Fundamental limits on dynamic inference from single cell snapshots

Caleb Weinreb[1], Samuel Wolock[1], Betsabeh K. Tusi[2], Merav Socolovsky[2], Allon M. Klein[1]

1. Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
2. Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA

Author for correspondence: A.M.K. (Allon_Klein@hms.harvard.edu).

Abstract

Single cell profiling methods are powerful tools for dissecting the molecular states of cells, but the destructive nature of these methods has made it difficult to measure single cell expression over time. When cell dynamics are asynchronous, they can form a continuous manifold in gene expression space whose structure is thought to encode the trajectory of a typical cell. This insight has spurred a proliferation of methods for single cell trajectory discovery that have successfully ordered cell states and identified differentiation branch-points. However, all attempts to infer dynamics from static snapshots of cell state face a common limitation: for any measured distribution of cells in high dimensional state space, there are multiple dynamics that could give rise to it, and by extension, multiple possibilities for underlying mechanisms of gene regulation. Here, we enumerate from first principles the aspects of gene expression dynamics that cannot be inferred from a static snapshot alone, but nonetheless have a profound influence on temporal ordering and fate probabilities of cells. On the basis of these unknowns, we identify assumptions necessary to constrain a unique solution for the dynamics and translate these constraints into a practical algorithmic approach, called Population Balance Analysis (PBA). At its core, PBA invokes a new method based on spectral graph theory for solving a certain class of high dimensional differential equation. We show the strengths and limitations of PBA using simulations and validate its accuracy on single cell profiles of hematopoietic progenitor cells. Altogether, these results provide a rigorous basis for dynamic interpretation of a gene expression continuum, and the pitfalls facing any method of dynamic inference. In doing so they clarify experimental designs to minimize these shortfalls.

Introduction

Over the past few years, technologies for making genome-scale high-dimensional measurements on single cells have transformed our ability to discover the constituent cell states of tissues(1). These measurements enable a molecular dissection of biological tissues at the single cell level, across development, differentiation, disease onset, or in response to external stimuli. The most mature of these technologies, single cell RNA sequencing (scRNA-Seq), can be applied at relatively low cost to thousands and even tens of thousands of cells to generate an 'atlas' of cell states in tissues, while also revealing transcriptional gene sets that define these states(2, 3). Rapidly maturing technologies are also enabling single cell measurements of the epigenome(4), the proteome(5, 6), and the spatial organization of chromatin(7).

A more ambitious goal of single cell analysis is to describe dynamic cell behaviors, and by extension, to reveal dynamic gene regulation. Since high-dimensional single cell measurements are destructive to cells, they reveal only static snapshots of cell state. However, it has been appreciated that dynamic progressions of cell state can be indirectly inferred from population snapshots by methods that fit a curve or a tree to the continuous distribution of cells in high dimensional state space. To date, a multitude of methods have been published to address the problem of 'trajectory reconstruction' from single cell data. These methods have ordered events in cell differentiation(8-12), cell cycle(13), regeneration, and perturbation response(14). The most advanced algorithms have addressed increasingly complex cell-state topologies including branching trajectories(15).

Unfortunately, all attempts to infer dynamics from static snapshots of cell state face a common limitation: for any measured distribution of cells in high dimensional state space, there are multiple dynamics that could give rise to it, and by extension, multiple possibilities for underlying mechanisms of gene regulation. These limitations can apply even when sampling from multiple time points is possible. Put differently, *any* computational method that reports a definite prediction for cell-state dynamics has made one choice among many about how to order observed cell states, whether or not the choice is made explicitly. To our knowledge, existing approaches rely on heuristic algorithms that do not explicitly state how bioinformatic decisions impact descriptions of biological dynamics. As such, the best methods for dynamic inference might be more accurately described as methods for non-linear dimensionality reduction, or 'manifold discovery': they robustly solve the problem of how to describe a static continuum of cell states using a small number of coordinates (often described as 'pseudo-time' coordinates), but they provide little or no guidance on how the observed static continuum (or 'pseudo-time') should be interpreted with respect to the many redundant dynamic processes that could give rise to it. Therefore, what assumptions must be made in dynamic inference – once the important task of manifold discovery is completed – remains an unsolved problem.

The difference between describing a manifold and describing its underlying biological dynamics becomes clear when considering the types of predictions one might make from data. Heuristic algorithms may be sufficient to provide an intuition for the biology, leaving the researchers to form hypotheses based on data exploration. However, quantitative predictions about cell behavior may require stronger forms of dynamical prediction. We would be curious to know, for example, how real time relates to progression along 'pseudo-time', and how we should think about cell dynamics in the absence of a clear linear or branching structure. Our intuition also tells us that single cell data could clarify how transcriptional programs influence cell fate bias in multi-lineage differentiation systems, but it is not yet clear how to extract such information in a principled manner. The ambiguity of the biological dynamics associated with manifold descriptions becomes even more important in studies seeking to infer mechanisms underlying those dynamics, since statements about mechanism necessarily entail specific hypotheses about cell trajectories. It follows that the limits on dynamic inference from single cell snapshots also affect attempts to reverse engineer gene regulatory networks(11) or to define “landscapes”(10) that confine cell dynamics in gene expression space.

Here we explore if one can derive a framework for inferring cell state dynamics from static snapshots that overcomes the above ambiguities by making explicit biological assumptions and identifying key fitting parameters that cannot be inferred from single cell data alone. With many algorithms now available for trajectory reconstruction from single cell data, our first focus is to define the limits of identifiability faced by *any* algorithm.

The second focus of this paper is to develop a practical algorithm for dynamic inference, which we call Population Balance Analysis (PBA). At one level, PBA provides a continuum description of cell states, just as existing methods, and it can similarly be used in a purely heuristic manner to order cell states. However, rather than focusing on manifold discovery, our approach builds instead on a first-principles biophysical description of stochastic gene expression. PBA therefore differs from existing algorithms in that it formally solves a problem of dynamic inference, and can thus be considered predictive of cell dynamics under clearly stated assumptions. Developing PBA required overcoming a computational challenge, which represents the major technical contribution of this work. In particular, the biophysical foundation of PBA is embodied by a diffusion-drift equation over high-dimensional space, which, though simple to define, cannot be practically solved using established computational tools. We therefore invoke a novel, asymptotically exact and highly efficient solution to diffusion-drift equations using recent innovations in spectral graph theory. The ubiquity of diffusion-drift equations in fields of quantitative biology, physics and chemistry suggest that applications of these methods may exist in other fields.

By making all biological assumptions explicit, PBA gains the ability to generate alternative predictions on system behavior for candidate hypotheses. For example, it assigns each transcriptional state a set of testable fate probabilities. The best fit to experimental data can then be used to identify which hypotheses accurately capture the systems behavior, and which do not. The explicit assumptions of PBA also clarify how to design experiments that optimize the quantitative accuracy of dynamic inference. We demonstrate the accuracy of PBA inference on simulated data. We then apply it to single cell RNA-seq data of hematopoietic progenitor cells (HPCs), reconciling scRNA-seq data on HPCs with fate assays made over the past few decades in this system. Extensive validation of novel PBA predictions in HPCs forms the subject of a second paper (Tusi, Wolock, Weinreb et al., *in submission*).

Results

A first-principles relationship of cell dynamics to static observations

When reconstructing a dynamic process from single cell snapshot data, cells are typically observed in a continuous spectrum of states owing to asynchrony in their dynamics. The goal is to reconstruct a set of rules governing possible dynamic trajectories in high-dimensional space that are compatible with the observed distribution of cell states. The inferred rules could represent a single curve or branching process in gene expression space, or they could reflect a more probabilistic view of gene expression dynamics. In some cases, multiple time points can be collected to add clarity to the temporal ordering of events. In other cases, a single time point could capture all stages of a dynamic process, such as in steady-state adult tissue turnover.

To develop a framework for dynamic reconstruction from first principles, we wish to identify a general, model-independent, mathematical formulation linking cell dynamics to static observations. One possible starting point is the *population balance equation*, (also known as the *flux balance law(16)*), which has the form:

$$\frac{\partial c}{\partial t} = -\nabla(c\boldsymbol{v}) + Rc \tag{1}$$

This partial differential equation provides a useful starting point for analysis, because it fully describes how the density of cells at a point in gene expression space depends on the speed and direction of travel of cells. Formally, Eq. (1) states that in each small region of gene expression space, the rate of change in the number of cells (left-hand side of the equation) equals the net cell flux into and out of the region (right-hand side) (1A). The equation introduces the cell density, $c(x,t)$, which is the distribution of cell states from which we sample a static snapshot of cells in an experiment. This density depends on the net average velocity, $\boldsymbol{v}(x)$, of the cells at point x, which is a feature of the dynamics that we wish to infer. Notably, being an average quantity, $\boldsymbol{v}$ is not necessarily a description of the dynamics of any individual cell, but it alone governs the form of the sampled cell density c. Eq. (1) also introduces a third variable: $R(x)$ is a rate of cell accumulation and loss at point x caused by the discrete phenomena of cell proliferation and cell death, and by entrance and exit from the tissue being isolated for analysis.

Though general, Eq. (1) nonetheless introduces some specific assumptions about the nature of cell state space. First, it approximates cell state attributes as continuous variables, though they may in fact represent discrete counts of molecules such as mRNAs or proteins. Second, it assumes that changes in cell state attributes are continuous in time, so that, for example, the sudden appearance or disappearance of many biomolecules at once cannot be described in this framework.

Multiple dynamic trajectories can generate the same high dimensional population snapshots

Given knowledge of the cell population density, $c(x,t)$, we hope to infer the underlying dynamics of cells by solving for the average velocity field $\boldsymbol{v}$ in Eq. (1). This approach falls short, however, because $\boldsymbol{v}$ is not fully determined by Eq. (1), and even if it were, knowing the average velocity of cells still leaves some ambiguity in the

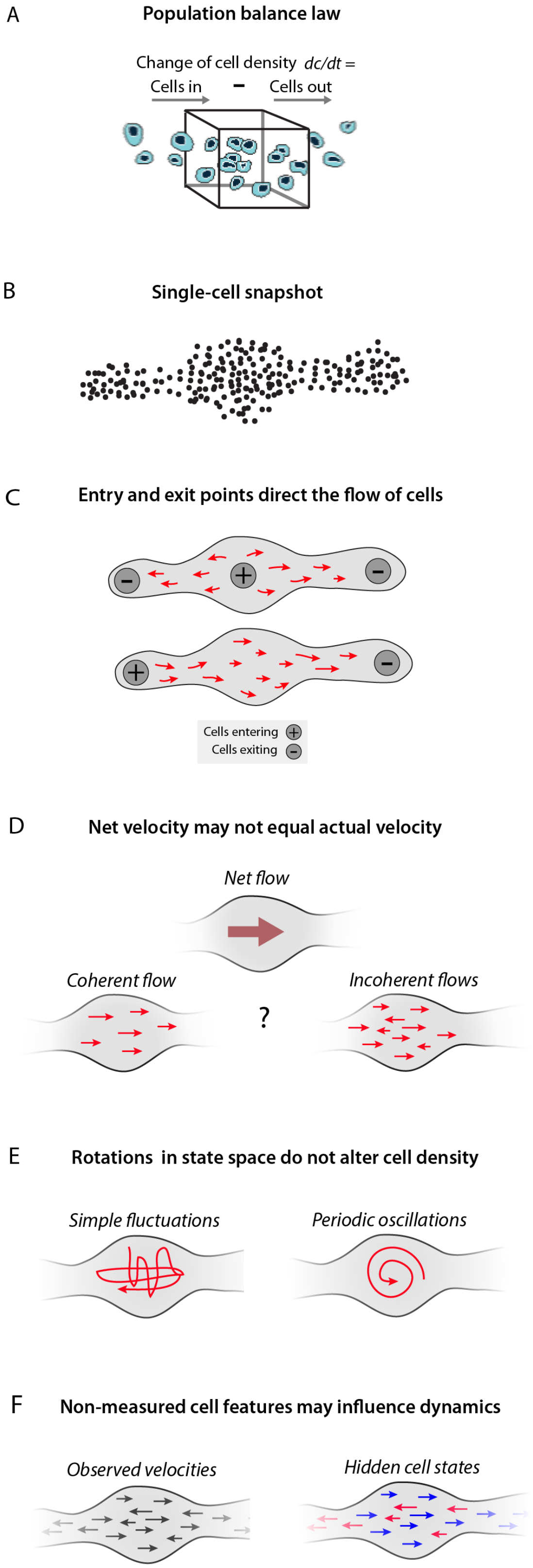

A
Population balance law
Change of cell density dc/dt =
Cells in − Cells out
B
Single-cell snapshot
C
Entry and exit points direct the flow of cells
Cells entering +
Cells exiting −
D
Net velocity may not equal actual velocity
Net flow
Coherent flow
?
Incoherent flows
E
Rotations in state space do not alter cell density
Simple fluctuations
Periodic oscillations
F
Non-measured cell features may influence dynamics
Observed velocities
Hidden cell states

Figure 1: Principles of dynamic inference from static snapshots. A starting point for inferring cell dynamics from high-dimensional snapshots is the population balance law (A), which states that in each small region of gene expression space, the rate of change in cell density equals the net cell flux into and out of the region. However, this law alone does not determine a unique solution for the dynamics because there are several phenomena that cannot be directly inferred from a static snapshot (B). For example, high-dimensional cell state measurements do not disclose the rates of cell entry and exit at different points in gene expression space (C), or the extent to which single-cell trajectories are coherent or stochastic (D). Static snapshots also cannot distinguish periodic oscillations of cell state from simple fluctuations that do not have a consistent direction and periodicity (E). Furthermore, there may be stable properties of a cell – such as epigenetic state – that affect its behavior but are not detectable by expression measurements alone (F).

specific trajectories of individual cells. This raises the question: does there exist a set of reasonable assumptions that constrain the dynamics to a unique solution? To explore this question, we enumerate the causes of non-uniqueness in cell state dynamics, using a cartoon to introduce each cause (detailed in Figure 1), as well as referring to their mathematical foundation in Eq. (1).

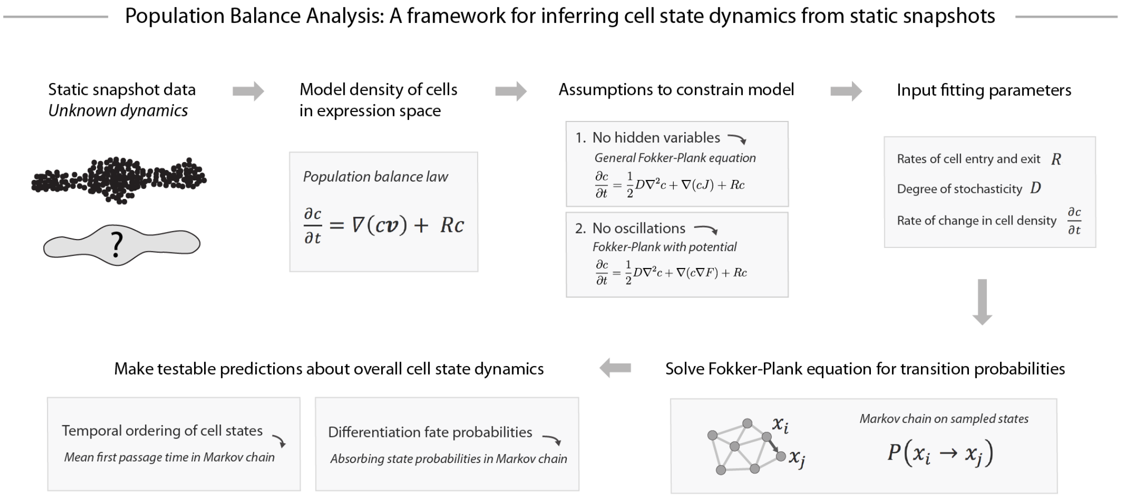
> 1) *Assumed cell entry and exit points strongly influence inferred dynamics:* For the same data, making different assumptions about the rates and location of cell entry and exit lead to fundamentally different inferences of the direction of cell progression in gene expression space, as illustrated in Figure 1C. Cells can enter a system by proliferation, by physically migrating into the tissue that is being analyzed, or more mundanely by up-regulating selection markers used for sample purification (e.g. cell surface marker expression). Similarly, cells exit observation by cell death, physical migration out of the tissue being studied, or by down-regulation of cell selection markers. These events could be associated with particular gene expression states, or could occur broadly. Referring to Eq. (1), this discussion is formally reflected in the need to assume a particular form for the rate field $R(x)$ when inferring dynamics $\boldsymbol{v}$ from the observed cell density c.
>
> 2) *Net velocity does not equal actual velocity:* A second unknown is the stochasticity in cell state dynamics, reflected in the degree to which cells in the same molecular state will follow different paths going forward. A net flow in gene expression space could result from imbalanced flows in many directions or from a single coherent flow in one direction (see Figure 1D). If the goal of trajectory analysis is to go beyond a description of what states exist and make predictions about the future behavior of cells (e.g. fate biases) given their current state, then it is necessary to account for the degree of such incoherence of dynamics. More or less stochastic cell behaviors also change inferences that might be made about underlying gene regulatory networks. Referring to Eq. (1), the net velocity field $\boldsymbol{v}$ reflects only the mean cell behavior, with individual cells deviating from the mean owing to stochastic gene expression.
>
> 3) *Rotations and oscillations in state space do not alter cell density:* Static snapshot data cannot distinguish periodic oscillations of cell state from simple fluctuations that do not have a consistent direction and periodicity (Figure 1E). As with incoherent motion above, predictive models may need to explicitly consider oscillatory behaviors. The inability to detect oscillations from snapshot data is formally reflected in Eq. (1) by invariance of the concentration c to the addition of arbitrary rotational velocity fields $\boldsymbol{u}$ satisfying $\nabla(c\boldsymbol{u}) = 0$.
>
> 4) *Hidden features of cell state can lead to a superposition of different dynamic processes:* Stable properties of cell state that are invisible to single cell expression measurements, such as chromatin state or tissue location, could nonetheless impact cell fate over multiple cell state transitions (Figure 1F). The existence of such long-term “hidden variables” would clearly compromise attempts to predict the future fate of a cell from its current gene expression state. Previously

published algorithms for trajectory inference do not consider long-term hidden variables. This choice is inescapable for any modeling approach based on single cell RNA-seq or mass-cytometry data, since these measurement modalities simply do not capture every feature of a cell's molecular state.

In summary, we have enumerated the reasons that no unique solution exists for dynamic inference. However, sensible predictions about dynamics can still be made by introducing certain explicit assumptions, as we now describe.

Construction of the Population Balance Analysis framework

To infer cell dynamics from an observed cell density c, we will make several assumptions that together are necessary and sufficient to constrain a unique solution given the unknowns outlined above (see Figure 2).



Population Balance Analysis: A framework for inferring cell state dynamics from static snapshots

Static snapshot data
Unknown dynamics

Model density of cells in expression space

Population balance law

$$\frac{\partial c}{\partial t} = \nabla(c\boldsymbol{v}) + Rc$$

Assumptions to constrain model

1. No hidden variables
General Fokker-Plank equation
$$\frac{\partial c}{\partial t} = \frac{1}{2}D\nabla^2 c + \nabla(cJ) + Rc$$

2. No oscillations
Fokker-Plank with potential
$$\frac{\partial c}{\partial t} = \frac{1}{2}D\nabla^2 c + \nabla(c\nabla F) + Rc$$

Input fitting parameters

Rates of cell entry and exit R

Degree of stochasticity D

Rate of change in cell density $\frac{\partial c}{\partial t}$

Solve Fokker-Plank equation for transition probabilities

Markov chain on sampled states

$$P(x_i \rightarrow x_j)$$

Make testable predictions about overall cell state dynamics

Temporal ordering of cell states
Mean first passage time in Markov chain

Differentiation fate probabilities
Absorbing state probabilities in Markov chain

Figure 2: Population Balance Analysis
Though many dynamics are consistent with a given static snapshot of cell states, appropriate assumptions and fitting parameters can constrain a unique solution. We have developed an approach for calculating this solution called Population Balance Analysis (PBA). The first step is applying the population balance law, which relates the observed cell density to the average velocity and entrance/exit rate of cells. Next, assuming that there are no hidden variables allows application of the Fokker-Plank equation, and assuming that there are no rotations introduces a potential landscape. The potential landscape is related to the cell density through fitting parameters that can be directly measured or inferred from prior knowledge. With these parameters and sampled cell states as input, the PBA algorithm outputs transition probabilities for each pair of observed states, which can then be used to compute dynamic properties such as temporal ordering and fate potential.

The Fokker-Plank equation models memory-less cell state dynamics

The first assumption is that there are no hidden variables, meaning the properties of the cell available for measurement (such as its mRNA content) fully encode a probability distribution over its possible future states. This assumption is made implicitly by all current approaches to trajectory analysis and cell fate prediction, and we reflect on its plausibility in the discussion.

An equivalent statement of this first assumption is that cell trajectories are memory-less with respect to their measured properties, i.e. past states of the cell do not affect its future states other than through having led to its present state. If so, Eq. (1) can be approximated as a Fokker-Plank equation, which can be thought of as the continuum approximation of the chemical master equation (CME) that specifies the discrete stochastic molecular interactions underlying gene regulatory networks in the cell(17). In the Fokker-Plank formalism, cell trajectories are modeled as biased random walks, with a deterministic component that reflects the reproducible aspects of cell state changes such as their differentiation through stereotypical sequences of states, and a stochastic component that reflects random fluctuations in cell state, partly driven by bursty gene-expression, fluctuations in cellular environment, and intrinsic noise from low molecular number processes.

Fokker-Plank equations, which represent special cases of the Population Balance equation [Eq. (1)], have been applied previously to low dimensional biological processes, such as differentiation with a handful of genes(18) or a one-parameter model of cell cycle progression(13). Here, we apply them to high-dimensional data. Although Fokker-Plank descriptions are necessarily approximations, their emergence from first-principles descriptions of transcriptional dynamics(17), and their ubiquity in describing chemical reaction systems(19), justify their use instead of the more general form of Eq. (1). Specifically, the Generalized Fokker-Plank approximation takes the form of Eq. (1) with velocity field, $\boldsymbol{v} = \boldsymbol{J} - \frac{1}{2} D \nabla \log c$, where the first term is a deterministic average velocity field, and the second term is a stochastic component of the velocity that follows Fickian diffusion with a diffusion matrix D (Figure 2). We assume here that D is isotropic and invariant across gene expression space. Though more complex forms of diffusion could better reflect reality, we propose that this simplification for D is sufficient to gain predictive power from single cell data in the absence of specific data to constrain it otherwise. The resulting Population Balance equation is thus,

$$\frac{\partial c}{\partial t} = \frac{1}{2} \nabla (D \nabla c) - \nabla (c\boldsymbol{J}) + Rc. \qquad (2)$$

Eq. (2) explains the rate of change of cell density ($\partial c / \partial t$) as a sum of three processes: (1) stochastic gene expression, $\frac{1}{2} \nabla (D \nabla c)$, which causes cells to diffuse out of high-density regions in gene expression space; (2) convergences (and divergences) of the mean velocity field, $\nabla (c\boldsymbol{J})$, which cause cells to accumulate (or escape) from certain gene expression states over time; and (3) as before, cell entry and exit rates, Rc, will cause certain cell states to gain or lose cells over time.

Potential landscapes define a minimal model for dynamic inference

Our second assumption is that there are no oscillatory gene expression dynamics, which would appear as rotations in gene expression space. Though oscillations certainly do exist in reality – for example, the cell cycle – it is impossible to establish their existence from static snapshots alone. One is therefore forced to make an a priori assumption about their existence, for example by specifically searching for signatures of a known oscillatory process in the data. For processes not known to be oscillatory, one can begin by making predictions of fate bias and temporal ordering while ignoring oscillatory phenomena. The utility of such predictions is supported by our analysis of single cell RNA-seq data in a later section.

In the Fokker-Plank formalism, the presumed absence of oscillations implies that the velocity field $\boldsymbol{J}$ is the gradient of a potential function F (i.e. $\boldsymbol{J} = -\nabla F$). The potential would define a landscape in gene expression space, with cells flowing towards minima in the landscape, akin to energy landscapes in descriptions of physical systems. Applying the potential landscape assumptions to Eq. (2) gives rise to the simplified diffusion-drift equation below, where the potential is represented by a function $F(x)$.

$$\frac{\partial c}{\partial t} = \frac{1}{2} D \nabla^2 c + \nabla(c \nabla F) + Rc. \tag{3}$$

A recipe for dynamic inference from first principles

Equation (3) represents our best attempt to relate an observed density of cell states (c) to an underlying set of dynamical rules, now represented by a potential landscape (F) rather than the exact velocity field $\boldsymbol{v}$. Crucially, we have in these first few results sections: explained why the net cell velocity $\boldsymbol{v}$ is inherently unknowable; clarified why the description provided by a potential field F is the best that *any* method could propose without further knowledge about the system; and identified critical fitting parameters (D, R and $\partial c/\partial t$), that are not revealed by single cell snapshot measurements, but are required for determining aspects of the dynamics such as temporal ordering of states and fate probabilities during differentiation. By starting from first principles, it becomes clear that these requirements are not limited to any particular algorithm; they afflict any method one might develop for trajectory inference.

The challenge is now to develop a practical approach that relates the fitting parameters D, R and $\partial c/\partial t$ to dynamic predictions through Eq. (3). In the following, we focus on steady-state systems where $\partial c/\partial t = 0$, and, for the cases we analyze, we use prior literature to estimate R. We report results for a range of values of D. Building on the work here, more elaborate approaches could be taken, for example determining R from direct measurements of cell division and cell loss rates or integrating data from multiple time points to estimate $\partial c/\partial t$, thus generalizing to non-steady-state systems.

Reducing to practice: solving the Population Balance equation with spectral graph theory

Equipped with single cell measurements and estimates for each fitting parameter, we now face two practical problems in using of Eq. (3) to infer cell dynamics: the first is that Eq. (3) is generally high-dimensional (reflecting the number of independent gene programs acting in a cell), but numerical solvers cannot solve diffusion equations on more than perhaps ten dimensions. Indeed, until now, studies that used diffusion-drift equations such as Eq. (3) to model trajectories(10, 13, 18) were limited to one or two dimensions, far below the intrinsic dimensionality of typical single cell RNA-seq data(20). The second practical problem is that we do not in fact measure the cell density *c*: we only sample a finite number of cells from this density in an experiment.

Overcoming these problems represents the main technical contribution of this paper. We drew on a recent theorem by Ting, Huang and Jordan in spectral graph theory(21) to extend diffusion-drift modeling to arbitrarily high dimension. The core technical insight is that an asymptotically-exact solution to Eq. (3) can be calculated on a nearest-neighbor graph constructed with sampled cells as nodes, rather than on a low-dimensional sub-space of gene expression as performed previously (e.g.(13)). Our approach, which we call Population Balance Analysis (PBA) actually *improves* in accuracy as dimensionality increases, rewarding high-dimensional measurements. We thus avoid conclusions based on low-dimensional simplifications of data, which may introduce distortions into the analysis. The supplement of this paper provides technical proofs and an efficient framework for PBA in any high-dimensional system.

The inputs to PBA are a list of sampled cell states $\boldsymbol{x} = (x_1, \dots x_N)$, an estimate $R = (R_1, \dots R_N)$ for the net rate of cell accumulation/loss at each state x_i, and an estimate for the diffusion parameter D. (We are assuming steady-state, so $\partial c/\partial t = 0$). The output of PBA is a discrete probabilistic process (Markov chain) that describes the transition probabilities between the states x_i. The analysis is asymptotically exact in the sense that – if a potential exists and the estimates for R and D are correct – then the inferred Markov chain will converge to the underlying continuous dynamical process in the limit of sampling many cells ($N \to \infty$) (Theory supplement, Theorem 4).

PBA computes the transition probabilities of the Markov chain using a simple algorithm, which at its core involves a single matrix inversion. Briefly:

1. Construct a *k*-nearest-neighbor (knn) graph *G*, with one node at each position $\boldsymbol{x}_i$ extending edges to the *k* nearest nodes in its local neighborhood. Calculate the graph Laplacian of *G*, denoted *L*.
2. Compute a potential $V = L^{+}R$, where L^{+} is the pseudo-inverse of L
3. To each edge $(x_i \to x_j)$, assign the transition probability

$$P(x_i \to x_j) \sim \begin{cases} e^{(V_i - V_j)/D} & \text{if } (x_i, x_j) \text{ is an edge in } G \\ 0 & \text{if } (x_i, x_j) \text{ is not an edge in } G \end{cases}$$

With the Markov chain available, it is possible to calculate the temporal ordering of states (via mean first passage time), and the fate biases of progenitor cells in a differentiation process (via absorbing state probabilities), by integrating across many trajectories (see Figure 2). These calculations are simple, generally requiring a single matrix inversion. Specific formulas are provided in the Theory Supplement Section 3. Code for implementing these and other aspects of PBA is available online at https://github.com/AllonKleinLab/PBA.

PBA accurately reconstructs dynamics of simulated differentiation processes

We tested PBA on a sequence of simulations, first using an explicit model of diffusion-drift process, and then moving on to direct simulations of gene regulatory networks. In the first simulation (Figure 3; Supplementary Figure 1), cells drift down a bifurcating potential landscape into two output lineages. Cell trajectories span a 50-dimensional gene expression space (two of which are shown in Fig. 3A). With 200 cells sampled from this simulated system (Figure 3B), PBA predicted cell fate probabilities and temporal ordering of the measured cells. PBA made very accurate predictions (Pearson correlation, *ρ > 0.96*, Fig. 1C-D) if provided with correct estimates of proliferation, loss and stochasticity (parameters *R* and *D*). Estimates of temporal ordering remained accurate with even 5-fold error in these parameters ($\rho > 0.93$), but predictions of fate bias degraded ($\rho > 0.77$; Supplementary Figure 2A-D). Thus even very rough knowledge of the entry/exit points in gene expression space is sufficient to generate a reasonable and quantitative description of the dynamics. Interestingly, PBA also remained predictive in the presence of implanted oscillations (Supplementary Figure 3, fate probability $\rho >$ 0.9; temporal ordering $\rho > 0.8$). In addition, the simulations confirmed the theoretical prediction that inference quality improves as the number of noisy genes (dimensions) increases, and as more cells are sampled: maximum accuracy in

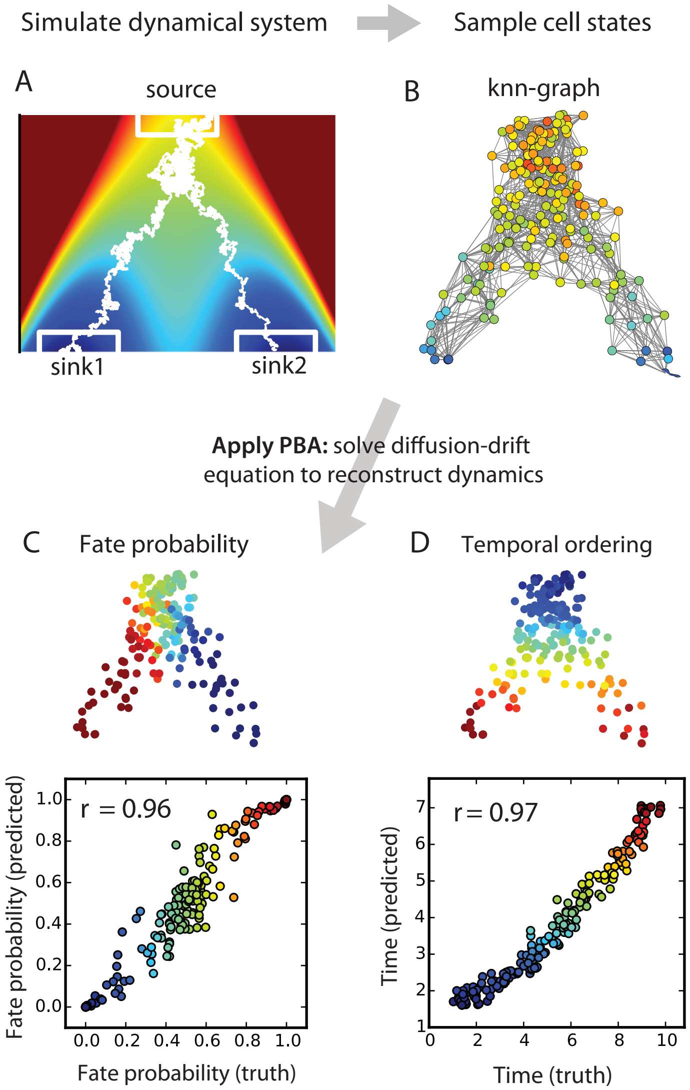


Figure 3: Demonstration of PBA on a simulated high-dimensional differentiation process

(A) Cells emerge from a proliferating bi-potent state (source) and differentiate into one of two fates (sinks 1 and 2) in a high-dimensional gene expression space, with two dimensions shown. Heat map colors show a potential field containing the cell trajectories. Example trajectories shown in white. (B) Static expression profiles sampled asynchronously through differentiation serve as the input to PBA, which reconstructs trajectories and accurately predicts future fate probabilities (C) and timing (D) of each cell.

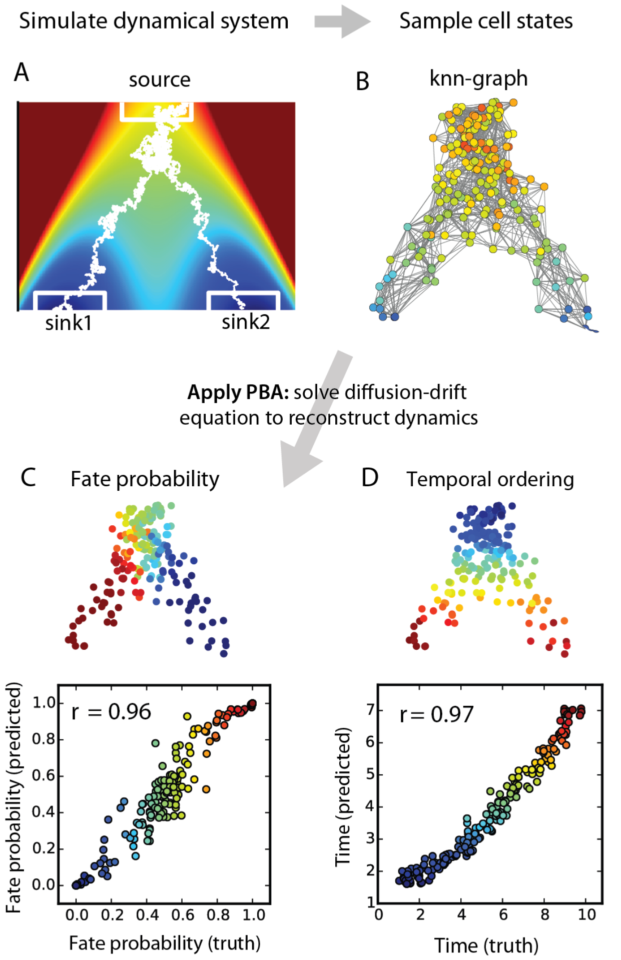


Figure 3: Demonstration of PBA on a simulated high-dimensional differentiation process
(A) Cells emerge from a proliferating bi-potent state (source) and differentiate into one of two fates (sinks 1 and 2) in a high-dimensional gene expression space, with two dimensions shown. Heat map colors show a potential field containing the cell trajectories. Example trajectories shown in white. (B) Static expression profiles sampled asynchronously through differentiation serve as the input to PBA, which reconstructs trajectories and accurately predicts future fate probabilities (C) and timing (D) of each cell.

this simple case was reached after ~100 cells and 20 dimensions (Supplementary Figure 2e-g). These simulations showcase the ability of PBA to not just *describe* continuum trajectories, but to additionally *predict* cell dynamics and by extension cell fate.

Having demonstrated the accuracy of PBA on an explicit model of a diffusion-drift process, we next tested its performance on gene expression dynamics arising from gene regulatory networks (GRNs) (Figure 4). As before, we simulated cell trajectories, obtained a static snapshot of cell states, and supplied PBA with this static snapshot as well as the parameter R encoding the location of entry and exit points. We began with a simple GRN representing a bi-stable switch, in which two genes repress each other and activate themselves (Figure 4A). Simulated trajectories from this GRN begin with both genes at an intermediate expression level, but quickly progress to a state where one gene dominates the other (Figure 4B). In addition to the two genes of the GRN, we included 48 uncorrelated noisy dimensions. With 500 cells sampled from this process, PBA predicted cell fate bias and temporal ordering very well (r>0.98 for fate bias and r > 0.89 for ordering; Figure 4C), though the precise accuracy depended on the assumed level of diffusion D (Supplementary Figure 4).

PBA assumes the absence of oscillations in gene expression space. Therefore, it is unclear how well PBA can infer cell trajectories that result from GRNs with oscillatory dynamics. We simulated an oscillatory GRN in the form of a “repressilator” circuit(22) with the addition of positive feedback loops that create two “escape routes” leading to alternative stable fixed points of the dynamics (Figure 4D). Simulated trajectories from this GRN begin with all genes oscillating, followed by a stochastic exit from the oscillation when one of the genes surpasses a threshold level (Figure 4E). With 500 cells sampled from this process, PBA was significantly less accurate than for the previous simulations (Figure 4F). Though PBA correctly identified which cells were fully committed to the two 'escape routes', it was entirely unable to resolve the fate biases of cells in the uncommitted oscillatory state. PBA also made poor predictions of mean first passage time, underestimating the amount of time that cells spent in the oscillatory state. Thus, when the assumptions of PBA are strongly violated, its prediction accuracy suffers.

PBA predictions of fate bias in hematopoiesis reconcile past experiments

The simulations suggest that PBA accurately reconstructs dynamics for systems that satisfy its underlying assumptions, but that it makes poor predictions when these assumptions are strongly violated. It remains unclear whether PBA will be accurate when applied to experimental data from real biological systems. We tested PBA on single cell gene expression measurements of 3,803 adult mouse hematopoietic progenitor cells (HPCs) from another study by our groups (Tusi, Wolock, Weinreb et al., *in submission*).

HPCs reside in the bone marrow and participate in steady-state production of blood and immune cells through a balance of self-renewal and multi-lineage differentiation. Fate commitment of HPCs is thought to occur through a series of hierarchical fate choices, investigated over the past four decades through live cell tracking, in vitro colony-forming assays and transplantation of defined sub-populations of HPCs(23). Depictions of the HPC hierarchy invoke a tree structure, with gradual lineage-restriction at branch points. However the precise tree remains controversial(24, 25), since existing measurements of fate potential reflect a patchwork of defined HPC subsets that may have internal heterogeneity(26) and provide only incomplete coverage of the full HPC pool. We asked whether PBA applied to single cell RNA profiling of HPCs could generate predictions consistent with experimental data, and possibly help resolve these controversies by providing a global map of approximate cell-fate biases of HPCs.

The single cell expression measurements – derived from mouse bone marrow cells expressing the progenitor marker Kit – represent a mixture of multipotent progenitors as well as cells expressing lineage commitment markers at various stages of maturity. Since PBA prescribes analysis of a k-nearest-neighbor (knn) graph of the cells, we developed an interactive knn visualization tool for single cell data exploration, called SPRING, (kleintools.hms.harvard.edu/tools/spring.html; (27)). The SPRING plot (Figure 5A) revealed a continuum of gene expression states that pinches off at different points to form several downstream lineages. Known marker genes (Supplementary Table 1) identified the graph endpoints as monocytic (Mo), granulocytic (G), dendritic (D), lymphoid (Ly), megakaryocytic (Mk), erythroid (Er) and basophil (Ba) progenitors (Supplementary Figure 5); we also identified cells in the graph expressing HSC markers. The lengths of the branches reflect the timing of Kit down-regulation and the abundance of each lineage.

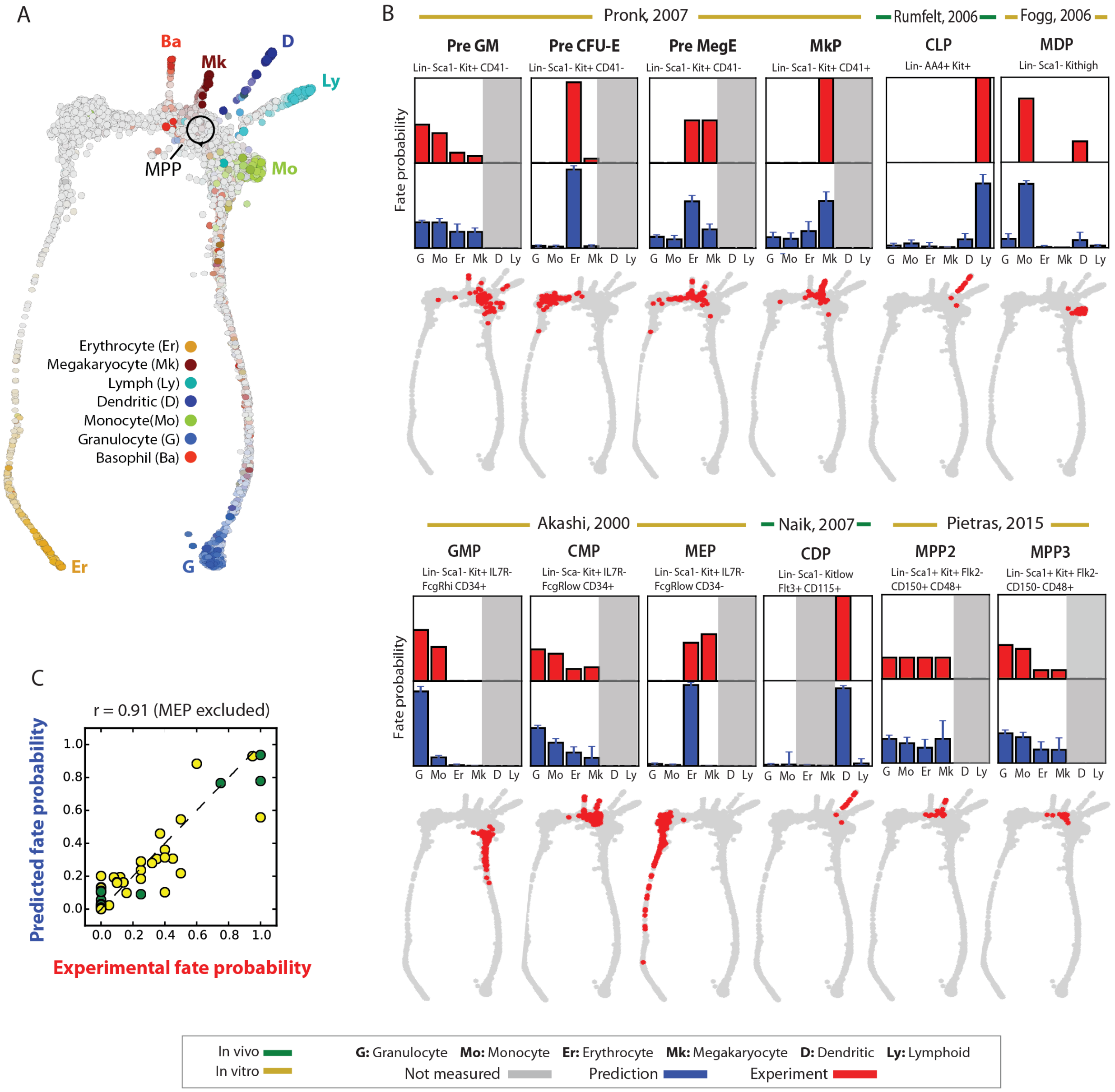


Figure 5: Population Balance reproduces known fate probabilities of hematopoietic progenitor cell (HPC) subpopulations from single cell data

(A) Single-cell profiles of 3,803 Kit+ HPCs reveal a continuum of gene expression states that pinches off at different points to form seven downstream lineages. Cells in this map are colored by marker gene expression and are laid out as a k-nearest-neighbor (knn) graph using SPRING, an interactive force-directed layout software. (B) PBA is applied to HPCs, and the predicted fate probabilities (blue bars) are compared to those observed experimentally (red bars) for reported HPC subpopulations (red dots on gray HPC map; identified using transcriptional similarity to existing microarray profiles for each reported subpopulation). Cell fates predicted by PBA but not measured experimentally are shaded gray. Error bars = 90% confidence intervals across 120 parameter combinations for the PBA pipeline. (C) Summary of comparisons made in (B); green points = in vivo measurements; yellow points = in vitro measurements.

For steady-state systems, PBA requires as fitting parameters an estimate of the diffusion strength D, and the net rates of cell entry and exit at each gene expression state (R). We estimated R using prior literature (see methods), and tested a range of values of D. All results that follow hold over the physiological range of PBA parameter values (Supplementary Figure 6).

We compared PBA results to previously reported fate probabilities by localizing reported cell types on the graph using published microarray profiles. Remarkably, in a panel of twelve progenitor cell populations from six previous papers(28-33) (Supplementary Table 2), the PBA predicted fate outcomes for each reported cell type (Figure 5B) closely matched fate probabilities measured in functional assays (defined as the proportion of clonogenic colonies containing a given terminal cell-type; see Supplementary Figure 5). The main qualitative disagreement between PBA predictions and experiment was in the behavior of Lin$^-$Sca1$^-$Kit$^+$IL7R$^-$FcgR^{low}CD34$^-$ HPCs, previously defined as megakaryocyte-erythroid precursors (MEP)(28). Our prediction was that these cells should lack megakaryocyte potential, which is indeed consistent with recent studies(24, 26, 34). Excluding these cells, our predicted fate probabilities matched experimental data with correlation ρ=0.91 (Fig. 5C). In a second paper (Tusi, Wolock, Weinreb et al., *in submission*), we test several novel predictions emerging from PBA in hematopoiesis.

Discussion

In this work we laid down a formal basis for the problem of dynamic inference from high-dimensional population snapshots. We invoked a conservation law, known as the law of Population Balance, as a general starting point for considering how a snapshot of cell states evolves over time. From there, we identified a number of features of the cell dynamics that are invisible in a static snapshot but must be known to reverse-engineer the dynamics. We argued that these features impose fundamental limits on the reconstruction of dynamics from single cell data faced by any algorithm. Accepting that all approaches face this limitation, we then attempted to develop an algorithm for dynamic inference that is explicit in the assumptions it makes about underlying gene regulation, as well as cell division and loss rates. The resulting method, PBA, was shown to provide not just an ordering of cell states, but also predictions of fate probabilities in simulations and in single cell data from hematopoietic progenitor cells of the mouse bone marrow. We also showed how PBA fails when its assumptions are violated, an important reminder of the limitations of all inference methods.

In developing PBA, we hoped that an algorithm with clear assumptions would help to clarify the ways in which data analysis might mislead us about the underlying biology. More practically, we hoped that the algorithm would suggest how to best design experiments to extract dynamic information from static measurements, and how to visualize single cell data to preserve aspects of the true dynamics. We discuss a number of points that follow from our analysis, along with a note about the technical underpinnings of PBA.

Experimental design for trajectory reconstruction from static snapshot measurements. We have shown that accurate dynamic inference requires knowledge of the density of cells in high dimensional state space, as well as the rates of cell entry and exit across the density. These requirements immediately suggest a set of principles for experimental design to optimize dynamic inference. First, to minimize distortions in the cell density in gene expression space it is useful to profile a single, broad population than many subpopulations fractionated in advance. Second, if cells of interest are sorted prior to analysis, it is best to minimize the number of sorting gates and enrichment steps, since each introduces an additional term to the entry/exit rates and subsequently a risk of distortion to the inferred dynamics. The HPC dataset analyzed in this paper was well suited for trajectory reconstruction because it included a single population, enriched using a single marker (Kit). This contrasts with previous single-cell RNA seq datasets of hematopoietic progenitors that included a composite of many sub-populations(35) or used complex FACS gates to exclude early progenitors(26).

Experimental methods for beating the limits of trajectory reconstruction. Given the inherent limits of trajectory reconstruction from single cell snapshots alone, orthogonal experimental data is required to disambiguate the true dynamics. In differentiation systems, pulse-chase experiments – where cells labeled in a given state are followed over time – could be used to infer rates of cell entry and exit by reporting on the flux of cells into different lineages. A previous study(36) quantified the transition rates between different FACS-defined hematopoietic compartments using a Tie2 – driven reporter to pulse label HSCs; the same assay could be coupled to singe-cell sequencing to enable direct fitting of the entry/exit parameter *R*. Clonal barcoding is another approach that would powerfully complement dynamical inference algorithms such as PBA: the dispersion of clones in high dimensional state space should constrain the stochasticity in the dynamics, allowing estimates of the diffusion constant *D* or even allowing consideration of non-uniform diffusion across gene expression space. Finally, it is well appreciated that live imaging of one or a few reporters could provide information on the significance of oscillatory behaviors that are not detectable in snapshot data.

How Population Balance Analysis could go wrong. To constrain a unique solution for trajectory reconstruction, PBA makes several strong assumptions, such as the absence of hidden variables and the absence of oscillations in gene expression space. Our simulations show that PBA is highly accurate for systems that meet these assumptions, but incorrectly infers dynamics for systems that break them. The agreement of prediction to fate commitment assays when we applied PBA to single cell profiles of hematopoietic progenitor cells suggests that, despite some sensitivity to assumptions, accurate inference is possible for complex differentiation systems. However, in cases where oscillatory dynamics strongly influence cell fate, or where hidden variables play a large role, single cell snapshot data could be misleading, and methods such as PBA may be ill suited.

In general, the impact of hidden variables on cell state dynamics remains unclear. Though there are many stable and possibly unobserved properties that impact a cell's behavior – including chromatin state; post-translational modifications; cellular localization of

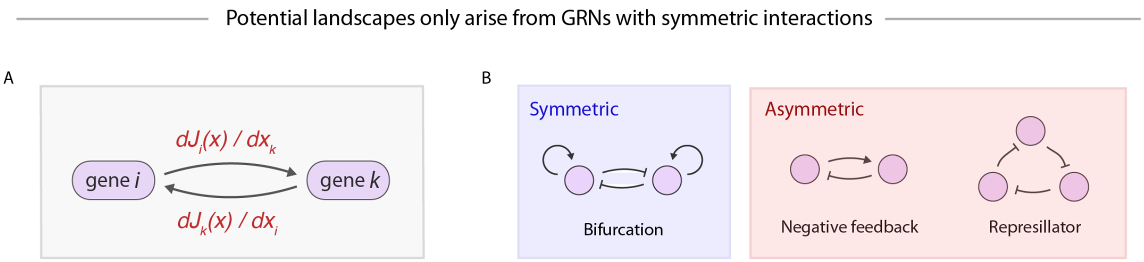


Figure 6: Potential landscapes arise from symmetric GRNs
In our formalism, inferences about the average velocity of cells can be interpreted as statements about an underlying gene regulatory network (GRN) (A). In this context, the assumption that there are no oscillatory expression dynamics implies that the underlying GRN has strictly symmetric interactions, which allows for some common gene regulatory motifs but rules out others (B).

proteins; metabolic state; and cellular micro-environment – it is possible that these properties percolate to some aspect of cell state that is observed, e.g. effecting a change in at least one gene measured by RNA-seq. By altering the observed state, such variables would thus not be hidden. For example, chromatin state exists in constant dialogue with transcriptional state, and is well reflected in mRNA content.

Though intuition suggests a minimal role for hidden variables, there may exist cases where they cannot be safely ignored. For example, a recent study(37) showed some evidence of clonally inherited biases in differentiation in HSCs that are not readily distinguished from each other by single cell RNA seq profiling. If true, the existence of different fate potentials for cells in similar transcriptional states would indicate that non-transcriptional factors, i.e. hidden variables in our study, could influence HSC behavior.

Fundamental limits on the inference of gene regulatory networks
One promise of single cell expression measurements is their possible use for reconstructing gene regulatory networks (GRNs) (2, 11). However, since any GRN model entails specific hypotheses about the gene expression trajectories of cells, efforts to infer GRNs from single cell data must also confront the limits of knowledge identified in our framework. In particular, GRN inference may benefit from an explicit consideration of cell entry and exit rates (embodied by R) and the rate of change in the cell density ($\partial c/\partial t$), as well as acknowledging the inability to distinguish oscillations from fluctuations.

Indeed, the inability to detect oscillations in single cell data, embodied in our framework by the use of a potential landscape, suggests severe limits on the types of underlying gene regulatory relationships that can be modeled. In fact, potential landscapes can only emerge from GRNs with strictly symmetric interactions, meaning every "arrow" between genes has an equal and opposite partner. This result follows from observing that the "arrows" in a GRN describe the influence of gene i on gene j, which is given by $\partial \boldsymbol{J}_i/\partial x_j$ (Figure 6A), where $\boldsymbol{J}$ is the deterministic component of average cell velocities (see Equation 2). The assumption of a potential landscape (i.e. $\boldsymbol{J} = -\nabla F$) then imposes symmetry on the GRN because $\partial J_i/\partial x_j = \partial J_j/\partial x_i = - \partial^2 F/\partial x_i \partial x_j$. Though a few well-known GRN motifs follow this symmetry rule – such as the "bistable switch" resulting from the mutual inhibition of two genes – many others do not, such as negative feedback loops and oscillators (Figure 6B). Potential landscapes are frequently invoked to explain gene expression dynamics(10, 38, 39), and we have shown them to be useful for predicting HPC fate outcomes in the context of PBA. It seems paradoxical that a tool that provides realistic phenomenological descriptions of gene expression dynamics reflects an entirely unrealistic picture for the underlying gene regulatory mechanisms. Resolving this paradox is an interesting direction for future work.

How should we visualize single cell data? At its core, the PBA algorithm performs dynamic inference by solving a diffusion-drift equation in high dimensions. This computation relies on a 2011 result in spectral graph theory by Ting Huang and Jordan(21) that describes the limiting behavior of *k*-nearest-neighbor graph Laplacians on sampled point clouds. Interestingly, several recent studies(8, 40, 41) have developed k-nearest neighbor graph-based representations of single cell data, and others have suggested embedding cells in diffusion maps(20, 42) on the basis of other similarity kernels. It has been unclear, until now, how to evaluate which of these different methods provides the most useful description of cell dynamics. Our technical results (Theorems 1-4 in the supplement) confirm that certain graph representations provide an asymptotically exact description of the cell state manifold on which dynamics unfold, suggesting them to be useful techniques for visualizing single cell data sets. Therefore PBA formally links dynamical modeling to choices of single cell data visualization.

References

1. Linnarsson S & Teichmann SA (2016) Single-cell genomics: coming of age. *Genome Biol* 17:97.
2. Klein AM*, et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161(5):1187-1201.
3. Macosko EZ*, et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161(5):1202-1214.
4. Buenrostro JD*, et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523(7561):486-490.
5. Lombard-Banek C, Moody SA, & Nemes P (2016) Single-Cell Mass Spectrometry for Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog (Xenopus) Embryo. *Angew Chem Int Ed Engl* 55(7):2454-2458.
6. Bendall SC*, et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332(6030):687-696.
7. Stevens TJ*, et al.* (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544(7648):59-64.

8. Bendall SC, *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157(3):714-725.
9. Macaulay IC, *et al.* (2016) Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell Rep* 14(4):966-977.
10. Marco E, *et al.* (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A* 111(52):E5643-5650.
11. Moignard V, *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* 33(3):269-276.
12. Shin J, *et al.* (2015) Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* 17(3):360-372.
13. Kafri R, *et al.* (2013) Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. *Nature* 494(7438):480-483.
14. Gaublomme JT, *et al.* (2015) Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* 163(6):1400-1412.
15. Setty M, *et al.* (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol*.
16. Ramkrishna D (2000) Chapter 1 - Introduction. *Population Balances*, (Academic Press, San Diego), pp 1-6.
17. Gillespie DT (2007) Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* 58:35-55.
18. Morris R, Sancho-Martinez I, Sharpee TO, & Izpisua Belmonte JC (2014) Mathematical approaches to modeling development and reprogramming. *Proc Natl Acad Sci U S A* 111(14):5076-5082.
19. Grima R, Thomas P, & Straube AV (2011) How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations? *J Chem Phys* 135(8):084103.
20. Angerer P, *et al.* (2015) destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*.
21. Ting D, Huang L, & Jordan M (2011) An Analysis of the Convergence of Graph Laplacians. *ArXiv e-prints*.
22. Elowitz MB & Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403(6767):335-338.
23. Eaves CJ (2015) Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood* 125(17):2605-2613.
24. Notta F, *et al.* (2016) Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* 351(6269):aab2116.
25. Ema H, Morita Y, & Suda T (2014) Heterogeneity and hierarchy of hematopoietic stem cells. *Exp Hematol* 42(2):74-82 e72.
26. Paul F, *et al.* (2015) Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* 163(7):1663-1677.
27. Weinreb C, Wolock S, & Klein A (2016) SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *bioRxiv*.

28. Akashi K, Traver D, Miyamoto T, & Weissman IL (2000) A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* 404(6774):193-197.
29. Fogg DK*, et al.* (2006) A clonogenic bone marrow progenitor specific for macrophages and dendritic cells. *Science* 311(5757):83-87.
30. Naik SH*, et al.* (2007) Development of plasmacytoid and conventional dendritic cell subtypes from single precursor cells derived in vitro and in vivo. *Nat Immunol* 8(11):1217-1226.
31. Pietras EM*, et al.* (2015) Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell* 17(1):35-46.
32. Pronk CJ*, et al.* (2007) Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell* 1(4):428-442.
33. Rumfelt LL, Zhou Y, Rowley BM, Shinton SA, & Hardy RR (2006) Lineage specification and plasticity in CD19- early B cell precursors. *J Exp Med* 203(3):675-687.
34. Psaila B*, et al.* (2016) Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol* 17:83.
35. Nestorowa S*, et al.* (2016) A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* 128(8):e20-31.
36. Busch K*, et al.* (2015) Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* 518(7540):542-546.
37. Yu VW*, et al.* (2017) Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell* 168(5):944-945.
38. Wang J, Xu L, Wang E, & Huang S (2010) The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys J* 99(1):29-39.
39. Moris N, Pina C, & Arias AM (2016) Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet* 17(11):693-703.
40. Samusik N, Good Z, Spitzer MH, Davis KL, & Nolan GP (2016) Automated mapping of phenotype space with single-cell data. *Nat Methods* 13(6):493-496.
41. Setty M*, et al.* (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 34(6):637-645.
42. Coifman RR*, et al.* (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci U S A* 102(21):7426-7431.

METHODS

1. Population balance analysis (PBA) source code and inputs

The core functions of PBA are implemented in python scripts on our github page: https://github.com/AllonKleinLab/PBA. The github page contains example files sufficient to reproduce the main calculations in this paper. In the theoretical supplement, we develop the rigorous foundations for PBA, provide detailed pseudo-code for the PBA algorithm and prove mathematically that it is asymptotically exact when sufficient cells are sampled and when PBA assumptions (see main text) are satisfied.

PBA was applied to simulated datasets and to experimental data from hematopoietic progenitor cells (HPCs) by calling PBA subroutines as follows:

$$\begin{aligned} V &= \text{compute\_potential}(X, R, k) \\ B &= \text{compute\_fate\_probabilities}(X, R, S, D, k) \\ T &= \text{compute\_mean\_first\_passage\_times}(X, R, D, k) \end{aligned}$$

In each case, the inputs to PBA are: a collection of single-cell expression profiles X, where X_{ij} is an expression matrix of gene j in cell i; prior estimates of the relative rates of proliferation and loss provided at each sampled gene expression state as a vector R of length n; the exit rates of cells into M terminal fates specified as a matrix S of size (n x M); and a diffusion constant (D) that reflects stochasticity in the dynamics. The number of neighbors of the nearest neighbor graph, k, is a fitting parameter, but results are not sensitive to the choice of k (see Supp. Fig. 2b). We used a range of k values for all analyses. The first output of PBA is a vector giving the values of the potential V at the n sampled expression states. V is then used to calculate a set of transition probabilities between sampled cells, from which we further derive terminal fate probabilities of each sampled cell provided as a matrix B of size (n x M), as well as the conditional mean first passage time between every pair of sampled cells provided as a matrix T of size (n x n).

For the simulated data, parameters R and D are determined in Methods sections (2-5). For the experimental data, we fitted R and D in Methods sections (7-8). For the analysis of HPCs, we normalized and reduced dimensionality of the raw expression data to generate a reduced matrix X as input for PBA (see Methods section 6).

Note that in general, R and D are partially redundant, since multiplying both by a common factor does not change the fate probabilities output by PBA.

2. Simulating a diffusion-drift process (Figure 3)

Data used to test PBA were generated from a simulated differentiation process in which initially bipotent cells choose one of two fates. In each simulation a single cell is generated in a gene expression state chosen uniformly at random in an m-dimensional box (the entry point) corresponding to an initial gene expression profile, with boundaries $-0.5<x_1<0.5$, $0.75<x_2<1$, and $0<x_{k>2}<1$ (Supp. Fig. 1). The cell gene expression profile is then updated over time using a Langevin simulation, meaning that a cell with initial position $x(t)$ is given a new position as follows:

$$x(t+\ \Delta t) = x(t) + \nabla F\big(x(t)\big)\Delta t + \sqrt{D\Delta t}\,\prod_{k=1}^{n} \xi_k \vec{e}_k,$$

where $\xi_1, \dots, \xi_m \sim \text{Gaussian}(0,1)$ are independent random variables sampled from a normal distribution, $\vec{e}_1,\dots,\vec{e}_m$ are unit vectors in each of the m dimensions, the simulation time step is $\Delta t = 0.001$ and $D = 0.05$. The mean gene expression velocity for cells in state x is the gradient of the m-dimensional potential field F, which defines a bifurcation in two dimensions with a quadratic basin in the remaining $m-2$ dimensions:

$$F(x) = \underbrace{\frac{x_1^4}{4} - \frac{x_2 x_1^2}{2} - \frac{x_2}{6}}_{bifurcation} + \underbrace{\sum_{i=3}^{m} x_i^2}_{\substack{quadratic\\ basin}}.$$

The simulation is terminated when a cell enters either of two box-shaped regions (the exit regions) where they were removed at rate $R = 5$ (boundaries shown in Supp. Fig. 1). This means that in a time step Δt, cells in an exit box are removed with probability $1 - e^{-R\Delta t} \approx 5\Delta t$. All simulations used m=50 dimensions, except for Supp. Fig. 2g, where m was varied from 2-50. The simulation was repeated to generate N cell trajectories, and each cell was then "sampled" at a time selected uniformly at random to generate mock single cell data set for PBA. All calculations were performed with N=200 sampled cells, except for Supp. Fig. 2e, where N was varied from 10-300.

For comparison of PBA predictions to the true dynamics (Figs. 1e-f and Supp. Figs. 2d-g), the "true fate probability" was defined for each sampled state x_i by carrying out a further 1000 Langevin simulations for each x_i as the initial condition, and recording the fraction of simulations terminating in exit box 2. The "true time since entry" was assigned to each x_i as the mean simulation time to reach x_i and its 5 nearest neighbors from a look-up table of 10,000 simulated trajectories.

3. Tests of PBA on simulated diffusion-drift process (Figure 3)

PBA was used to predict fate probability and time since entry for each sampled state x_i. PBA takes as input the entire point cloud $\{x_i\}$, as well as prior estimates of the entry/exit rates R_i at each point. For the main test of PBA (Figure 3) we used the true R_i values ($R_i = 5$ for cells in the entry-box; $R_i = -5$ for cells in the exit-boxes; $R_i = 0$ otherwise). For the robustness test in Supp. Fig. 2a-b we used false assumptions about entry/exit rates as indicated in the figure panels. For the robustness test in Supp. Fig. 2c-d we used false assumptions about the diffusion constant D as indicated in the figure panels. Changing $D \rightarrow D'$ is equivalent to scaling R uniformly, $R \rightarrow R' = (D/D')R$.

4. Testing the effect of gene oscillations on PBA predictions for diffusion-drift simulation (Supplementary Figure 3)

PBA assumes that the gene expression dynamics that give rise to a given set of sampled points $\{x_i\}$ is the gradient of a potential. However, this solution is not unique. The PBA solution implicitly assumes there are no rotations in gene expression space: a rotational field would not change the static density of cell states, and so it is invisible in a single cell sampling experiment. The effect of rotations on PBA predictions was tested by implanting rotational fields into the above simulations (Supp. Fig. 3). The rotational field used was

$$\vec{f}(x-\bar{x}) = (-x_2, x_1, 0, \ldots, 0) * \mathcal{N}(\sqrt{x^2+y^2}; \mu = 0, \sigma = 0.2)$$

where $\bar{x}$ is the center-point of the rotational field and $\mathcal{N}$ denotes a normal distribution. Langevin simulations were repeated as described above after adding this velocity field to the potential gradient velocity field. PBA predictions were repeated as described above to generate the results shown in Supp. Fig. 3.

7. Tests of PBA on simulated GRNs (Figure 4)

We used the Gillespie algorithm(1) to generate molecular counts for the simulations of gene regulatory networks (GRNs) in Figure 4. In every case, we supplemented the simulated counts with additional noisy dimensions (values drawn from a Gaussian), so that the total dimensionality of the data was always 50.

For the GRN in Figure 4A, we implemented the following stochastic chemical reactions (x_1 represents the green node, x_2 represents the blue node).

1) $(x_1, x_2) \rightarrow (x_1 - 1, x_2)$; rate = $0.003 * x_1$
2) $(x_1, x_2) \rightarrow (x_1, x_2 - 1)$; rate = $0.003 * x_2$
3) $(x_1, x_2) \rightarrow (x_1 + 1, x_2)$; rate = $hill(0.01 * x_1, 4) - hill(0.003 * x_2, 4)$
4) $(x_1, x_2) \rightarrow (x_1, x_2 + 1)$; rate = $hill(0.01 * x_2, 4) - hill(0.003 * x_1, 4)$
5) Simulation end; rate = $\delta(x_1 x_2 < 0.1) * 0.01$

For the GRN in Figure 4D, we implemented the following stochastic chemical reactions, where the variables x_i correspond to the colors in the figure as follows (x_1, red; x_2, green; x_3, blue; x_4, black; x_5, yellow)

1) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1 - 1, x_2, x_3, x_4, x_5)$; rate = $0.005 * x_1$
2) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2 - 1, x_3, x_4, x_5)$; rate = $0.005 * x_2$
3) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2, x_3 - 1, x_4, x_5)$; rate = $0.005 * x_3$
4) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2, x_3, x_4 - 1, x_5)$; rate = $0.01 * x_4$
5) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2, x_3, x_4, x_5 - 1)$; rate = $0.01 * x_5$
6) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1 + 1, x_2, x_3, x_4, x_5)$; rate = $1 - hill(0.1 * x_3, 2) + hill(0.025 * x_4, 2) - hill(0.025 * x_5, 4)$

7) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2 + 1, x_3, x_4, x_5)$; rate $=1 - hill(0.1 * x_1, 2) + hill(0.025 * \ x_5, 4)$
8) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2, x_3, x_4 + 1, x_5)$; rate = $1 - hill(0.1 * x_2, 2) - hill(0.025 * \ x_4, 2)$
9) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2, x_3, x_4 + 1, x_5)$; rate = $hill(0.013 * x_1, 8) + hill(0.025 * \ x_4, 2)$
10) $(x_1, x_2, x_3, x_4, x_5) \rightarrow (x_1, x_2, x_3, x_4, x_5 + 1)$; rate = $hill(0.013 * x_2, 2) + hill(0.025 * \ x_5, 4)$
11) Simulation end; rate = $0.002 * (hill(0.005 * x_4, 2) + hill(0.005 * x_5, 2)$

6. Data processing and normalization of single-cell RNA-seq data

Single-cell gene expression data from adult mouse bone marrow cells expressing Kit are reported and processed in another paper from our groups (Tusi, Wolock, Weinreb et al., *in submission*) Recapping in brief, reads were mapped as described in (2) to produce a (cell x gene) matrix of unique molecular identifier (UMI) counts that served as the starting point for the analysis in this paper.

Data was filtered to remove cells with < 1000 total UMIs. Visualization of the remaining cells in tSNE revealed three aberrant clusters of cells: one cluster strongly expressed mitochondrial genes and likely contained to stressed cells; the other two clusters co-expressed markers for distinct mature lineages (erythrocyte/macrophage and erythrocyte/granulocyte) and likely contained doublets. We removed all three aberrant clusters, resulting in 3803 cells.

Single cell data was then prepared for PBA by normalizing the total gene expression counts in each cell as described in (2). Genes with mean expression > 0.05 across the data set, and Fano Factor > 2, were then used to perform principal components analysis down to p dimensions, for p = 40, 50, 60, 70, 80, 90. When applying PBA, we also used a range of graph neighbor connectivities k (k=10 – 30). In Figure 5, we report medians and confidence intervals of fate probabilities for all 120 combinations of p and k.

7. Determining entry and exit parameters (R) for PBA analysis of HPCs

To apply PBA to hematopoietic differentiation, we estimated the entry/exit rates R from considerations of the proliferation rate and exit rates of Kit+ HPCs as follows. In adult hematopoiesis, all progenitors including HSCs express Kit, but eventually down-regulate it as they terminally differentiate. Thus, no cells enter the experimental system other than through proliferation of existing Kit+ HPCs, but there is a steady outflow (exit) owing to down-regulation of Kit as cells differentiate. We encoded this exit as negative R values for the top 10 cells with highest marker gene expression for each of the seven terminal lineages (Supplementary Figure 5 and Supplementary Table 1). We assigned different magnitudes of R for each of the seven lineages using a fitting procedure (see next paragraph). All remaining cells were assigned a uniform positive value of R, corresponding to a uniform proliferation rate, based on recent studies (3, 4) that found roughly similar growth rates across hematopoietic progenitor compartments. The magnitude of the growth rate was chosen so that $\sum R_i = 0$, reflecting a steady state in the total number of cells.

The flux of cells down-regulating Kit for each lineage varies widely between different hematopoietic lineages. This impacts PBA because it directly sets the relative magnitude of R for each lineage, although the simulations indicate that predictions do not require very accurate flux estimates. Because the flux of Kit+ cells from each lineage is not generally known, we fitted the seven fluxes by requiring that PBA reproduce measured fate probabilities of hematopoietic stem cells (HSCs). We performed a separate fitting for each of the studies shown in Figure 5 (see Supplementary Figure 7). When a study did not report fate probabilities for HSCs, we assumed a uniform distribution. We identified HSCs in our data by comparison to a microarray profile of HSCs, as described in Methods section 9.

8. Determining the diffusion rate (D) for PBA analysis of HPCs

The diffusion rate (*D*) controls the level of stochasticity in the PBA model. The exact value of D cannot be directly measured, but it is possible to constrain D using known quantities. We defined a physiologically plausible range by scanning through different values of D and checking the number of PBA-predicted multipotent cells for each value (see Supp. Fig. 7). We used prior literature (see https://www.immgen.org/, (4) and Methods section 9) to estimate that 2-20% of Kit+ bone marrow cells are multipotent. We defined a cell in our dataset as multipotent (for a given value of *D*) if it satisfied $P(\text{fate}) > 1/14$ for all 7 fates.

9. Validation of PBA-predicted fate probabilities for HPC

To validate PBA predictions of HPC fate probability, we compared them to the fate probabilities of 12 HPC subsets measured in previous studies (Table 2; Supplementary Figure 6). For each of the 12 cell surface marker-defined hematopoietic compartments, we used a published microarray profile to search for similar cells in our own dataset using a naïve Bayesian classifier, implemented as follows.

The Bayesian classifier assigns cells to microarray profiles based on the Likelihood of each microarray profile for each cell, with the Likelihood calculated by assuming that individual mRNA molecules in each cell are multinomially sampled with the probability of each gene proportional to the microarray expression value for that gene. Consider a matrix E of mRNA counts (UMIs) with n rows (for cells) and g columns (for genes), and also a matrix M with m rows (for microarray profiles) and g columns for genes. M was quantile normalized and then each microarray profile was normalized to sum to one. E was previously normalized in Methods section 5. The $(n \times m)$ matrix S_{ij} giving the Likelihood of each microarray profile j for each cell i is,

$$S_{ij} = Z_i \prod_{k=1}^{g} M_{jk}^{E_{ik}}$$

where Z_i is a normalization constant that ensures $\sum_i S_{ij} = 1$.

We assigned N_j cells with highest log-Likelihoods to each microarray profile j, with N_j determined from prior literature to reflect the abundance of each cell type among HPCs (see Supp. Table 2). Previous studies only provide abundance ratios between cell compartments, so we estimated N_j values by first estimating the number of ST-HSCs in our data, and then multiplying this value by the relative of abundance of each compartment compared to ST-HSCs. We estimated that the number of ST-HSCs in our data set was $N = 5$, reasoning that: (1) 1% of adult bone marrow is Kit+ (i.e. in our dataset); (2) the proportion of HSCs in adult bone marrow is 1-2 in 100,000 (5) and thus 1-2 in every 1,000 Kit+ cells is an ST-HSC; (3) our dataset contains approximately 5000 cells. Final assignments are indicated on the knn graphs in Figure 5.

References

1. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81(25):2340-2361.
2. Klein AM*, et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161(5):1187-1201.
3. Busch K*, et al.* (2015) Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* 518(7540):542-546.
4. Pietras EM*, et al.* (2015) Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell* 17(1):35-46.
5. Filip S*, et al.* (2014) The peripheral chimerism of bone marrow-derived stem cells after transplantation: regeneration of gastrointestinal tissues in lethally irradiated mice. *J Cell Mol Med* 18(5):832-843.

Table 1: Marker genes used to identify the most mature cells in each lineage.

Name	Marker genes
Erythrocyte (Er)	Hbb-bt, Hba-a2, Hba-a1, Alas2, Bpgm
Megakaryocyte (Mk)	Pf4, Itga2b, Vwf, Mef2c
Granulocyte (G)	Lcn2, S100a8, Ltf, Lyz2, S100a9
Monocyte (Mo)	Csf1r, Ly6c2, Ccr2, Glipr1
Lymph (Ly)	Cd79a, Igll1, Vpreb3, Vpreb1, Lef1
Dendritic (D)	H2-Aa, Cd74, H2-Eb1, H2-Ab1, Cst3
Basophil (Ba)	Ifitm1, Ly6e, Srgn

Table 2: Summary of 12 HPC subpopulations with microarray profiles and fate assays from previous papers, used to validate PBA

*NM = not measured

Name	Lineage Potential[1] (%) Er-Mk-G-Mo-Ly-D	Microarray reference	Microarray accession codes (replicates)	Differentiation assay	Differentiation reference	Figure in original reference	Num. cells[2]
PreMegE	50-50-0-0-NM-NM	Pronk, 2007	GSM2076(82-86)	Agar	Pronk, 2007	Figure 3	12[3]
PreCFU-E	95-5-0-0-NM-NM	Pronk, 2007	GSM2076(90-91)	Agar	Pronk, 2007	Figure 3	12[3]
MkP	0-100-0-0-NM-NM	Pronk, 2007	GSM2076(87-89)	Agar	Pronk, 2007	Figure 3	12[3]
Pre GM	12-8-45-35-NM-NM	Pronk, 2007	GSM2076(79-81)	Agar	Pronk, 2007	Figure 3	12[3]
CMP	14-16-37-32-NM-NM	Teng, 2008	GSM7911(17-18)	Methylcellulose	Akashi, 2000	Figure 1	12[4]
MEP	45-55-0-0-NM-NM	Teng, 2008	GSM7911(08-09)	Methylcellulose	Akashi, 2000	Figure 1	30[4]
GMP	0-0-60-40-NM-NM	Teng, 2008	GSM7911(19-21)	Methylcellulose	Akashi, 2000	Figure 1	18[4]
CDP	NM-NM-0-0-0-100	Teng, 2008	GSM7911(14-16)	In vivo	Naik, 2007	Supp Fig 1	3[3]
CLP	0-0-0-0-100-0	Teng, 2008	GSM5383(48-50)	In vivo	Rumfelt, 2006	Figure 2	3[4]
MDP	0-0-0-75-0-25	Teng, 2008	GSM7911(05-07)	S17 stroma	Fogg, 2006	Figure 2B	12[4]
MPP2	25-25-25-25-NM-NM	Pietras, 2015	GSM16746(29-30)	Methylcellulose	Pietras, 2015	Figure 2	3[5]
MPP3	10-10-40-40-NM-NM	Pietras, 2015	GSM16746(31-33)	Methylcellulose	Pietras, 2015	Figure 2	3[5]

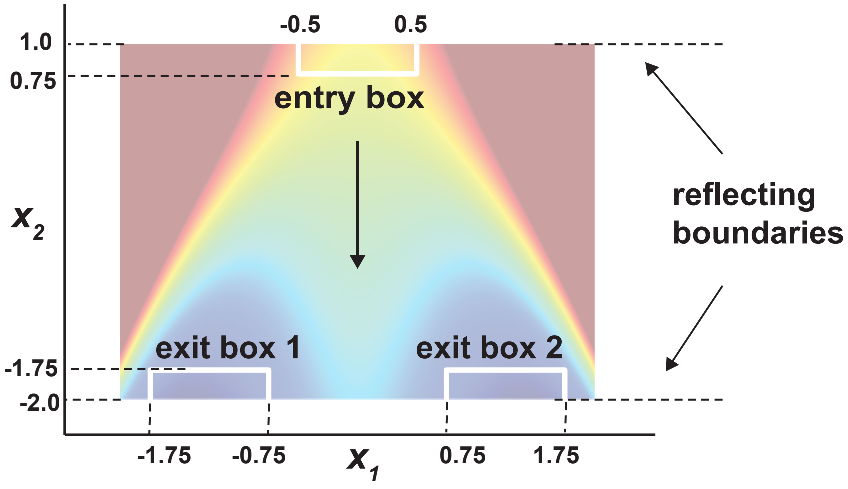
[1] Lineage potential refers to the proportion of colonies/mice that produce a terminal cell-type when inoculated with the given progenitor population. Potentials are normalized to add up to one. When measurements were made for HSCs, the potentials were renormalized so that HSC would have uniform potential across cell types.

[2] All cell numbers represent a ratio with respect to short-term stem cells (ST-HSC). When data was not available for a specific progenitor population, we used data from a population with the same functional potential, or otherwise made a conservative guess.

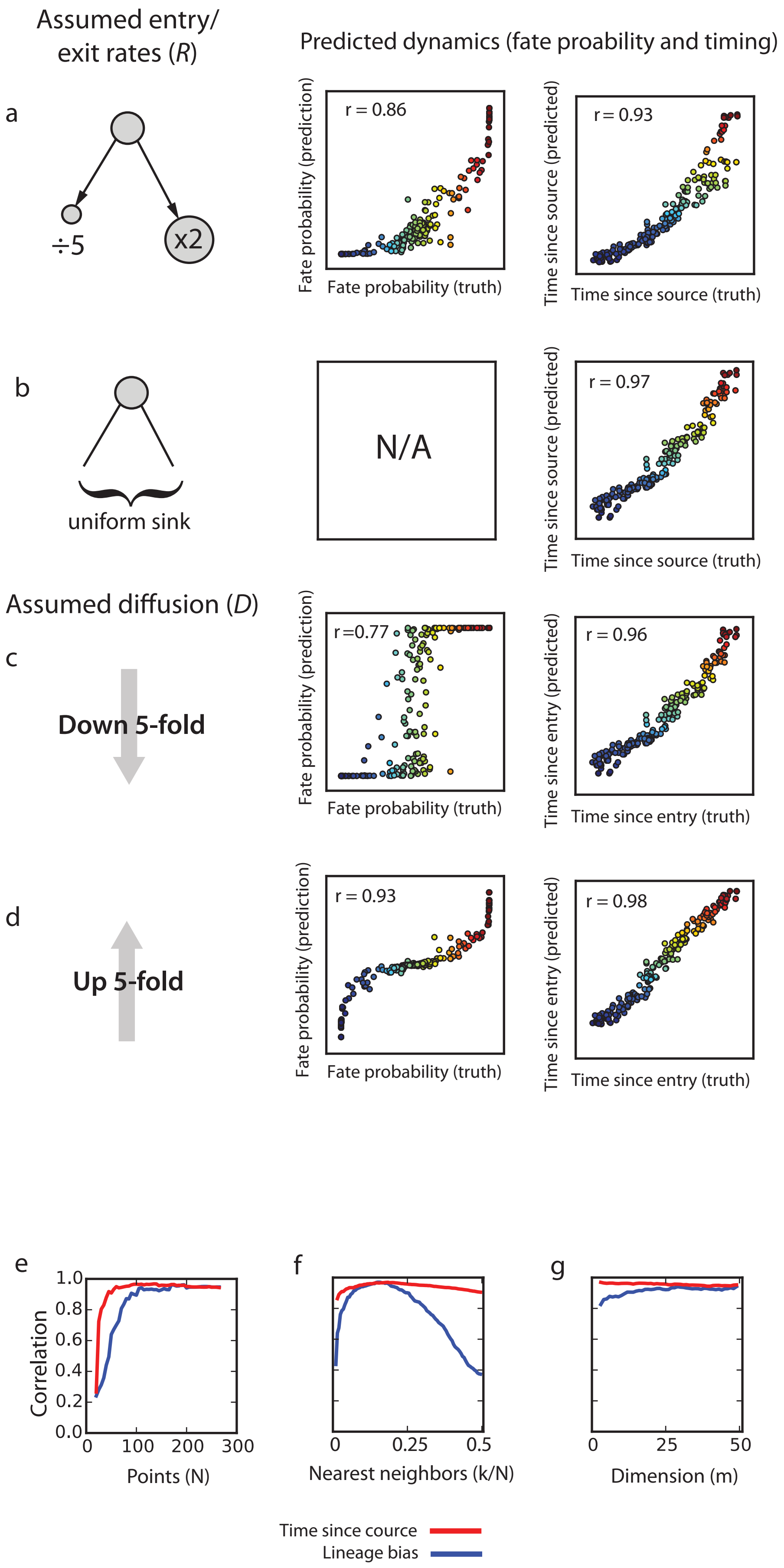
[3] Conservative guess

[4] (Busch, 2015)

[5] (Pietras, 2015)

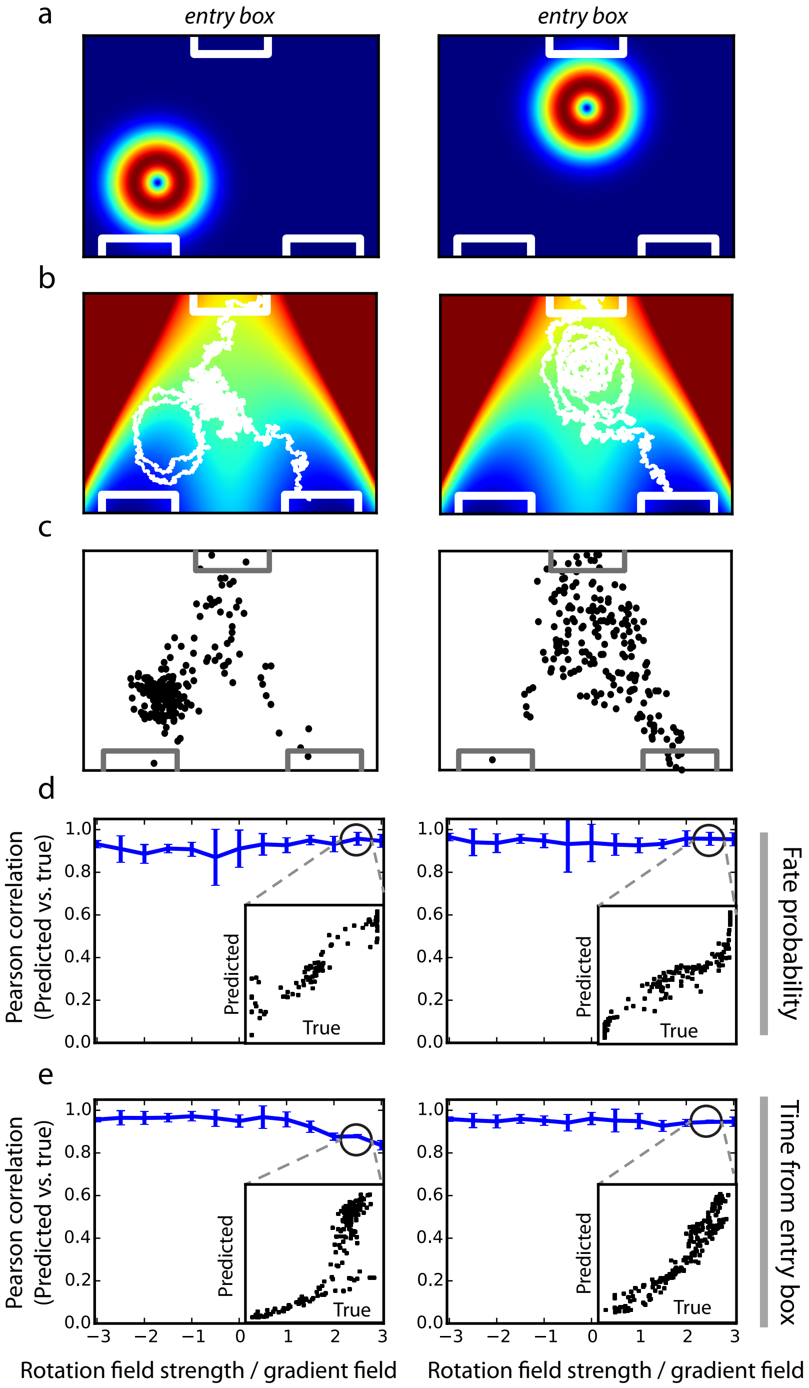


Supplementary Figure 1: Entry/exit boundaries for a simulation of lineage bifurcation. Figure supporting Fig. 3 and Methods section 2, showing the entry/exit locations in the first two dimensions of the m-dimensional gene expression simulation. Each cell was generated at simulation time t=0 in a gene expression state chosen uniformly at random in the indicated m-dimensional source box. Cells exit through one of the two sink boxes.



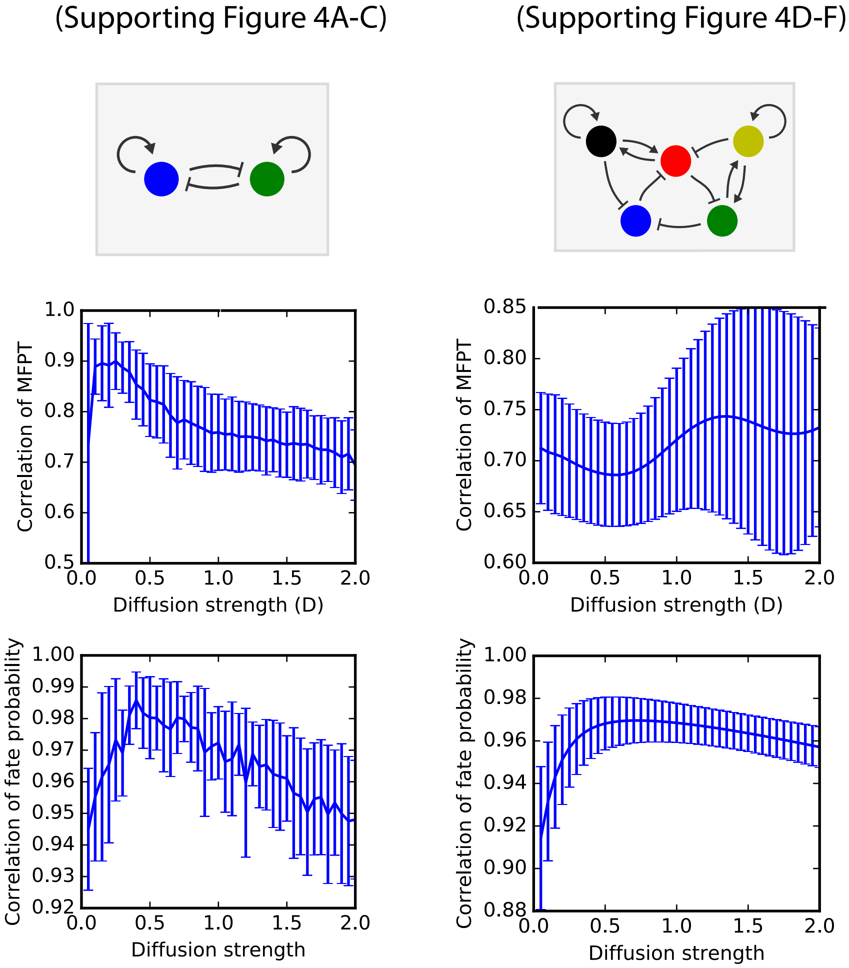
Supplementary Figure 2: Robustness tests for the PBA algorithm.

(a-d), Comparison of PBA predictions to “true” (simulated) fate bias and temporal order under imprecise assumptions about the entry/exit and diffusion parameters, D and R. These analyses also showed that: (a) imprecisely estimating the exit rates between two fates with a ten-fold error skews estimated fate probabilities but maintains high correlations; (b), treating every point as an exit does not diminish the accuracy of predicted temporal ordering; (c),decreasing the assumed diffusion rate predicts fate commitment to occur prematurely, causing PBA to under-estimate the number of bi-potent cells; (d), increasing the assumed diffusion rate has the opposing effect, leading to over-estimate the number of bi-potent cells. (a-c), Pearson correlation between “true” and “predicted” values of fate bias and temporal order for a range of algorithm parameter values: (e), the number of cells sampled; (f), number of graph neighbors *k* (measured as fraction of total graph size); (g), simulation dimensionality *m* (i.e. number of independent genes per cell). For each case, the relevant parameter is varied while keeping the other parameters fixed (*N*=200, *k*=20, *m*=50). In general, inference of temporal order is more accurate than fate probability.

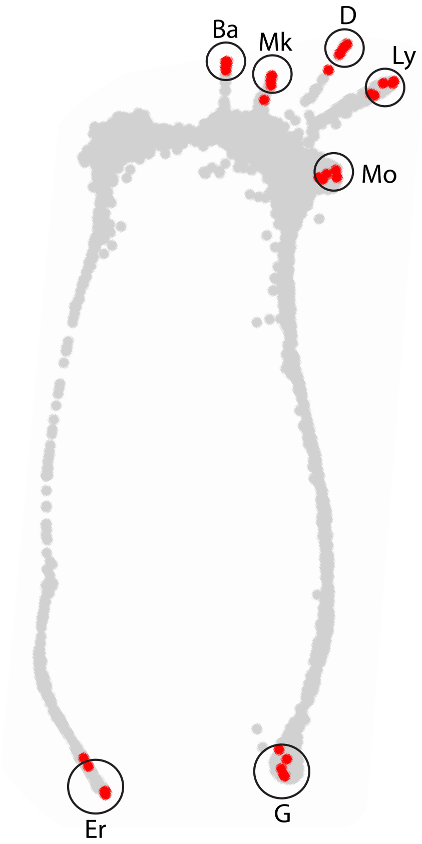


Supplementary Figure 3: Testing PBA robustness to gene expression oscillations.

PBA models gene expression dynamics as a diffusion-drift process down a potential landscape. This model makes an implicit assumption that no oscillations exist, since potential fields are irrotational. We measured the error that could be introduced by this assumption, by implanting a rotational gene expression field into the simulated fate bifurcation at two different points (left and right column), shown in (a). (b), Example simulated cell trajectories in presence of a rotational field; (c), location of sampled cells in the first two simulated gene expression dimensions; (d-e), Fidelity of PBA dynamic predictions measured by the Pearson correlation between “true” (simulated) and predicted quantities. Despite violating the assumptions of PBA, the oscillations did not significantly impact accuracy for fate probability (d) and timing (e). Blue curves indicate the mean correlation value as a function of the rotational field strength measured relative to the gradient field strength. Error bars indicate 90% confidence intervals from 10 independent trials. Panel insets show comparisons at the indicated rotational field strengths.



Supplementary Figure 4: Accuracy of PBA for a range of diffusion strengths in the GRN simulations shown in Figure 4. The optimal diffusion parameter value was used in Figure 4.

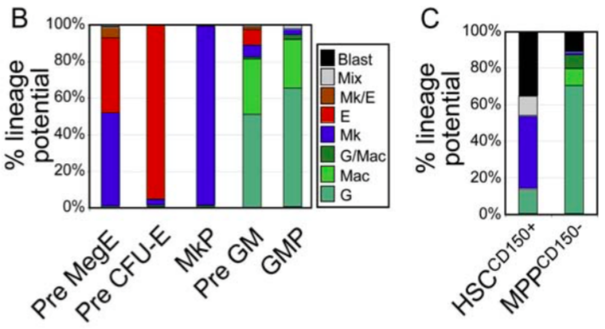


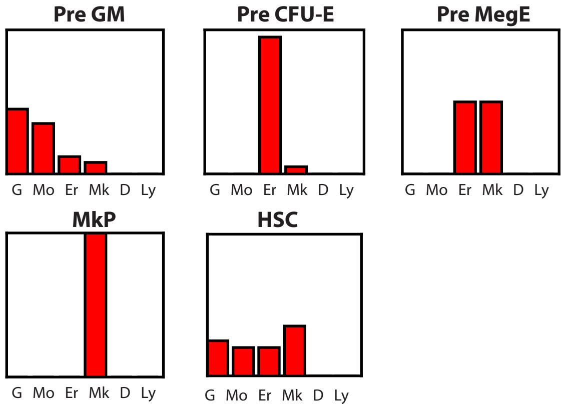
Supplementary Figure 5: Matching cell subsets to known progenitor subpopulations.

Identification of the endpoints of each lineage represented in our dataset, which occur when Kit is down-regulated. For each of seven fates, we identified endpoints as the 10 cells (red dots) with highest standardized (z-score) expression of known marker genes (Supplementary Table 1).

Pronk, 2007

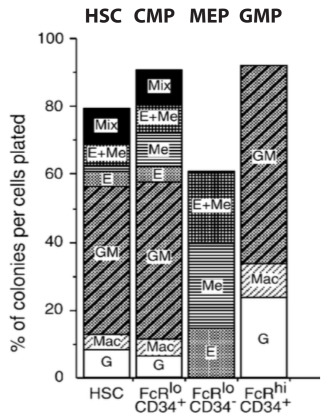
Figure 2

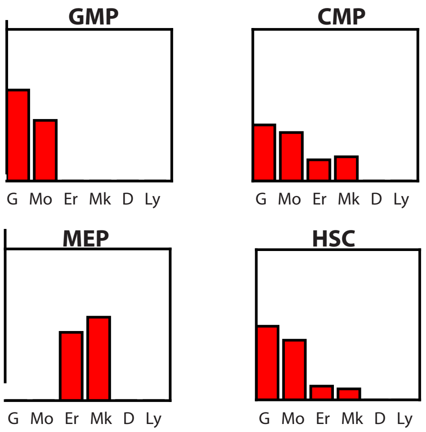




Akashi, 2000

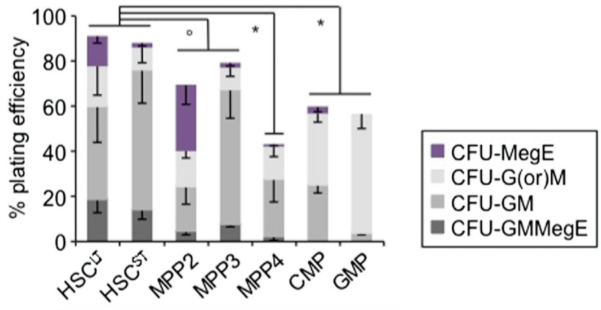
Figure 1C

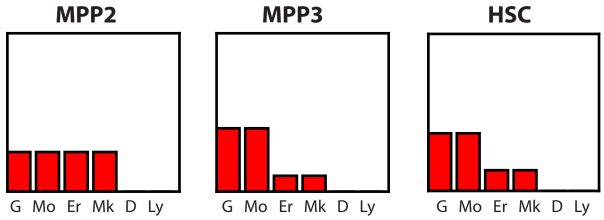




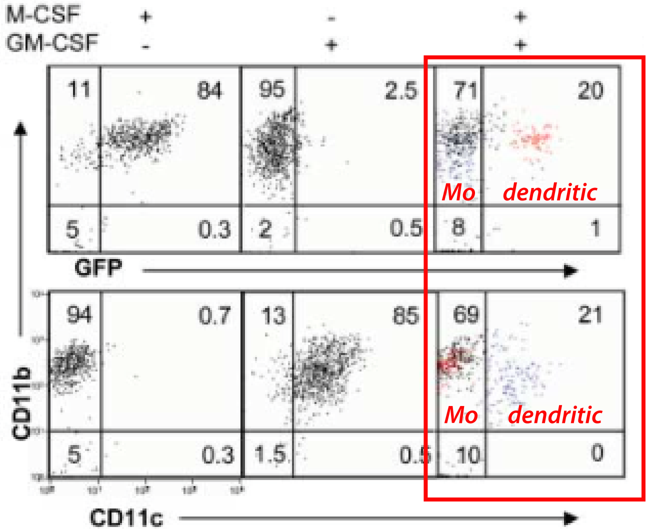
Pietras, 2015

Figure 2C

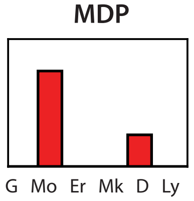




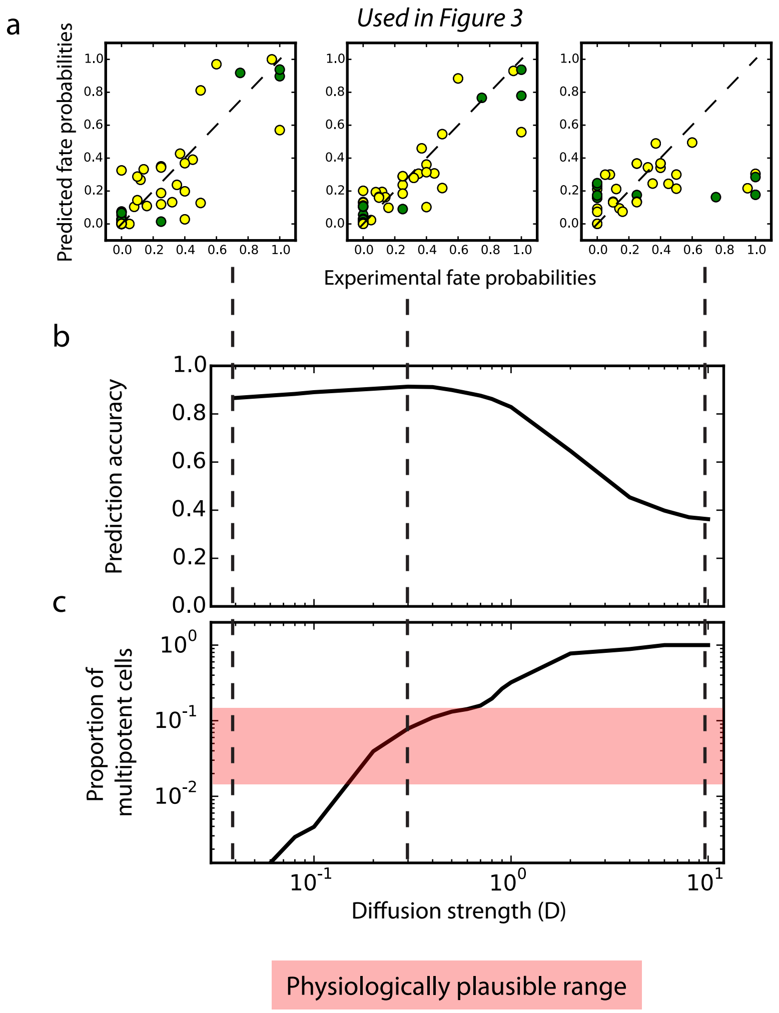
Rumfelt, 2006



inducers for both monocyte and dendritic



Supplementary Figure 6



Supplementary Figure 7: The PBA diffusion parameter (D) is constrained by subpopulation fate probabilities and by the fraction of multipotent cells.

(a), The diffusion constant (D) sets the stochasticity of the PBA model and impacts the predicted fate probabilities for HPCs. (b), A systematic scan of *D* values shows that prediction accuracy remains high over a broad range of *D* values. (c), The PBA-predicted proportion of multipotent cells plotted as a function of *D.* The physiological range is highlighted (pink) (see Methods). This analysis reveals a narrow range of physiologically plausible D values that includes the point of maximum prediction accuracy. Dashed lines relate the panels in (a) to values of *D* plotted in (b,c).