PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Rare-event sampling of epigenetic landscapes and phenotype transitions

Margaret J. Tse[1], Brian K. Chu[1], Cameron P. Gallivan[1], Elizabeth L. Read[1,2]*

1 Department of Chemical Engineering & Materials Science, University of California, Irvine, Irvine, California, United States of America, **2** Department of Molecular Biology & Biochemistry, University of California, Irvine, Irvine, California, United States of America

* elread@uci.edu

Abstract

Stochastic simulation has been a powerful tool for studying the dynamics of gene regulatory networks, particularly in terms of understanding how cell-phenotype stability and fate-transitions are impacted by noisy gene expression. However, gene networks often have dynamics characterized by multiple attractors. Stochastic simulation is often inefficient for such systems, because most of the simulation time is spent waiting for rare, barrier-crossing events to occur. We present a rare-event simulation-based method for computing epigenetic landscapes and phenotype-transitions in metastable gene networks. Our computational pipeline was inspired by studies of metastability and barrier-crossing in protein folding, and provides an automated means of computing and visualizing essential stationary and dynamic information that is generally inaccessible to conventional simulation. Applied to a network model of pluripotency in Embryonic Stem Cells, our simulations revealed rare phenotypes and approximately Markovian transitions among phenotype-states, occurring with a broad range of timescales. The relative probabilities of phenotypes and the transition paths linking pluripotency and differentiation are sensitive to global kinetic parameters governing transcription factor-DNA binding kinetics. Our approach significantly expands the capability of stochastic simulation to investigate gene regulatory network dynamics, which may help guide rational cell reprogramming strategies. Our approach is also generalizable to other types of molecular networks and stochastic dynamics frameworks.

Author summary

Cell phenotypes are controlled by complex interactions between genes, proteins, and other molecules within a cell, along with signals from the cell's environment. Gene regulatory networks (GRNs) describe these interactions mathematically. In principle, a GRN model can produce a map of possible cell phenotypes and phenotype-transitions, potentially informing experimental strategies for controlling cell phenotypes. Such a map could have a profound impact on many medical fields, ranging from stem cell therapies to wound healing. However, analytical solution of GRN models is virtually

OPEN ACCESS

Citation: Tse MJ, Chu BK, Gallivan CP, Read EL (2018) Rare-event sampling of epigenetic landscapes and phenotype transitions. PLoS Comput Biol 14(8): e1006336. https://doi.org/10.1371/journal.pcbi.1006336

Editor: Ilya Ioshikhes, Ottawa University, CANADA

Received: December 9, 2017

Accepted: June 29, 2018

Published: August 3, 2018

Data Availability Statement: All software and data used to generate the figures and tables in this paper, including the supporting information, is available from https://github.com/Read-Lab-UCI/Rare-Event-Sampling-Gene-Networks.

Funding: The research was supported in part by the National Science Foundation [https://www.nsf.gov DMS 1715455 to ELR]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

impossible, except for the smallest networks. Instead, time course trajectories of GRN dynamics can be simulated using specialized algorithms. However, these methods suffer from the difficulty of studying rare events, such as the spontaneous transitions between cell phenotypes that can occur in Embryonic Stem Cells or cancer cells. In this paper, we present a method to expand current stochastic simulation algorithms for the sampling of rare phenotypes and phenotype-transitions. The output of the computational pipeline is a simplified network of a few stable phenotypes, linked by potential transitions with quantified probabilities. This simplified network gives an intuitive representation of cell phenotype-transition dynamics, which could be useful for understanding how molecular processes impact cellular responses and aid interpretation of experimental data.

This is a *PLoS Computational Biology* Methods paper.

Introduction

In multicellular organisms, differentiation of pluripotent stem cells into tissue-specific cells was traditionally considered to be an irreversible process. The discovery of cell reprogramming revealed that the identity of a cell is not irreversibly stable, but rather plastic and amenable to control by perturbation of gene regulatory interactions—for example, through over-expression of key transcription factors [1]. Cellular plasticity has also been observed in other contexts, where cells appear to spontaneously transition among phenotypically distinct states. For example, in embryonic stem cells, expression levels of key transcription factors show dynamic heterogeneity, which is thought to enable diversification of the population prior to lineage commitment [2–6]. This heterogeneity may result at least in part from stochastic state-transitions between functionally distinct, metastable subpopulations [4, 7–9]. Stochastic state-transitions have also been proposed to play a role in cancer, by enabling cancer stem cells to arise *de novo* from non-stem subpopulations [10], or by enabling cells to reversibly transition to a drug-tolerant phenotype [11]. In microbial systems, stochastic phenotype switching has been identified as a survival mechanism for populations subjected to fluctuating environments [12, 13].

Mathematical modeling has provided a basis for understanding how gene regulatory mechanisms and network interactions control cellular identity, stability, and phenotype-transitions. These approaches yield a quantitative means of reinterpreting the long-standing conceptual framework known as Waddington's epigenetic landscape [14–17]. In a mathematical framework, the "valleys" in the landscape that stabilize cell identities within distinct lineages correspond to attractor basins of a high-dimensional nonlinear dynamical system [18]. The nonlinearity results from positive feedback in transcriptional regulation and epigenetic barriers to chromatin remodeling, for example. These feedback mechanisms give rise to multiple, stable (or metastable) phenotype-states accessible to a given genome. Given the "bursty" nature of gene expression and ever-present molecular fluctuations in the cell [19, 20], an active area of research is in modeling the effects of so-called intrinsic noise on gene regulatory network (GRN) dynamics. These mathematical models support the idea that intrinsic noise can drive stochastic phenotype-transitions [21–25], which, though likely to be exceedingly rare in general cellular contexts, may explain the heterogeneity observed in embryonic stem cells where epigenetic barriers appear to be lowered [26].

Mathematical models of GRN dynamics that treat stochastic molecular processes are often formulated as probabilistic Master Equations, in which the system evolves probabilistically over a discrete state-space of molecular species and configurations according to a defined set of biochemical reaction rules. Another common framework is that of a coupled system of ODEs describing the expression levels of genes in the network, with the inclusion of additive noise terms. The Master Equation framework is well-suited to studying how "local" stochastic molecular events (e.g., transcription factors interacting with DNA or chromatin state-transitions near promoters) impact "global" dynamics of phenotype stability and state-switching [23–25, 27, 28]. These molecular fluctuations affecting promoter activity have been shown to significantly impact the structure of epigenetic landscapes, motivating the use of Master Equation-based approaches. That is, the number and stability of phenotype-states accessible to a given GRN varies depending on the kinetic parameters governing these fluctuations [23, 24, 29]. Furthermore, ODE or "mean-field" models that average over these fluctuations can show qualitatively different landscape features [30–32].

Master Equation approaches face the well-known challenge of the "Curse-of-Dimensionality", as solving them requires enumeration of a state-space that grows exponentially with the number of molecular species in the network. For this reason, discrete stochastic models of GRNs are often studied by stochastic Monte Carlo simulation, via the Gillespie algorithm [33]. However, stochastic simulation can also be problematic: in systems with metastability, such as GRNs, stochastic simulation becomes highly inefficient. Transitions between metastable states are rare events (i.e., rare relative to the timescale of fluctuations within a metastable attractor basin), and thus difficult or impossible to observe. Often, these rare events are precisely the events of interest, such as in GRNs where infrequent state-transitions represent critical cell-fate transitions.

Rare-event sampling algorithms are designed to overcome these challenges, by redirecting computational resources towards events of interest, while maintaining statistical accuracy to global system dynamics [34, 35]. In this work, we present a rare-event simulation-based method for computing and analyzing epigenetic landscapes of stochastic GRN models. We combine rare-event methods with coarse-graining and analysis by Transition Path Theory—adopted from the field of Molecular Dynamics of protein folding [36]–and show that this unified framework provides an automated approach to map epigenetic landscapes and transition dynamics in complex GRNs. The method quantifies the number of metastable phenotype-states accessible to a GRN, calculates the rates of transitioning among phenotypes, and computes the likely paths by which transitions among phenotypes occur. We apply the method to a model of pluripotency in mouse Embryonic Stem Cells. Our results reveal rare sub-populations and transitions in the network, demonstrate how global landscape structure depends on kinetic parameters, and reveal irreversibility in paths of differentiation and reprogramming. Our approach is not limited to gene regulatory networks; it is generalizable to other stochastic dynamics frameworks and is thus a potentially powerful tool for computing global dynamic landscapes in areas such as signal-transduction, population dynamics, and evolutionary dynamics.

Methods

A graphical overview of the computational pipeline presented in this paper can be found in Fig 1.

Gene regulatory network models

We demonstrate the rare-event sampling method for two representative GRN models. A small, two-gene network serves as a model system to validate the simulations. We then apply

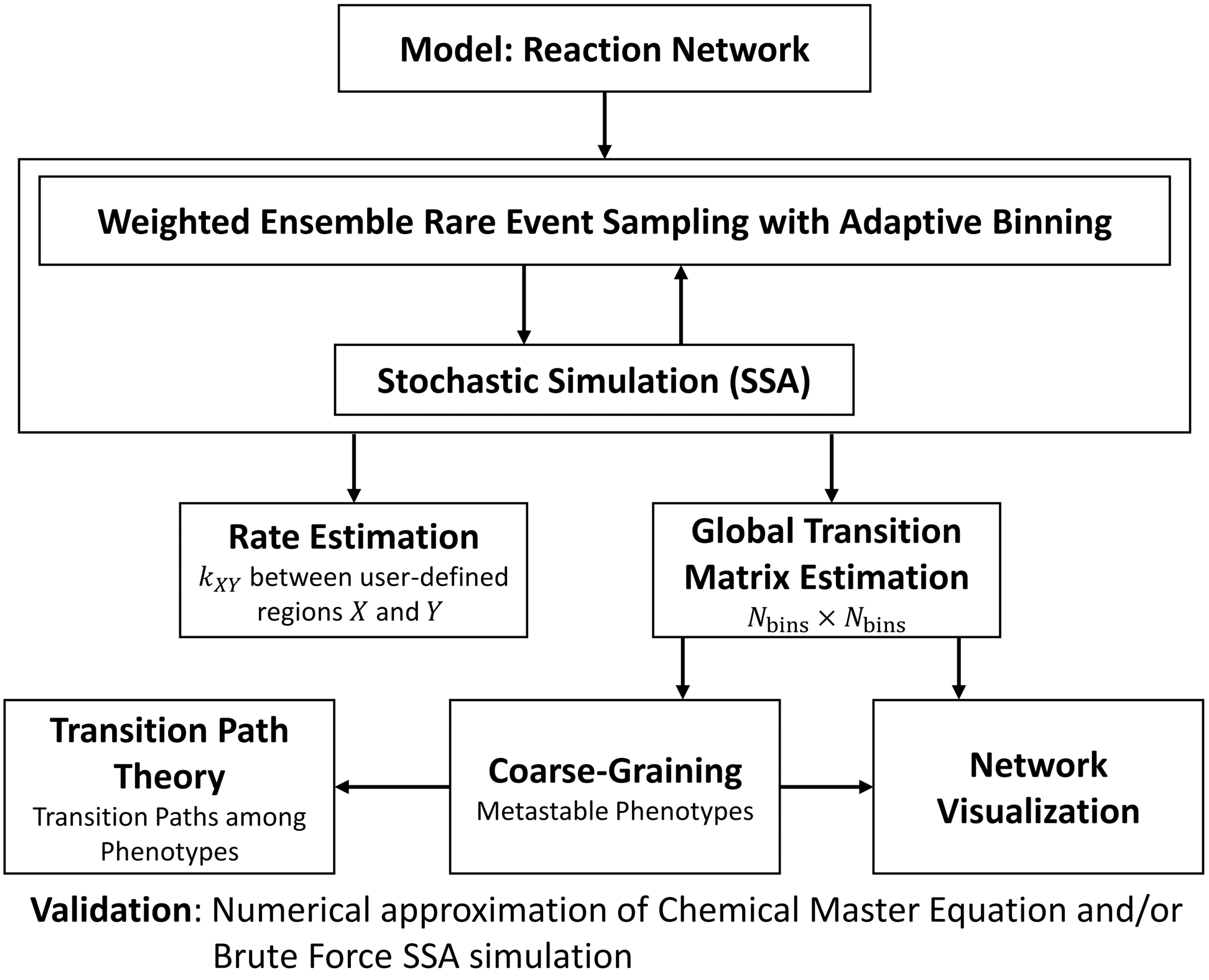

Model: Reaction Network
Weighted Ensemble Rare Event Sampling with Adaptive Binning
Stochastic Simulation (SSA)
Rate Estimation
k_{XY} between user-defined regions X and Y
Global Transition Matrix Estimation
$N_{\text{bins}} \times N_{\text{bins}}$
Transition Path Theory
Transition Paths among Phenotypes
Coarse-Graining
Metastable Phenotypes
Network Visualization
Validation: Numerical approximation of Chemical Master Equation and/or Brute Force SSA simulation

Fig 1. Computational pipeline for rare-event sampling of epigenetic landscapes and phenotype transitions. The input to the computational pipeline is a reaction network model of gene regulatory network dynamics. Stochastic simulations are performed using SSA [33] and Weighted Ensemble rare-event sampling [45]. The WE method can be run in two modes: *Rate Mode* computes the rate of transitioning between two user-defined regions of interest with high accuracy. *Transitio n-Matrix Mode* computes the pairwise transition probabilities among N_{bins} adaptively defined sampling bins that span the system state-space. Further visualization and analysis of the transition-matrix can be performed, including automatic designation of metastable phenotypes via the coarse-graining framework [42] and identification of likely transition paths [36].

https://doi.org/10.1371/journal.pcbi.1006336.g001

the method to a more complex model of pluripotency in mouse Embryonic Stem Cells (mESCs).

Exclusive mutual inhibition, self-activation model. The Exclusive Mutual Inhibition, Self-Activation (ExMISA) model is a two-gene network representing an archetypal motif at cell-fate branch points [37, 38]. Each gene, denoted generically as *A* or *B*, encodes a transcription factor that activates its own transcription and represses transcription of the other gene. We adopt previous conventions [22, 23, 39] for stochastic GRN dynamic models. The full list of biochemical reactions and parameters can be found in the Supplement, S1 File and S1 Table. The model encompasses stochastic birth/death processes for transcription factor production and degradation, and stochastic binding and unbinding of transcription factors to DNA regulatory/promoter regions; the binding-states of these regions governs the production rate. Each transcription factor is assumed to bind to DNA as a homodimer, giving cooperative regulation (explicit dimerization reactions are neglected, such that the transcription-factor-binding reaction is modeled as tri-molecular). In the "exclusive" network variant, transcription factors compete for binding sites on DNA (only one transcription factor dimer can be bound to a gene's promoter at a time). The discrete state-vector, which completely describes the state of the system, is given by $\mathbf{x} = [A_{ij}, B_{ij}, n_a, n_b]$. A_{ij} and B_{ij} represent the three possible promoter binding-states for each gene (i.e., A/B_{00}, A/B_{10}, A/B_{01} denote unbound, activator-bound, or repressor-bound states). The copy-numbers of expressed protein transcription factors are denoted by n_a and n_b for products of gene A and B, respectively, and may in principle take any nonnegative integer value. All processes related to transcription, translation, and assembly are subsumed into a single protein birth reaction. For genes in state A/B_{ij}, this production occurs with rate constant g_{ij}. The production rate is high when the promoter is bound by the activator (its own product). Otherwise, if unbound or repressor-bound, a low "basal" rate of expression is assumed, i.e. $g_{00} = g_{01} < g_{10}$. Degradation of protein products occurs with rate k, and stochastic binding/unbinding of transcription factors to DNA occur with h and f, respectively. The model is symmetric, with equivalent parameters for the two genes.

We studied a parameter set (S1 Table) in the regime of slow DNA-binding kinetics, in contrast to the so-called "adiabatic" regime where binding/unbinding of regulators to promoters occurs quickly relative to protein production and degradation. We adopt this regime here for two reasons. First, it has recently been suggested that the slow- or moderate-binding regime is likely to be more accurate in eukaryotic systems, where complex, slow-timescale changes in chromatin structure accompany binding events [27, 28, 40]. Second, the number of metastable states in GRNs appears to generally increase in the slow-binding regime, due to distinct combinations of relatively stable promoter configurations [41]. Therefore, this regime presents a test-case to develop enhanced sampling techniques that can efficiently traverse multiple system barriers and reconstruct complex, multi-modal dynamics.

Pluripotency network model. The pluripotency network model of mESCs was developed by Zhang and Wolynes [28] on the basis of experimental literature and previous models. The 8-gene network shares the same stochastic reaction framework as the ExMISA model. The genes (NANOG, OCT4, SOX2, GCNF, KLF4, PBX1, GATA6, and CDX2) suppress and activate each other through homo- and heterodimers of their encoded transcription factors (OCT4 and SOX2 form a heterodimer; all other regulatory interactions occur via homodimers). Binding of transcription factors to promoters is not exclusive. The model has five kinetic parameters: g_{on}, g_{off}, h, f, and k, corresponding to the rate of gene expression in the activated state, the rate of gene expression in the un-activated state, binding of transcription factors to DNA, unbinding of transcription factors from DNA, and transcription factor degradation (or exit from the nucleus). Genes are expressed at the basal rate g_{off} except when bound by at least one activator and no repressor, in which case they are expressed with rate g_{on}. The exception to this logic rule is NANOG, which must be bound by the the KLF4 and PBX1 transcription factor homodimers and the heterodimer OCT4-SOX2 to be activated. Overall, these interactions lead to a total of 396 biochemical reactions, with a total of 88 "species" (counting 80 distinct gene promoter configurations and 8 protein species). The complete logic rules and list of reaction rate parameters can be found in the Supplement (S1 File, S2 and S3 Tables).

Theoretical background: The chemical master equation and stochastic transition-matrix

The mathematical framework of the network models is the discrete Chemical Master Equation (CME) [33], which gives the time-evolution of the probability to observe the system in a given state. In vector-matrix form, the CME can be written

$$\frac{d\mathbf{p}(\mathbf{x},t)}{dt} = \mathbf{K}\mathbf{p}(\mathbf{x},t) \tag{1}$$

where $\mathbf{p}(\mathbf{x}, t)$ is the probability over the system state-space ($\mathbf{x}$) at time t, and $\mathbf{K}$ is the reaction rate-matrix containing stochastic reaction propensities (diagonal elements $k_{jj} = -\sum_i k_{ij}$, i.e., columns sum to 0). Eq 1 assumes a well-mixed system of reacting species, and assumes that the technically infinite state-space described by $\mathbf{x}$ (containing molecular species numbers/configurations) may be limited to some finite number of "reachable" states, (i.e., with non-negligible probability) for an enumeration of N states of the system, $\mathbf{K} \in \mathbb{R}^{N\times N}$. The steady-state probability $\pi(\mathbf{x}) \equiv \mathbf{p}(\mathbf{x}, t \to \infty)$ over N states satisfies

$$\mathbf{K}\pi(\mathbf{x}) = \mathbf{0}. \tag{2}$$

Thus, $\pi(\mathbf{x})$ can be obtained from $\mathbf{K}$ as the normalized right-eigenvector corresponding to the zero-eigenvalue.

It is sometimes desirable to work with the time-dependent stochastic transition-matrix $\mathbf{T}(\tau)$ rather than the time-independent stochastic rate matrix $\mathbf{K}$ [42]. For example, $\mathbf{T}(\tau)$ may be more amenable to estimation by sampling (as we demonstrate in this work for the pluripotency network, for which $\mathbf{K}$ is impractical to enumerate). For a CME with rate matrix $\mathbf{K}$, $\mathbf{T}(\tau)$ is given by

$$\mathbf{T}(\tau) = \exp(\tau\mathbf{K}^{\mathrm{T}}) \tag{3}$$

where exp denotes the matrix exponential. $\mathbf{T}(\tau) \in \mathbb{R}^{N\times N}_{0\leq x\leq 1}$ then gives the conditional probability for the system to transition between each pair of states within a lagtime τ. That is, the elements T_{ij} give the probability that the system, if found in state i, will then be found in state j at a time τ later, and rows sum to 1. Using $\mathbf{T}(\tau)$, the evolution of probability over discrete intervals of the lagtime τ is given by the Chapman-Kolmogorov equation:

$$\mathbf{p}^{T}(\mathbf{x}, t + k\tau) = \mathbf{p}^{\mathrm{T}}(\mathbf{x}, t)\mathbf{T}^{k}(\tau). \tag{4}$$

Eigenvectors corresponding to dominant eigenvalues of the stochastic transition-matrix are associated with slow system processes. By Perron-Frobenius, for an irreducible stochastic matrix $\mathbf{T}(\tau)$ with eigenvalues λ_i, there exists $\lambda_1 = 1$, and all other eigenvalues satisfy $|\lambda_i| < 1$. Analogous to Eq (2) for $\mathbf{K}$, the steady-state probability can be obtained directly from $\mathbf{T}(\tau)$ according to $\pi^{T}(\mathbf{x}) = \pi^{T}(\mathbf{x})\mathbf{T}(\tau)$, i.e., as the normalized left-eigenvector corresponding to λ_1. Eigenvalues λ_i are related to global system timescales t_i by

$$t_i = -\frac{\tau}{\ln|\lambda_i(\tau)|}, \tag{5}$$

(with t_1 giving the infinite-time, stationary result) [42]. Additionally, the Mean First Passage Time for transitions from an individual state i to a region Y (MFPT$_{i,Y}$, where Y may be an individual state or a set of states) can be computed using the matrix elements $T_{i,j}$ by [43, 44]:

$$\mathrm{MFPT}_{i,Y} = \begin{cases} 0 & i \in Y \\ 1 + \sum_{j\notin Y} T_{i,j}\mathrm{MFPT}_{j,Y} & i \notin Y \end{cases}. \tag{6}$$

MFPT$_{i,Y}$ is defined as the expected time for the system to reach Y for the first time, having started in state i. The MFPTs may be computed by solving the linear system in Eq 6. Eq 6 computes the MFPT as a dimensionless quantity, the expected number of "steps" (of duration τ)

required for the transition; multiplication by τ gives the MFPT in units of time. The MFPT starting from a region X (i.e., a set of states, rather than an individual state) and ending in a region Y is given by the stationary-probability-weighted sum:

$$\mathrm{MFPT}_{X,Y} = \sum_{i\in X} \frac{\pi_i \mathrm{MFPT}_{i,Y}}{\sum_{j\in X}\pi_j}. \tag{7}$$

Weighted ensemble stochastic simulation

Stochastic reaction kinetics can be simulated by the Stochastic Simulation Algorithm (SSA) [33], which produces numerically exact realizations of the CME (Eq 1). Simulation circumvents the need for enumerating the exceedingly large system state-spaces typical of gene network models, but suffers from inefficiency due to rare events. The Weighted Ensemble (WE) rare-event sampling algorithm [45] redistributes computational resources from high-probability regions of state-space to low-probability regions, which tend to be under-sampled in conventional simulation. The method thereby reduces computational effort in sampling rare transitions and improves accuracy of estimating probability density in, e.g., barrier-regions or tails of distributions. The method can be applied to any stochastic dynamics framework; in recent years, it has been widely applied to atom-scale Molecular Dynamics. Details of the methodology are discussed in a recent review [35] and references therein. Both WE and a related method, Forward Flux Sampling, have been applied previously to the study of 2-gene networks [46, 47].

Briefly, the algorithm works as follows: state-space is divided up into bins that span transitions of interest. The number of bins, N_{bins}, is typically $\mathcal{O}(100)$, and a variety of binning procedures can be used (we use an adaptive procedure described below). Initially, a single simulation trajectory, or "replica", is assigned a weight of 1 and allowed to freely move within and between bins for a user-defined lagtime τ_{WE}. After each iteration of τ_{WE}, a splitting and culling procedure divides and/or combines replicas and their associated weights in such a way as to reach and maintain an equal target number of weighted replicas, M_{targ}, in each bin. Over the course of the simulation, the combined weights of the replicas in a bin (averaged over successive iterations) will evolve toward the probability of the system to reside in that bin. By maintaining the same number of replicas in each bin (M_{targ}), with weights proportional to probability, the algorithm devotes comparable computational time to low- and high-probability regions. Effectively, the algorithm computes long-time processes on the basis of many short-time simulated trajectories.

Adaptive binning procedure. As with other enhanced sampling methods, the WE algorithm requires dividing of state-space into defined sampling regions or "bins". For high-dimensional systems, discretization poses a challenge because, for an N-dimensional, evenly spaced grid, the number of required sampling bins increases exponentially with the number of degrees of freedom. To address this challenge, a variety of Voronoi-polyhedra-based procedures have been developed [48–50]. These methods balance the need to focus simulation toward regions with non-negligible probability, while still enabling capture of rare transitions of interest. In addition to efficiently discretizing high-dimensional spaces, the methods have the benefit of requiring little to no *a priori* knowledge of system dynamics (e.g., of the locations of regions of interest, or of appropriate progress coordinates for transitions). We utilize an adaptive binning procedure from ref. [50]. Each bin (of user-defined number N_{bins}) is a Voronoi polyhedron with a generating node; the bin is defined as the region of state-space encompassing all points closer to the generating node than to nodes of any other region. After each lagtime τ_{WE}, new Voronoi regions are generated by successively selecting N_{bins} node-positions from the current replica positions in a way that maximizes the Euclidean distance between them. By this procedure, over the course of the simulation, bins spread to encompass all areas of state-space reached by any simulated trajectory. After sufficient iterations, the bin positions stop spreading to new areas but continue to fluctuate. The procedure is shown by representative simulations in S1 Fig.

Computation of transition rates. One important output of WE sampling is the quantitative rate of transitions between regions of interest, which may be difficult or impossible to estimate from conventional simulation. WE sampling may be run in different modes, depending on whether the sought-after information concerns a specific transition of interest, or a more global picture of system dynamics, i.e., encompassing approximate rates of transitions among many system states. We term the two modes "rate" mode and "global transition-matrix" mode. The former can deliver a more accurate estimate for a particular state-transition, while the latter can yield a more comprehensive, but approximate, measure of global system dynamics.

In rate mode, the user specifies two regions of interest, X and Y, The flux of probability into/out of regions of interest can be estimated by recording the amount of weight transferred at the end of each simulation iteration. The mean first passage time of transitions from X to Y ($\mathrm{MFPT}_{X,Y}$) is given in general by the inverse of probability flux from X to Y. In practice, we apply a "labeling" scheme [51, 52], where each replica is labeled as belonging to either set $\mathcal{S}_X$ or $\mathcal{S}_Y$ according to its history, i.e., whether it most recently visited region X or Y, respectively. The summed weight of all replicas in $\mathcal{S}_X$ is given by $P_{\mathcal{S}_X}$, and $P_{\mathcal{S}_X} + P_{\mathcal{S}_Y} = 1$ satisfies probability conservation. Then,

$$\mathrm{MFPT}_{X,Y} = \frac{\overline{P}^{SS}_{\mathcal{S}_X}}{\overline{\Phi}^{SS}(Y|\mathcal{S}_X)} \tag{8}$$

where $\overline{\Phi}^{\mathrm{SS}}(Y|\mathcal{S}_X)$ is the average probability flux from $\mathcal{S}_X$ into Y at steady-state, which is measured by the weight of $\mathcal{S}_X$-labeled replicas entering Y during the simulation after convergence to steady-state. The labeling scheme enables accurate estimates, including for non-Markovian transitions. For Markovian transitions well-described by a single rate-constant, $k_{X,Y} = 1/\mathrm{MFPT}_{X,Y}$.

Computation of network transition-matrix. Running WE in transition-matrix mode enables visualization and analysis of global system dynamics on the basis of a single simulation, and requires no designation of regions of interest. In this mode, the previously-converged Voronoi bins are fixed, and simulations are used to estimate a coarse-grained stochastic transition-matrix $\tilde{\mathbf{T}}(\tau)$ of size $N_{\mathrm{bins}} \times N_{\mathrm{bins}}$. The coarse-grained $\tilde{\mathbf{T}}(\tau)$ approximates the true dynamics over the full state-space, as given by $\mathbf{T}(\tau)$. Thus, the procedure enables estimation of the global transition-matrix (and subsequent analysis) in systems where enumeration of states is not feasible. To estimate $\tilde{\mathbf{T}}(\tau)$, the weight transferred between bins is recorded at each iteration, and the elements of the transition-matrix are estimated according to [51]:

$$\tilde{T}_{i,j} = \frac{\langle w_{i,j}\rangle_2}{\langle w_i\rangle} \tag{9}$$

where $\langle w_{i,j}\rangle_2$ is the average weight transferred from bin i to bin j over the iteration time τ_{WE} (counting only after at least 2 transitions, and averaging over multiple iterations) and $\langle w_i\rangle$ is the average population (summed weight) in bin i. By construction, this is a row-stochastic transition-matrix with state-space "resolution" determined by N_{bins} (each state in the full state-space sampled by the simulation is assigned to its nearest neighboring Voronoi node). The lag-time τ of the transition-matrix corresponds to the sampled WE-time τ_{WE}. However, use of $\tilde{\mathbf{T}}(\tau)$ to compute system dynamics imposes a Markovian approximation, by which equilibration of replicas within bins is assumed to be rapid on the timescale of τ, and hops between states (i.e. bins) are memoryless. As such, while this mode of simulation has the advantage of acquiring a holistic view of global system dynamics, it has the disadvantage of introducing a Markovian approximation.

Coarse-graining procedure to classify phenotype-states

While the sampled $N_{\mathrm{bins}} \times N_{\mathrm{bins}}$ transition-matrix provides a global approximation of the epigenetic landscape and state-transitions, we apply a method to further coarse-grain dynamics, known as the Markov State Model framework [29, 36, 42]. This automated procedure produces a highly simplified representation of global dynamics in terms of a few (generally < 10) clustered sets and the transitions among them. Such highly-reduced models can be beneficial in terms of human intuition of system dynamics, comparison to experiments, and—in this application—automated designation of dynamic phenotype-states. The method utilizes the concept of metastability, i.e., system states that experience relatively fast transitions among them are clustered together into the same coarse-grained set. Collectively, the coarse sets experience relatively rare inter-cluster transitions and frequent intra-cluster transitions. We employ the metastability concept as a definition of cell phenotype, reasoning that a phenotype should be a relatively stable attribute of a cell, and stochastic inter-phenotype transitions should be relatively rare. In practice, we employ the Markov State Model framework to further reduce the sampled row-stochastic transition-matrix $\tilde{\mathbf{T}}(\tau)$ from size $N_{\mathrm{bins}} \times N_{\mathrm{bins}}$ down to $C \times C$, where C is the number of coarse-grained clusters chosen by the user. As the Markov State Model (MSM) is itself a stochastic transition-matrix on a coarse-grained space, it implies a more severe Markovian approximation. It provides a way to describe global system dynamics in a highly simplified way while maintaining high accuracy to the slowest system dynamics as sampled by $\tilde{\mathbf{T}}(\tau)$. In previous work, we demonstrated the application of this coarse-graining approach to automatically designate phenotypes in small gene networks [29]; here, we extend the applicability of the coarse-graining to large, complex networks by combining it with rare-event sampling.

The coarse-graining procedure is a spectral clustering method based on the Perron Cluster Cluster Analysis (PCCA+) algorithm [53], which optimizes the (nearly)-block-diagonal structure of $\tilde{\mathbf{T}}(\tau)$ for systems with metastability. The signature of such metastability is a separation-of-timescales for intra- and inter-basin dynamics, which may be seen as gaps in the eigenvalue spectrum [42]. As noted above, $\mathbf{T}(\tau)$ (or its sampled counterpart, $\tilde{\mathbf{T}}(\tau)$) has $\lambda_1 = 1$, corresponding to the infinite time-limit. If a set of m dominant eigenvalues exists, such that for decreasing eigenvalues $\lambda_i \lessapprox 1$, $i \in \{2, \ldots, m\}$, and a gap is present, $\lambda_j \ll \lambda_m$ for $j > m$, this indicates the presence of m slow-timescale processes in the system, and further indicates that $\tilde{\mathbf{T}}(\tau)$ may be re-ordered to give m nearly-uncoupled blocks. In practice, the algorithm attempts to find a coarse-graining onto C clusters, where C may be user-defined, or may be determined algorithmically, e.g., according to the spectral gap [53]. Here, we choose C clusters, where the last significant gap in the spectrum is seen between λ_C and λ_{C+1}. For the GRNs studied here, this corresponds to choosing C such that $\lambda_C/\lambda_{C+1} > 10$.

Transition path analysis. The coarse-grained model of system dynamics given by the MSM enables estimation of the ensemble of dominant transition paths among phenotypes, along with their relative probabilities. We adopt methods from Transition Path Theory according to Noe, *et al.* [36] (details therein). Briefly, $\tilde{\mathbf{T}}(\tau)$ can be used to compute the effective flux of trajectories, along any edge in the coarse-grained network, contributing to transitions between states X and Y (where these designated states correspond to one or more coarse-grained phenotype-states produced by the MSM). A pathway decomposition algorithm on the matrix of effective fluxes for $X \rightarrow Y$ transitions then yields a set of dominant pathways and the relative contribution of each to the overall flux. Each state in the MSM is analogous to a cell phenotype, and transition path analysis is used to identify parallel phenotype transition paths and the relative rates of transitioning between phenotypes.

Visualization of epigenetic landscapes

Both the sampled transition-matrix $\tilde{\mathbf{T}}(\tau)$ and the coarse-grained MSM encode stationary and dynamic information about global dynamics—that is, they quantify the epigenetic landscape. For visualization, we use Gephi graph visualization software [54] using the Force Atlas algorithm. Every circle (or node) in the graph corresponds to a sampling bin or to a coarse-grained phenotype, and the area of a circle is proportional to its relative steady state probability according to $\ln(\gamma P_{SS})$, where P_{SS} is the steady state probability of the node and γ is a constant chosen to improve visibility of low probability regions of the landscape. Lines between circles (edges) correspond to transitions between sampling regions or coarse-grained phenotype. Their thickness and coloring correspond to their relative transition probability and source state, respectively.

Validation: Numerical solution of the chemical master equation

To validate the simulation method, we compare the simulated dynamics to the numerical solution to the CME. We choose the parameters of the ExMISA model in such a way as to restrict the effective state-space, so that a numerical solution of the CME is tractable. Building the reaction rate matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$ requires enumeration of N system states. In general, if a system of S molecular species has a maximum copy number per species of n_{max}, then $N \approx n_{max}^{S}$. In the ExMISA model, the state-vector is given by $\mathbf{x} = [A_{ij}, B_{ij}, n_a, n_b]$. For enumeration, we neglect states with protein copy-numbers larger than a cutoff value which exceeds g_{10}/k (corresponding to the average number of transcription factors maintained in the system from a gene while in its active state). For example, with model parameters $g_{10} = 18$ and $k = 1$, we truncate at $n_{a,max} = n_{b,max} = 41$ and assume that probability flux between states with $n_a, n_b \leq 41$ and states with $n_a, n_b > 41$ is assumed to be 0 (i.e., the boundaries of the state-space are reflective). Including the gene-binding states, this gives $N = 3 \times 3 \times 42 \times 42 = 15876$ states. This size is tractable for complete solution of the CME using matrix methods in MATLAB [55]. This truncation of the state-space introduces a small approximation error (see S2 Fig).

The pluripotency network has 8 genes with copy numbers of $\mathcal{O}(10^3)$ (determined by the parameters $g_{on}/k = 3900$). The number of distinct binding-promoter states for each gene are 16, 32, 8, 8, 2, 8, 4, and 2 for GATA6, NANOG, CDX2, OCT4, SOX2, KLF4, GCNF, PBX1, respectively (see S2 Table). Together these combinations enumerate a state-space of $N > 10^{30} \approx 1000^8 \times 16 \times 32 \times 8 \times 8 \times 2 \times 8 \times 4 \times 2$. This size precludes solution of the CME, and we instead estimate the dynamics by WE sampling. Where possible, we validate the WE-sampling results by "conventional", i.e., by direct simulation using SSA.

Validation of coarse-grained models. To check the validity of the coarse-grained MSM as a representation of the global dynamics, we use the Chapman-Kolmogorov test to compare the relaxation curves of the coarse-grained system to those found through direct SSA following Eq 4 [42]. If the coarse-graining is appropriate, the relaxation curves of the MSM probabilities will match the relaxation profile of long conventional (direct SSA) simulations initiated within each coarse-grained phenotype. Transition paths through the coarse-grained phenotype network are validated, where possible, against conventional SSA simulation.

Implementation and software

Stochastic Gillespie (SSA) simulations were carried out using BioNetGen [56]. WE sampling was implemented with in-house software code written in MATLAB. Simulations were run on the high performance computing cluster (HPC) at the University of California, Irvine, and parallelization of BioNetGen SSA simulations was performed using the Sun Grid Engine scheduler. The coarse-graining procedure and transition path analysis was implemented in python scripts, adapted from MSMBuilder [57] and Pyemma [43], respectively. Transition-matrix and MSM visualization was carried out using Gephi software and the Force Atlas layout [54]. All simulation parameters can be found in the supplement S4 Table. Pseudo-code for the adaptive binning procedure can be found in S2 File and software can be found in https://github.com/Read-Lab-UCI/Rare-Event-Sampling-Gene-Networks.

Results

Rare states and transitions in gene regulatory networks are accessible by rare-event sampling

We first apply the computational pipeline to a small two-gene model (the exclusive Mutual Inhibition, Self-Activation model, ExMISA, see Methods), exhibiting an archetypal motif for cell fate-decisions [37, 38]. The model is tractable for computation of full, discrete stochastic dynamics to within a small approximation error using matrix methods. Thus, the model provides a numerical benchmark for assessing the accuracy of the simulation method, before extension to larger systems where solution of the Chemical Master Equation (CME) is intractable. For the chosen parameters, the ExMISA model shows four peaks in the steady-state probability distribution (projected onto protein copy numbers, n_a and n_b). Peaks in probability correspond to basins in the so-called quasipotential landscape, defined by $U = -\ln(\pi(\mathbf{x}))$ (Fig 2). The four peaks/basins corresponds to four possible combinations of binarized A/B gene expression: hi/hi, hi/lo, lo/hi, and lo/lo. These four phenotype-states arise due to the combination of balanced repression and self-activation in the network, and the slow kinetic parameters (Supplementary S1 Table) for transcription factor binding and unbinding to promoters that effect changes in individual gene-activity states between low and high expression rates [29, 58].

The WE-based simulation method enabled estimation of global dynamics of the ExMISA model. By redistributing computational resources from relatively high-probability to low-probability regions (see Methods), the WE method enabled uniform sampling of the quasipotential landscape, i.e., mapping basins (high-probability regions) along with high barriers (low probability regions) (Fig 2a). The simulation estimated individual steady-state bin-probabilities as low as 1.3×10^{-6} and showed good global agreement with the numerical CME benchmark (see Fig 2 and Supplement, S3 Fig).

In addition to sampling global dynamics, the WE method can be used to estimate rate constants for individual, rare transitions of interest. The Mean First Passage Time of the global network switch from the center of one polarized phenotype-state to another, i.e., $\mathrm{MFPT}_{X \rightarrow Y}$ from protein a/b expression level hi/lo to lo/hi was estimated from WE to be 1.82×10^{5} (units of k^{-1}) (see S6 Table), in agreement with the CME result.

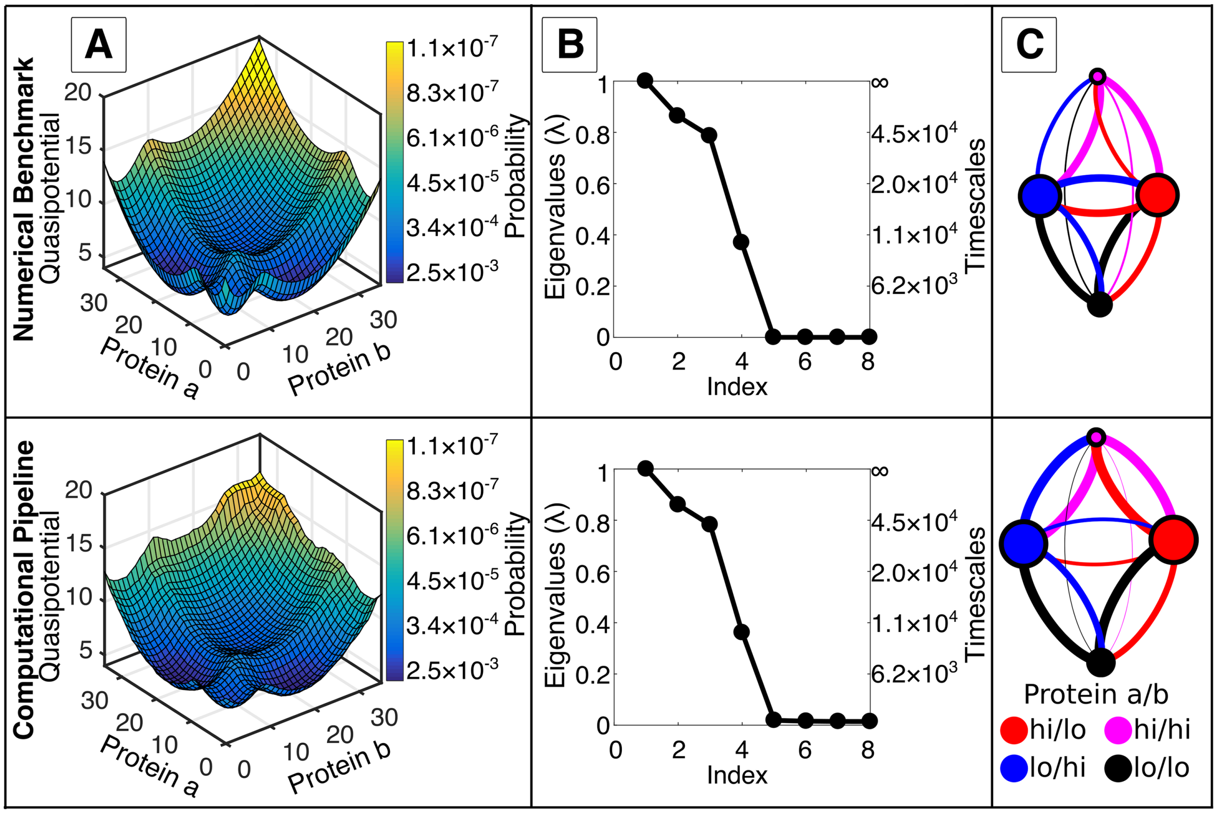


Fig 2. Simulation results show good agreement with a theoretical benchmark for the 2-gene ExMISA (mutual inhibition, self-activation) cell-decision circuit. The Chemical Master Equation for the 2-gene model, ExMISA, was solved numerically (see Methods) (top) and compared to simulation results from the computational pipeline presented in this paper (bottom). Shown for each are the Quasipotential Landscape (A), Eigenvalue Spectrum (B), and Markov State Model (C). (A) Quasipotential landscapes of the ExMISA network projected onto the two protein coordinates. Deep blue regions denote low potential (high probability) and yellow denote high potential (low probability). The four visible basins in both correspond to combinations of lo/hi expression for the two genes *A* and *B*. (For both rows, quasipotential surfaces estimated over discrete states/bins are smoothed for visualization). WE sampling captured both the basin structure and low probability edge and barrier regions. (B) Eigenvalue spectra and corresponding computed global transition timescales. Gaps in the eigenvalue spectrum indicate separation of timescales, i.e., the presence of metastability. C) Four-phenotype coarse-grained models automatically generated from the clustering algorithm (see Methods). Each colored circle represents a cell phenotype, sized proportionally to its probability. Edges are inter-phenotype transitions (colored by source-state, with width proportional to probability). The full CME and simulation pipeline identify similar metastable phenotype networks (see S11 Fig for details).

https://doi.org/10.1371/journal.pcbi.1006336.g002

Phenotype transitions can be approximated by Markovian jumps, enabling construction of coarse-grained models

A network transition-matrix $\tilde{\mathbf{T}}(\tau)$ over sampled bins ($N_{\text{bins}} = 300$) was constructed from WE sampling for ExMISA and used for subsequent analysis of global system dynamics. By comparison, a full network transition-matrix $\mathbf{T}(\tau)$ over the enumerated system state-space was constructed from the CME ($N = 15876$, see Methods). The full, computed ($\mathbf{T}(\tau)$) and simulated ($\tilde{\mathbf{T}}(\tau)$) transition-matrices showed qualitatively similar eigenvalue spectra with four dominant eigenvalues, indicating the presence of metastability (separation-of-timescales between intra-basin and inter-basin transitions) (Fig 2b). The slow system-timescales predicted by the full CME model corresponding to eigenvalues $\lambda_2, \lambda_3, \lambda_4$ were $t_2, t_3, t_4 = 6.8 \times 10^4, 4.2 \times 10^4, 1.0 \times 10^4$ respectively, in units of k^{-1} where k is the protein degradation rate (the Perron eigenvalue $\lambda_1 = 1$ is associated with the infinite-time (stationary) distribution). The corresponding values given by the WE-simulated $\tilde{\mathbf{T}}(\tau)$ were 6.1×10^4, 3.5×10^4, 9.4×10^3, respectively. These numbers demonstrate how the sampled $\tilde{\mathbf{T}}(\tau)$ enables global approximation of slow system timescales to $< 20\%$ relative error. Error in these values (relative to the slowest timescales implied by the true eigenvalues) depends on both "spectral" (lagtime) and discretization error, i.e., improvements can be achieved only with a larger number of bins (finer discretization) and/or longer lagtime [42] (see S4 Fig). In contrast, WE sampling in "rate mode" (see Methods) enabled highly accurate estimation of $\text{MFPT}_{X\rightarrow Y}$ to within 2% error (S6 Table).

According to the Markov State Model framework, the presence of timescale separation indicates that a simplified model, retaining a few coarse-grained metastable states with Markovian transitions among them, can reasonably approximate the full system dynamics. Using this approach, we label the metastable sets as *phenotypes* accessible to the network, reasoning that a useful classification of cell phenotypes should be one that gives relatively stable, rather than transient, cell types. We apply the Markov State Model coarse-graining procedure to both the full $\mathbf{T}(\tau)$ and simulated $\tilde{\mathbf{T}}(\tau)$, yielding similar results. The coarse sets (or metastable phenotype-states) in the reduced models for both cases are generated automatically, and map directly onto the four basins seen in the quasipotential landscape (i.e., the gene A/B expression hi/hi, hi/lo, lo/hi, and lo/lo cell phenotypes). The reduced models are visualized by network graphs, in which node sizes are proportional to steady-state probability, and the thicknesses and lengths of edges are proportional to the transition probability between them (on lagtime τ) (Fig 2c). Some discrepancies can be seen visually in the network graphs. These discrepancies likely result in part from the slightly different mappings of the full state-space onto the four clusters (see S11 Fig for details), which could in turn result from the distance-metric-based binning, which is relatively insensitive to changes in promoter configuration. Numerical values for the reduced models can be found in S5 Table. The network graph can be considered to be an alternative representation of the global epigenetic landscape, which contains both stationary and dynamic information. (In contrast, the epigenetic landscape plotted as a quasipotential function does not explicitly contain dynamic information, due to non-gradient dynamics [16]).

Validation of the coarse-grained model can be carried out according to the Chapman-Kolmogorov test [42], which tests how well the relaxation dynamics initialized in the metastable phenotypes approximate the dynamics that are predicted either by the full model (CME) or simulated trajectories. According to this test, relaxation dynamics out of metastable phenotypes from WE sampling was predicted with relative error values between 0.02 and 0.12 for all phenotypes (S5 Fig). Together, these results indicate (i) that a Markovian model of phenotype transitions is a good approximation of the full system dynamics for the ExMISA model, and (ii) that the WE-simulation based computational pipeline predicts a quantitatively similar coarse-grained phenotype-network to the full CME model.

The method maps the epigenetic landscape and identifies dominant phenotypes in a pluripotency network model

We apply the computational pipeline to a pluripotent fate-decision network from mouse Embryonic Stem Cells (mESCs) introduced by Zhang *et al.* [28] (Fig 3A). The network comprises eight interacting genes: NANOG, GATA6, CDX2, SOX2, OCT4, GCNF, and PBX1. Three of these genes, NANOG, SOX2, and OCT4 have been suggested to maintain pluripotency [59], and NANOG inhibits the expression of differentiation markers [60]. The GATA6 and CDX2 genes have been used in experiments as markers of differentiation, with the GATA6 transcription factor being a marker of the primitive endoderm cell lineage, and the CDX2 transcription factor being a marker of the trophectoderm lineage [61].

Using the WE-based computational pipeline, we estimate $\tilde{\mathbf{T}}(\tau)$ with a resolution of $N_{\text{bins}} = 250$. To visualize the global landscape as a graph network at this resolution, we plot the

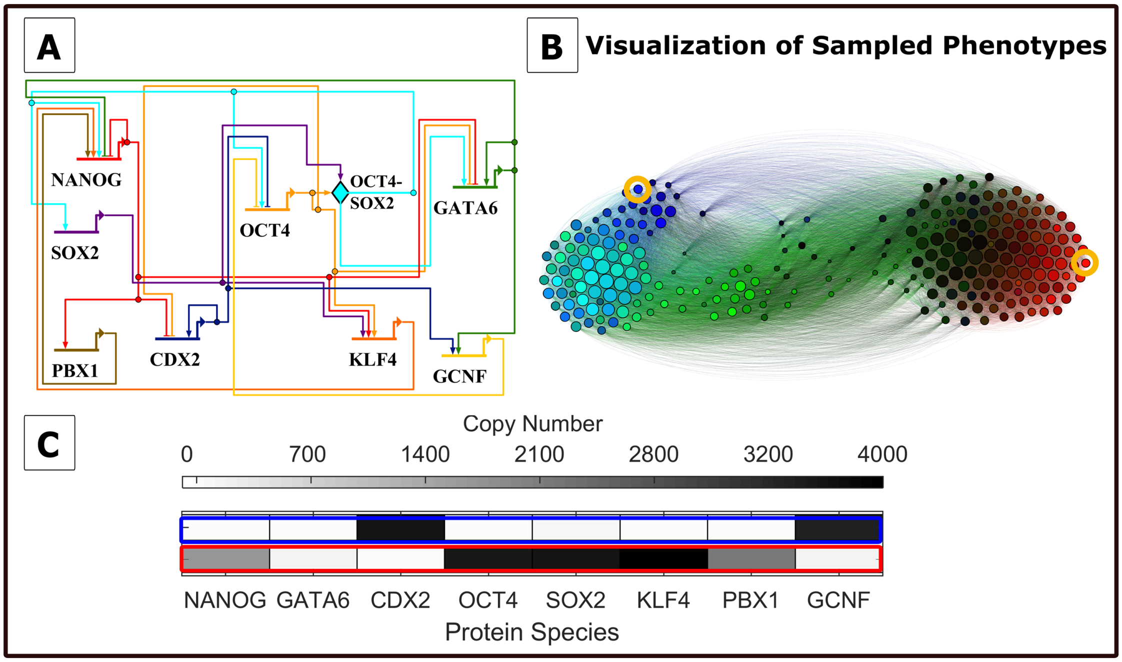
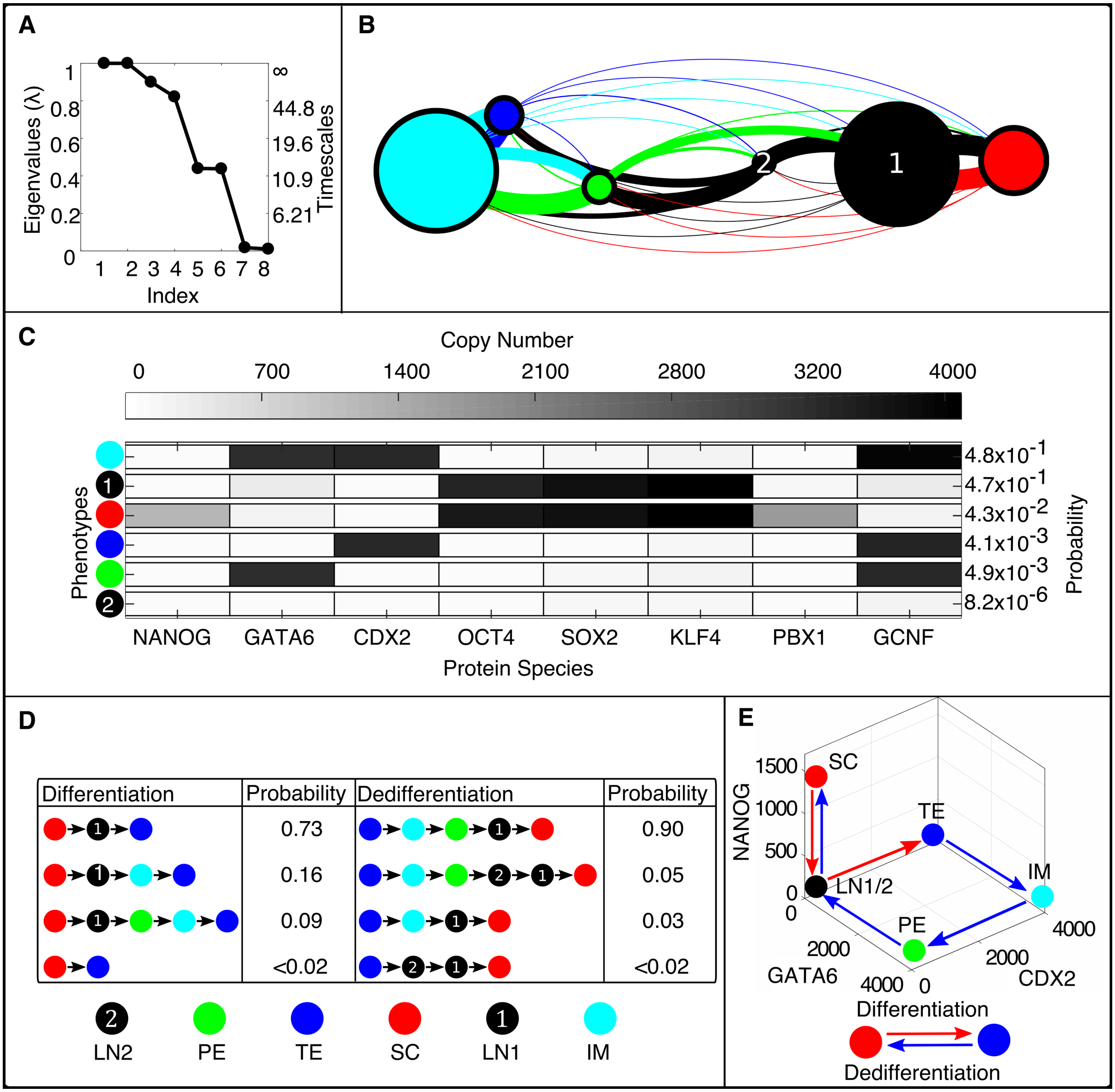


Fig 3. Pluripotency network model and simulation results (Parameter Set I). A)Wiring diagram for the eight-gene pluripotency network model, adapted from [28]. Arrowheads represent positive interactions, while flat lines denote repression. B) Simulation results: state-transition graph of sampled network states. Circles represent aggregate gene-expression states sampled during the Weighted Ensemble simulation. Circle areas are proportional to the steady-state probability π_i in each state according to $\ln(\gamma\pi_i)$ with scaling factor $\gamma = 3.4$. States are colored according to the gene expression levels of three of the genes; red, green, and blue correspond to high NANOG, GATA6, and CDX2 expression respectively, while black corresponds to low or no gene expression. Edges connecting the states indicate possible state-transitions, colored according to the originating state. The graph is produced using Gephi [54] using a force-directed layout algorithm (Force Atlas), therefore short inter-state distances reflect higher probability of transitioning. C) Full protein compositions of two representative states, with either high CDX2 expression (blue) or high NANOG expression (red). States in (C) correspond to yellow circles in (B).

https://doi.org/10.1371/journal.pcbi.1006336.g003

converged $\tilde{\mathbf{T}}(\tau)$ using a force-directed automated graph layout [54] (Fig 3B). The barbell shape of the network reflects the broad antagonism between pluripotency and differentiation genes, which is a general feature of the overall network topology. At the same time, each "pole" comprises multiple distinct patterns of gene expression (seen in the graph as different colors with full compositions in Fig 3C), hinting at the existence of multiple phenotypes associated with both pluripotency and lineage-specification. Moreover, the network representation reveals numerous links between pluripotent and differentiated states, pointing to both direct and indirect transitions, through a network of relatively transient intermediate states.

To further analyze the global dynamics of the pluripotency network, we apply the Markov State Model coarse-graining framework. The simulated $\tilde{\mathbf{T}}(\tau)$ shows gaps in the eigenvalue spectrum after four and after six eigenvalues (Fig 4a). The corresponding approximate time-scales are given by $t_2, t_3, t_4, t_5, t_6 = 1.1 \times 10^5, 95, 51, 12, 12$ (k^{-1}), respectively. These values, though only approximate, indicate the presence of a single long timescale process (t_2) corresponding to transfer between differentiated and pluripotent states, while transitions within those basins (t_3, etc.) occur at least four orders of magnitude more quickly. Applying the coarse-graining algorithm to achieve six clusters results in a reduced model (Fig 4b), with the clusters representing metastable phenotypes. The phenotypes can largely be distinguished in the subspace of NANOG, GATA6, and CDX2 expression levels; the differentiated phenotypes show expression of either GATA6 (primitive endoderm, PE), CDX2 (trophectoderm, TE), or



Differentiation	Probability	Dedifferentiation	Probability
🔴→❶→🔵	0.73	🔵→⚪→🟢→❶→🔴	0.90
🔴→❶→⚪→🔵	0.16	🔵→⚪→🟢→❷→❶→🔴	0.05
🔴→❶→🟢→⚪→🔵	0.09	🔵→⚪→❶→🔴	0.03
🔴→🔵	<0.02	🔵→❷→❶→🔴	<0.02

Fig 4. Simulation results for the pluripotency network (Parameter Set I). The Computational Pipeline Uncovers Six Metastable Phenotypes and Irreversible Phenotype Transitions. A) Computed eigenvalue spectrum and global timescales indicating the presence of metastability in the network. The gap in the eigenvalue spectrum after the sixth eigenvalue suggests that a partitioning can be found into six metastable phenotypes. B) The coarse-grained network showing six algorithmically-identified phenotypes designated as Low NANOG 1 (LN1), Low NANOG 2 (LN2), Stem Cell (SC), Primitive Endoderm (PE), Trophectoderm (TE), and the Intermediate Cell (IM) state. C) The averaged gene expression levels (copy numbers) of each transcription factor for each phenotype and their respective steady-state probabilities. D) The four most probable transition pathways from the SC state to the TE state (differentiation) and from the TE state to the SC state (dedifferentiation). E) The highest probability transition paths projected onto three protein coordinates, NANOG, GATA6, and CDX2. Differentiation from SC to TE is visibly irreversible, i.e., the system returns by a separate route.

https://doi.org/10.1371/journal.pcbi.1006336.g004

both (denoted an intermediate cell type, IM). Phenotypes associated with pluripotency do not express high levels of GATA6 or CDX2, and may express high levels of NANOG (stem cell, SC). The coarse-grained model reveals two separate pluripotent phenotypes that are low in NANOG expression: one which expresses other pluripotent factors OCT4, SOX2, and KLF4 ("Low NANOG 1" LN1), and one which has low expression of all factors ("Low NANOG 2" LN2) (Fig 4c). Overall, these phenotypes broadly match experimentally-determined categories, coincide with steady-states of the stochastic model computed previously by a CME-approximation method [28], and coincide with phenotype-states identified in related pluripotency GRN models [62]. The steady-state probabilities associated with the phenotypes are highly nonuniform, with 95% of the population divided nearly evenly between the IM and LN1 phenotypes, which are associated with differentiation and pluripotency, respectively. The LN2 state is rarest, comprising only 8×10^{-4}% of the population, and was not identified previously [28]. Together, these results indicate that the clustering method identifies both common and exceedingly rare phenotypes in the *in silico* cell population modeled by simulation trajectories. Furthermore, the automated method identifies both expected phenotypes and one novel (albeit low probability) phenotype.

The method reveals multiple, irreversible pathways for phenotype transitions in the pluripotency network

Previously, Markov State Models constructed on the basis of Molecular Dynamics simulations were used to analyze the ensemble of distinct pathways of protein-folding [36]. Here, we utilize the coarse-grained model of phenotype transitions in the pluripotency GRN in a similar manner, to analyze pathways of cell differentiation and dedifferentiation. Using Transition Path Theory, the method identifies the pathways that carry the greatest fraction of net probability flux, among sequences associated with successful SC→TE transitions (and reverse) (Fig 4d and 4e). Transition paths between the stem cell (SC) and PE phenotypes can be found in S6 Fig. For Parameter Set I, the method identifies three pathways encompassing $>$ 98% of the probability flux for both forward and reverse transitions. While the SC→ TE transition is most likely to occur directly through the LN1 state (i.e., NANOG expression will shut off, followed by turning on CDX2), the reverse transition shows a different route through the IM and PE states (i.e., GATA6 expression turns on, then CDX2 turns off, then GATA6 turns off, and finally NANOG turns on).

Dynamic analysis of the coarse-grained model, including analysis of transition paths, relies on the Markovian approximation for inter-phenotype transitions. In the pluripotency network, stochastic transitions between pluripotency (SC, LN1, LN2) and differentiation (TE, IM, PE) basins are infrequent relative to transitions within those basins, justifying the Markovian assumption, since the system equilibrates within those basins much more rapidly than inter-basin transitions occur. However, the Markovian assumption may be less accurate for describing intra-basin transitions between phenotypes, which occur much more frequently. Despite the coarse-grained model encompassing transitions on highly disparate timescales, the qualitative results of transition path analysis were validated by collected conventional simulation trajectories (not subject to any Markovian assumption), which identified the same dominant transition paths (S7 Fig). Overall, these results indicate that a stochastic excursion of a cell from the SC to TE phenotypes and back maps a cycle in gene-expression space, echoing previous studies indicating nonequilibrium dynamics in GRNs [16, 23]. The results further indicate that the Markov State Model, while a highly coarse-grained approximation, can provide an accurate estimation of inter-phenotype transition dynamics.

Table 1. Computed mean first passage times (MFPTs) of phenotype transitions in the pluripotency network. MFPTs are shown for transitions between the pluripotency (high NANOG) state (SC) and low NANOG expression states (LN(1)) (left columns) and for transitioning between the pluripotency state (SC) and the trophectoderm state (TE) (right columns), in units of the inverse transcription factor decay rate, k^{-1}. Transitions for Parameter Set I were computed using the WE method in rate mode while transitions for Parameter Set II were estimated from the sampled transition matrix. The definitions of SC and LN(1) are analogous to the high NANOG production (N^{hi}) and low NANOG production (N^{lo}) transitions measured in experiments [8, 9]. Increasing the adiabaticity (i.e., the rates of DNA-(un)binding, h, f), leads to rarer inter-phenotype transitions. The simulations also show that, within the same gene network for a given parameter set, inter-phenotype transition times span four orders of magnitude.

Transition	SC → LN(1)	LN(1) → SC	SC → TE	TE → SC
Parameter Set I(f = 10)	1.71×10^1	1.94×10^2	1.36×10^5	2.70×10^5
Parameter Set II(f = 50)	7.71×10^4	1.28×10^4	8.13×10^8	5.82×10^9

https://doi.org/10.1371/journal.pcbi.1006336.t001

Cell phenotype landscape and transition dynamics are sensitive to kinetic parameters

We applied the computational pipeline to the pluripotency network using two different rate parameters sets (see S1 File), which differ in rates of transcription factor binding and unbinding to DNA. In line with previous studies [23, 24, 29], we found that increasing the so-called adiabaticity (i.e., increasing h and f, or the rates of TF-binding relative to protein production and degradation, Parameter Set II) led generally to rarer inter-phenotype transitions (see Table 1). For example, in Parameter Set I, the Mean First Passage Time (MFPT) for transitions from SC → TE was calculated to be 1.36×10^5 in units of k^{-1}, as compared to 8.13×10^8 for Parameter Set II. The MFPTs of the reverse transition TE → SC for each set were 2.70×10^5 and 5.82×10^9, respectively (see Table 1 and S7 Table). These differences in magnitude broadly reflect that moving toward the adiabatic regime leads to increased epigenetic barriers between phenotypes.

In addition to generally slowing transitions, the increased adiabaticity of Parameter Set II gives rise to an epigenetic landscape structure that is distinct from that of Parameter Set I, with altered steady-state phenotype probabilities (Fig 5a). The eigenvalue spectrum shows qualitatively distinct features as well, with a gap after five values (Fig 6a). As such, the Markov State Model framework identifies five dominant phenotypes in the network, which correspond broadly to those of Parameter Set I, except that only a single Low-NANOG (LN) phenotype is

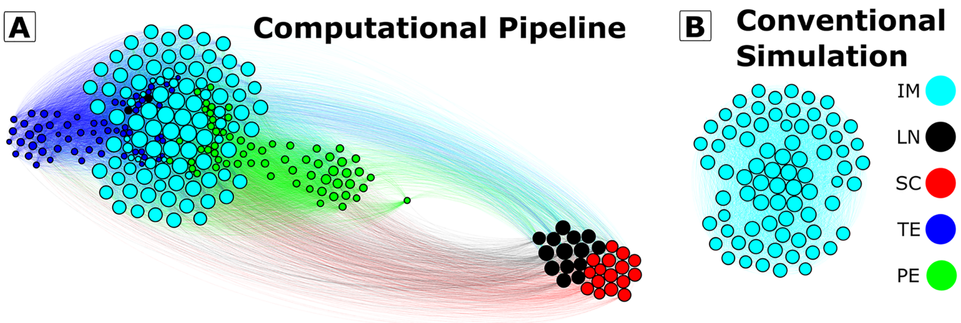


Fig 5. The rare-event sampling pipeline makes rare states and transitions accessible to simulation. A) The global state-transition graph computed with the computational pipeline for the Pluripotency Network with rare transitions (Parameter Set II). The states are colored according to the coarse-grained (algorithmically-identified) phenotypes. In this parameter regime (f = 50) the differentiated (TE, PE, IM) and pluripotent phenotypes are cleanly separated, reflecting exceedingly rare transitions between the two phenotypes (O(10^9), see Table 1). (B) States visited in conventional SSA simulation (using the same initialization, definitions, and placement as in (A)). In the conventional simulation, a transition out of the IM phenotype was never observed.

https://doi.org/10.1371/journal.pcbi.1006336.g005

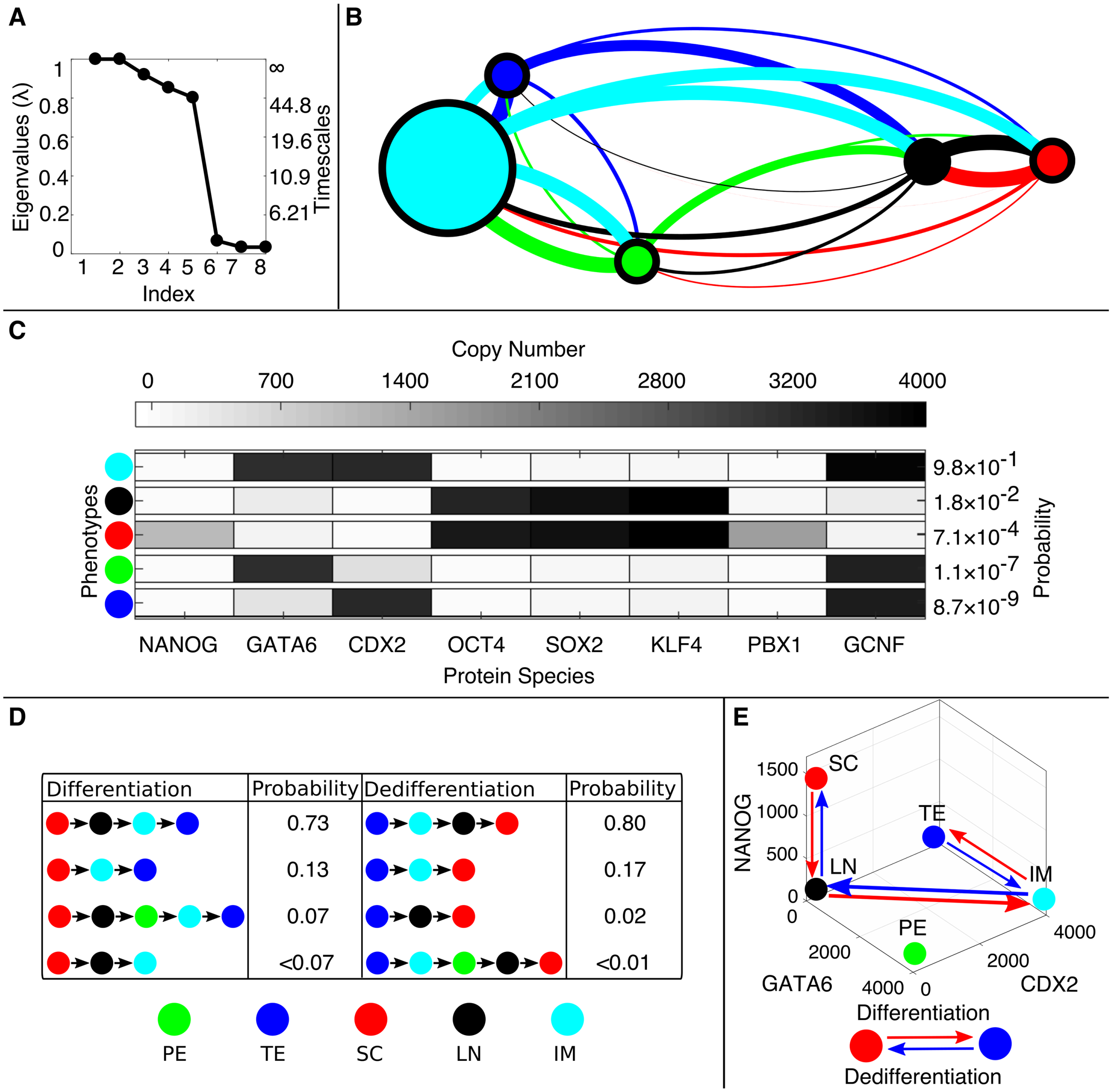


Fig 6. Simulation results for the pluripotency network (Parameter Set II). Changing DNA-Binding Kinetics Alters the Epigenetic Landscape. A) Computed eigenvalue spectrum and global timescales. B) The coarse-grained Markov State Model showing five phenotypes corresponding to the LN1, SC, PPE, TE, and IM phenotypes of Parameter Set I. The majority of the steady state probability is in the IM phenotype (0.98). C) The gene expression levels for each phenotype and their respective steady-state probabilities. D) The four most probable differentiation pathways between SC and TE phenotypes. E)The dominant pathways of (de)differentiation projected onto the GATA6, CDX2, and NANOG coordinates. The change in DNA-binding kinetics shows different transition dynamics from Parameter Set I. Here, the forward and reverse paths are the same.

https://doi.org/10.1371/journal.pcbi.1006336.g006

identified (Fig 6b). Most of the steady-state probability is contained in the IM state (Fig 6c). In addition to altering the transition rates and relative phenotype probabilities, the kinetic parameters altered the dynamics of differentiation and dedifferentiation. The two likeliest pathways of forward (and reverse) SC → TE transitions follow the same route through LN and IM phenotypes (Fig 6d and 6e). Alternative differentiation pathways of forward (and reverse) SC → PE transitions can be found in S9 Fig. These results indicate that, while the same GRN model with different kinetic parameters may give rise to qualitatively similar phenotypes, they differ in quantitative stationary and dynamic features, including relative steady-state probabilities, transition times, and likeliest transition pathways.

Efficiency of rare-event sampling compared to conventional SSA

Rare phenotype transitions can be difficult to observe with conventional SSA simulation. We compared simulated landscapes (based on estimated $\tilde{\mathbf{T}}(\tau)$) from the computational pipeline for the Pluripotency network (Parameter Set II) to those obtained from an equivalent (large) number of SSA simulation steps (Fig 5a and 5b). This comparison revealed that the WE-based method uncovers multiple phenotypes and associated transitions that are effectively invisible to conventional simulation due to the rarity of exiting metastable basins.

Quantitative estimates of efficiency gains for WE have been based on comparing the number of simulation steps required to estimate a desired quantity (such as a rate constant) using WE versus conventional simulation [47]. Treating $\tilde{\mathbf{T}}(\tau)$ as the desired output (as it contains holistic dynamic information for the system), we estimate the efficiency gain of our pipeline by computing:

$$E = \frac{\text{Sim. steps to estimate } \tilde{\mathbf{T}}(\tau), \text{Conv.}}{\text{Sim. steps to estimate } \tilde{\mathbf{T}}(\tau), \text{WE}}. \qquad (10)$$

The denominator of Eq 10 is given by $N_{\text{bins}} \times N_{\text{iterations}} \times \tau \times M_{\text{targ}}$, thus accounting for all individual replica-steps in the total WE simulation time. The numerator is computed by asking how many steps of a conventional simulated trajectory are required to estimate $\tilde{\mathbf{T}}(\tau)$. It is generally prohibitive to collect enough conventional simulation steps to estimate $\tilde{\mathbf{T}}(\tau)$ to a similar resolution as WE. However, given a $\tilde{\mathbf{T}}(\tau)$ estimated from WE, it is in principle possible to estimate how many steps would be necessary to achieve the same $\tilde{\mathbf{T}}(\tau)$ by conventional simulation. We used an approximate, conservative estimate given by:

$$[\text{Sim. steps to estimate } \tilde{\mathbf{T}}(\tau), \text{Conv.}] \gtrapprox \tau \sum_i \left(P_{5\%,i}\{T_{ij}\}\right)^{-1}, \qquad (11)$$

where $P_{5\%,i}$ denotes the 5th percentile over nonzero elements of row i. Justification of Eq 11 is given in the Supplement, S3 File. Briefly, Eq 11 reflects the fact that the required simulation time should be dominated by the rare transitions (i.e., the smaller elements of $\tilde{\mathbf{T}}(\tau)$), while attempting to avoid over-dependence on individual estimates of small T_{ij}, which generally have unknown error. The error versus simulation time in WE- and Conv.-estimated $\tilde{\mathbf{T}}(\tau)$ are plotted in S12 Fig.

According to Eq 11, we estimate that our pipeline provided efficiency gains of 2 for ExMISA (Fig 2), 900 for Pluripotency Parameter Set I (Fig 3), and 1×10^6 for Parameter Set II (Fig 6). These numbers show that the pipeline can afford a significant speedup over conventional simulation in providing global dynamic information. The numbers further show that the efficiency gain is most pronounced for the Pluripotency network with exceedingly rare inter-phenotype transitions.

Discussion

In this work, we present a method for efficient, automated computation of epigenetic landscapes, metastable phenotypes, and phenotype-transition dynamics of stochastic GRN models. Our computational pipeline was inspired by studies of metastability and barrier-crossing in Molecular Dynamics, and our application of the pipeline to cell-scale networks addresses a number of current challenges for stochastic GRN dynamics. First, it overcomes the curse-of-dimensionality of complex models, by leveraging available rule-based modeling tools for stochastic biochemical networks [56]. Second, it overcomes the challenge of efficiently simulating stochastic systems with rare events, by using enhanced Weighted Ensemble rare-event sampling [45]. Third, it addresses the challenge of extracting and interpreting essential dynamics of complex systems on the basis of simulated trajectories, by using the Markov State Model framework [36] to automatically generate a compact, approximate representation of global system dynamics. Combining these tools into a unified pipeline provides an automated means of computing and visualizing essential stationary and dynamic properties of stochastic GRNs, including the number and identities (i.e. state-space mapping) of metastable phenotypes, their steady-state probabilities, and most-likely pathways of inter-phenotype transitions and their transition rates. By advancing the capability to compute and interpret hypothesized or experimentally-derived stochastic GRN models, the method can yield insight into how "local" stochastic, molecular processes involved in epigenetic regulation affect "global" dynamics such as phenotypic stability and fate-transitions in cells. Moreover, it can help close the gap between dynamic, molecular-detailed models of gene regulation and cell-population level experimental data, to inform rational cell reprogramming strategies.

Insights from the pluripotency network simulations

We used the pluripotency network as a model system to develop and demonstrate the simulation approach, but the results also yielded biological insights. For example, the simulations revealed a hierarchical structure of the epigenetic landscape. The network—exhibiting 5-6 metastable phenotypes—occupies a limited subspace from the vast possible gene combinations (e.g., $2^8 = 256$ possible distinct on/off combinations of gene expression states). The dominant feature of the global landscape is a high barrier/slow timescale between pluripotent and differentiated phenotypes. Within each of these categories, further sub-states were identified. The model revealed multi-timescale dynamics of phenotype transitions; the pluripotency network showed relatively rapid transitions between phenotype-states that differed in the expression-level (high vs. low) of a single gene, e.g. the high NANOG to low NANOG transition, whereas phenotype transitions involving a change in expression level of seven genes, e.g. the SC macrostate to the TE macrostate, occurred five orders of magnitude more slowly on average.

While the accessible phenotypes appear broadly similar across parameter sets, the relative stability and transition dynamics among phenotypes were sensitive to kinetic parameters governing transcription factor binding/unbinding. A global change in these parameters (affecting all individual transcription factor-DNA interactions equally) changed the shape of the landscape, altering the relative steady-state probabilities of different phenotypes and the likely transition pathways linking them. The DNA binding parameters capture the local epigenetic mechanisms that enable/disable transcription factors from accessing regulatory elements. A global rate change nevertheless has a varying influence on different genes because the number of regulators differs, as does the molecular logic by which activators and repressors exert combinatorial control on different genes. These results echo findings that global modification of chromatin regulators often have lineage-specific effects [63]. These results highlight both the need for, and the challenge of, informing cell reprogramming strategies with quantitative network models, as they suggest that the dynamic response of cellular networks to perturbations is governed by the detailed kinetics of molecular regulatory mechanisms, which are generally difficult to parameterize.

Dynamic definition of cell phenotype

The Markov State Model framework implicitly imposes a dynamic definition of cell phenotypes; the number of phenotypes was determined using spectral gap-analysis, and the coarse-graining algorithm automatically identified metastable aggregates (i.e., grouped sampled network states into larger clusters). This is different from the classifications of phenotypes that are generally used in analyzing experimental data, where gene expression or marker levels are often used to categorize cells. However, experiments have also revealed the potential need for a dynamic definition of cell phenotype, based not only on single-timepoint measurements of gene expression or phenotype-markers, but also on information from past or future timepoints [4, 8]. For example, Filipczyk *et al.* [8] identified distinct subpopulations within a compartment of NANOG-negative cells in mESCS, which differed in their propensity to re-express NANOG. At the same time, fluctuations between low- and high-NANOG expressing cells were not necessarily associated with any functional state change. The Markov State Model approach, based on kinetic/dynamic coarse-graining, thus provides a quantitative approach for classifying phenotype-states that is both completely generalizable rather than *ad hoc* (it requires no *a priori* knowledge or designation of markers/genes) and is in line with these recent experiments revealing the need for a dynamic definition of phenotype.

Timescales of stochastic phenotype transitions

Markovian transitions (i.e., memoryless "hops") among cell phenotypes have been observed experimentally: examples include transitions among phenotypes in cancer cells, as measured by flow cytometry [10], and among pluripotency-states in mESCs, as measured by time-lapse microscopy of fluctuating gene expression [7–9]. The compact nature of these data-inferred networks—showing hops among a limited set of broad phenotypes—suggests that the computed MSM framework advanced in this study provides an appropriate level of resolution at which to analyze GRN dynamics and may serve as a useful tool for comparing models to experimental data.

Experimental studies have quantified the timescales of Markovian transitions between NANOG-high and NANOG-low states in mESCs [8, 9]. From Hormoz *et al.*, the probability of transitioning from NANOG-high to NANOG-low in mESCs is 0.02 per cell cycle, while that of the reverse transition is 0.08. These values represent a relatively rapid transition rate, since NANOG expression is known to be particularly dynamic [60]. Similarly, plasticity has been observed in cancer cells where quantitative estimates of stochastic cell transitions between a stem cell cancer cell phenotype to a basal cancer cell phenotype were observed to be roughly on the order of 0.01 to 0.1 per cell cycle [10]. We can translate our model results to approximate biological timescales: the degradation rate, which sets the timeunit for model results (i.e., k is taken to be 1) was experimentally determined to be on the order of a few hours (in the E14 mouse embryonic stem cell line, the half-lives of NANOG, OCT4, and SOX2 are approximately 4.7, >6, and 1.6 hours, respectively [64]). Assuming that degradation is unimolecular, $k = \ln(2)/t_{[NANOG]1/2}$, and the half-life of NANOG, $t_{[NANOG]1/2} = 5$ hours, the degradation rate is $k = 0.1$. Using a mESC cell cycle time of 12 hours [65], the simulations for Parameter Set I then predict NANOG-high to NANOG-low transitions occurring with a rate of 0.03 per cell cycle, and of 3×10^{-3} for the reverse. For Parameter Set II, the computed rates were 8×10^{-6} and 5×10^{-5}, respectively. Comparison of these computed and experimental rates of NANOG transitions indicates that Parameter Set I ($f = 10$) is more in line with experimental observations, while Parameter Set II ($f = 50$) gives transition rates that are three orders of magnitude too slow. These results are in agreement with previous findings from theoretical studies that GRNs in pluripotency networks operate in a so-called "weakly-adiabatic" regime [24, 27, 28], in which the timescale of DNA-binding by transcription factors is on the order of transcription factor production and degradation.

Comparison to other models and computational approaches

A number of theoretical studies have elucidated dynamics of stochastic molecular-detailed GRN models (i.e., models that include molecular fluctuations and regulatory mechanisms, in contrast to Boolean models [66]). These studies have largely focused on small 1- or 2-gene motifs[[21–25, 32, 39]]. In the limit of slow DNA-binding/unbinding, it was shown that the stationary distribution of the stochastic model can be solved exactly [41]. Recent years have seen extension of stochastic methods to studies of more complex, experimentally derived GRN models encompassing $\mathcal{O}(10)$ genes. For example, determination of global dynamic properties of such networks has been achieved by combining information from long stochastic simulations of discrete models [27, 62], or of continuum SDE models, in combination with path integral approaches [58, 67]. The pluripotency network studied herein was developed by Zhang and Wolynes [28]; in their work, the authors developed a continuum approximation to the Chemical Master Equation that enabled quantitative construction of the epigenetic landscape. Here, we present an alternative approach that is unique in two major aspects: (1) the use of stochastic simulations (i.e., SSA [33]), which is enabled by use of the WE rare-event sampling algorithm, and (2) the automated Markov State Model framework for designating phenotypes and constructing a coarse-grained view of the epigenetic landscape. While we utilize a different framework (that of coarse-grained, discrete stochastic models) from Zhang and Wolynes to approximate and interpret dynamics, our results are broadly consistent with theirs. For example, the dominant identified phenotypes we found are the same as in their work (the only exception being the exceedingly rare LN2 phenotype identified by the coarse-graining algorithm for Parameter Set I).

Current challenges and future directions

Our approach is uniquely suited to extracting global dynamics information for stochastic systems with metastability, using simulations. An advantage of this approach is that both the WE and coarse-graining algorithms are“dynamics-agnostic” [47], meaning that they can be applied to any type of stochastic dynamics framework. In the context of computational biology, our pipeline could be extended to other types of stochastic biochemical systems, such as systems with hybrid discrete-continuum dynamics [68], systems with spatial heterogeneity [69], or multi-level models [70]. In addition to this flexibility, simulation-based methods have the advantage of being able to leverage existing, widely-used open-source packages, which in turn facilitate model specification and model sharing. For example, BioNetGen [56] can interpret models specified in the Systems Biology Markup Language [71].

Several challenges and potential weaknesses with the pipeline exist, both with regard to sampling rare events, and in determining an appropriate coarse-grained model. Potential challenges with the WE algorithm itself have been described elsewhere [35, 69], and include the difficulty of determining a binning that captures slow degrees of freedom and the existence of time-correlations between sampled iterations of the simulation, which can impede unbiased sampling. The Voronoi-based binning procedure we employ here is related to a number of similar approaches [24, 48–50], and has the advantage of effectively tiling a high-dimensional space without the need for *a priori* knowledge. However, in practice, according to others and our own studies, the method is effective up to about 10 degrees of freedom. Therefore, in larger gene networks (as in other complex systems) an ongoing challenge will be to identify optimal binning methods to effectively partition slow degrees of freedom and thus enable efficient enhanced sampling. New adaptive partitioning methods could also have the effect of improving the accuracy of coarse-grained Markov models, as finer partitioning of transition regions has been found to reduce errors in the Markovian approximation [42]. Additional improvements to efficiency, which could aid in scaling the method to larger networks, could be achieved in the future by using alternatives to the direct SSA algorithm (see e.g., [72]) or improved parallelization techniques.

Supporting information

S1 File. Description of network models, kinetic parameters, and weighted ensemble parameters.
(PDF)

S2 File. Pseudo-code for the computational pipeline.
(PDF)

S3 File. Details of efficiency gain estimate.
(PDF)

S1 Table. ExMISA network parameters.
(PDF)

S2 Table. Pluripotency network.
(PDF)

S3 Table. Pluripotency network parameters.
(PDF)

S4 Table. Weighted ensemble simulation parameters.
(PDF)

S5 Table. Transition matrices of metastable phenotype clusters (MSMs).
(PDF)

S6 Table. Computed mean first passage times in the ExMISA network—Comparison of different methods.
(PDF)

S7 Table. Computed mean first passage times of inter-phenotype transitions in the pluripotency network (Parameter Set I).
(PDF)

S1 Fig. Movement of Voronoi Centers during weighted ensemble sampling.
(PDF)

S2 Fig. Error in computed steady-state probability as a function of *N*, the number of protein states retained in the state-space truncation.
(PDF)

S3 Fig. Convergence of the flux of the transition between the polarized phenotype-states in the ExMISA network.
(PDF)

S4 Fig. Convergence of the slowest implied timescale t_2 with increasing number of sampling regions (bins) and increasing lagtime τ.
(PDF)

S5 Fig. The Chapman-Kolmogorov test on the four Markov State Model phenotypes of the sampled ExMISA network.
(PDF)

S6 Fig. Pathway decomposition for the SC → PE transition for $f = 10$.
(PDF)

S7 Fig. Validation of the SC → TE transition pathway calculated through weighted ensemble simulation.
(PDF)

S8 Fig. Reproducibility of the weighted ensemble sampling of the pluripotency network.
(PDF)

S9 Fig. Pathway decomposition for the SC → PE transition for $f = 50$.
(PDF)

S10 Fig. Convergence of the flux of the TE → SC transition in the pluripotency network with $f = 10$.
(PDF)

S11 Fig. Difference in Coarse-Grained clustering for the 2-gene ExMISA cell decision network studied through the numerical benchmark (top) and the WE sampling pipeline (bottom).
(PDF)

S12 Fig. Plotted errors in sampled $\tilde{\mathrm{T}}(\tau)$ for ExMISA.
(PDF)

Acknowledgments

We thank the administrators of the University of California, Irvine High-Performance Computing Cluster and we thank Jun Allard for helpful discussions.

Author Contributions

Conceptualization: Margaret J. Tse, Elizabeth L. Read.

Data curation: Margaret J. Tse.

Formal analysis: Margaret J. Tse, Brian K. Chu, Elizabeth L. Read.

Investigation: Margaret J. Tse, Brian K. Chu.

Methodology: Margaret J. Tse.

Project administration: Elizabeth L. Read.

Resources: Elizabeth L. Read.

Software: Margaret J. Tse, Brian K. Chu, Cameron P. Gallivan, Elizabeth L. Read.

Supervision: Elizabeth L. Read.

Validation: Margaret J. Tse, Cameron P. Gallivan.

Visualization: Margaret J. Tse.

Writing – original draft: Margaret J. Tse, Elizabeth L. Read.

Writing – review & editing: Margaret J. Tse, Brian K. Chu, Elizabeth L. Read.

References

1. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell. 2007; 131(5):861–872. https://doi.org/10.1016/j.cell.2007.11.019 PMID: 18035408
2. Abranches E, Guedes AMV, Moravec M, Maamar H, Svoboda P, Raj A, et al. Stochastic NANOG fluctuations allow mouse embryonic stem cells to explore pluripotency. Development. 2014; 141(14): 2770–2779. https://doi.org/10.1242/dev.108910 PMID: 25005472
3. Dietrich JE, Hiiragi T. Stochastic patterning in the mouse pre-implantation embryo. Development. 2007; 134(23):4219–4231. https://doi.org/10.1242/dev.003798 PMID: 17978007
4. Kalmar T, Lim C, Hayward P, Muñoz-Descalzo S, Nichols J, Garcia-Ojalvo J, et al. Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells. PLoS Biology. 2009; 7(7):e1000149. https://doi.org/10.1371/journal.pbio.1000149 PMID: 19582141
5. Singh AM, Hamazaki T, Hankowski KE, Terada N. A heterogeneous expression pattern for Nanog in embryonic stem cells. Stem Cells (Dayton, Ohio). 2007; 25(10):2534–2542. https://doi.org/10.1634/stemcells.2007-0126
6. Ohnishi Y, Huber W, Tsumura A, Kang M, Xenopoulos P, Kurimoto K, et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nature cell biology. 2014; 16(1):27–37. https://doi.org/10.1038/ncb2881 PMID: 24292013
7. Singer ZS, Yong J, Tischler J, Hackett JA, Altinok A, Surani MA, et al. Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. Molecular Cell. 2014; 55(2):319–331. https://doi.org/10.1016/j.molcel.2014.06.029 PMID: 25038413
8. Filipczyk A, Marr C, Hastreiter S, Feigelman J, Schwarzfischer M, Hoppe PS, et al. Network plasticity of pluripotency transcription factors in embryonic stem cells. Nature Cell Biology. 2015; 17(10): 1235–1246. https://doi.org/10.1038/ncb3237 PMID: 26389663
9. Hormoz S, Singer ZS, Linton JM, Antebi YE, Shraiman BI, Elowitz MB. Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. Cell Systems. 2016; 3(5): 419–433.e8. https://doi.org/10.1016/j.cels.2016.10.015 PMID: 27883889
10. Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, et al. Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. Cell. 2011; 146(4):633–644. https://doi.org/10.1016/j.cell.2011.07.026 PMID: 21854987
11. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, et al. A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. Cell. 2010; 141(1):69–80. https://doi.org/10.1016/j.cell.2010.02.027 PMID: 20371346
12. Acar M, Mettetal JT, van Oudenaarden A. Stochastic switching as a survival strategy in fluctuating environments. Nature Genetics. 2008; 40(4):471–475. https://doi.org/10.1038/ng.110 PMID: 18362885
13. Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S. Bacterial Persistence as a Phenotypic Switch. Science. 2004; 305(5690):1622–1625. https://doi.org/10.1126/science.1099390 PMID: 15308767
14. Waddington C, Kacser H. The Strategy of the Genes. Routledge; 1957.
15. Bhattacharya S, Zhang Q, Andersen ME. A deterministic map of Waddington's epigenetic landscape for cell fate specification. BMC Systems Biology. 2011; 5:85. https://doi.org/10.1186/1752-0509-5-85 PMID: 21619617
16. Wang J, Zhang K, Xu L, Wang E. Quantifying the Waddington landscape and biological paths for development and differentiation. Proceedings Of The National Academy Of Sciences Of The United States Of America. 2011; 108(21536909):8257–62. https://doi.org/10.1073/pnas.1017017108 PMID: 21536909
17. Huang S. The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology. 2012; 34(2):149–157. https://doi.org/10.1002/bies.201100031
18. Huang S, Eichler G, Bar-Yam Y, Ingber DE. Cell fates as high-dimensional attractor states of a complex gene regulatory network. Physical Review Letters. 2005; 94(12):128701. https://doi.org/10.1103/PhysRevLett.94.128701 PMID: 15903968
19. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic Gene Expression in a Single Cell. Science. 2002; 297(5584):1183–1186. https://doi.org/10.1126/science.1070919 PMID: 12183631
20. Kærn M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics. 2005; 6(6):451–464. https://doi.org/10.1038/nrg1615 PMID: 15883588

21. Aurell E, Sneppen K. Epigenetics as a First Exit Problem. Physical Review Letters. 2002; 88(4). https://doi.org/10.1103/PhysRevLett.88.048101 PMID: 11801174

22. Sasai M, Wolynes PG. Stochastic gene expression as a many-body problem. Proceedings of the National Academy of Sciences. 2003; 100(5):2374–2379. https://doi.org/10.1073/pnas.2627987100

23. Feng H, Wang J. A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes. Sci Rep. 2012; 2:550. https://doi.org/10.1038/srep00550 PMID: 22870379

24. Tse MJ, Chu BK, Roy M, Read EL. DNA-Binding Kinetics Determines the Mechanism of Noise-Induced Switching in Gene Networks. Biophysical Journal. 2015; 109(8):1746–1757. https://doi.org/10.1016/j.bpj.2015.08.035 PMID: 26488666

25. Ge H, Qian H, Xie XS. Stochastic Phenotype Transition of a Single Cell in an Intermediate Region of Gene State Switching. Physical Review Letters. 2015; 114(7):078101. https://doi.org/10.1103/PhysRevLett.114.078101 PMID: 25763973

26. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008; 453(7194):544–547. https://doi.org/10.1038/nature06965 PMID: 18497826

27. Sasai M, Kawabata Y, Makishi K, Itoh K, Terada TP. Time Scales in Epigenetic Dynamics and Phenotypic Heterogeneity of Embryonic Stem Cells. PLOS Computational Biology. 2013; 9(12):e1003380. https://doi.org/10.1371/journal.pcbi.1003380 PMID: 24348228

28. Zhang B, Wolynes PG. Stem cell differentiation as a many-body problem. Proceedings of the National Academy of Sciences. 2014; 111(28):10185–10190. https://doi.org/10.1073/pnas.1408561111

29. Chu BK, Tse MJ, Sato RR, Read EL. Markov State Models of gene regulatory networks. BMC Systems Biology. 2017; 11:14. https://doi.org/10.1186/s12918-017-0394-4 PMID: 28166778

30. Lipshtat A, Loinger A, Balaban NQ, Biham O. Genetic Toggle Switch without Cooperative Binding. Physical Review Letters. 2006; 96(18):188101. https://doi.org/10.1103/PhysRevLett.96.188101 PMID: 16712399

31. Schultz D, Walczak AM, Onuchic JN, Wolynes PG. Extinction and resurrection in gene networks. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(49): 19165–19170. https://doi.org/10.1073/pnas.0810366105 PMID: 19033463

32. Ma R, Wang J, Hou Z, Liu H. Small-Number Effects: A Third Stable State in a Genetic Bistable Toggle Switch. Physical Review Letters. 2012; 109(24):248107. https://doi.org/10.1103/PhysRevLett.109.248107 PMID: 23368390

33. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 1977; 81(25):2340–2361. https://doi.org/10.1021/j100540a008

34. Allen R, Valeriani C, ten Wolde P. Forward flux sampling for rare event simulations. Journal Of Physics-Condensed Matter. 2009; 21(000271268400004):463102. https://doi.org/10.1088/0953-8984/21/46/463102

35. Zuckerman DM, Chong LT. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. Annual Review of Biophysics. 2017; 46(1):43–57. https://doi.org/10.1146/annurev-biophys-070816-033834 PMID: 28301772

36. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences. 2009; 106(45):19011–19016. https://doi.org/10.1073/pnas.0905466106

37. Huang S. Reprogramming cell fates: reconciling rarity with robustness. BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology. 2009; 31(5):546–560. https://doi.org/10.1002/bies.200800189

38. Graf T, Enver T. Forcing cells to change lineages. Nature. 2009; 462(7273):587–594. https://doi.org/10.1038/nature08533 PMID: 19956253

39. Kepler TB, Elston TC. Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations. Biophysical Journal. 2001; 81(6):3116–3136. https://doi.org/10.1016/S0006-3495(01)75949-8 PMID: 11720979

40. Lin YT, Buchler NE. Efficient analysis of stochastic gene dynamics in the non-adiabatic regime using piecewise deterministic Markov processes. Journal of The Royal Society Interface. 2018; 15(138). https://doi.org/10.1098/rsif.2017.0804

41. Al-Radhawi MA, Del Vecchio D, Sontag ED. Multi-modality in gene regulatory networks with slow gene binding. arXiv preprint arXiv:170502330. 2017;.

42. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, et al. Markov models of molecular kinetics: Generation and validation. The Journal of Chemical Physics. 2011; 134(17):174105. https://doi.org/10.1063/1.3565032 PMID: 21548671

43. Scherer MK, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, et al. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. Journal of Chemical Theory and Computation. 2015; 11(11):5525–5542. https://doi.org/10.1021/acs.jctc.5b00743 PMID: 26574340

44. Hoel PG, Port SC, Stone CJ. Introduction to Stochastic Processes. Waveland Press; 1986.

45. Huber G, Kim S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. Biophysical Journal. 1996; 70(8770190):97–110. https://doi.org/10.1016/S0006-3495(96)79552-8 PMID: 8770190

46. Allen RJ, Warren PB, ten Wolde PR. Sampling Rare Switching Events in Biochemical Networks. Physical Review Letters. 2005; 94(1). https://doi.org/10.1103/PhysRevLett.94.018104

47. Donovan R, Sedgewick A, Faeder J, Zuckerman D. Efficient stochastic simulation of chemical kinetics networks using a weighted ensemble of trajectories. The Journal of chemical physics. 2013; 139(24070313):115105. https://doi.org/10.1063/1.4821167 PMID: 24070313

48. Dickson A, Warmflash A, Dinner AR. Nonequilibrium umbrella sampling in spaces of many order parameters. The Journal of Chemical Physics. 2009; 130(7):074104. https://doi.org/10.1063/1.3070677 PMID: 19239281

49. Dickson A, Brooks CL. WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. The Journal of Physical Chemistry B. 2014; 118(13):3532–3542. https://doi.org/10.1021/jp411479c PMID: 24490961

50. Zhang B, Jasnow D, Zuckerman D. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. The Journal of chemical physics. 2010; 132(20136305):054107. https://doi.org/10.1063/1.3306345 PMID: 20136305

51. Suárez E, Lettieri S, Zwier MC, Stringer CA, Subramanian SR, Chong LT, et al. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. Journal of Chemical Theory and Computation. 2014; 10(7):2658–2667. https://doi.org/10.1021/ct401065r PMID: 25246856

52. Dickson A, Warmflash A, Dinner AR. Separating forward and backward pathways in nonequilibrium umbrella sampling. The Journal of Chemical Physics. 2009; 131(15):154104. https://doi.org/10.1063/1.3244561 PMID: 20568844

53. Röblitz S, Weber M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. Advances in Data Analysis and Classification. 2013; 7(2):147–179. https://doi.org/10.1007/s11634-013-0134-6

54. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks; 2009. Available from: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

55. MATLAB and Parallel Computing Toolbox Release 2012b;.

56. Faeder JR, Blinov ML, Hlavacek WS. Rule-Based Modeling of Biochemical Systems with BioNetGen. In: Systems Biology. Methods in Molecular Biology. Humana Press; 2009. p. 113–167. Available from: https://link.springer.com/protocol/10.1007/978-1-59745-525-1_5.

57. Harrigan MP, Sultan MM, Hernández CX, Husic BE, Eastman P, Schwantes CR, et al. MSMBuilder: Statistical Models for Biomolecular Dynamics. Biophysical Journal. 2017; 112(1):10–15. https://doi.org/10.1016/j.bpj.2016.10.042 PMID: 28076801

58. Wang P, Song C, Zhang H, Wu Z, Tian XJ, Xing J. Epigenetic state network approach for describing cell phenotypic transitions. Interface Focus. 2014; 4(3). https://doi.org/10.1098/rsfs.2013.0068

59. Chambers I, Silva J, Colby D, Nichols J, Nijmeijer B, Robertson M, et al. Nanog safeguards pluripotency and mediates germline development. Nature. 2007; 450(7173):1230–1234. https://doi.org/10.1038/nature06403 PMID: 18097409

60. Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, et al. Nanog Is the Gateway to the Pluripotent Ground State. Cell. 2009; 138(4):722–737. https://doi.org/10.1016/j.cell.2009.07.039 PMID: 19703398

61. Hay DC, Sutherland L, Clark J, Burdon T. Oct-4 Knockdown Induces Similar Patterns of Endoderm and Trophoblast Differentiation Markers in Human and Mouse Embryonic Stem Cells. STEM CELLS. 2004; 22(2):225–235. https://doi.org/10.1634/stemcells.22-2-225 PMID: 14990861

62. Li C, Wang J. Quantifying Waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. Journal of The Royal Society Interface. 2013; 10(89):20130787. https://doi.org/10.1098/rsif.2013.0787

63. Constantinides PG, Jones PA, Gevers W. Functional striated muscle cells from non-myoblast precursors following 5-azacytidine treatment. Nature. 1977; 267(5609):364–6. https://doi.org/10.1038/267364a0 PMID: 68440

64. Abranches E, Bekman E, Henrique D. Generation and Characterization of a Novel Mouse Embryonic Stem Cell Line with a Dynamic Reporter of Nanog Expression. PLOS ONE. 2013; 8(3):e59928. https://doi.org/10.1371/journal.pone.0059928 PMID: 23527287

65. Wakayama T, Rodriguez I, Perry ACF, Yanagimachi R, Mombaerts P. Mice cloned from embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96(26):14984–14989. https://doi.org/10.1073/pnas.96.26.14984 PMID: 10611324

66. Chang R, Shoemaker R, Wang W. Systematic Search for Recipes to Generate Induced Pluripotent Stem Cells. PLoS Computational Biology. 2011; 7(12). https://doi.org/10.1371/journal.pcbi.1002300

67. Li C, Wang J. Quantifying the Landscape for Development and Cancer from a Core Cancer Stem Cell Circuit. Cancer Research. 2015; 75(13):2607–2618. https://doi.org/10.1158/0008-5472.CAN-15-0079 PMID: 25972342

68. Hepp B, Gupta A, Khammash M. Adaptive hybrid simulations for multiscale stochastic reaction networks. The Journal of Chemical Physics. 2015; 142(3):034118. https://doi.org/10.1063/1.4905196 PMID: 25612700

69. Donovan RM, Tapia JJ, Sullivan DP, Faeder JR, Murphy RF, Dittrich M, et al. Unbiased Rare Event Sampling in Spatial Stochastic Systems Biology Models Using a Weighted Ensemble of Trajectories. PLoS Computational Biology. 2016; 12(2). https://doi.org/10.1371/journal.pcbi.1004611 PMID: 26845334

70. Maus C, Rybacki S, Uhrmacher AM. Rule-based multi-level modeling of cell biological systems. BMC Systems Biology. 2011; 5:166. https://doi.org/10.1186/1752-0509-5-166 PMID: 22005019

71. Harris LA, Hogg JS, Tapia JJ, Sekar JA, Gupta S, Korsunsky I, et al. BioNetGen 2.2: advances in rule-based modeling. Bioinformatics. 2016; 32(21):3366–3368. https://doi.org/10.1093/bioinformatics/btw469 PMID: 27402907

72. Gillespie DT, Petzold LR. Improved leap-size selection for accelerated stochastic simulation. The Journal of Chemical Physics. 2003; 119(16):8229–8234. https://doi.org/10.1063/1.1613254