dot{x} = -2x^3 + x/2 + a + \eta$) that undergoes a (imperfect pitchfork) bifurcation as a function of a control parameter (a) and is stochastic, controlled by the amplitude of the Langevin term, η. Fig. 2C displays the system before ($a < 0$), at ($a = 0$), and after ($a > 0$) the bifurcation. The slope of the potential function, drawn in black, determines the deterministic features of the dynamics of the system. In particular, parameter regimes ($a \ll 0$ and $a \gg 0$) where the potential has highly convex curvature exhibit stable fixed points, while parameter regimes near the bifurcation ($a = 0$), that have much flatter curvature exhibit instability. This dramatic reduction in curvature (mathematically, the curvature goes to 0 at the bifurcation itself) is the geometric signature of a bifurcation. Stochastic simulations of the system (drawn as open circles in Fig. 2C – color corresponds to the value of the control parameter) as a function of the control parameter demonstrate that owing to the reduction in curvature of the underlying potential, the data is spread maximally at the bifurcation, and narrows on either side of it.

The above toy model extends naturally to higher dimensions and captures the essence of the idea put forward in this paper. Summarizing the mathematical argument below, if a complex high-dimensional dynamical system undergoes a bifurcation, then in its vicinity there must be, by definition, some direction in the high-dimensional space whose curvature reduces dramatically, or softens. Consequentially, at the bifurcation point, the fluctuations in the system will be greatly enhanced along that soft direction. Generically, this direction need not point along any single dynamical variable (for example individual genes in a gene interaction network) of the original systems.

We now elaborate on the mathematical details of our central argument. Generically, the landscape's curvature can be obtained from diagonalizing $\mathbf{J}$, such that, if

$$\mathbf{J} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} \qquad [2]$$

where Λ is a diagonal matrix of eigenvalues $\{\lambda_1, \lambda_2 \dots \lambda_{n_g}\}$ and $\mathbf{P}^T$ is the square matrix of eigenvectors $\{\vec{p}_1, \vec{p}_2 \dots \vec{p}_{n_g}\}$, then λ_i is the curvature of the landscape in the $\vec{p}_i$ direction. $\lambda_i < 0 \; \forall \; i$ indicates that the landscape is convex in all directions, while $\lambda_d = \max(\Lambda) \to 0$ indicates that the landscape is flat in the $\vec{p}_d$ direction, allowing for a fixed point exchange, or bifurcation. $\lambda_d \to 0$ also considerably simplifies Eq. (1), such that it can be shown

$$\lim_{\lambda_d \to 0} C_{ij} \propto \frac{\vec{p}_d^{\,i} \vec{p}_d^{\,j}}{2\lambda_d} \qquad [3]$$

(details in Secn. 5B). The effect of $\lambda_d \to 0$ on $\mathbf{C}$ is most clearly recognized by rewriting $\mathbf{C}$ in terms of its eigen-decomposition,

$$C_{ij} = \sum_{k=1}^{n_g} \omega_k \vec{s}_k^{\,i} \vec{s}_k^{\,j} \qquad [4]$$

where $\{\omega_1, \omega_2, \dots, \omega_{n_g}\}$ are its eigenvalues, and $\{\vec{s}_1, \vec{s}_2, \dots, \vec{s}_{n_g}\}$ are its eigenvectors. Since all $\vec{s}_i$, by definition, normalize to 1, Eq. (4) can only equal Eq. (3) if for at least one ω (ω_1, without loss of generality)

$$\lim_{\lambda_d \to 0} \omega_1 = \infty. \qquad [5]$$

Thus, the covariance diverges along the principal direction $\vec{s}_1$, as illustrated for the toy model in Fig. 2C. Furthermore, since the direction of the landscape flattening (x in the toy model) is the same as the direction of the covariance expansion, it can be shown, by equating Eq. (3) to Eq. (4) (where the $k = 1$ term will dominate), that

$$\lim_{\lambda_d \to 0} \vec{s}_1 = \pm \vec{p}_d \qquad [6]$$

(details in Secn. 5C). Lastly, we note a direct result of Eq. (3) is that at a bifurcation the Pearson's correlation coefficient $R_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$ of the data along axes i and j, where their corresponding loadings on the eigen vector are non-zero, becomes maximal, as

$$\lim_{\lambda_d \to 0} \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} = \pm 1. \qquad [7]$$

There is thus a secure correspondence between the fluctuations of a system and the underlying geometric landscape that determines its dynamics. In particular, three specific changes to the transcriptomic covariance data, Eq. (5)-7, that can be determined from observations of the dynamics alone can inform us as to the salient features of the system, its bifurcations, even when we have no direct access to the generative model for the dynamics, or its corresponding underlying geometry. Notably, these features only rely on $\mathbf{G}$ being at steady state in the vicinity of an attractor, and do not rely on special circumstances, such as the Jacobian being symmetric, or the noise being of a particular nature. Landscape geometry and gene dynamics are two sides of the same phenomena, and we leverage one to learn about the other.

We first demonstrate that these theoretical results apply to high-dimensional noisy dynamical systems in the context of a simulated gene regulatory network, and that we can leverage them to detect and assess bifurcations in the underlying developmental landscape from observations of state-variables alone. Following this, we demonstrate the power of this approach by directly applying it to scRNA-seq data for neutrophil lineage in the hematopoietic system.

3. Results

A. Covariance analysis recovers salient features of high-dimensional *in silico* gene regulatory networks. To better understand our mathematical framework in the context of scRNA-seq data, where the large number of discordant genes and biological noise may obfuscate the predicted covariance signal indicative of a bifurcation, and cell fate changes may take different geometric forms, we tested the framework on a noisy, high-dimensional, gene-regulatory network (GRN), whose dynamics are governed by a set of explicit ordinary differential equations,$\dot{\mathbf{G}} = \mathbf{F}(\mathbf{G})$, simulated with Poissonian noise (details in Secn. 6). In our model, cell fate transitions result from two mutually inhibiting "driver" genes, g_1 and g_2 via their dynamics

$$\begin{aligned} \dot{g}_1 &= \frac{m_1}{(1+g_2^2)} - k_D g_1 \\ \dot{g}_2 &= \frac{m_2}{(1+g_1^2)} - k_D g_2 \end{aligned} \qquad [8]$$

where k_D are their degradation rate, and $m_{1,2}$ determine the scales of their synthesis (35). This simple system is illustrated in Fig. 3A. Varying the control parameter m_1 yields a saddle-node bifurcation in gene-expression while varying k_D yields a pitchfork bifurcation (Fig. S1). Similar networks have been analyzed to provide insight into gene-inhibition and activation (36) in a diversity of biological systems, such as the lac-operon (21) and cell-cycle control (37).

As GRNs typically involve hundreds of genes, we include an additional $n_g - 2$ genes in the network that respond variably to the driver genes, according to

$$\dot{g}_i = \frac{\alpha_i g_{d(i)}^2 + 1 - \alpha_i}{(1+g_{d(i)}^2)} - k_i g_i \qquad [9]$$

where $i \in [3, n_g]$. k_i is the degradation rate of ith gene, g_i. The variable $d(i)$ is set to 1 if g_i responds to g_1 and 2 if it responds to g_2. $\alpha_i \in [0, 1]$ is the strength of the connection between the responder g_i and its driver gene, $g_{d(i)}$ ($\alpha_i = 0$ yields full inhibition and $\alpha_i = 1$ yields full activation). Thus each of these responder genes is connected to one of the two driver genes, to greater or lesser extents, as indicated in Fig. 3A. Though this GRN can be made more complex by including feedback from the responding genes to the two driver genes, or increasing the number of driver genes themselves, this simple model provides an interpretable demonstration of our proposed scheme.

We simulated Eq. (8)-Eq. (9) for a fixed number of genes (n_g), statistical replicates, or cells, (n_c), noise scale (s), duration (N_t), and timestep (δt) for different values of the control parameters (m_1, k_D) (details in Secn. 6). We define the $[n_c \times n_g]$ transcriptomic matrix $\mathbf{G}(m_1, k_D)$ once the system has reached steady-state in the simulation. We observed that

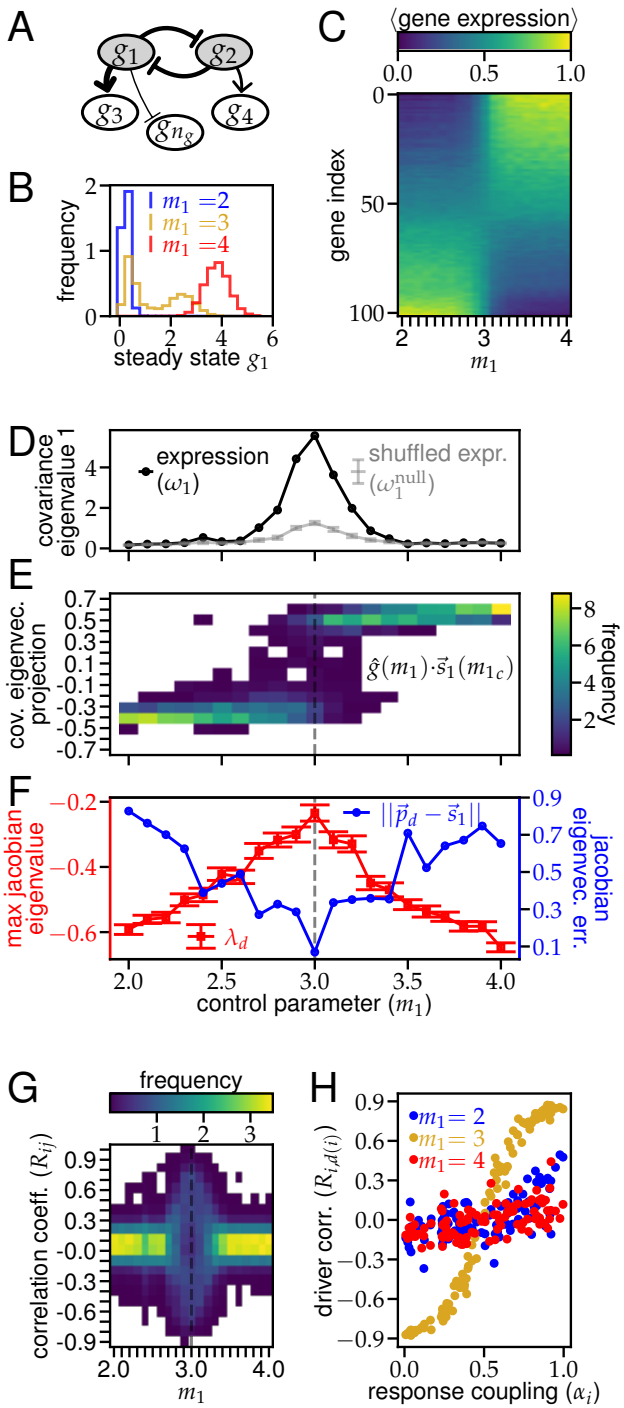

Fig. 3. Analysis of a gene regulatory network around a saddle-node bifurcation. (A) Schematic of the GRN for 5 of the 102 genes. Undisplayed nodes have a unidirectional arrow stemming from either g_1 or g_2. (B) Distributions of g_1 at steady state (details in Secn. 6) for three values of m_1. (C)-(G) GRN observables as a function of the bifurcating variable m_1 evaluated over a distribution of 100 cells (Secn. 6). (C) Average final expression for each gene. Expression of driver genes, $g_{1,2}$, (bottom and top rows, respectively) are min-max normalized. Response genes are sorted by their corresponding driver ($d(i)$) and activation level (α_i). (D) Largest eigenvalue of covariance matrix. (E) Distribution of normalized gene expression for each cell projected onto the bifurcating axis. (F) Red squares: largest eigenvalue of the Jacobian matrix. Error bars are SEM. Blue circles: Euclidean distance from the corresponding Jacobian eigenvector ($\vec{p}_d$) to the principle covariance eigenvector $\vec{s}_1$. (G) Distribution of Pearson's correlation coefficients for all gene pairs. Each column integrates to 1. (H) Correlation coefficients between responding genes and their driver as a function of the coupling coefficient α (Eq. (9)) at three values of m_1.

the steady state distributions for individual genes (for example, g_1, shown in Fig. 3B) shift their mean as the control parameter, m_1, is varied, and exhibit bimodality at the bifurcation point, $m_1 = m_{1c} = 3$, as expected for saddle-node bifurcations (Fig. 1B). We verified that N_t was sufficiently large by averaging $\mathbf{G}(m_1)$ across cells, and observing that individual genes discontinuously, but predictably, switch their expression at m_{1c} (Fig. 3C; genes sorted by $d(g_i)$ and α_i) compared to the continuous and unpredictable transitions observed with low N_t (Fig. S2A).

Having verified that our model simulates a situation where the systems undergoes a high dimensional saddle-node bifurcation driven by a 2-gene driver core, we use it to examine the effects of noise and a large number of responding genes on the theoretical predictions described in Secn. 2 (Eq. (5),6,7). As predicted, we found that $\omega_1(m_1)$, the largest eigenvalue of the covariance of $\mathbf{G}(m_1)$, is maximal at the critical value m_{1c} (darker line in Fig. 3D), and the increase is significantly larger than can be obtained from a null distribution (lighter line in Fig. 3D). We generated this null through a marginal-resampling approach, which explictly destroys any correlation between genes, of $\mathbf{G}(m_1)$ (details in Secn. S2). This contrast between the data and the null can be understood by considering the bimodality of the transcriptomic distribution at the bifurcation. Far from the bifurcation, the transcriptomic distribution is unimodal, and all ω_i scale with the noise scale s, which is undirected, and therefore unaffected by resampling, yielding $\omega_1 \sim \langle \omega_1^{\text{null}} \rangle$ (Fig. S3 left and right panels). However, at the saddle-node bifurcation, the transcriptomic distribution is bimodal, so ω_1 scales with the distance between the two modes (Fig. S3 center-top); marginal resampling of transcriptomes at the bifurcation yields new modes and the increased dimensionality of the bifurcation diminishes ω_1^{null}, compared to ω_1 (Fig. S3 center-bottom). While Fig. S3 only demonstrates the bifurcation bimodality in $g_{1,2}$, the full transcriptomic bimodality can be visualized by computing $\hat{g}(m_1) \cdot \vec{s}_1(m_{1c})$, the normalized projection of each cell's transcriptome along the principal covariance eigenvector. The distribution of this projection is densely centered around different fixed points to the right and left of m_{1c}, but widens significantly at m_{1c} as there is non-zero probability for both transcriptomic modes (Fig. 3E).

Since, in this example, we have an explicit generative model ($\mathbf{F}(\dot{\mathbf{G}})$ given by Eq. (8)-9), we can validate that just as m_{1c} resembles a bifurcation from analysis of the covariance matrix, it also resembles a bifurcation of the full, noisy GRN, from analysis of the Jacobian. We show that the maximum eigenvalue (λ_d) of the Jacobian ($\mathbf{J}(m_1) = \frac{\partial \dot{\vec{g}}}{\partial \vec{g}}|_{\vec{g}(m_1)}$) for this network approaches 0 as $m_1 \to m_{1c}$ (Fig. 3F). We also show that at m_{1c}, the direction of maximal covariance is given by the corresponding eigenvector of the Jacobian ($\vec{p}_d$), as the Euclidean distance between $\vec{s}_1$, the principal eigenvector of the covariance, and $\vec{p}_d$ approaches 0 as $m_c \to m_{1c}$ (Fig. 3F). Thus, while the finite system size (n_c) prevents, or regularizes, ω_1 from diverging completely, and $|\vec{s}_1 - \vec{p}_d| > 0$, ω_1 is still at its largest and the eigenvectors are in closest correspondence, at the bifurcation.

While the covariance eigen decomposition provides insight into the timing and direction of a bifurcation, our theory predicts that the (Pearson) correlation coefficients between genes may help determine which genetic relationships are most critical for the dynamics at the bifurcation. We found that for low and high m_1, when the network only has one fixed point, the distribution of R_{ij} is strongly centered around 0 (Fig. 3G). However, at the bifurcation, this distribution spreads out to ± 1 as predicted in Eq. (7) (Fig. 3G). To determine if the

gene pairs that yielded large R_{ij} corresponded with critical gene relationships in our network, we plot $R_{i,d(i)}$ the correlation between all responder genes and their drivers, sorted by their connection strength α_i. We demonstrate that these correlation coefficients were much more strongly indicative of the responder-driver dependency (α_i) at the bifurcation (Fig. 3H, green) than far away from the bifurcation (Fig. 3H, red and blue). Again, the model makes explicit that although the correspondence between geometry and dynamics is not universal, it provides a bridge in the vicinity of a bifurcation. Thus, entries of a correlation matrix with high magnitude at a bifurcation may be reliable indicators of mechanistic gene-regulatory features.

We further used our model to understand Eq. (5) in the context of the pitchfork bifurcation induced by varying k_D (Fig. S4A). Unlike the example of a saddle-node bifurcation, we observed that ω_1 does not peak at the bifurcation parameter $k_{Dc} = 0.5$, but rather begins to increase (Fig. S4B). This feature directly follows our interpretation that ω_1 is the distance between the two modes of the transcriptomic distribution. While the bimodality of the saddle-node bifurcation results from the discontinuous transition between states, the bimodality of a pitchfork bifurcation emerges continuously from its root and becomes more pronounced as the control parameter is increased. Therefore, the distance between the modes (ω_1) increases with the control parameter. By clustering the cells according to their transcriptomic mode, or branches, we are able to recover the bifurcation signature predicted by Eq. (5) (Fig. S4C), but we note that precise clustering requires prior knowledge (e.g., of how many clusters there are).

In this example, we have demonstrated the applicability and power of our theoretical infrastructure to analyze a high-dimensional and noisy dynamical system undergoing a variety of bifurcations, uncovering its crucial aspects, including its location, direction in gene space, and influential genetic relationships. These calculations are also comptuationally simple; while covariance matrices can be cumbersome to compute for large numbers of genes and cells, reduced singular value decomposition can be used to quickly determine its largest eigenvalue and eigenvector, which is all our approach requires. Notably, our results only apply if the system is measured at steady state; in the absence of which there is no reason to anticipate clear divergences in the distribution of eigenvalues, transient bimodality, or equivalence between the covariance and Jacobian principal directions (Fig. S2B-D).

B. Covariance analysis pinpoints a bifurcation in Hematopoietic stem cell development. Having verified that gene-gene covariance can be used to identify and classify a bifurcation in a simulated genetic context, we applied our analysis framework to a recently published scRNA-seq data set of mouse hematopoietic stem cell (HSC) differentiation (18).In this experiment, HSCs were isolated *in vitro*, barcoded, plated in a media that supports multilineage differentiation (day 0), and subsequently sampled for single-cell sequencing using inDrops (38, 39) on days 2, 4, and 6. The resultant transcriptomic matrix (25,289 genes in 130,887 individual cells) was visualized in $2D$ via the SPRING method (28) (Fig. 4(A)), in which the matrix was filtered to only include highly variable, non-cell cycle genes, and (x, y) coordinates are determined such that each cell is optimally placed closest to its 4 nearest neighbors in the space of the top 50 gene-wise principal components (PC) (28). Each cell was then ascribed to one of 11 different cell types (annotations in Fig. 4(A)) based on its position in the SPRING plot and expression of cell-type specific marker genes (18). Cells that belonged to the developmental transition from multipotent progenitor (MPP) to neutrophil were identified by recategorizing cells as a cell-label distribution, and ranking cells by their similarity to fully committed

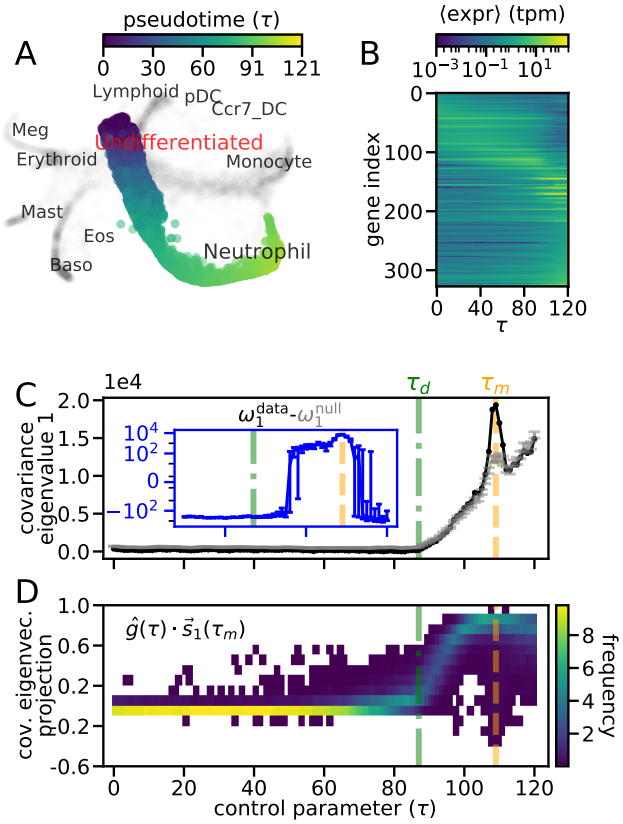

Fig. 4. Covariance analysis of temporal scRNA-seq data shows signatures of developmental bifurcations. (A) SPRING visualization for each cell (point) in an *in-vitro* scRNA-seq experiment of mouse hematopoeitic stem cell differentiation (18). Cells in the pseudotime trajectory analyzed are colored accordingly (blue to yellow) while others are gray. SPRING coordinates, cluster labels, pseudotimes and a similar visualization were first reported in Ref. (18). (B)-(D) Observables of neutrophil trajectory as a function of pseudotime calculated for each of 121 bins of pseudotemporally adjacent cells. All bins had 1000 cells except for the last one which had 1310 cells, and had a 50% overlap with neighboring bins. (B) Average gene expression in pseudotemporal bins for highly expressed ($\max(\langle \text{expr} \rangle) > 1$) and highly varying ($\text{coeff. of var.}(\langle \text{expr} \rangle) > 0.5$) genes. (C) Largest covariance eigenvalue (black) compared with a statistical null (gray, details in Secn. S2). Error bars of null are one SD. (D) Distribution of normalized gene expression for each cell projected onto the convariance eigenvector at τ_m. Dashed lines indicate bifurcation pseudotimes.

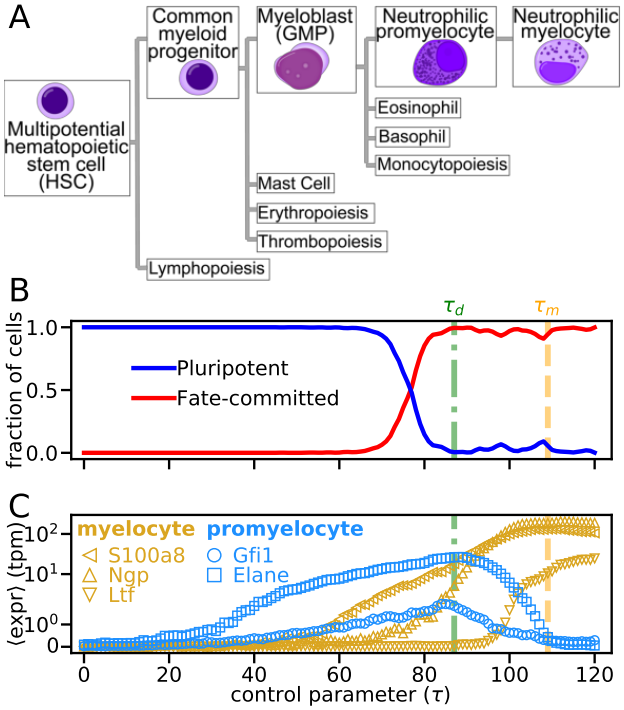

Fig. 5. Detected bifurcations correspond to biologically characterized developmental transitions. (A) Schematic of neutrophil development, beginning from hematopoietic stem cells and ending at the neutrophil myelocyte, a committed neutrophil progenitor. Lines indicate naturally occuring progeny, other than the cell type itself. Subsequent neutrophil-committed fates (neutrophil metamyelocyte, band cells, and neutrophils) are not shown. Cell type images by Mikael Häggström, and A. Rad, used with permission. (B) Fraction of cells in each cell type, based on annotated clustering in Ref. (18). (C) Average expression of promyelocyte (blue) and myelocyte (gold) marker genes (18). SEM error bars are smaller than symbols.

neutrophils (details in Secn. S3). The 61,310 cells identified as belonging to the neutrophil transition were sorted into a neutrophil pseudotime trajectory (Fig. 4A) by ranking cells according to their similarity with the earliest pluripotent cells (Secn. S3). This data-specific pseudotime algorithm was validated via the clonal barcodes, by ensuring that the MPP cells in the trajectory included neutrophil clones, and via the sequencing time, by ensuring that cells collected earlier were ranked earlier in the trajectory. We found that expression of hundreds of highly expressed and variable genes in these cells varied temporally along this trajectory, with large groups of cells either monotonically increasing or decreasing (Fig. 4B).

There were several features of this trajectory that make it ideal for applying our analysis framework: it is robust to the systematic, temporal controls of sequencing-time and cellular bar-codes, and included a large number of cells, enabling statistically reliable covariance and correlation measurements. Finally, this trajectory is part of hematopoiesis, a well-characterized developmental process, allowing us to compare our findings against past work.

To determine if the transitions from HSC to neutrophil were due to bifurcations in transcriptomic space, we split the neutrophil trajectory into overlapping bins of 1000 cells (last bin had 1310; details in Secn. S4 and Fig. S5) and applied our covariance analysis to the transcriptomic matrix at each bin $\mathbf{G}(\tau)$. We found that the largest eigenvalue of the covariance matrix ($\omega_1(\tau)$, dark line in Fig. 4C) exhibited very little variation for $\tau < \tau_d = 85$, but began

to increase at τ_d, and exhibited a significant spike at $\tau_m = 109$. To determine if these ω_1 changes were statistically significant, we computed the corresponding statistical null (ω_1^{null}, lighter line in Fig. 4(C), details in Secn. S2) and found that the large peak at τ_m was easily distinguishable from the null. We first focus our attention at the dynamics at τ_m, following which we will address those observed at τ_d. As this pattern of a statistically significant spike following near-constant ω_1 echoed the observed behavior of a saddle-node bifurcation in our toy model (Fig. 3D), we speculated that at τ_m there was a one-to-one transcriptomic state transition. To further visualize the multimodality due to this one-to-one state transition, we computed the projection of normalized gene expression onto the first covariance eigenvector at the bifurcation ($\vec{s}_1(\tau_m)$). We found (Fig. 4D) that similarly to the saddle-node bifurcation (Fig. 3E) the distribution of this projection was widest at the transition point, τ_m, providing further evidence that the cells switched from one transcriptomic state to another at τ_m, and that at τ_m the transcriptomic state was multimodal.

We focus now on our observations in proximity to τ_d. As the increase in ω_1 at τ_d strongly resembled our toy model's pitchfork bifurcation (Fig. S4B), as well as the proliferation of cell fates seen in high-resolution time-course scRNA-seq experiments (14), we sought to determine if the increase of ω_1 at τ_d was also due to transcriptomic state changes. While ω_1^{data} is not obviously well-separated from ω_1^{null} during the increase, the difference between them switched from negative to positive shortly after τ_d (Fig. 4C, inset), indicating a regime crossover. Furthermore, despite gene expression being normalized per cell throughout the dataset, the distribution of each gene's expression (across cells) begins to significantly shift toward higher values at τ_d (Fig. S5C). These findings suggest that with additional sequencing or prior knowledge, the genomic and geometric details of a one-to-many transcriptomic state transition at τ_d may be distinguishable.

To determine if the transcriptomic state transitions we identified had biological significance, we compared our findings against the tree of cell fates that has been identified for neutrophil development (Fig. 5A). This tree includes multiple one-to-many cell fate changes, such as the cell fate decision at which a granulocyte monocyte progenitor (GMP, or myeloblast), can be induced to divide into the progenitors of any of four terminal fates: neutrophil, monocyte, eosonophil, and basophil, as well as multiple one-to-one cell fate changes, such as the maturation step where a neutrophil promyelocyte is induced to become a neutrophil myelocyte (18, 40, 41). We found that τ_d corresponded well with the crossover point between pluripotent and neutrophil-commitment labels (Fig. 5B) identified in previous cell fate clustering (18), as well as the point in pseudotime at which promeylocyte marker genes are maximal (Fig. 5C). This finding suggests that the increase in ω_1 beginning at τ_d is a consequence of the cell fate decision that occurs between GMP and promyelocyte. Specifically, the resemblance of this increase to the unclustered pitchfork bifurcation (Fig. S4B) suggests that while cells in this trajectory were expected to only include the neutrophil lineage of GMP, they may in fact include other GMP lineages, such as eosonophils or basophils.

Conversely, the spike in ω_1 identified at τ_m correponds well with the switchover from high expression of promyelocyte marker genes (identified in Ref. (18)) to myelocyte marker genes (Fig. 5C). This biochemical result matches our geometric interpretation that the spike in ω_1 at τ_d indicates a one-to-one fate transition, as the promyelocyte-to-myelocyte cell fate transition is a maturation step of committed neutrophil progenitors, and not a decision between two

fates.

Thus, by using Eq. (5) to quantify the geometry of neutrophil development, we were able to recover the known GMP-neutrophil cell fate decision, qualify the trajectory as likely including other lineages, and pinpoint a maturation step in neutrophil development. Broadly, these results qualify the notions portrayed in Waddington's landscape, as they show that not only do cell fate decisions correspond to (one-to-many) transcriptomic state changes, but that even maturation fate changes can correspond to discontinous transitions between transcriptomic states. Interestingly, the state changes at both τ_m and τ_d appear to occur in similar directions, as revealed by the fact that the distribution of gene expression projected onto $\vec{s}_1(\tau_m)$ begins to widen at τ_d (Fig. 4D), suggesting that there may be generic transcriptomic modes that more readily enable state changes. Because the promyelocyte-to-myelocyte transition identified at τ_m is distinct from its null, we use this bifurcation to further test the our mathematical framework (Eq. (7)) and present new hypotheses for mechanistic aspects of neutrophil development.

C. Correlation analysis identifies genetic contributor in myeloid differentiation. As correlation analysis at a bifurcation correctly ranked important gene relationships in our example saddle-node bifurcation (Fig. 3G-H), we used $R_{ij}(\tau_m)$, the Pearson's correlation coefficient between genes i and j to identify pairwise gene relationships that are important for the promyelocyte-to-myelocyte transition. We verified that the distribution of R_{ij} widened at τ_m, relative to nearby τ (Fig. 6A; see Secn. S4 for PCC calculation details) as predicted for one-to-one state transitions (Eq. (7)) and observed in the saddle-node bifurcation example (Fig. 3G)).

We found that the genes in the tails of the $R_{ij}(\tau_m)$ distribution (details in Secn. S4) formed a well connected correlation network, comprising only 2 connected components. By clustering the genes based on their various cellular functions, we determined that the larger component (28/31 genes) showed features of a developmental gene regulatory network. The genes in this cluster included markers that broadly classify neutrophils, as well as markers that more specifically classify neutrophilic promyelocytes, some of which (S100a6, S100a8, S100a9, and Ngp) had previously been used to classify these cells (18), and others (Sirpa, Ccl6) which had not. In particular, the S100a9 gene was found to be highly negatively-correlated with the majority of other genes in the network, suggesting that it is pivotal for the observed bifurcation. This cluster also included genes involved in the mechanisms that drive fate changes, such as signaling and transmembrane processes necessary for cell-cell interactions, as well as housekeeping and metabolic genes responsible for maintaining basic cellular processes (Fig. 6B). Interestingly, the smaller component (3/31 genes) exclusively comprised mitochondria regulators, indicating that background, highly covariable expression relationships may be difficult to filter out exclusively from correlation matrix analysis (Fig. 6B).

We also found that the gene expression patterns within the clusters reflected their importance in the promyelocyte-to-myelocyte transition (Fig. S6). Neutrophil and myelocyte cell fate markers, as well as genes implicated in signaling, appeared to switch between low and high expression at τ_m, indicating that they have expression patterns tied to specific cell fates. Conversely, metabolic, membrane, and housekeeping genes nearly exclusively exhibit local peaks in expression at τ_m, indicating that they are transiently important for driving, or guiding, the cell-fate transition but do not pertain to a specific fate. The mitochondrial cluster, as well as some miscellaneous and membrane genes (H2d1, Anxa4) exhibited gradual increases

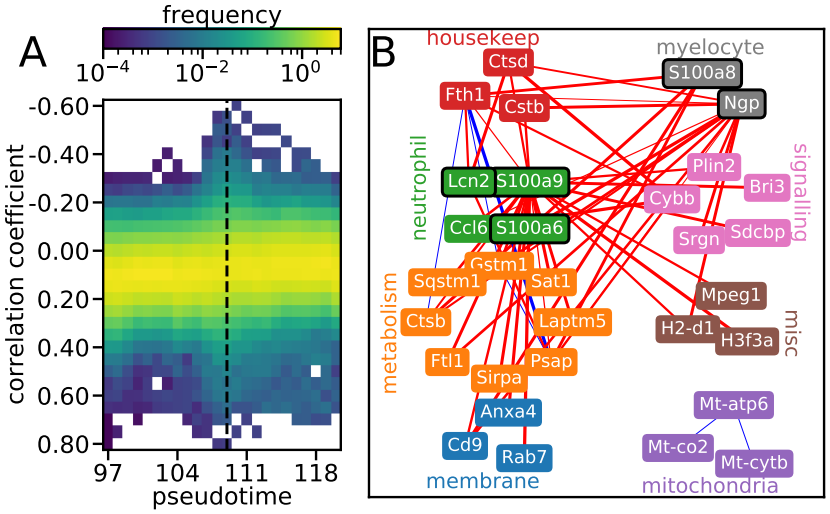

Fig. 6. Correlation analysis at bifurcation reveals critical genes for developmental transition. (A) Distribution of correlation coefficients between highly coincidentally expressed gene pairs (details in Secn. S4) as function of pseudotime. Columns integrate to 1. Dashed line indicates bifurcation bin. (B) Correlation graph for gene pairs that have $R_{ij}(\tau_m) < -0.35$ or $R_{ij}(\tau_m) > 0.65$, grouped by cellular function. Edge color indicates positive correlation (blue) or negative (red) and thickness is proportional to correlation magnitude (separate scale for positive and negative correlation.)

or decreases in expression, indicating that they may have developmental patterns that are generically tied to cell fate specification, or distance from pluripotency, but not specifically to the promyelocyte-myelocyte transition.

Lastly, since the principal covariance eigenvector $\vec{s}_1$ is equivalent to the soft mode of the Jacobian at a bifurcation (Eq. (6) and Fig. 3F), we examined how the $\vec{s}_1$ loadings of the network genes varied around τ_m to determine if they implied a deeper mechanistic interpretation (Fig. S7). We found that for nearly all genes in the developmental cluster, loadings were locally maximal at τ_m, suggesting that their relationships with other genes are most important at τ_m, even if their absolute expression is tied to a specific fate (e.g., neutrophil markers). Interestingly, some marker genes (Ngp, S100a8, S100a9) appeared to transition between high and low loading values at τ_m, indicating that they may maintain the specific cell fates of promyelocyte or myelocyte through their interactions with other genes. The loadings also appeared to reiterate the lack of importance of genes in the mitochondrial cluster for the promyelocyte-to-myelocyte fate change, as their $\vec{s}_1$ loadings were independent of pseudotime.

In summary, by applying our mathematical framework, that genetic relationships closely tied to a fate change would be highly correlated at a bifurcation (Eq. (7)), and that a bifurcation's principal eigenvector is mechanistically informative (Eq. (6)) we recovered genes that were known to be important in neutrophil development, and could place them in context of a larger, newly hypothesized gene regulatory network in data-driven manner. We verified the importance of many of these genes by examining their expression in a pseudotime window around the promyelocyte-to-myelocyte fate change, and are thus able to predict the extent to which individual genes are either transiently important for a cell fate transition, or characteristic of a specific cell fate. The eigenvector loadings were particularly useful in determining whether gene pairs are highly correlated at a bifurcation because of their importance to other genes, maintaining specific cell fates, or pluripotency. Thus, we claim that the mathematical framework presented here can be extremely elucidative of the pairwise gene relationships that underly cell fate changes.

4. Discussion

A singular challenge in understanding cellular fate transitions using transcriptomics has been dimensionality: cell fates are a low-dimensional, functional description, a valley in Waddington's landscape, while gene-expression profiles are points in a myriad-dimensional space– how can gene-expression possibly show the geometry of development? Here, we have introduced a robust, statistical, model-free approach for mapping transcriptomic changes onto fate decisions based on the first-principles of fluctuative-dissipative systems, and shown that, in fact, the high dimensionality of transcriptomes is a catalyst, not a hindrance, to understanding cell development, as gene-pair correlations reveal exactly when and how the cellular state space bifurcates. We applied this approach to a scRNA-seq transcriptomic trajectory encompassing a sequence of fate decisions from pluripotent stem cells to neutrophils, and found that two of these fate decisions corresponded well with bifurcations in gene-expression state space, exemplifying how neutrophil development could be mapped onto a low-dimension cell-fate (Waddington) landscape.

Our finding that a transcriptomic trajectory can have distinct geometric signatures, including durations during which the principal covariance eigenvalue is constant or spikes, has considerable consequences for understanding scRNA-seq datasets that capture developmental transitions. That we saw any consistent behavior in the principal covariance eigenvalue lends significant support to our initial assumption that cell fate modifiers operate at a much slower rate than transcriptomic modifiers, because if these occured on similar timescales, no statistical signature would be evident, let alone those that align well with our current developmental understanding of the system. Additionally, while previous statistical analyses of scRNA-seq data found that development appears as monotonic proliferation of cell fates (14), our focus on a single developmental trajectory enabled distinguishing multiple developmental epochs, including durations of development during which cell fates did not undergo qualitative changes, proliferated (the GMP-to-promyelocyte transition), and changed state (the promyelocyte-to-myelocyte transition). Finally, our evidence of bifurcations starkly contrasts with scRNA-seq visualizations that show gene expression varying smoothly along a developmental path, and underscores the importance of understanding noise and non-linear dependencies when using transcriptomic profiles to classify a cell's fate (15).

Our analysis of the data in Ref. (18) also yielded intriguing implications regarding the specific geometry of neutrophil development. In particular, some of the known cell fate changes in neutrophil development were not distinguishable in the covariance eigenvalue trajectory (e.g. from Common myeloid progenitors [CMP] to GMP), which indicates that these changes are less discontinuous than the GMP-to-promyelocyte or promyelocyte-to-myelocyte transitions. This could mean, for example, that even when CMPs differentiate to GMPs, the transition lacks committment, and is dependent on a sustained developmental signal, whereas once cells transition from GMP to promyelocyte, they are committed to becoming neutrophils regardless of an external signal. Furthermore, in projecting gene expression onto the principal covariance eigenvector at the promyelocyte-to-myelocyte fate transition (Fig. 4E), it became apparent that direction along which that fate change happened was well aligned with the direction of the GMP-to-promyelocyte transition. This result may be a sign of distinct, soft directions in transcriptomic space along which cell fates are most likely to change.

Aside from these geometric implications, pinpointing the bifurcation in pseudotime may

also enable efficent identification of the genes and molecular mechanisms that drive a cell fate transition. As shown, important pairwise gene relationships are explicitly highlighted in the correlation matrix at a bifurcation. While these correlations lack directionality, and therefore cannot be used to infer all aspects of mechanism, they may aid in building regulatory network models when combined with prior protein-interaction data or new experimental perturbations (42). Additionally, since the principal covariance eigenvector is equivalent to the Jacobian eigenvector at the bifurcation, and the Jacobian directly reflects gene dynamics, the eigenvector may reveal critical dynamic information or constrain an inferred global Jacobian (43).

While we focus here on scRNA-seq data, our approach is broadly applicable, and could, in principle, aid in determining other aspects of developmental bifurcations, such as the genomic structural modifications necessary for fate transitions from single-cell ATAC-seq data (44, 45). Additionally, it may be possible to incorporate our covariance analysis into other indications of pseudotime rank, such as cellular barcodes and low-dimensional distance, to constrain developmental trajectories along bifurcative paths. Our analysis was only possible because scRNA-seq experiments can now measure the expression of tens of thousands of genes in hundreds of thousands of cells, enabling accurate covariance measurements. That we found bifurcative events in this data implies that there are low dimensional, non-linear dynamical systems at play, and that sufficient biological sampling will reveal the knobs to controllably tilt developmental landscapes.

Materials and Methods

Instructions and Python code for reproducing all figures in this manuscript are available at github.com/simfreed/sc_bifurc_figs .

5. Derivation of results in Secn. 2

A. Continuous time Lyapunov equation for transcriptomic matrices. Let $\mathbf{G}$ be the steady state transcriptomic matrix at a single developmental time with n_c rows (cells) and n_g columns (genes) (Fig. 2A) and $\mathbf{F}$ be a set of differential equations describing the molecular interactions that generate $\mathbf{G}$, such that

$$\dot{\mathbf{G}} = \mathbf{F}(\mathbf{G}) \qquad [10]$$

where $\dot{\mathbf{G}}$ is the derivative of $\mathbf{G}$ with respect to time. Since all cells (columns) in $\mathbf{G}$ are at steady state at the same developmental time τ, we assume (for the purpose of contradiction) that they are all statistical replicates of the same transcriptomic state, $\vec{g}^*$, and the full matrix, $\mathbf{G}$, is therefore in the vicinity of the hyperbolic fixed point

$$E(\mathbf{G}) = \mathbf{G}^* = \vec{g}^* \otimes 1_{n_c}^T \qquad [11]$$

where 1_{n_c} is a vector of n_c ones and E denotes the expectation operator. The dynamics of $\mathbf{G}$ can be by linearized via the distance to the fixed point $\mathbf{X} = \mathbf{G} - \mathbf{G}^*$, such that

$$\dot{\mathbf{G}} = \mathbf{F}(\mathbf{G}^* + \mathbf{X}) \approx \mathbf{F}(\mathbf{G}^*) + \sum_{i=1}^{n_g} \frac{\partial \mathbf{G}}{\partial \vec{G}_i} \mathbf{X_i} = \mathbf{J}\mathbf{X} \qquad [12]$$

where $\mathbf{J} = \frac{d\dot{\mathbf{G}}}{d\mathbf{G}} \Big|_{\mathbf{G}^*}$ is the Jacobian of $\mathbf{G}$ and we have used the fact that at steady state, $\mathbf{F}(\mathbf{G}^*) = 0$.

If $\mathbf{F}$ is stochastic and Markovian, then the dynamics of $\mathbf{X}$ can be described as a discretized Ornstein-Uhlenbeck (OU) process,

$$\mathbf{X}_{t+\Delta t} = \mathbf{X}_t + \Delta t \mathbf{J}\mathbf{X}_t + \sqrt{\Delta t}\zeta_t \qquad [13]$$

where Δt is the molecular interaction timescale and $\zeta_{t;i,j}$ is sampled from $N(0, \sigma_i)$ where σ_i is the variance of gene i. The gene-gene covariance matrix can then be defined as

$$\mathbf{C} = E\left((\mathbf{G} - E(\mathbf{G}))^T(\mathbf{G} - E(\mathbf{G})\right) = E(\mathbf{X}^T\mathbf{X}) \qquad [14]$$

where the superscript T denotes transpose and we have approximated $E(\mathbf{G}) = \mathbf{G}^*$. The stationary condition for an OU process, (i.e., that $\partial\mathbf{C}/\partial t = 0$) then yields

$$\begin{aligned}\frac{\partial \mathbf{C}}{\partial t} &= \lim_{\Delta t \to 0} \frac{E(\mathbf{X}_{t+\Delta t}\mathbf{X}_{t+\Delta t}^T) - E(\mathbf{X}_t\mathbf{X}_t^T)}{\Delta t} \\ &= \mathbf{J}\mathbf{C} + \mathbf{C}\mathbf{J}^T + \mathbf{D} = 0 \end{aligned} \qquad [15]$$

where $\mathbf{D} = E(\zeta_t\zeta_t^T)$ and we have used the fact that $E(\zeta_t)$,$E(\mathbf{X}_t)$,$E(\zeta_t\mathbf{X}_t)$, and $E(\mathbf{X}^T\zeta^T)$ are all 0 (26).

B. Covariance at bifurcation. If $\mathbf{J}$ is diagonalizable, such that,

$$\mathbf{J} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} \qquad [16]$$

where Λ is a diagonal matrix of eigenvalues $(\lambda_1, \lambda_2 \ldots \lambda_{n_g})$ and $\mathbf{P}^T$ is the square matrix of eigenvectors $(\vec{p}_1, \vec{p}_2 \ldots \vec{p}_{n_g})$, then Eq. (15), often referred to as the continuous-time Lyapunov (CL) equation, can be used to qualitatively assess $\mathbf{G}$. Left-multiplying Eq. (15) by $\mathbf{P}^{-1}$ and right-multiplying by $(\mathbf{P}^\dagger)^{-1}$, yields

$$\lambda\tilde{\mathbf{C}} + \tilde{\mathbf{C}}\lambda^\dagger + \tilde{\mathbf{D}} = 0 \qquad [17]$$

where the † superscript indicates conjugate transpose, $\tilde{\mathbf{C}} = \mathbf{P}^{-1}\mathbf{C}(\mathbf{P}^\dagger)^{-1}$ and $\tilde{\mathbf{D}} = \mathbf{P}^{-1}\mathbf{D}(\mathbf{P}^\dagger)^{-1}$. Since $\mathbf{\Lambda}$ is diagonal, Eq. (17) can be rewritten elementwise,

$$\begin{aligned}\lambda_i\tilde{C}_{ij} + \bar{\lambda}_j\tilde{C}_{ij} + \tilde{D}_{ij} = 0 \\ \rightarrow \tilde{C}_{ij} = \frac{-\tilde{D}_{ij}}{\lambda_i + \bar{\lambda}_j}\end{aligned} \qquad [18]$$

which can be substituted to yield an expression for elements of the covariance

$$C_{ij} = \sum_{k=1}^{n_g}\sum_{l=1}^{n_g} P_{ik}\left(\frac{-\tilde{D}_{kl}}{\lambda_k + \bar{\lambda}_l}\right)P_{lj}. \qquad [19]$$

since $\mathbf{C} = \mathbf{P}\tilde{\mathbf{C}}\mathbf{P}^\dagger$. At a bifurcation, $\max(\mathbf{\Lambda}) = \lambda_d \rightarrow 0$, so the $k = l = d$ term in Eq. (19) becomes dominant, and

$$C_{ij} = \left(\frac{-\tilde{D}_{dd}}{2\lambda_d}\right)\vec{p}_d^{\,i}\vec{p}_d^{\,j} \qquad [20]$$

where $\vec{p}_d$ is the dth column of $\mathbf{P}$.

C. Bifurcation eigenvector equivalence. Since $\mathbf{C}$ is real and symmetric, the eigenvalue decomposition can be written as a single sum,

$$\begin{aligned} \mathbf{C} &= \mathbf{S}\mathbf{\Omega}\mathbf{S}^T \\ \rightarrow C_{ij} &= \sum_{k=1}^{n_g} \omega_i \vec{s}_k^i \vec{s}_k^j \end{aligned} \qquad [21]$$

where $\{\omega_1, \omega_2, \ldots, \omega_{n_g}\}$ are its eigenvalues, and $\{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_{n_g}\}$ are its eigenvectors, which are normalized to 1. For large n_g, $\mathbf{\Omega}$ and $\mathbf{S}$ are signficantly easier to compute than $\mathbf{C}$ as they can be obtained from the singular value decomposition of $\mathbf{X} = \mathbf{G} - E(\mathbf{G})$. For Eq. (21) to be equivalent to Eq. (20) at a bifurcation, at least one eigenvalue $\omega_i \rightarrow \infty$, which we may, without loss of generality, refer to as ω_1. If $\omega_1 \gg \omega_i$ for $i \in [2 \ldots n_g]$ then the $k = 1$ dominates the sum in Eq. (21), and by equating with Eq. (20) we obtain

$$\begin{aligned} \vec{s}_1^i \vec{s}_1^j &= \left(\frac{-\tilde{D}_{dd}}{2\lambda_d \omega_1}\right) \vec{p}_d^i \vec{p}_d^j \\ \rightarrow \frac{\vec{s}_1^j}{\vec{s}_1^i} &= \left(\frac{-\tilde{D}_{dd}}{2\lambda_d \omega_1}\right) \frac{\vec{p}_d^i \vec{p}_d^j}{(\vec{s}_1^i)^2} = \frac{\vec{p}_d^i \vec{p}_d^j}{(\vec{p}_d^i)^2} = \frac{\vec{p}_d^j}{\vec{p}_d^i} \\ \rightarrow 1 = \sum_{j=1}^{n_g} (\vec{s}_1^j)^2 &= \sum_{j=1}^{n_g} \left(\frac{\vec{s}_1^i}{\vec{p}_d^i} \vec{p}_d^j\right)^2 = \left(\frac{\vec{s}_1^i}{\vec{p}_d^i}\right)^2 \\ \rightarrow \vec{s}_1^i &= \pm \vec{p}_d^i \\ \rightarrow \vec{s}_1^j &= \vec{p}_d^j \vec{s}_1^i / \vec{p}_d^i = \pm \vec{p}_d^j \\ \rightarrow \vec{s}_1 &= \pm \vec{p}_d \end{aligned} \qquad [22]$$

where we have used the fact that $\vec{s}_1$ and $\vec{p}_d$ both normalize to 1.

6. Simulation methodology

To explore our analysis framework on a more biologically relevant gene network (Fig. 3A) we utilized a Focker-Plank simulation method. For each of the $N_c = 100$ cells (N_c chosen by examining how many cells were neccesary to accurately detect bifurcations in the neutrophil data Fig. S5A-B) the expression of gene i, $(g_i(t; m_1, m_2, k_D)$ is initalized uniformly randomly in the interval (0,4]. The expression at subsequent timesteps $(g_i(t + \Delta t))$ is sampled from a Gaussian distribution $N(\mu, \sigma)$ where

$$\mu_i = g_i(t) + \dot{g}_i(g_i(t))\Delta t \qquad [23]$$

$$\sigma_i = \begin{cases} \sqrt{\frac{\Delta t}{s}\left(\frac{m_i}{1+g_{3-i}^2} + k_D g_i\right)} & i = 1, 2 \\ \sqrt{\frac{\Delta t}{s}\left(\frac{\alpha_i g_{d(i)}^2 + 1 + \alpha_i}{(1+g_{d(i)}^2)} + k_i g_i\right)} & i = 3 \ldots n_g \end{cases} \qquad [24]$$

and bounded to be non-negative. The simulation ran for $N_t = 50,000$ timesteps, with $\Delta t = 0.01$ as that was sufficient for equilibration Fig. 3, and the last timestep of each simulation is the steady state expression $\mathbf{G}$. In the saddle-node example (Fig. 3) the remaining parameters were $k_D = 1, m_2 = 3, m_1 \in [2, 4], s = 20$ while in the pitchfork example (Fig. S4) $m_{1,2} = 1, k_D \in [0.24, 5], s = 200$.

ACKNOWLEDGMENTS. We thank Richard Carthew, Yogesh Goyal, and Karna Gowda for reviewing the manuscript and providing suggestions. This work was supported by the NSF-Simons Center

for Quantitative Biology at Northwestern University and Simons Foundation (597491, MM). SG acknowledges University of Toronto's Medicine by Design initiative, which receives funding from the Canada First Research Excellence Fund, and NSERC Discovery Grant MM is a Simons Foundation Investigator.

1. C Waddington, The strategy of the genes (1957).
2. S Huang, G Eichler, Y Bar-Yam, DE Ingber, Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. review letters* **94**, 128701 (2005).
3. F Corson, ED Siggia, Geometry, epistasis, and developmental patterning. *Proc. Natl. Acad. Sci.* **109**, 5568–5575 (2012).
4. JMW Slack, , et al., *From egg to embryo: regional specification in early development.* (Cambridge University Press), (1991).
5. E Camacho-Aguilar, A Warmflash, DA Rand, Quantifying cell transitions in c. elegans with data-fitted landscape models. *bioRxiv* (2021).
6. SA Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. *J. theoretical biology* **22**, 437–467 (1969).
7. S Huang, YP Guo, G May, T Enver, Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. biology* **305**, 695–713 (2007).
8. BR Weston, L Li, JJ Tyson, Mathematical analysis of cytokine-induced differentiation of granulocyte-monocyte progenitor cells. *Front. immunology* **9**, 2048 (2018).
9. E Marco, et al., Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci.* **111**, E5643–E5650 (2014).
10. M Setty, et al., Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. biotechnology* **34**, 637–645 (2016).
11. BK Tusi, et al., Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
12. LA Furchtgott, S Melton, V Menon, S Ramanathan, Discovering sparse transcription factor codes for cell states and state transitions during development. *Elife* **6**, e20488 (2017).
13. X Qiu, et al., Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell systems* **10**, 265–274 (2020).
14. M Nitzan, MP Brenner, Revealing lineage-related signals in single-cell gene expression using random matrix theory. *Proc. Natl. Acad. Sci.* **118** (2021).
15. N Moris, C Pina, AM Arias, Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).
16. JA Farrell, et al., Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360** (2018).
17. JA Briggs, et al., The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360** (2018).
18. C Weinreb, A Rodriguez-Fraticelli, FD Camargo, AM Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367** (2020).
19. C Trapnell, et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. biotechnology* **32**, 381 (2014).
20. K Street, et al., Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 477 (2018).
21. EM Ozbudak, M Thattai, HN Lim, BI Shraiman, A Van Oudenaarden, Multistability in the lactose utilization network of escherichia coli. *Nature* **427**, 737 (2004).
22. M Shamir, Y Bar-On, R Phillips, R Milo, Snapshot: timescales in cell biology. *Cell* **164**, 1302–1302 (2016).
23. L Chen, R Liu, ZP Liu, M Li, K Aihara, Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. reports* **2**, 1–8 (2012).
24. M Mojtahedi, et al., Cell fate decision as high-dimensional critical state transition. *PLoS biology* **14**, e2000640 (2016).
25. JE Ferrell Jr, Bistability, bifurcations, and waddington's epigenetic landscape. *Curr. biology* **22**, R458–R466 (2012).
26. M Oku, K Aihara, On the covariance matrix of the stationary distribution of a noisy dynamical system. *Nonlinear Theory Its Appl. IEICE* **9**, 166–184 (2018).
27. EaD Amir, et al., visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. biotechnology* **31**, 545–552 (2013).
28. C Weinreb, S Wolock, AM Klein, Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).
29. MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 550 (2014).
30. N Rosenfeld, JW Young, U Alon, PS Swain, MB Elowitz, Gene regulation at the single-cell level. *science* **307**, 1962–1965 (2005).
31. A Raj, A Van Oudenaarden, Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
32. T Gregor, DW Tank, EF Wieschaus, W Bialek, Probing the limits to positional information. *Cell* **130**, 153–164 (2007).
33. G Tkačik, T Gregor, W Bialek, The role of input noise in transcriptional regulation. *PloS one* **3**, e2774 (2008).
34. NG Van Kampen, *Stochastic processes in physics and chemistry.* (Elsevier) Vol. 1, (1992).
35. TS Gardner, CR Cantor, JJ Collins, Construction of a genetic toggle switch in escherichia coli. *Nature* **403**, 339–342 (2000).
36. CP Gallivan, H Ren, EL Read, Analysis of single-cell gene pair coexpression landscapes by stochastic kinetic modeling reveals gene-pair interactions in development. *Front. genetics* **10**, 1387 (2020).
37. B Novak, JJ Tyson, Numerical analysis of a comprehensive model of m-phase control in xenopus oocyte extracts and intact embryos. *J. cell science* **106**, 1153–1168 (1993).
38. AM Klein, et al., Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
39. LW Plasschaert, et al., A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
40. N Borregaard, Neutrophils, from marrow to microbes. *Immunity* **33**, 657–670 (2010).
41. R Ostuni, G Natoli, MA Cassatella, N Tamassia, Epigenetic regulation of neutrophil development and function. *Semin. Immunol.* **28**, 83–93 (2016) SI: Neutrophils in innate and adaptive immunity.
42. X Sun, B Länger, W Weckwerth, Challenges of inversely estimating jacobian from metabolomics data. *Front. bioengineering biotechnology* **3**, 188 (2015).
43. T Nägele, et al., Solving the differential biochemical jacobian from metabolomics covariance data. *PloS one* **9**, e92299 (2014).
44. JD Buenrostro, PG Giresi, LC Zaba, HY Chang, WJ Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat. methods* **10**, 1213 (2013).
45. G Jia, et al., Single cell rna-seq and atac-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. communications* **9**, 1–17 (2018).
46. CR Harris, et al., Array programming with NumPy. *Nature* **585**, 357–362 (2020).
47. M Jacomy, T Venturini, S Heymann, M Bastian, Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one* **9**, e98679 (2014).

7. Supporting Information

S1. Bifurcations possibilities from two mutually inhibiting genes. At steady state, Eq. (8) satisfies the quintic polynomial

$$g_1 = \frac{m_1/k_D}{\left(\frac{m_2/k_D}{g_1^2+1}\right)^2 + 1} \qquad \text{[S1]}$$

which, depending on the parameter values, can have one real solution that is an attractor (e.g., if $m_{1,2} = 1$ and $k_D = 1$) or three real solutions, two attractors (nodes) and one repellor (saddle) (e.g., $m_{1,2} = 1, k_D = 1/3$). By examining the null clines,

$$g_1(g_2) = \frac{m_1/k_D}{g_2^2 + 1} \qquad \text{[S2]}$$

$$g_1(g_2) = \sqrt{\frac{m_2}{k_D g_2} - 1} \qquad \text{[S3]}$$

it can be deduced that varying m_1, while fixing τ and m_2 can yield a saddle-node bifurcation, as Eq. (S2) moves vertically while Eq. (S3) does not, allowing for either node to merge with the saddle (Fig. S1A). Conversely, varying k_D, while fixing $m_{1,2}$ and m_2, can yield a pitchfork bifurcation, as both null clines move, such that above the bifurcation value, all three real solutions remain (Fig. S1B). Solving Eq. (S1) computationally via the Python function numpy.roots and plotting the real solutions (Fig. S1C-D) yields the bifurcations used in Fig. 3 and Fig. S4 (46).

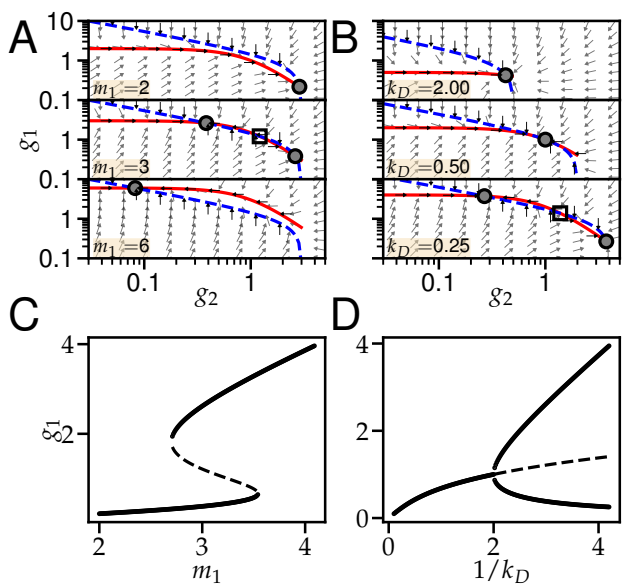

Fig. S1. Analysis of Eq. (8). (A-B) Phase planes for different parameter sets yield a saddle-node bifurcation (A) or pitchfork bifurcation (B). Solid red line is given by Eq. (S2) while dashed blue line is given by Eq. (S3). Open squares are saddles while closed circles are nodes. Arrow angles are given by $tan^{-1}(\dot{g}_1/\dot{g}_2)$ and are uniform length. (C-D) Solutions to Eq. (S1) while varying m_1 (C) or k_D (D). Solid lines are nodes and dashed lines are saddles. In (A,C) $k_D = 1, m_2 = 3$ and in (B,D) $m_{1,2} = 1$.

S2. Resampling principal eigenvalue. Given the transcriptomic matrix $\mathbf{G} = \left\{\vec{g_1}^T, \vec{g_2}^T, \ldots \vec{g_{n_g}}^T\right\}$, where $\vec{g}_i = \{G_{1,i}, G_{2,i}, \ldots, G_{n_c,i}\}$ and $G_{i,j}$ is the expression of the jth gene in the ith cell, we generate a

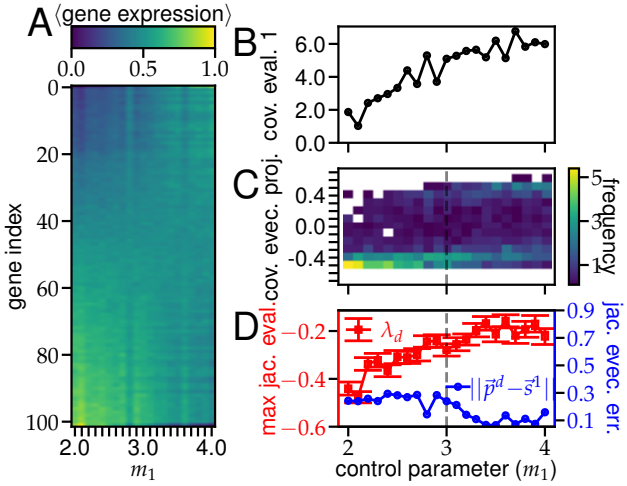

Fig. S2. Unequilibrated saddle-node bifurcation analysis. (A)-(D) Corresponding plots for Fig. 3 (C)-(F), respectively, when the simulation is runs for a $N_t = 500$ iterations rather than $N_t = 50000$.

null sample $\mathbf{G}^{\text{null}}$ by drawing each of its entries $G^{\text{null}}_{i,j}$ randomly, with replacement, from $\vec{g}_i$. In Fig. S5,3,4, we compute the principal covariance eigenvalue ω_1^{null} for each of $n_s = 20$ samples, and compare this null distribution against the principal covariance eigenvalue of the data ω_1. This resampling technique has little impact on ω_1 for unimodal distributions as the scale of ω_1 is still determined by the system's noise (Fig. S3 left and right), but significantly decreases ω_1 for multimodal distributions (Fig. S3 center) since the structure of the multimodality is scrambled; thus we found it was an effective method for determining if a spike in ω_1 is due to multimodality or increased noise.

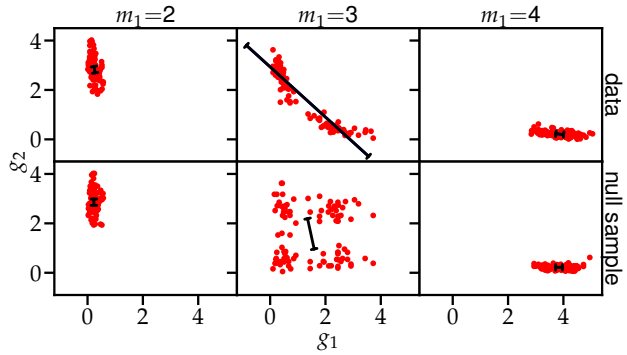

Fig. S3. Effect of resampling on principal eigenvector. Red dots are cells for the data (top) and a single marginal resampling (bottom) before the bifurcation (left), at the bifurcation (center) and after the bifurcation (right). Direction of the black lines corresponds to the principal eigenvector and length corresponds to the principal eigenvalue.

S3. Algorithm for generating the pseudotime labels in Weinreb et al. SPRING (x-y) positions, cell type annotations, and pseudotime ranks for the data presented in Fig. 4A-B were downloaded from https://github.com/AllonKleinLab/paper-data/tree/master/Lineage_tracing_on_transcriptional_landscapes_links_state_to_fate_during_differentiation. The algorithms to generate these values are described in detail in Ref. (18) (Supplementary Materials) and recapitulated here for completeness. Given the full *in-vitro*

hematopoiesis transcriptomic matrix (all cells and all genes), the SPRING positions in Fig. 4A plot were generated using the following procedure.

1. A filtered transcriptomic matrix was generated which did not include genes that
   (a) had low variability as determined via the filter_genes function with parameters (85,3,3) from https://github.com/AllonKleinLab/SPRING_dev/blob/master/data_prep/spring_helper.py (28).
   (b) correlated highly (R>0.1) across all cells with any of the following cell cycle genes: Ube2c, Hmgb2, Hmgn2, Tuba1b,Ccnb1, Tubb5, Top2a, and Tubb4b.
2. The top 50 principal components (PC) of the filtered transcriptomic matrix were computed.
3. 40,000 of the cells were selected randomly, and a k-nearest-neighbors (KNN) graph between those cells was constructed using the top 50 PC of the filtered transcriptomic matrix and k=4.
4. X-Y positions of these 40,000 cells were generated using the ForceAtlas2 algorithm with 500 steps (47).
5. Positions for each of the remaining 90,887 cells were computed as the average position of their 40 nearest neighbors (in the 50-PC space) among the initial 40,000 cells.

Cells were annotated with their cell types (cluster annotation in Fig. 4A) based on their position in the SPRING plot and their expression (terminal cell fates) or lack of expression (pluripotent) of pre-selected marker genes. Specifically the marker genes used to determine if cells were neutrophils were S100a9, Itgb2l, Elane, Fcnb, Mpo, Prtn3, S100a6, S100a8, Lcn2, and Lrg1.

Neutrophil pseudotime rank was then determined by smoothly interpolating between cells in the pluripotent and neutrophil clusters. The interpolation method used throughout this procedure is an iterative, diffusive process defined as

$$S_0(\mathbf{X}, b, i, k) = \vec{x}_i \qquad \text{[S4]}$$
$$S_n(\mathbf{X}, b, i, k) = bS_{n-1}(\mathbf{X}, b, i, k) + \frac{1-b}{k} \sum_{j \in K_k(i)} S_{n-1}(\mathbf{X}, b, j, k)$$

where $\vec{x}_i$ is a vector quantity defined for cell i, $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{n_c}\}$ is the matrix of this quantity for all cells, $K_k(i)$ are the cell indices of the k nearest neighbors of cell i, $n > 0$ is the number of iterations, and b is the neighbor weight (low b and high n both yield high diffusion) (11). The pseudotime ranking procedure is:

1. Cells are identified to be part of the neutrophil trajectory
   (a) Let $\vec{t}_i$ be an indicator vector for the cell type of i; i.e. $t_{ij} = 1$ if cell i is type j and 0 otherwise. Let $\mathbf{T} = \left\{\vec{t}_1, \vec{t}_2, \ldots, \vec{t}_{n_c}\right\}$ be the corresponding matrix for all cells.
   (b) Let $\mathbf{K}_{100}$ be the k-nearest-neighbor graph between cells for $k = 100$ using the top 50 PC.
   (c) Let $\hat{t}_i = S_{250}(\mathbf{T}, 0.1, i, 100)$ be the smooth cell type indicator.
   (d) Let $z_i = \sum_j a_j \hat{t}_{ij}$ be the weighted average cell type $\hat{t}_i$ where the weights for each cell type (j) are

$$a_j = \begin{cases} 0.1 & \text{if neutrophil or pluripotent} \\ -2 & \text{if megakaryocyte} \\ -1 & \text{otherwise} \end{cases} \qquad \text{[S5]}$$

(e) Let $\vec{c}_i$ be a neutrophil trajectory indicator such that $\vec{c}_i = \{1\}$ if $z_i > Q_{0.6}(z)$ and $\{0\}$ otherwise, where $Q_{0.6}(z)$ is the 60th quantile of z. Let $\mathbf{C} = \{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{n_c}\}$.

(f) Let $\hat{c}_i = S_{50}(\mathbf{C}, 0.1, i, 100)$ be the smoothed neutrophil trajectory indicator.

(g) Cells were considered part of the neutrophil trajectory if $\hat{c}_i > Q_{0.6}(\hat{c})$ where $Q_{0.6}(\hat{c})$ is the 60th percentile of $\hat{c}$.

2. The $61,310$ cells identified as part of the neutrophil trajectory are sorted

(a) Let $\vec{p}_i = \{1\}$ if a cell in the trajectory is pluripotent and 0 otherwise; i.e., it is an indicator for pluripotency. $\mathbf{P} = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{n_c}\}$ is the corresponding matrix for all cells in the trajectory.

(b) Let $\hat{p}_i = S_{300}(\mathbf{P}, 0.1, i, 100)$ be the smoothed pluripotency indicator.

(c) The pseudotime of cell i is the rank (largest to smallest) of $\hat{p}_i$ among all $\hat{p}$.

S4. Computing correlation coefficients. A significant challenge in accurately measuring $R_{ij}(\tau)$, the Pearson's correlation coefficient between genes i and j at pseudotime τ (Fig. 6) is that the low read-depth of scRNA-seq experiments results in many genes having 0 or very few reads in many cells, which leads to systematically high correlations (e.g., if the read counts for both gene i and j are almost all 0, but there is one cell in which both are highly expressed). To mitigate this issue, we: (a) used large bins of 1000 or more cells for all experimental calculations (even though smaller bins are sufficient to detect bifurcations, as shown in Fig. S5A-B), (b) only use cells that have non-zero read count of both genes i and j (effectively viewing 0 reads as a lack of information, rather than a measurement) and (c) only calculate correlations between gene pairs that had 400 or more non-zero read count cells to ensure that correlations are not spuriously increased by low read counts. While this filter yields a significant reduction in correlations measured (an average of $5.5 * 10^4$ per pseudotime window, out of a possible $3.2 * 10^8$), it is independent of pseudotime, and therefore, with respect to the widening of the $R_{ij}(\tau)$ distribution at τ_m (Fig. 6A), can be thought of as a random statistical sample of correlation coefficients for which there is high confidence. Genes used to form the network in Fig. 6B had $R_{ij}(\tau_m) > 0.65$, or $R_{ij}(\tau_m) < -0.35$ as determined via inspection of Fig. 6A.

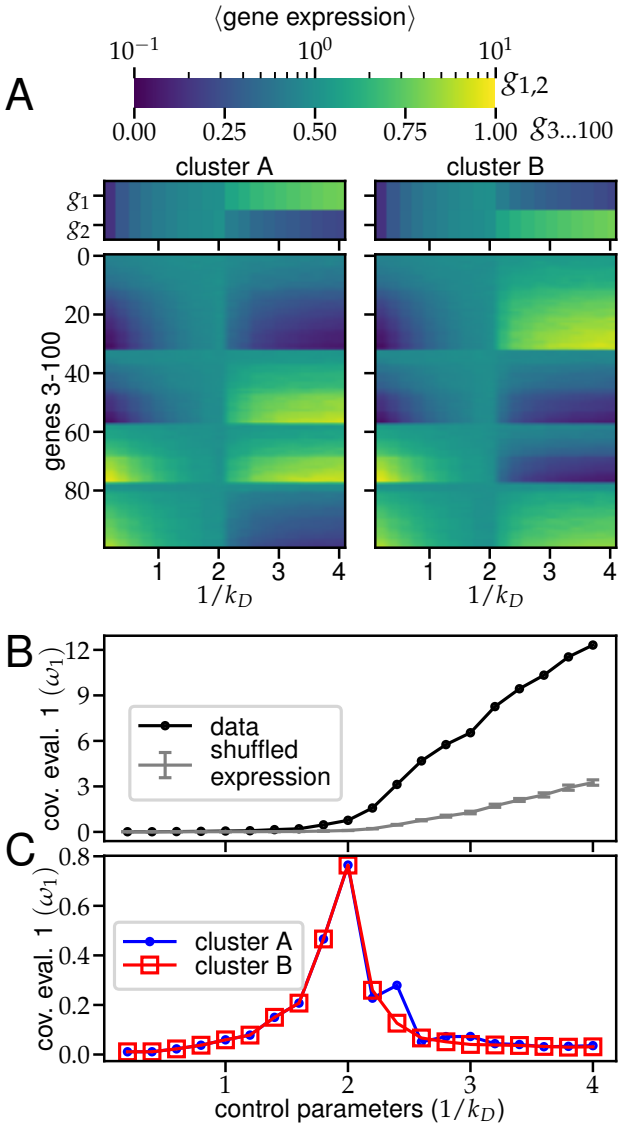

Fig. S4. Pitchfork bifurcation analysis. (A) Gene expression as function of the bifurcation variable τ, separated by cluster. (B) ω_1 and null for unclustered data. (C) ω_1 for clusters.

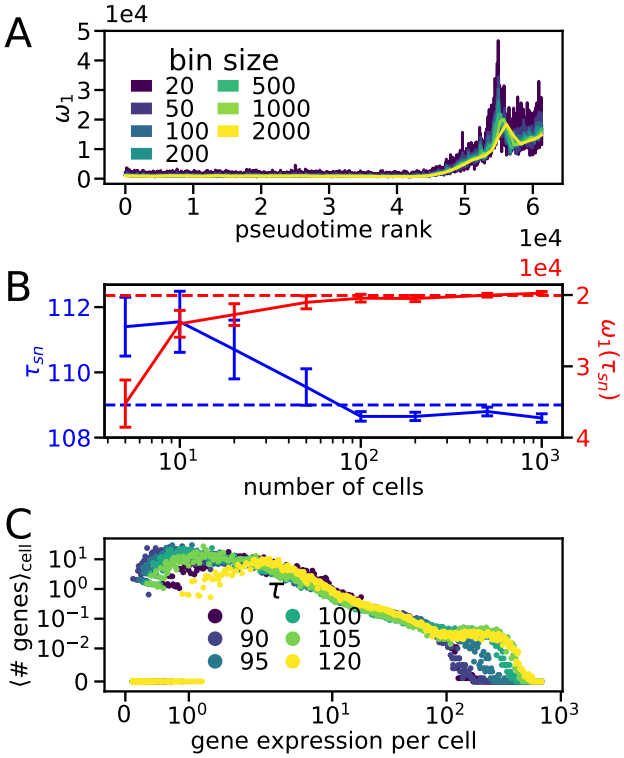

Fig. S5. Distributional properties of the neutrophil pseudotime trajectory. (A) Effect of varying the bin size on principal covariance eigenvalue. (B) Effect of undersampling a bin of 1000 cells on the detected saddle-node bifurcation time and magnitude. (C) Distribution of gene expression during the trajectory.

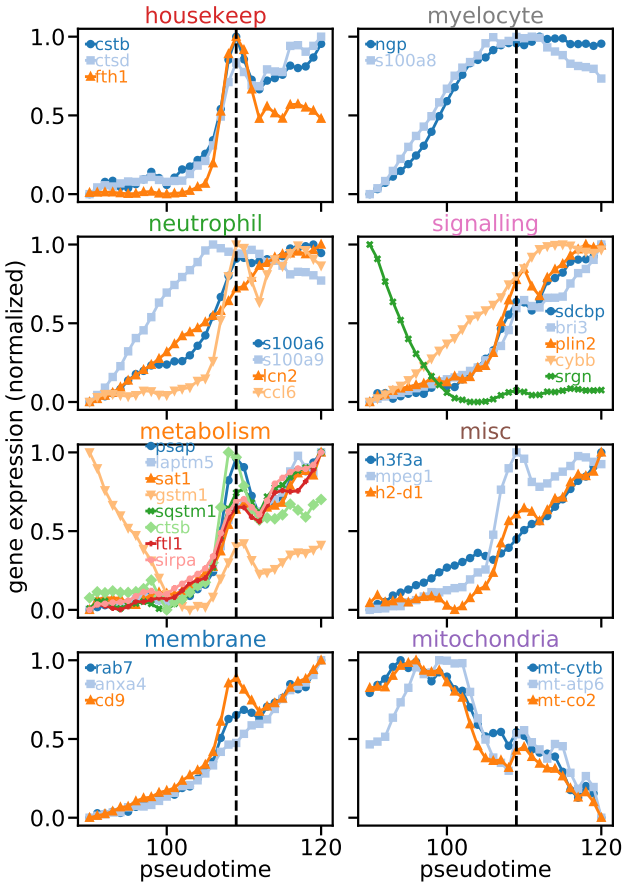

Fig. S6. Gene expression of gene groups in Fig. 6, averaged over all cells in bin that have non-zero expression for that gene as a function of pseudotime, min-max normalized in the pseudotime window shown.

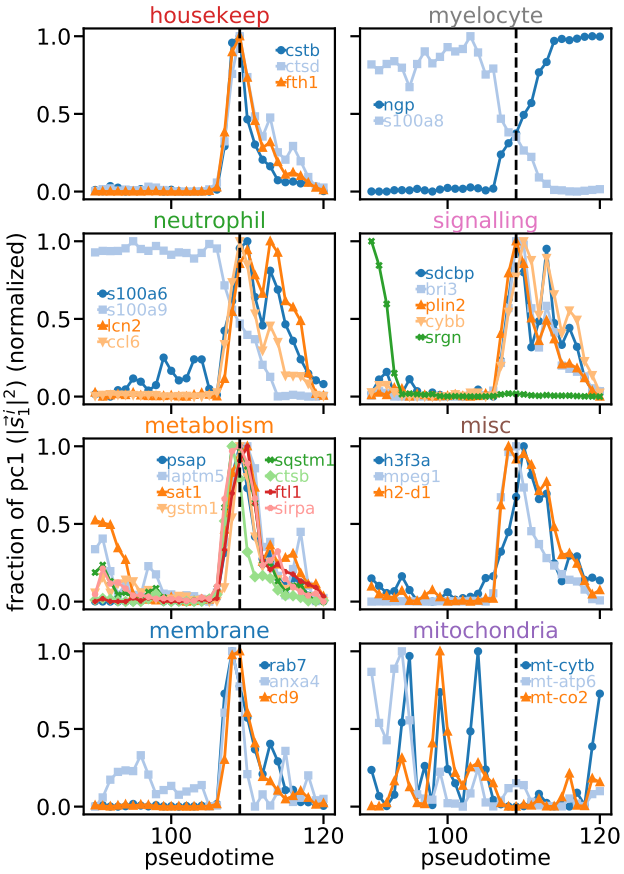

Fig. S7. Fractional weight of first covariance eigenvector for gene groups in Fig. 6, as a function of pseudotime, min-max normalized in the pseudotime window shown.