A dynamical systems treatment of transcriptomic trajectories in 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. It remains unclear, however, whether the dynamical systems framework can go beyond a word-based caricature of the system and provide sharp quantitative insights that further our understanding of differentiation. Single-cell RNA sequencing (scRNA-seq), which measures the distributions of possible transcriptional states in large populations of differentiating cells, provides an alternate view, in which development is marked by the individual concentration variations of a myriad of genes. Here, starting from formal mathematical derivations, we challenge these transcriptomic trajectories to a rigorous statistical evaluation of whether they display signatures consistent with bifurcations. After pinpointing bifurcations along transcriptomic trajectories of the neutrophil branch of hematopoeitic differentiation we are able to further leverage the primitive features of a linear instability to identify the single-direction in gene expression space along which the bifurcation unfolds and identify possible gene contributors. This scheme identifies transcription factors long viewed to play a crucial role in the process of neutrophil differentiation in addition to identifying a host of other novel genetic players. Most broadly speaking, we provide evidence that, though very high-dimensional, a bifurcating dynamical systems formalism might be appropriate for the process of cellular differentiation and that it can be leveraged to provide insights. Ambitiously, our work attempts to take a step beyond data-analysis and towards the construction of falsifiable mathematical models that describe the dynamics of the entire transcriptome.

1. Introduction

During development and tissue regeneration, it is envisioned that 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 they depend on significantly fewer 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 have been 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 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 individual 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

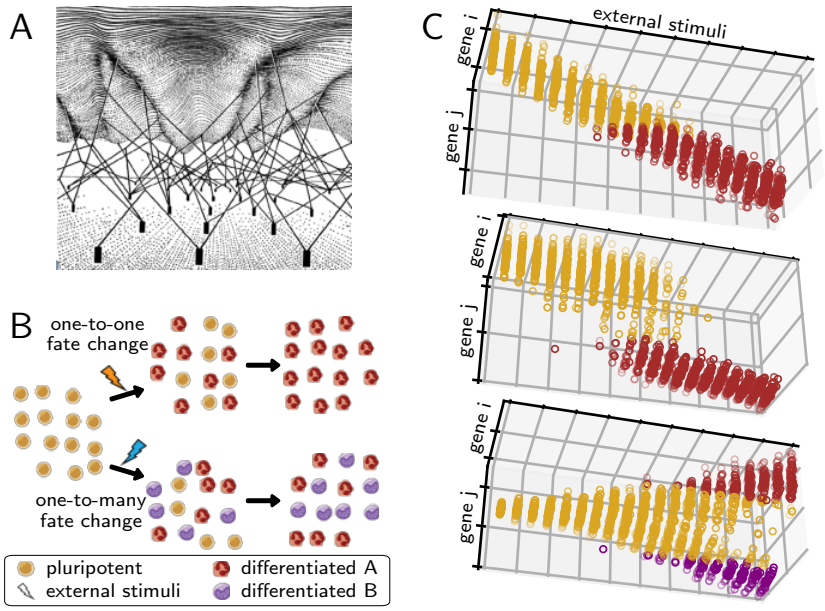

Fig. 1. Cell type differentiation as a dynamical process. (A) Waddington's landscape of cell fate commitment in which cell fates are represented as valleys, committment barriers as hills, and gene activity as pegs underneath that control the heights of hills and valleys. Adapted from Refs. (1, 17). (B) 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).

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). While these studies generally characterize cell fate decisions as bifurcations of an underlying developmental landscape, other studies model cell fate decisions as noise-induced state transitions on non-bifurcating developmental landscapes, by parameterizing cells' dynamics and stochasticity, and use these models to infer lineage relationships and state transition probabilities (14–16).

In the face of these contrasting views, it remains unclear when, during development, transcriptomes undergo bifurcations. Even more abstractly, the strict geometric structure of bifurcations become less constraining as the dimensionality of the system grows, perhaps leading us to anticipate that the vast dimensionality of the transcriptome precludes the possibility of formal bifurcations being present, let alone found. However, the audacious hope of the perspective put forward in this study is that the same feature that permits $\sim 10^{23}$ atoms present in a straw to collude together to buckle at a critical load, the interactions amongst them, is present at a sufficient level within the transcriptome to display such collective phenomena. The central question guiding this study is therefore what statistical signatures must be present to confirm the presence of a dynamical system bifurcating and, once found, what additional insights does a bifurcation imply. Our goal is not to put forward a data-analysis tool, but instead to formally investigate the extent to which a bifurcating dynamical systems view of the transcriptome is an accurate and predictive mathematical description of the observed dynamics.

How is it possible to address these more abstract questions if single-cell RNA sequencing doesn't usually involve a high-temporal-resolution measurement? To circumvent this shortcoming of the measurement, and leveraging the inherent cellular variability in a population of differentiating cells, multiple groups have formulated pseudotime algorithms, parametric curve fitting tools that use the transcriptome similarity of cells to determine their relative ordering in developmental time, and thereby infer transcriptomic trajectories that span developmental decisions (10, 18, 19). The central assumption, which finds increasing empirical support (20–23), is that the cell-level variability in a differentiating population is primarily along the path of cell-fate specification. Thus, stitching together the transcriptomes of a single population of cells assayed at the same lab time permits a dynamical measurement.

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, or bifurcation parameter(s) steadily varies, permitting a dynamical systems style investigation of possible bifurcations in the collective transcriptomic state. The dynamical systems view seems suitable due to the qualitative 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 (24), 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. This separation of timescales enables the possibility for cells to have steady transcriptomic states, and is therefore necessary for there to be bifurcations in development, since a bifurcation is a qualitative augmentation of the steady state solutions, or branches of a dynamical system.

Here, we layout and demonstrate a statistical formalism for detecting and interrogating bifurcations in developmental fate transitions directly from transcriptomic pseudotime trajectories. Contrasting previous studies (7–10), we do not assume any specific mathematical form for the underlying genetic interactions, nor do we assume the shape, or even existence of an underlying cell-fate landscape (Fig. 1B) (25, 26) since it is not our goal to discern a specific model. Instead we rigorously query whether the necessary statistical signatures of a bifurcations are present in a developmental timecourse. We build on and compare to similar styles of approach, which have been used to detect signatures and molecular mechanisms of disease (25, 27), analyze differentiation processes in temporal and pseudotemporal gene expression data (26, 28), and characterize reversibility in saddle-node bifurcations (29). 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 (30); and a cell fate change that is due to a one-to many-state transition like those 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

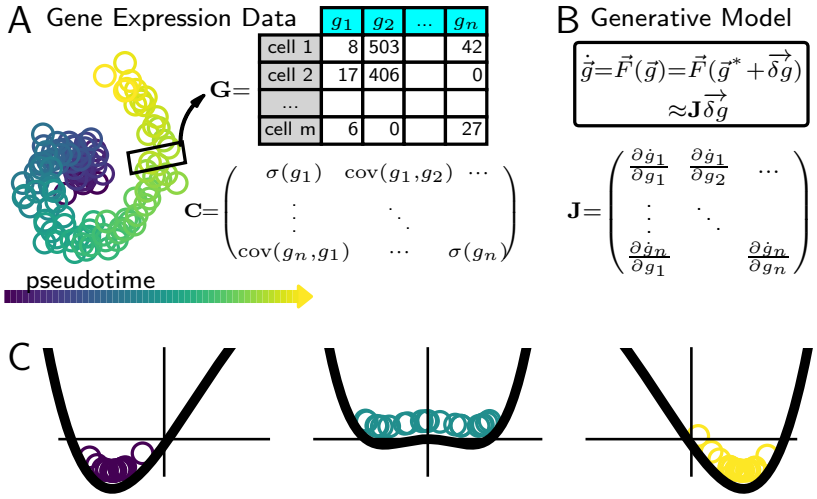

Fig. 2. Relating the Jacobian to the Covariance at a bifurcation. (A) Schematic of a single-cell RNA-seq dataset arranged by each cell's developmental (pseudo-) time. (Left) Visualization of dataset in two collective gene-expression dimensions. (Right) Gene expression matrix at the pseudotime indicated by the 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). Dark lines show the underlying landscapes $\int \dot{x}dx$.

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.

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

In this section we outline the theoretical foundations that enables our attempt to formally assess the presence and essential properties of bifurcations in high-dimensional transcriptomic dynamics. Summarizing our key insights, we layout a conceptual framework tailored to the analysis of transcriptomic data that, despite the absence of a mathematical model for the dynamics, can reveal and investigate the system's most salient dynamical features, its bifurcations, from transcriptomic trajectories alone (26, 31).

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 (32), and k-nearest neighbor maps (14), have been used to visualize this data and determine pairwise distances between cells based on their transcriptomes (for example, the methods described in Secn. S4.1). 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, 18, 19) (Secn. S4.1). 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, $\mathbf{G}(\tau+\Delta\tau)$ (33). While such an analysis may reveal which genes exhibit dynamic signals, it will not reveal if the transcriptome's dynamics have undergone a bifurcation, 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 (34–37),(15, 38) and (2) occur at significantly faster timescales (seconds to minutes) than the timescales over which transitions in cellular fates are observed (hours to days) (4, 24). A consequence of these assumptions (details in Secn. 5A) 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 (39),

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

where $\mathbf{D}$ is the expected noise amplitude for individual genes and their interactions (derivation in Secn. 5A) (31). 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 ($a < 0$), near ($a = 0$), 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, a dynamical system's local geometry can be obtained from diagonalizing $\mathbf{J}$, such that,

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

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 $\{\vec{p}_1, \vec{p}_2 \dots \vec{p}_{n_g}\}$. $\lambda_i < 0 \; \forall \; i$ indicate a single, stable transcriptomic state, in the same way that the highly convex curvature in our toy model (Fig. 2C) indicate a single fixed point. Conversely, $\lambda_d = \max(\Lambda) \to 0$ indicates a bifurcation in the $\vec{p}_d$ direction, just as the landscape of the toy model flattens out along the horizontal axis at $a = 0$ to enable the fixed point exchange. $\lambda_d \to 0$ 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,just as the direction along which the variance expands in the toy model (x) is the same as the direction along which the landscape flattens, it can be shown, by equating Eq. (3) to Eq. (4) (where the $k = 1$ term will dominate; see Secn. 5C), that

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

meaning, $\vec{s}_1$ is the direction of the bifurcation! 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 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.

The expansion of correlation coefficients between crucial system features at a bifurcation (Eq. (7)) has previously been used to infer the networks that drive critical transitions (sometimes referred to as dynamic network biomarkers, or DNBs) (25, 27–29). More recently these bifurcation analysis methods have been used to identify critical developmental times from transcriptomic time-series (26, 40) and pseudotime trajectories (28). A notable difference in our analysis is our focus on the covariance eigen-decomposition, which directly reflects the dimensionality reduction that occurs at a bifurcation, rather than specific entries of the correlation matrix, which may require additional analyses to identify (see Secn. S1).

We first use theoretical models of noisy, high dimensional gene networks to concretely demonstrate how the approach outlined above (outlined in Secn. 7) can be leveraged to detect and assess bifurcations of an underlying dynamical system from observations of state-variables alone, bolstering previous applications. Following this, we emphasize the power of this approach by directly applying it to scRNA-seq data for the neutrophil lineage in the hematopoietic system.

3. Results

A. Covariance analysis recovers salient features of a high-dimensional *in silico* gene regulatory network. 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 (41). 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 (42) in a diversity of biological systems, such as the lac-operon (43) and cell-cycle control (44).

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

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

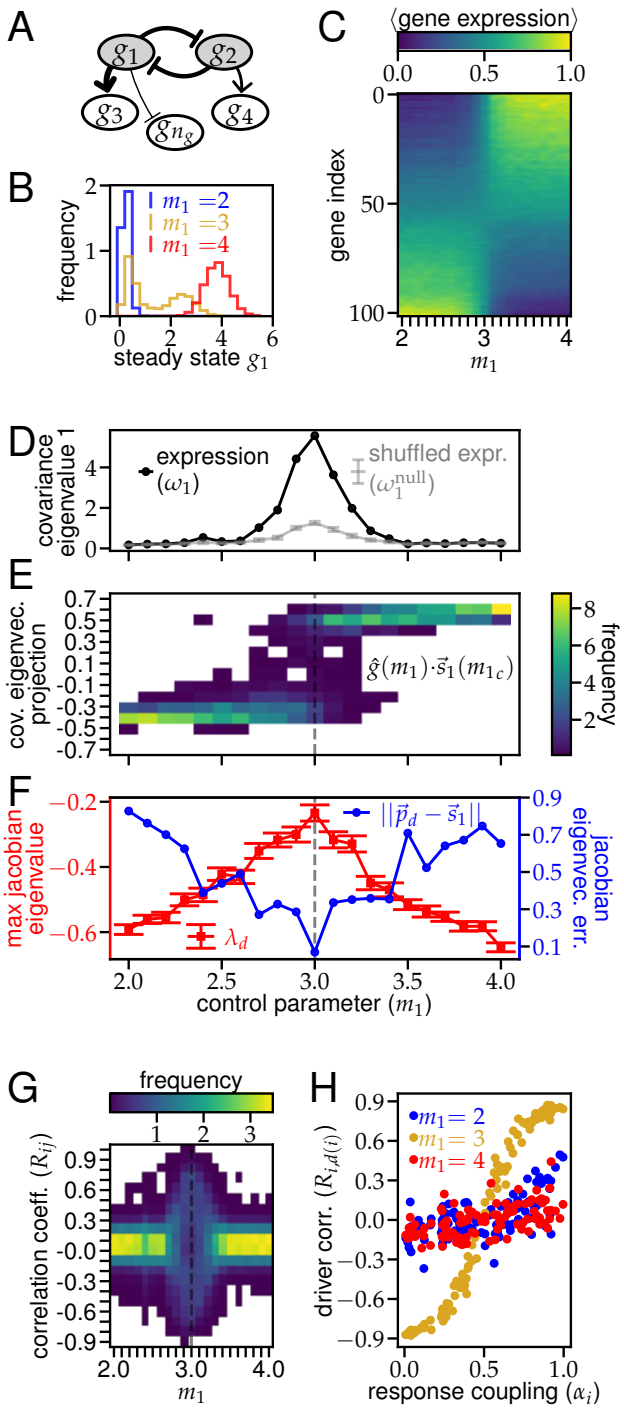

Fig. 3. Analysis of a gene regulatory network around a saddle-node bifurcation. (A) Schematic of the GRN for 5 of the 102 genes. Undisplayed nodes have a unidirectional arrow stemming from either g_1 or g_2. (B) Distributions of g_1 at steady state (see Secn. 6) for three values of m_1. (C)-(G) GRN observables 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 shifted to have 0 min. (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. (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. Each column in (E) and (G) integrates to 1.

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 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 system that 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 explicitly destroys any correlation between genes of $\mathbf{G}(m_1)$ (details in Secn. S3). 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, negative, 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 from below 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 , 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, the 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, owing to the high-dimensionality of the system, in the vicinity of a bifurcation, a form of dimensionality reduction emerges which enables gleaning geometric characteristics directly from the dynamics. 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 corresponds to 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 the theoretical infrastructure outlined above to analyze a high-dimensional and noisy dynamical system undergoing a variety of bifurcations, by uncovering its crucial aspects, including its location, direction in gene space, and influential genetic relationships. These calculations are also computationally 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).

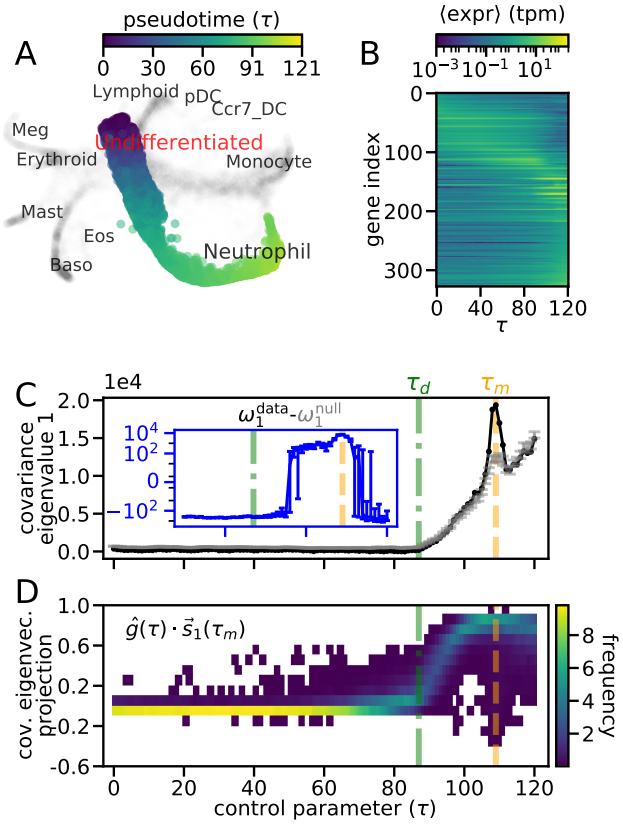

Fig. 4. Covariance analysis of temporal scRNA-seq data shows signatures of developmental bifurcations. (A) SPRING visualization for each cell (point) in an *in-vitro* scRNA-seq experiment of mouse hematopoeitic stem cell differentiation (23). Cells in the pseudotime trajectory analyzed are colored accordingly (blue to yellow) while others are gray. SPRING coordinates, cluster labels, pseudotimes and a similar visualization were first reported in Ref. (23). (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. S3), shifted to have 0 min. 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.

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 (23).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 (45, 46) on days 2, 4, and 6. The resultant transcriptomic matrix (25,289 genes in 130,887 individual cells) was filtered to only include highly variable, non-cell cycle genes, and visualized in $2D$ via the SPRING method (14) (Fig. 4(A)), in which the (x, y) coordinates of each cell are determined by optimally placing each cell closest to its 4 nearest neighbors in the space of the top 50 gene-wise principal components (PC) (14). 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 (23). 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 neutrophils (details in Secn. S4.1). 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. S4.1). 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. S5 and Fig. S5) and applied our covariance analysis to the full, row (cell)-normalized 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. S3) 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

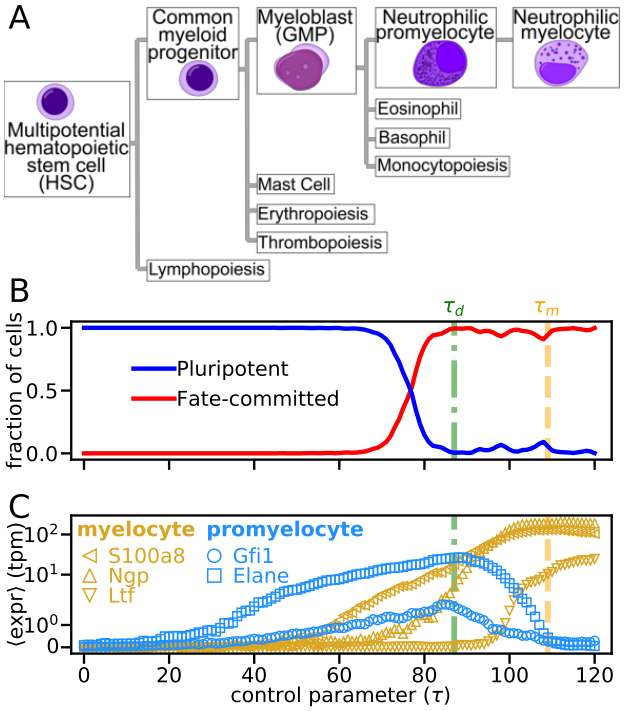

Fig. 5. Detected bifurcations correspond to biologically characterized developmental transitions. (A) Schematic of neutrophil development, beginning from hematopoietic stem cells and ending at the neutrophil myelocyte, a committed neutrophil progenitor. Lines indicate naturally occuring progeny, other 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. (23). (C) Average expression of promyelocyte (blue) and myelocyte (gold) marker genes (23). SEM error bars are smaller than symbols.

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. To verify that this temporal trend was not limited to the diffusion-based pseudotime algorithm used in Ref. (14), we recalculated pseudotime via Slingshot (19), a generic spline-based pseudotime inference tool, and found the same rise and peak of ω_1 (see Secn. S4.2 and Fig. S6A-B).

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 (47), 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 for neutrophil development (Fig. 5A). We found (Fig. 5B) that τ_d, the moment ω_1 begins to increase, corresponded well with the moment in pseudotime that cells switched between the endpoints of this tree, from not expressing any terminal-fate marker genes to primarily expressing neutrophil cell-fate markers (23). At a more granular developmental level, our covariance analysis highlights specific transitions between intermediate neutrophil progenitor states (Fig. 5A). These transitions include: (a) one-to-many cell fate changes (i.e., decisions) like the transition between a granulocyte monocyte progenitor (GMP, or myeloblast), and any of its four terminal fates: neutrophil, monocyte, eosonophil, and basophil, as well as (b) one-to-one cell fate changes (i.e., maturation) like the transition between promyelocyte and myelocyte (23, 48, 49). τ_d corresponds well with the point in pseudotime that promeylocyte marker genes are maximal (Fig. 5C and Fig. S6C), suggesting a connection between τ_d and the one-to-many change from GMP to promyelocyte. This can be understood in light of other one-to-many state transitions, like a pitchfork bifurcation (Fig. S4A), where ω_1 increases steadily if different branches are left unclustered (Fig. S4B). Thus, the increase of ω_1 at τ_d suggests that while cells in the neutrophil trajectory were expected to only include the neutrophil lineage branch of GMP, they may in fact include other GMP lineages, such as eosonophils or basophils. Conversely, τ_m correponds well with the point in pseudotime that myelocyte marker genes are maximal, and promyelocyte genes have reduced expression (Fig. 5C and Fig. S6C), suggesting that τ_m denotes the transition point between these two cell fates. This marker-gene analysis matches our geometric interpretation, that an ω_1 spike indicates a one-to-one state transition, as the promyelocyte-to-myelocyte cell fate transition is a maturation step of committed neutrophil progenitors, and not a decision between multiple 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 in mice.

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), and has been shown to reveal gene-pair relationships that are critical for state transitions in temporal RNA-seq data (26, 40), 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. S5 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. S5) formed a well connected correlation network, comprising only 2 connected components (Fig. 6B). 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 (23), 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. 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). Thus, bifurcation detection enables inferring gene regulatory networks for a specific moment in development from a physically principled, dynamical systems approach.

We also found that the gene expression patterns within the clusters reflected their importance in the promyelocyte-to-myelocyte transition (Fig. S7). 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 or decreases in expression, indicating that they may have developmental patterns that are

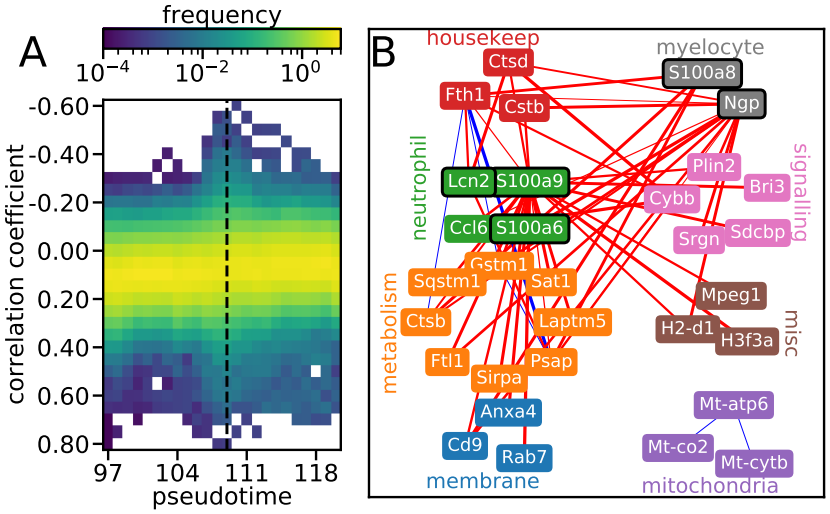

Fig. 6. Correlation analysis at bifurcation reveals critical genes for developmental transition. (A) Distribution of correlation coefficients between highly coincidentally expressed gene pairs (details in Secn. S5) as function of pseudotime. Columns integrate to 1. Dashed line indicates bifurcation bin. (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.)

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. S8). 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 for mice, 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 not only helps elucidate the dynamics of the transcriptome at the scale of a cell's fate, but even provides clues at the scale of molecular interactions for the underlying genetic relationships that enable cell-scale changes.

4. Discussion

It is now both appealing and common place to loosely graft dynamical systems language to the dynamics of cellular differentiation. In particular, it is appealing to allude to bifurcations of an underlying landscapes as driving qualitative transitions in accessible states. But what are the necessary statistical signatures that must be present in transcriptomic trajectory data such that a bifurcating dynamical systems point of view is an accurate description? And, if such signatures are present, what further theory-based quantitative insights can be had that go beyond word-based analogies?

One of the foundational principles of the study of bifurcations is the concept of linear stability, and its loss. The theory outlines that a generic way for a stable solution of the steady-state equations of motion to lose stability is through the change of sign of an eigenvalue/s of the Jacobian. In this context, the system is "free" in the direction of the corresponding eigenvector/s. Naturally, the system isn't unbounded, and non-linearities emerge to stabilize a new stable state of the system. In this study, we show that covariances of the state-variables are sufficient to assess whether linear stability is being lost, through leveraging the continuous-time Lyapunov equation. Our central result, based on the analysis of a restricted portion of the transcriptomic trajectories present during the process of hematopoiesis, is that the requisite statistical signatures of a bifurcation through a loss of linear stability are detectable and present during development. Once detected, additional features, such as the unstable eigenvector and the correlations between the state-variables can be identified. From a data-analysis point of view we demonstrate a linear statistical treatment of pseudotime-based transcriptomic trajectories, which is both computationally trivial and statistically robust. As the ability to construct increasingly accurate trajectories improves, our methodology should only become more robust and precise at characterizing critical moments in development. From a more theoretical point of view, our work takes a step toward inferring mathematical models for key developmental decisions at a whole genome scale to mirror the several beautiful dynamical systems biology inferred from imaging studies (43)[NEED CITATIONS].

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 developmental trajectories appear as monotonic proliferations of cell fates (47), 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). Our evidence of bifurcations also 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 (17).

While our results have been derived using bifurcation theory, such that the fixed-point

stability for developmental dynamical systems changes as a function of developmental time, they do not preclude the possibility that cell-fate transitions may occur due to stochasticity (15, 38). To explore if stochasticity may have contributed to neutrophil development, we simulated a noise-induced transition using the same model gene regulatory network (see Secn. S6 for details). We found that the principal covariance eigenvalue exhibited step-like dynamics as the stochasticity was varied (Fig. S9), unlike either the one-to-one (Fig. 4C) or one-to-many (Fig. S4B) bifurcation examples previously explored. As this step-like signature for the principal covariance eigenvalue did not resemble the steady increase and spike in the neutrophil trajectory (Fig. 4C), while bifurcations did, our results suggest that neutrophil differentiation likely occurs due to a loss of stability between fixed points, rather than noise-induced transitions.

Our analysis of the data in Ref. (23) also yielded intriguing implications regarding the specific geometry of neutrophil development in mice. 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 (50). 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 (51).

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 (52, 53). A particularly fascinating direction to pursue in the future is a comparative study of developmentally and/or evolutionary related trajectories and the nature of modifications of the bifurcations. 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

$$\dot{\mathbf{G}} = \mathbf{F}(\mathbf{G}) \tag{10}$$

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

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

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

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

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,

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

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

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

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

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

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 (31).

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

$$\mathbf{J} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} \tag{16}$$

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 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

$$\lambda\tilde{\mathbf{C}} + \tilde{\mathbf{C}}\lambda^\dagger + \tilde{\mathbf{D}} = 0 \tag{17}$$

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,

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

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

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

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

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

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,

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

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, 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 $i \in [2 \dots n_g]$ then the $k = 1$ dominates the sum in Eq. (21), and by equating with Eq. (20) we obtain

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

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 $N_t = 1e7$ timesteps, with $\Delta t = 0.01$ as that was sufficient for equilibration Fig. 3, at a noise scale of $1/s = 0.05$, 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 $k_D = 1, m_2 = 3, m_1 \in [2, 4]$ while in the pitchfork example (Fig. S4) $m_{1,2} = 1, k_D \in [0.24, 5]$.

7. Analysis pipeline

Eq. (5)-7 imply an analysis pipeline characterizing bifurcations in high-dimensional temporal data, which we use in this paper:

1. Obtain highly sampled temporal data

   Caveat: for data types such as scRNA-seq, where frequent sampling is difficult, and samples may include realizations from many different times, time may be inferrable, using, for example pseudotime inference (see Secn. S4.1)

2. Bin the data along the temporal axis (see Secn. S5).

3. Compute the largest eigenvalue of the covariance matrix (ω_1) in each bin (e.g., using an off-the-shelf PCA function).

4. Evaluate if a bifurcation occurs by comparing ω_1 with a suitable null (see Secn. S3).

   (a) Spike indicates a one-to-one bifurcation.

   (b) Steady increase indicates a one-to-many bifurcation.

5. If a bifurcation is detected (e.g., at τ_c) compute and examine the correlation matrix $\mathbf{R}(\tau_c)$, as entries of $\mathbf{R}(\tau_c)$ that are large in magnitude are critical to the bifurcation at τ_c.

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. C Weinreb, S Wolock, AM Klein, Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).
15. P Zhou, S Wang, T Li, Q Nie, Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *bioRxiv* (2021).
16. M Lange, et al., Cellrank for directed single-cell fate mapping. *Nat. Methods* (2022).
17. N Moris, C Pina, AM Arias, Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).
18. 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).
19. K Street, et al., Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 477 (2018).
20. JA Briggs, et al., The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360** (2018).
21. J Delile, et al., Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development* **146**, dev173807 (2019).
22. S Nowotschin, et al., The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
23. C Weinreb, A Rodriguez-Fraticelli, FD Camargo, AM Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367** (2020).
24. M Shamir, Y Bar-On, R Phillips, R Milo, Snapshot: timescales in cell biology. *Cell* **164**, 1302–1302 (2016).
25. 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).
26. M Mojtahedi, et al., Cell fate decision as high-dimensional critical state transition. *PLoS biology* **14**, e2000640 (2016).
27. R Liu, et al., Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci. reports* **2**, 1–9 (2012).
28. Z Chen, et al., A branch point on differentiation trajectory is the bifurcating event revealed by dynamical network biomarker analysis of single-cell data. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* **17**, 366–375 (2020).
29. JH Li, FXF Ye, H Qian, S Huang, Time-dependent saddle–node bifurcation: Breaking time and the point of no return in a non-autonomous model of critical transitions. *Phys. D: Nonlinear Phenom.* **395**, 7–14 (2019).
30. JE Ferrell Jr, Bistability, bifurcations, and waddington's epigenetic landscape. *Curr. biology* **22**, R458–R466 (2012).
31. 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).
32. 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).
33. MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 550 (2014).
34. N Rosenfeld, JW Young, U Alon, PS Swain, MB Elowitz, Gene regulation at the single-cell level. *science* **307**, 1962–1965 (2005).
35. A Raj, A Van Oudenaarden, Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
36. T Gregor, DW Tank, EF Wieschaus, W Bialek, Probing the limits to positional information. *Cell* **130**, 153–164 (2007).
37. G Tkačik, T Gregor, W Bialek, The role of input noise in transcriptional regulation. *PloS one* **3**, e2774 (2008).
38. C Weinreb, S Wolock, BK Tusi, M Socolovsky, AM Klein, Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci.* **115**, E2467–E2476 (2018).
39. NG Van Kampen, *Stochastic processes in physics and chemistry*. (Elsevier) Vol. 1, (1992).
40. R Bargaje, et al., Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc. Natl. Acad. Sci.* **114**, 2271–2276 (2017).
41. TS Gardner, CR Cantor, JJ Collins, Construction of a genetic toggle switch in escherichia coli. *Nature* **403**, 339–342 (2000).
42. 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).
43. EM Ozbudak, M Thattai, HN Lim, BI Shraiman, A Van Oudenaarden, Multistability in the lactose utilization network of escherichia coli. *Nature* **427**, 737 (2004).
44. 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).
45. AM Klein, et al., Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
46. LW Plasschaert, et al., A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
47. M Nitzan, MP Brenner, Revealing lineage-related signals in single-cell gene expression using random matrix theory. *Proc. Natl. Acad. Sci.* **118** (2021).
48. N Borregaard, Neutrophils, from marrow to microbes. *Immunity* **33**, 657–670 (2010).
49. 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.
50. X Sun, B Länger, W Weckwerth, Challenges of inversely estimating jacobian from metabolomics data. *Front. bioengineering biotechnology* **3**, 188 (2015).
51. T Nägele, et al., Solving the differential biochemical jacobian from metabolomics covariance data. *PloS one* **9**, e92299 (2014).
52. 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).
53. 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).
54. CR Harris, et al., Array programming with NumPy. *Nature* **585**, 357–362 (2020).
55. 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).

8. Supporting Information

S1. Methodological relationship to Dynamical Network Biomarkers. Chen et al. (25), developed the concept of a dynamical network biomarker (DNB), a group of genes that drive a critical transition and are detectable from high dimensional gene expression datasets. In particular, they define an indicator function

$$I = \frac{SD_d \cdot |PCC_d|}{|PCC_o|} \qquad \text{[S1]}$$

where SD_d is the average standard deviation of genes in the DNB, PCC_d is the average correlation coefficient between genes in the DNB, and PCC_o is the average correlation coefficient between genes in the DNB and genes outside the DNB (25). At a critical state transition, or bifurcation, I is predicted to diverge, because SD_d and $|PCC_d|$ become large, while $|PCC_o|$ becomes small. Mathematically, the genes in the DNB correspond to those that have non-zero weight in the direction of the transition, i.e., $\vec{p}_d^i \neq 0$, where $\vec{p}_d$ is the principal eigenvector of the Jacobian, while genes outside of the DNB have $\vec{p}_d^i = 0$. This prediction is qualitatively similar, but not the same as Eq. (5). In particular, while both SD_d and ω_1 increase at a bifurcation, they are not equivalent, as SD_d measures the variance of each individual gene, while ω_1 measures the variance across all genes, and therefore accounts for corrections to the total variance due to covariances between genes in the network. Therefore, for bifurcation detection, we focus solely on ω_1, instead of incorporating correlations into the indicator as in Eq. (S1).

As for determining which gene relationships are critical for the bifurcation, we take a similar approach to Refs. (25, 26), in focusing on the correlations that approach ± 1 at the bifurcation. This is justified via Eq. (7), which yields that $R_{ij} \to \pm 1$ if $\vec{p}_d^i \neq 0$ and $\vec{p}_d^j \neq 0$. Interestingly, while we derived Eq. (7) via the eigendecomposition of the covariance matrix, Refs. (25, 26) derived the same result form the covariance matrix itself, providing additional support to this method.

S2. Bifurcations possibilities from two mutually inhibiting genes. At steady state, Eq. (8) satisfies the quintic polynomial

$$g_1 = \frac{m_1/k_D}{\left(\frac{m_2/k_D}{g_1^2+1}\right)^2 + 1} \qquad \text{[S2]}$$

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,

$$g_1(g_2) = \frac{m_1/k_D}{g_2^2 + 1} \qquad \text{[S3]}$$

$$g_1(g_2) = \sqrt{\frac{m_2}{k_D g_2} - 1} \qquad \text{[S4]}$$

it can be deduced that varying m_1, while fixing τ and m_2 can yield a saddle-node bifurcation, as Eq. (S3) moves vertically while Eq. (S4) 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. (S2) 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 (54).

S3. 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 null sample $\mathbf{G}^{\text{null}}$ by drawing each of its entries $G_{i,j}^{\text{null}}$ 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

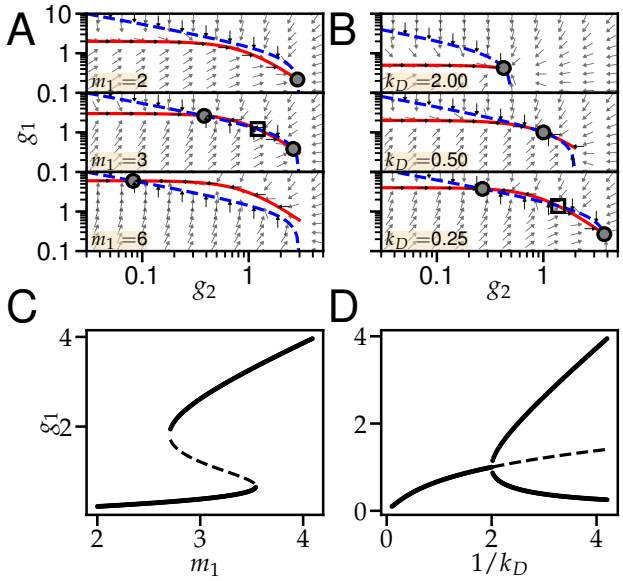

Fig. S1. Analysis of Eq. (8). (A-B) Phase planes for different parameter sets yield a saddle-node bifurcation (A) or pitchfork bifurcation (B). Solid red line is given by Eq. (S3) while dashed blue line is given by Eq. (S4). Open squares are saddles while closed circles are nodes. Arrow angles are given by $tan^{-1}(\dot{g}_1/\dot{g}_2)$ and are uniform length. (C-D) Solutions to Eq. (S2) 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) $m_{1,2} = 1$.

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.

S4. Pseudotime inference.

S4.1. 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. (23) (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 (14).
   (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.

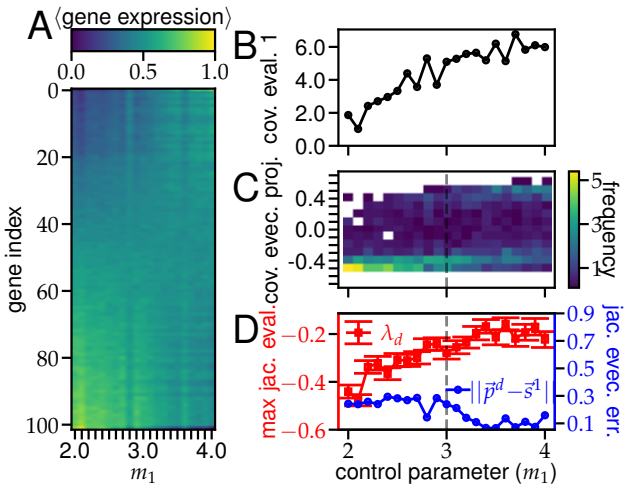

Fig. S2. Unequilibrated saddle-node bifurcation analysis. (A)-(D) Corresponding plots for Fig. 3 (C)-(F), respectively, when the simulation is runs for a $N_t = 500$ iterations rather than $N_t = 50000$.

4. X-Y positions of these 40,000 cells were generated using the ForceAtlas2 algorithm with 500 steps (55).

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

$$S_0(\mathbf{X}, b, i, k) = \vec{x}_i \qquad [S5]$$
$$S_n(\mathbf{X}, b, i, k) = bS_{n-1}(\mathbf{X}, b, i, k) + \frac{1-b}{k} \sum_{j \in K_k(i)} S_{n-1}(\mathbf{X}, b, j, k)$$

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 $\mathbf{T} = \{\vec{t}_1, \vec{t}_2, \ldots, \vec{t}_{n_c}\}$ 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.

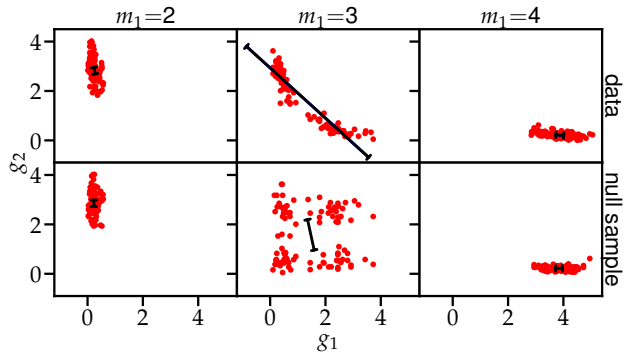

Fig. S3. Effect of resampling on principal eigenvector. Red dots are cells for the data (top) and a single marginal resampling (bottom) before the bifurcation (left), at the bifurcation (center) and after the bifurcation (right). Direction of the black lines corresponds to the principal eigenvector and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

(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

$$a_j = \begin{cases} 0.1 & \text{if neutrophil or pluripotent} \\ -2 & \text{if megakaryocyte} \\ -1 & \text{otherwise} \end{cases} \qquad [S6]$$

(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 $\hat{p}$.

S4.2. Pseudotime inferrence in the absence of metadata. To test if the neutrophil bifurcation characterization was dependent on the choice of pseudotime algorithm, we used the Slingshot algorithm (19) to compute the pseudotime of each cell for its trajectory from the undifferentiated cluster to each of the terminal fate clusters. The input to Slingshot were the cells' cluster labels and their SPRING coordinates, and the output was a probability, or weight, that a cell belonged to each undifferentiated-to-terminal-fate trajectory, as well as its pseudotime along that trajectory. In Fig. S6A, we show the pseudotime of all cells that had $weight > 0$ for belonging to the trajectory that led from undifferentiated cells toward neutrophils. Unlike the pseudotime method described in Secn. S4.1, the origin of the trajectory does not coincide with the earliest sequenced cells, as time of sequencing and clonal barcode data could not be input to Slingshot. Nevertheless, we obtain a clear bifurcation signature in the principal covariance eigenvalue (Fig. S6B) at the point where promyelocyte gene expression decreases to 0 and myelocyte marker gene expression become maximal (Fig. S6C). This result supports our belief that the bifurcation characterization does not depend on the specific pseudotime calculation.

S5. 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.

S6. Noise induced transitions. To determine if a non-bifurcating noise-induced transition model (15, 38) could yield a similar covariance eigenvalue signature to our observations in the neutrophil trajectory (Fig. 4C), we ran the 100 gene network model (Fig. 3A) in a regime of the dynamical system that had two fixed-points ($m_{1,2} = 1$, $k_D = 1/3$) at varying noise scales s (see Fig. S1 and Eq. (24) for details). To ensure a transition, we initialized all cells to populate the fixed point with higher g_1. We found that for low noise values ($1/s \leq 0.01$) the cells stayed near their initial fixed point, yielding a unimodal distribution for g_1 (Fig. S9A) and low principal covariance eigenvalue (Fig. S9B) while for high noise values ($1/s \geq 0.02$) the cells visited both fixed points, yielding a bimodal distribution for g_1, and a high principal covariance eigenvalue. As the covariance eigenvalue signature of the neutrophil trajectory did not include these switch-like dynamics, we suggest that in the neutrophil trajectory, the state transitions result from the underlying dynamical systems bifurcating rather than the scale of biological noise overcoming a fixed potential barrier.

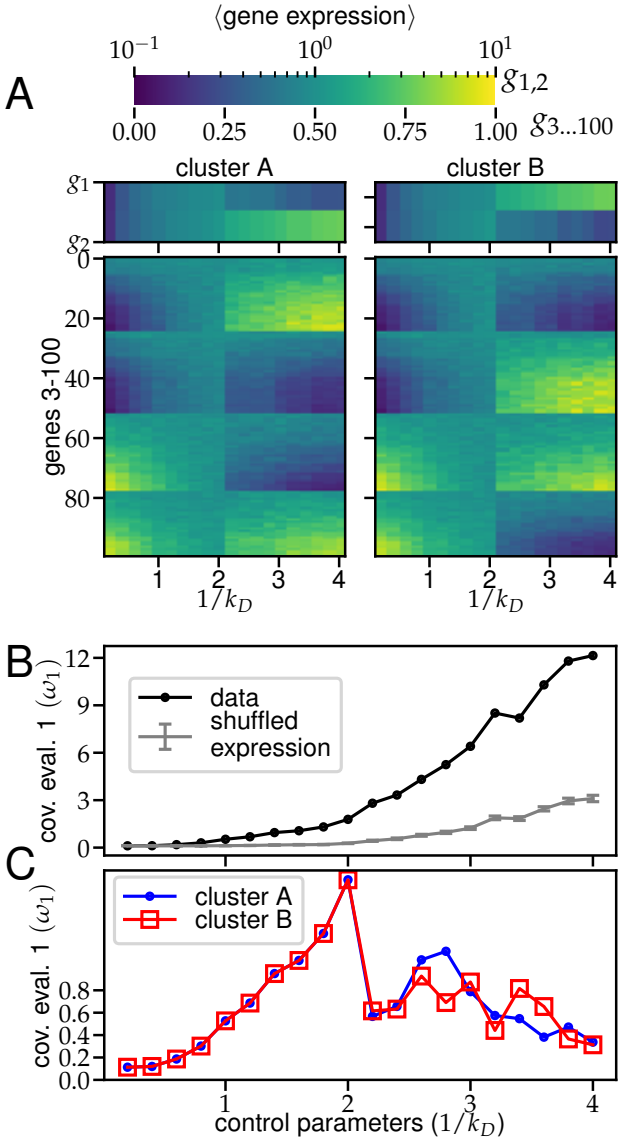

Fig. S4. Pitchfork bifurcation analysis. (A) Gene expression as function of the bifurcation variable τ, separated by cluster. (B) ω_1 and null for unclustered data. (C) ω_1 for clusters.

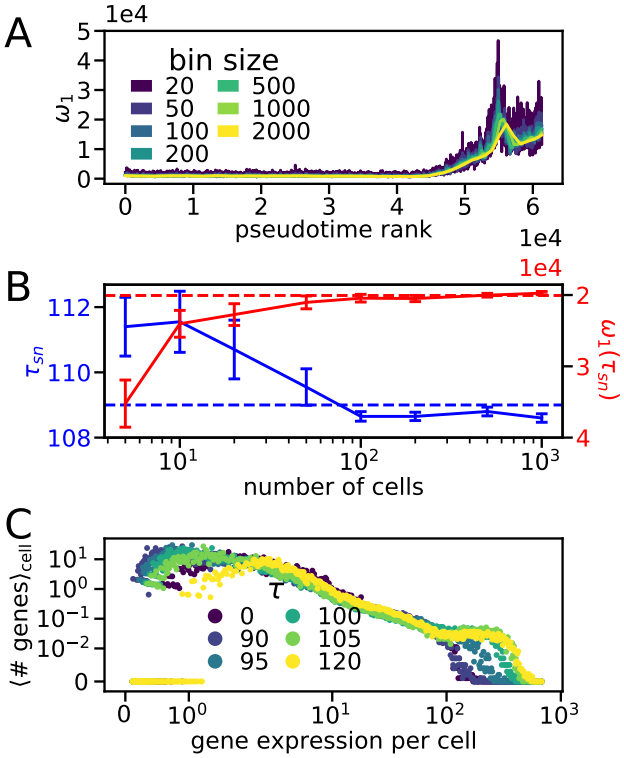

Fig. S5. Distributional properties of the neutrophil pseudotime trajectory. (A) Effect of varying the bin size on principal covariance eigenvalue. (B) Effect of undersampling a bin of 1000 cells on the detected saddle-node bifurcation time and magnitude. (C) Distribution of gene expression during the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

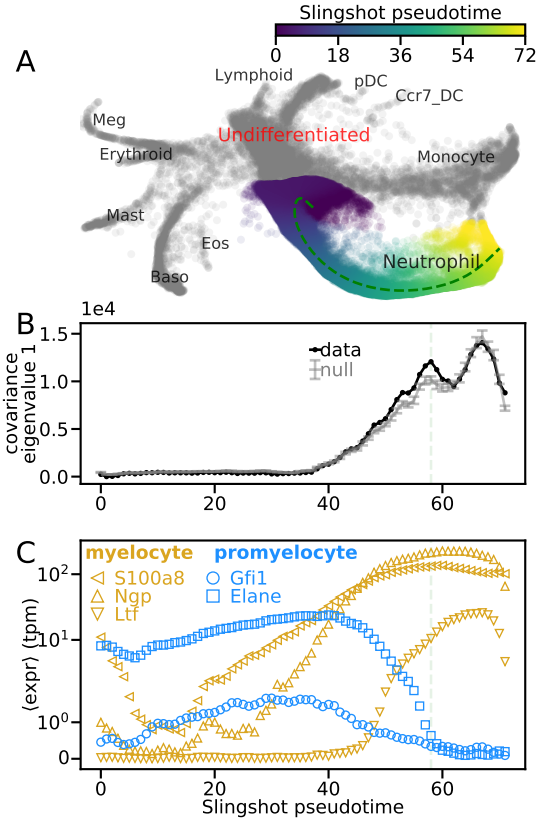

Fig. S6. Bifurcation characterization using Slingshot pseudotime algorithm. (A) Neutrophil development obtained by applying Slingshot to hematopoiesis scRNA-seq data (23). Principal curves were approximated to 1000 points by setting $\mathrm{approx_points} = 1000$ in the $\mathrm{slingshot}$ function as increasing $\mathrm{approx_points}$ further did not affect results. (B) Largest covariance eigenvalue (black) compared with a statistical null (gray, details in $\mathrm{Secn.\ S3}$) in each 1000 cell pseudotemporal bin, shifted to have 0 min, using the Slingshot pseudotime ordering. Error bars of null are one SD. (C) Average expression of promyelocyte (blue) and myelocyte (gold) marker genes in Slingshot pseudotemporal bins (23). SEM error bars are smaller than symbols. Light green line in (B-C) indicates peak of bifurcation window.

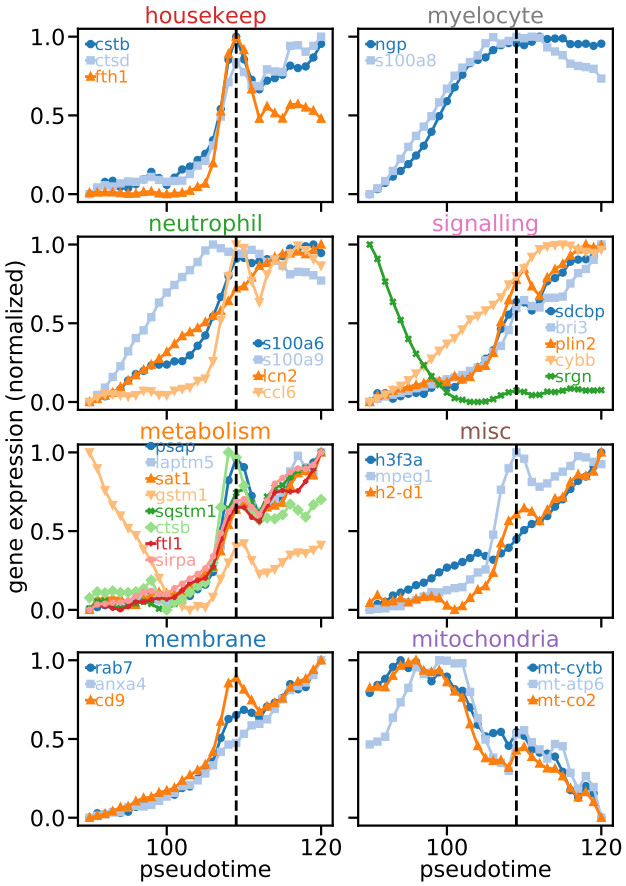

Fig. S7. Gene expression of gene groups in Fig. 6, averaged over all cells in bin that have non-zero expression for that gene as a function of pseudotime, min-max normalized in the pseudotime window shown.

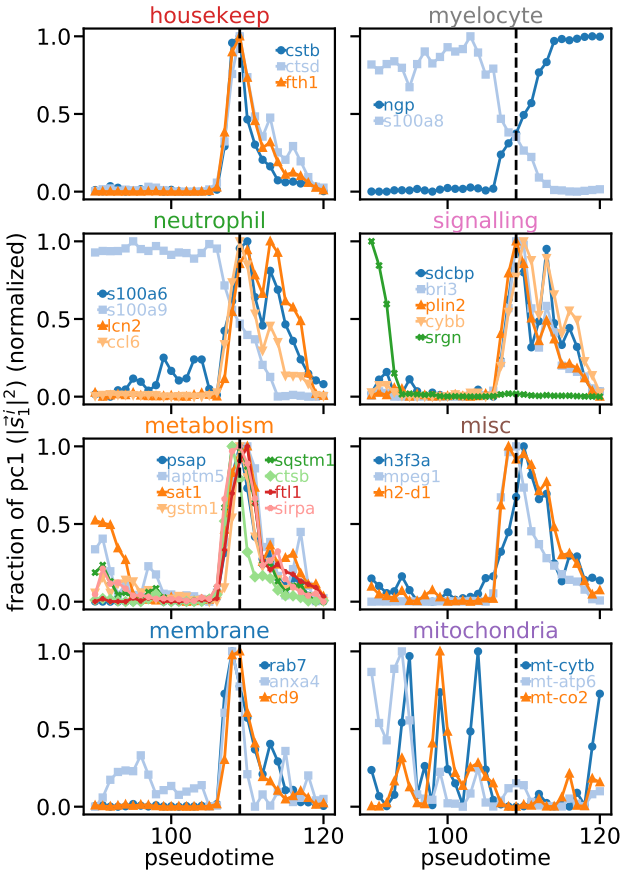

Fig. S8. Fractional weight of first covariance eigenvector for gene groups in Fig. 6, as a function of pseudotime, min-max normalized in the pseudotime window shown.

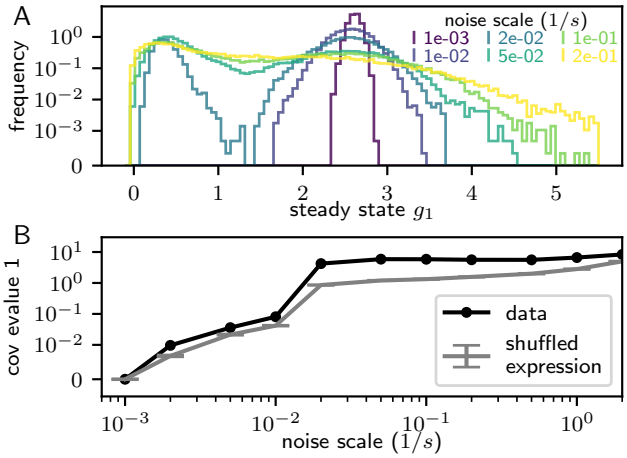

Fig. S9. Covariance eigenvalue signature for a noise induced transition. (A) Steady state distributions for the expression of g_1 at varying noise scales. (B) Principal covariance eigenvalue as a function of noise.