the FEBS Journal

Multi-scale modeling of GMP differentiation based on single-cell genealogies

Carsten Marr[1,*], Michael Strasser[1,*], Michael Schwarzfischer[1], Timm Schroeder[2] and Fabian J. Theis[1,3]

1 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg-Munich, Germany
2 Stem Cell Dynamics Research Unit, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg-Munich, Germany
3 Institute for Mathematical Sciences, Technische Universität München, Garching, Germany

Keywords
approximate Bayesian computing; branching process; differentiation; hematopoiesis; stochastic gene expression model

Correspondence
Carsten Marr, Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg-Munich, Germany Neuherberg-Munich, Germany
Fax: +49 89 3187 3585
Tel: +49 89 3187 3642
E-mail: carsten.marr@helmholtz-muenchen.de

*These authors contributed equally to this work.

(Received 7 December 2011, revised 21 May 2012, accepted 1 June 2012)

doi:10.1111/j.1742-4658.2012.08664.x

Hematopoiesis is often pictured as a hierarchy of branching decisions, giving rise to all mature blood cell types from stepwise differentiation of a single cell, the hematopoietic stem cell. Various aspects of this process have been modeled using various experimental and theoretical techniques on different scales. Here we integrate the more common population-based approach with a single-cell resolved molecular differentiation model to study the possibility of inferring mechanistic knowledge of the differentiation process. We focus on a sub-module of hematopoiesis: differentiation of granulocyte–monocyte progenitors (GMPs) to granulocytes or monocytes. Within a branching process model, we infer the differentiation probability of GMPs from the experimentally quantified heterogeneity of colony assays under permissive conditions where both granulocytes and monocytes can emerge. We compare the predictions with the differentiation probability in genealogies determined from single-cell time-lapse microscopy. In contrast to the branching process model, we found that the differentiation probability as determined by differentiation marker onset increases with the generation of the cell within the genealogy. To study this feature from a molecular perspective, we established a stochastic toggle switch model, in which the intrinsic lineage decision is executed using two antagonistic transcription factors. We identified parameter regimes that allow for both time-dependent and time-independent differentiation probabilities. Finally, we infer parameters for which the model matches experimentally observed differentiation probabilities via approximate Bayesian computing. These parameters suggest different timescales in the dynamics of granulocyte and monocyte differentiation. Thus we provide a multi-scale picture of cell differentiation in murine GMPs, and illustrate the need for single-cell time-resolved observations of cellular decisions.

Introduction

The generation of blood cells involves a complex, highly regulated set of processes. According to the current paradigm, hematopoiesis may be pictured as a tree made up from a concatenation of branching decisions [1]. Starting from a hematopoietic stem cell, all mature blood cell types are generated via progenitor

Abbreviations
CMP, common myeloid progenitor; G, granulocyte; GMP, granulocyte–monocyte progenitor; M, monocyte; MPP, multipotent progenitor.

states, where the lineage potential is reduced in each differentiation step. To replenish granulocytes (G) and monocytes (M), for example (blood cells with important function in immune response and phagocytosis [2]), a hematopoietic stem cell differentiates to a multipotent progenitor (MPP), a common myeloid progenitor (CMP) and a granulocyte–monocyte progenitor (GMP), before the final lineage decision between G and M is made (see Fig. 1A). Because of easy experimental access, the blood system has been studied for many decades. Although details of the hierarchical differentiation have been revealed, together with the increasing specificity of cell state markers, the tight balance between blood cell numbers and the inter-regulation of the involved processes are far from understood.

To obtain mechanistic insights into this dynamic equilibrium, various aspects of hematopoiesis have been modeled on very different scales [3,4]. Almost 50 years ago, growth of murine spleen colony-forming cells was interpreted within the framework of branching processes [5]. Recent advances of the method include multi-type branching models [6] and the incorporation of correlation between cells [7]. The dynamics of lineage specification has been modeled using competing lineage propensities [8], from the perspective of dynamic systems [9], and using a compartment model with a focus on cell–cell communication [10,11], for example. On a meso-scale, we recently constructed and validated a Boolean model of 11 myeloid transcription factors based on literature evidence [12].

On a smaller scale and with a focus on the intrinsic mechanism, hematopoietic branching decisions have been modeled, for example with a gene toggle switch of two mutually repressing genes. This approach allows molecular details to be linked to an attractor landscape that determines the dynamics and stable states of the differentiating cells [13]. In particular, the myeloid lineage decision of CMPs has been scrutinized in a number of studies [14–18]. From a mechanistic perspective, deterministic [19,20] and stochastic [21–25] toggle switch models have revealed various aspects of the possible dynamics and the emerging attractor landscape. At the molecular scale, we recently showed that inclusion of mRNA levels in a stochastic toggle switch model leads to multi-attractor dynamics, a feature that is only observable with a considerably detailed description of gene expression [26].

Here, we study the differentiation probability of a GMP using a multi-scale modeling approach, and compare it to experimental evidence. First we show what conclusions can be drawn from a standard colony assay of sorted undifferentiated GMPs. Using a branching process description, we can solve a recursive

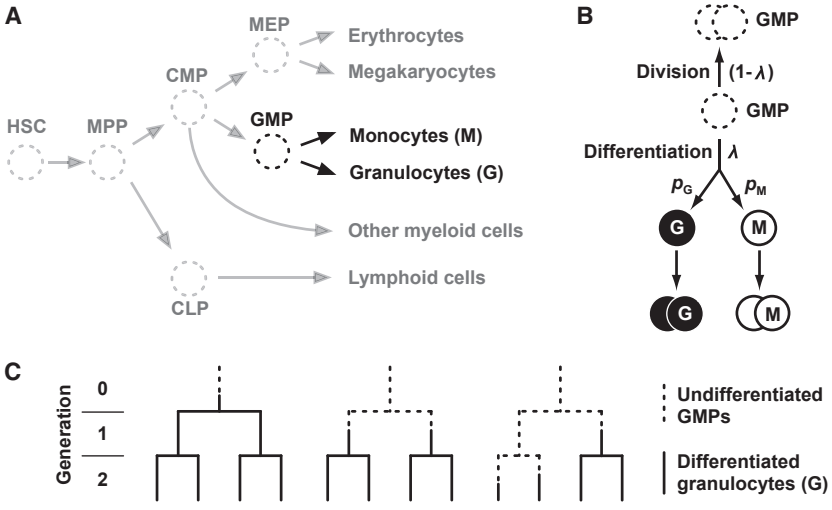

Fig. 1. Hematopoietic differentiation and branching process model. (A) Hematopoiesis is currently viewed as a hierarchy of differentiation processes [27]. From a hematopoietic stem cell (HSC), mature blood cells are replenished via a series of progenitor cells with limited potential. MPP, multipotent progenitor; CMP, common myeloid progenitors; MEP, megakaryocyte–erythrocyte progenitor; GMP, granulocyte–monocyte progenitor; CLP, common lymphoid progenitor. (B) Branching process model of GMP differentiation. In one cell cycle, a GMP will either differentiate (with probability λ) or divide (with probability 1 – λ). If it differentiates, it will decide on either the granulocyte (G) or the monocyte (M) lineage, with probabilities p_G and p_M, respectively. In the chosen lineage (G or M), cells keep dividing. (C) The branching process leads to genealogies with distinct probabilities for homogenous and heterogenous trees. The probabilities for the three homogeneous G trees shown are $\lambda \cdot p_G$, $(1-\lambda)(\lambda \cdot p_G)^2$ and $2(1-\lambda)^2(\lambda \cdot p_G)^3$, respectively. Dashed lines represent undifferentiated GMPs; solid lines represent differentiated granulocytes. Time runs downwards in units of generations.

equation that describes the probability for a homogeneously differentiated tree. Based on three observables (the relative abundance of homogeneous G and M colonies and the abundance of heterogeneous colonies), we can infer three probabilities: the probability of differentiating before division, which we assume to be time-independent, and the probabilities of choosing one or the other lineage. In contrast to the genealogies emerging from this model, we find that the differentiation probability in the genealogies from single-cell time-lapse microscopy data rises with time.

We then demonstrate how a molecular toggle switch model with stochastic gene expression can induce time dependence in the differentiation probability. This model operates on a smaller scale, and explicitly considers the mutual inhibition of the two transcription factors of the toggle switch. However, the molecular identities of the transcription factors are still under debate. After making a couple of assumptions about the molecular interactions of the associated genes, we infer parameters using approximate Bayesian computing, and discover different time scales for the granulocyte and monocyte differentiation dynamics.

Results

GMP differentiation as a branching process

We first describe the time evolution of a GMP as a branching process [28]: a GMP differentiates with probability λ into a granulocyte (G) or a monocyte (M), or it divides with probability $1 - \lambda$ into two identical GMPs (see Fig. 1B). A differentiating GMP chooses the G and M lineage with probabilities p_G and $p_M = 1 - p_G$, respectively. Once differentiated, granulocytes and monocytes keep dividing. The assumptions on which this stochastic model is based are (a) independent decisions of GMPs, (b) a time-constant differentiation probability $\lambda(t) = \lambda$, (c) symmetric divisions so that a GMP divides into two identical progenitors, and (d) irreversible transitions to the differentiated states. Note that we neglect cell death, and the time unit in the branching model is 'cell cycle' or 'generation'. Thus, it operates temporally on a coarse-grained level, as neither cell-cycle effects nor molecular mechanisms are taken into account.

Using this branching process model, we can generate genealogies and quantify their probability (see Fig. 1C). For now, we focus on homogeneous trees, i.e. trees where all leaves (cells with no descendants) differentiate towards the same lineage. The simplest way to create a homogeneous G tree, for example, is if the first cell in generation 0 differentiates before it divides. The probability for this tree is simply $\lambda \cdot p_G$. Also, the first cell may divide before it differentiates, but both daughters differentiate before they divide again, leading to a probability of $(1 - \lambda)(\lambda \cdot p_G)^2$. Alternatively, one of the two daughters may not differentiate but both granddaughters do. The probability for this is $2(1 - \lambda)^2(\lambda \cdot p_G)^3$, where the factor 2 accounts for the two symmetrical genealogies.

We can easily calculate the probabilities of differentiating cells and emerging genealogies within the branching process formalism. Let us clarify our nomenclature by comparing the differentiation process with a simple Bernoulli trial, in which balls are drawn out of a box containing black and white balls. The 'success rate' for drawing a black ball in each trial is constant (if drawn balls are replaced) and corresponds to the relative number of black balls in the box. The probability of drawing the first black ball after exactly n trials follows a geometric distribution. Similarly, the probability that a cell differentiates exactly in generation g, termed $P(g)$, follows a geometric distribution, where the success rate is the differentiation probability λ:

$$P(g) = \lambda(1 - \lambda)^g,\ g \in \mathbb{N}_0 \tag{1}$$

We are interested in the probability R_G for a tree to be homogeneously differentiated towards G at all leaves. We can formulate this by assuming that either the first cell differentiates to G, with probability $\lambda \cdot p_G$, or it first divides, with probability $(1 - \lambda)$, and both daughters differentiate at some point later, with probability $(R_G)^2$. This gives rise to the recursive equation

$$R_G = \lambda \cdot p_G + (1 - \lambda)(R_G)^2 \tag{2}$$

which is similar to the classical formulation of the Galton–Watson process [29]. A symmetric equation holds for R_M. Both can be solved, yielding

$$R_G = \frac{1 + \sqrt{1 - 4(1 - \lambda)\lambda \cdot p_G}}{2(1 - \lambda)} \tag{3}$$

$$R_M = \frac{1 + \sqrt{1 - 4(1 - \lambda)\lambda \cdot p_M}}{2(1 - \lambda)} \tag{4}$$

Finally, the probability of observing an heterogeneous tree with both G and M cells is

$$R_{GM} = 1 - R_G - R_M \tag{5}$$

Solving eqns (3)–(5) for the differentiation probability λ yields

$$\lambda = \frac{R_G(R_G - 1) + R_M(R_M - 1)}{(R_G)^2 + (R_M)^2 - 1} \quad (6)$$

Similarly, we can determine the probabilities towards the one and the other lineage using

$$p_G = R_G \frac{R_G(1 - R_M) + (R_M)^2 - 1}{R_G(R_G - 1) + R_M(R_M - 1)} \quad (7)$$

$$p_M = 1 - p_G \quad (8)$$

We can use eqns (6)–(8) to calculate λ from the fraction of heterogeneous (R_{GM}), homogenous G (R_G) and homogeneous M (R_M) colonies in colony assay experiments. Here, cells are plated at clonal density, meaning that every single cell gives rise to a single clearly distinguishable colony (see Fig. 2A). A colony assay of sorted GMPs performed previously [30] revealed a high sorting purity: only < 3% of the colonies contain erythrocytes and/or megakaryocytes or blast cells. Excluding the < 3% other colonies, 55 ± 7% of colonies are homogeneous G, 23 ± 6% of colonies are homogeneous M, and 22 ± 5% of colonies are mixed GM colonies (mean ± SD) (see inset in Fig. 2B). Using these numbers, we find that the differentiation probability λ of our model is 66 ± 10%. GMPs choose the G lineage with $p_G = 0.67 \pm 0.05$, and the M lineage with $p_M = 0.33 \pm 0.05$ (see Fig. 2B) under permissive culture conditions (interleukins 3 and 6, and stem cell factor (SCF)) that allow GMPs to differentiate into both lineages.

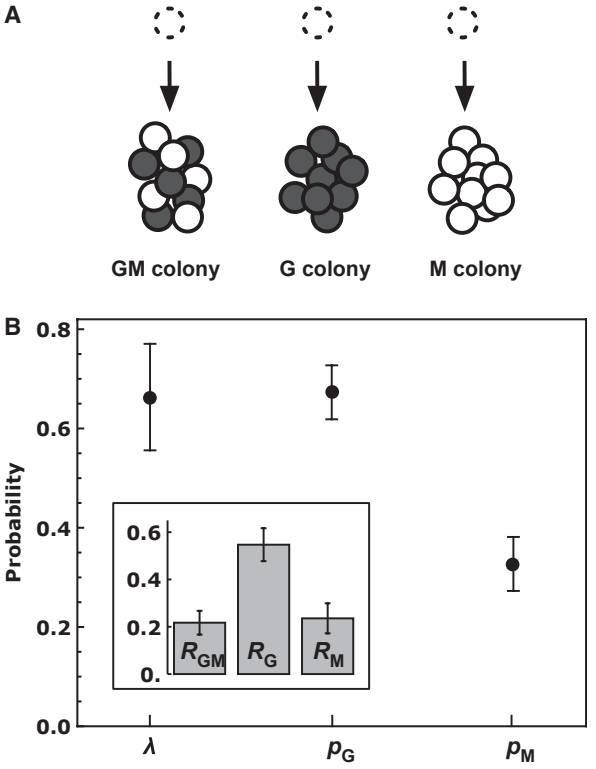

Fig. 2. Differentiation probability inferred from colony assay data. (A) In the colony assay used in this analysis [30], each progenitor cell gives rise to a single, distinguishable colony. The relative frequency of colonies containing cells of lineages other than G or M was < 3%, and these have been ignored. (B) The differentiation probability λ of the branching process model, inferred from the colony assay data, is 0.66 ± 0.10% (mean ± SD). Moreover, we can determine the probability of choosing the granulocyte lineage ($p_G = 0.67 \pm 0.05$) or the monocytic lineage ($p_M = 0.33 \pm 0.05$) under permissive culture conditions (interleukins 3 and 6, and SCF) allowing differentiation to both G and M. Inset: The relative frequencies of GM, G and M have been re-scaled. Only 22 ± 5% of the colonies are heterogeneous, while 23 ± 6% are homogeneously M, and 55 ± 7% are homogeneously G. The error bars correspond to one standard deviation as determined from five replicates of the colony assay.

GMP differentiation probability from single-cell time-lapse microscopy data

We now compare the probabilities derived from the colony assay with tree data from single-cell time-lapse microscopy. In previous experiments [30], GMPs were sorted with a purity > 95%. Sorted cells have been imaged, tracked [32] and analyzed for up to 5 days under permissive conditions (interleukins 3 and 6, and SCF), allowing differentiation into both granulocytes and monocytes. To monitor the loss of the GMP state (see trees in Fig. 3), LysM::GFP mice [33], which express enhanced GFP from the *lysozymeM* locus as a marker for uni-lineage commitment, were used in the experiments. The onset of a fluorescent signal in a cell is annotated in the data, and indicates irreversible loss of bi-potency. All cells are tracked until marker onset, if they have not died or been lost to tracking previously. After marker onset, tracking is normally stopped after a couple of time points, but occasionally is continued to the next generations (see Fig. 3A for three examples of tracked trees with marker annotation). The marker onset times for all cells in all trees with fluorescent signal (100 of 143 trees) are shown in Fig. 3A. The number of possible onsets per tree is given by the number of cells present, and thus increases with time and the number of branches. Interestingly, many trees show late and synchronous marker onset. This contradicts the expectation from the branching process model, where the probability for a synchronous marker onset in generation g is

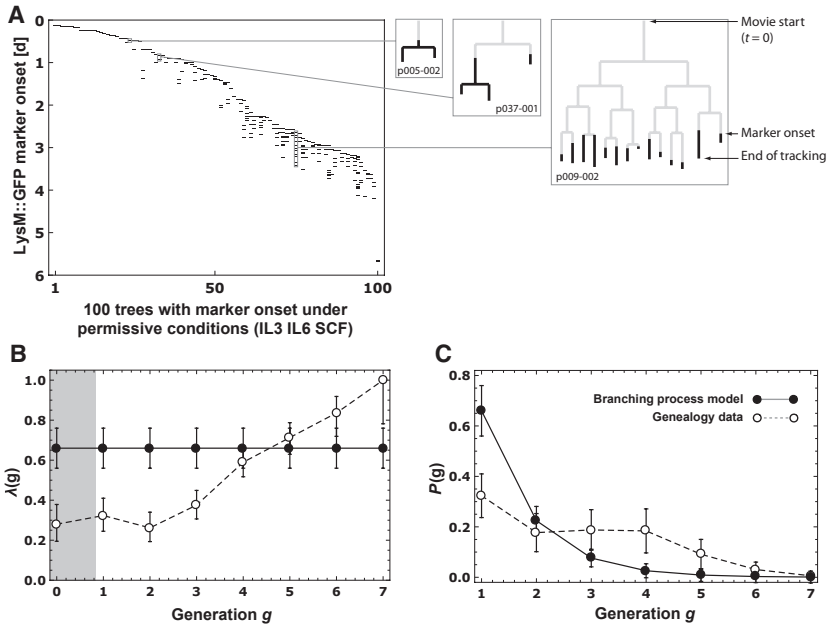

Fig. 3. Analysis of GMP genealogies from single-cell time-lapse microscopy. (A) LysM::GFP marks the loss of bi-potency. We show the onset time in days for 100 trees under permissive culture conditions (interleukins 3 and 6, and SCF) allowing differentiation of both G and M cells [30]. In each tree, multiple marker onsets were observed, depending on the number of leaves. We observe more late and synchronous onsets than expected from the branching process model with the parameters inferred from the colony assay data. In the three tree examples shown on the right, marker onset is indicated by the change from gray to black lines. Each tree is identified by its position and its tree number. (B) We assume that LysM::GFP is an instant marker of differentiation. The differentiation probability $\lambda(g)$ inferred from the genealogy data (open circles) depends on the generation within a tree. This is in contrast to the time-independent λ of the branching process model (filled circles). Error bars indicate 95% confidence intervals, and were determined using the Clopper–Pearson method [31]. (C) The probability that a cell differentiates exactly in generation g, $P(g)$, decreases less sharply than predicted by the branching process model. To calculate the errors, we propagated the larger of the two confidence intervals. Note that cells in generation 0 have not been tracked from birth and a bias towards undifferentiated cells may apply. Thus this data point is disregarded.

$\lambda^{2g}\ (1-\lambda)^{2g-1}$ for $g \geq 1$. For example, we found 13 of 100 trees with a synchronous onset in generation 2, but the probability of this occurring is below 1% if we consider the branching process model with $\lambda = 0.66$ as inferred from the colony assay data.

We thus wished to analyze the differentiation probability λ for different generations in genealogies under permissive conditions. Below, we use LysM::GFP marker onset as an instant reporter for differentiation. Limitations of this assumption are considered in the Discussion.

We divide the number of differentiating cells in generation g (as indicated by LysM::GFP marker onset) by the total number of cells in generation g. In contrast to the branching process model, where we assumed λ to be constant over generations, we observe an increase of λ with generation g (Fig. 3B). If we disregard generation 0 (as cells have not been tracked from birth but from the start of the experiment, and thus a bias towards LysM::GFP-negative cells may apply), λ increases from $\sim 35\%$ in generations 1, 2 and 3, to 100% in generation 7, when all cells differentiate (see Fig. 3B). This generation-dependent differentiation probability results in a less sharp decrease of the probability density $P(g)$ for a cell to differentiate in generation g. Figure 3C shows that the probability of differentiating in generation 5 is still well above 5%, which is in stark contrast to the predictions from the branching process model.

The differentiation probability $\lambda(g)$ defined above describes the probability that a cell differentiates in generation g given that it reaches generation g. In biomedically motivated survival analysis and reliability theory in engineering, analogous concepts are called hazard function and failure rate, respectively [34]. Time-dependent hazard functions and failure rates emerge quite naturally from aging or erosion, for example. In the case of GMP differentiation, a time-dependent differentiation probability may occur for a number of reasons: the medium conditions may change over time, cell–cell signaling may have an effect as the cell density increases, or an inherent program may increasingly force the cells to differentiate. However, the assumption of a generation-independent λ in our

branching model requires re-setting of the GMP after each division to its default progenitor state. Below we describe a mechanistic, molecular model of the differentiation process, and show that a time-dependent differentiation probability emerges naturally in this model.

Molecular toggle switch model

The molecular details of GMP differentiation are still under debate. The most detailed contribution comes, to the best of our knowledge, from Laslo *et al.* [35]. Here, the authors proposed that mutual antagonism between the transcription factor Gfi-1 and the integrated monocytic factor EgrNab (encoded by the genes Egr-1, Egr-2 and Nab, which have redundant molecular functions) mediates the lineage choice of GMPs. Previous analyses suggested a pivotal role for the transcription factors PU.1 and C/EBPα, which are both required for generation of GMPs [36,37]. One hypothesis links the ratio between PU.1 and C/EBPα to a primary cell-fate decision [38], and there is also evidence that other factors are involved in the differentiation process [2,12]. Although it has been unambiguously shown that cytokines can affect the lineage decision [30], the intrinsic toggle switch appears to be determined by the antagonistic Gfi-1 and EgrNab [39].

Assuming that two antagonistic transcription factors control the intrinsic GMP lineage decision that drives differentiation towards granulocytes and monocytes under permissive conditions, we established a chemical reaction kinetics model of a toggle switch involving a granulocytic transcription factor (potentially Gfi-1) and a monocytic transcription factor (potentially EgrNab). We describe the mutual inhibition of the two respective genes using a one-stage model of gene expression, where transcription and translation are lumped together, and with mutual inhibition being realised as DNA–protein binding (see Fig. 4A). The model may be represented as a set of biochemical reactions and a set of reaction rates. The seven reactions describing the synthesis and binding of the granulocytic transcription factor (indexed with G), with symmetric reactions (but potentially different rates) for the monocytic transcription factor (indexed with M) are:

$$\mathrm{DNA_G} \xrightarrow{\alpha_G} \mathrm{DNA_G} + \mathrm{Protein_G} \tag{9}$$

$$\mathrm{Protein} \xrightarrow{\delta_G} \varnothing \tag{10}$$

$$2 \cdot \mathrm{Protein_G} \xrightarrow{\kappa_G^+} \mathrm{Dimer_G} \tag{11}$$

$$\mathrm{Dimer_G} \xrightarrow{\kappa_G^-} 2 \cdot \mathrm{Protein_G} \tag{12}$$

$$\mathrm{Dimer_G} \xrightarrow{\delta_G} \varnothing \tag{13}$$

$$\mathrm{Dimer_G} + \mathrm{DNA_M} \xrightarrow{\tau_G^+} \mathrm{DNA_M^{bound}} \tag{14}$$

$$\mathrm{DNA_M^{bound}} \xrightarrow{\tau_G^-} \mathrm{Dimer_G} + \mathrm{DNA_M} \tag{15}$$

Reaction (9) corresponds to protein synthesis from an unbound promoter. Transcription and translation are aggregated into a single reaction. In general, one should consider extending the above system by explicit transcription and translation reactions, if no clear time-scale separation in the decay rates of mRNA and protein applies. In addition to being more detailed, such a two-stage gene expression model [40] can induce two undecided and two decided attractors [26]. Here, we chose the one-stage model, not only for the sake of simplicity but also because the simpler model reflects the type of observed data more naturally, where we are unable to discriminate between one or two undecided GMP states.

Proteins can either degrade, as shown in reaction (10), or form homodimers, as shown in reaction (11). Homodimers dissociate into two monomers in reaction (12) or are degraded in reaction (13). We assume that the homodimers are degraded at the same rate as monomers. Reactions (14) and (15) describe the binding and unbinding of a homodimer to the antagonistic gene, and thereby transition from an active to an inactive promoter and vice versa. Bound DNA cannot be transcribed, and only dimers can inactivate the promoter. Note that cooperative binding via dimerization is solely included to stabilize the system: Non-cooperative one-stage switches quickly forget a differentiation decision [41] and are therefore inadequate in the context of lineage choice.

The above reactions specify the possible transitions between all states $x \in (\mathbb{N})_0^8$ in the state space of the system. The master equation [42] describes the time evolution of the probability for being in state x, $\mathcal{P}(x)$, as

$$\frac{\mathrm{d}\mathcal{P}(x)}{\mathrm{d}t} = \sum_{x'} [w_{xx'}\mathcal{P}(x') - w_{x'x}\mathcal{P}(x)] \tag{16}$$

The first term considers transitions from states x' with rate $w_{xx'}$ to state x, while the second term accounts for transitions from x to all other possible states x' with transition rates $w_{x'x}$. For complex systems, the master equation cannot be easily solved. However, it can be simulated using Gillespie's algorithm [43]. After a transient time, the simulation leads

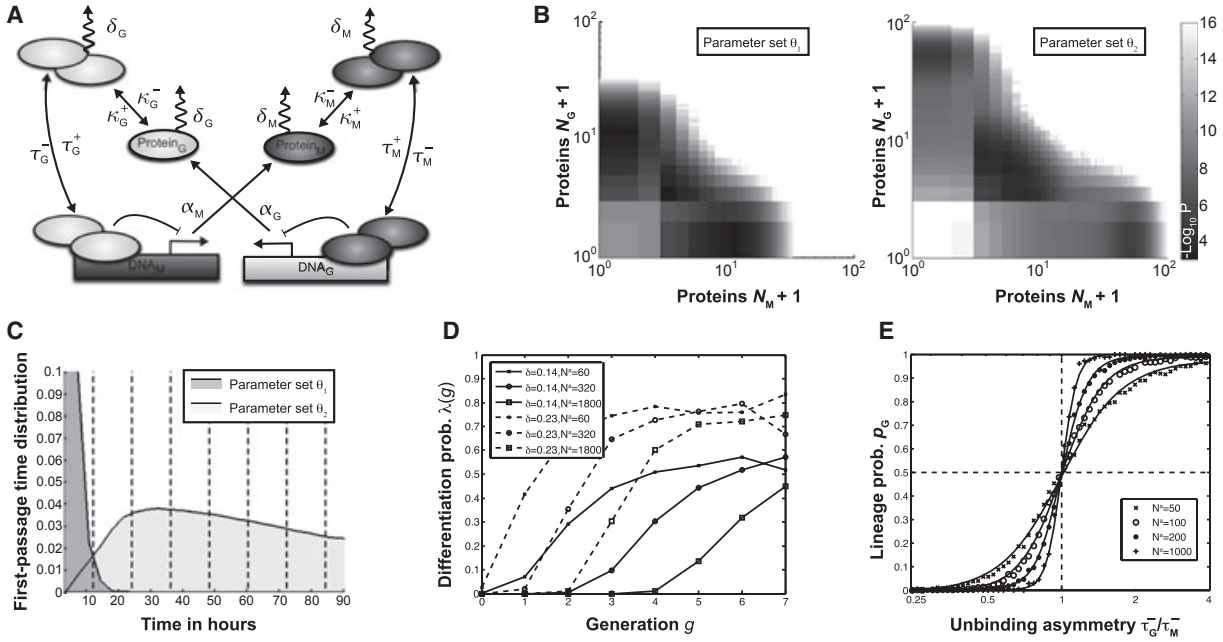

Fig. 4. Molecular toggle switch model. (A) The complex interactions between monocytic and granulocytic factors are simplified as two mutually inhibiting transcription factors, referred to by the indices G and M. DNA_G is transcribed to $Protein_G$, which can bind in a dimerized form to the promoter at DNA_M and inhibit its transcription. The same reactions apply for the monocytic transcription factor. (B) Plot of the resulting state space, with three clearly discernible attractors appearing as regions with high probability (black) for the two parameter sets θ_1 and θ_2. The eight-dimensional state space is projected onto the N_G, N_M plane, showing the total amount of either protein in the system. The central attractor represents the undifferentiated GMP, while the distant attractors represent differentiated cells. The $-\log_{10}$ probability of the steady state is indicated by shades of gray as defined on the right. (C) Probability distributions of the first-passage time from the central attractor to one of the two decided attractors for the two parameter sets θ_1 and θ_2 used in (B). Dashed vertical lines represent the mean cell-cycle length for GMPs, t_{cc} ($\sim$ 12 h), as inferred from the genealogies. (D) Generation dependence of the differentiation probability $\lambda(g)$ for various parameters of the molecular model. The protein decay rate δ and the total protein amount of the dominating transcription factor in the decided attractor, N^a, determine the position and the slope of onset of λ. (E) Differentiation bias p_G as a function of the binding strength of the granulocytic homodimer with, Hill functions fitted to the data. The sensitivity of the switch to asymmetry in the binding rates increases with the number of proteins in the system. The parameter sets used are given in Table S1.

to a quasi-steady state, where the probabilities in state space no longer change: $d\mathcal{P}(x)/dt = 0$. For appropriate parameters (see Doc. S4 and Table S1), we find that the molecular model (Fig. 4A) induces three regions of high probability in state space, referred to as attractors of the system (see Fig. 4B). In the spirit of Waddington's landscape [44], one can associate different cell fates with these attractors. The central attractor represents the undifferentiated GMP state, as both lineage-determining factors are present only in small amounts, neutralizing each other. The two distant attractors represent the differentiated granulocyte and monocyte cell fates, where the corresponding lineage-determining factor is abundant but its antagonist is completely absent. These are referred to as decided attractors.

The two parameter sets θ_1 and θ_2 (see Table S1) result in different steady-state distributions, as shown in Fig. 4B, and thus different first-passage times, as shown in Fig. 4C. The first-passage time is defined as the time required to leave the central attractor and reach any of the decided attractors of the system [45], and can be calculated from the simulations of the toggle switch model (see Docs S1 and S2). Figure 4C shows the first-passage time distribution for the two parameter sets θ_1 and θ_2, resembling an exponential distribution and a γ distribution, respectively. These are two common distributions for first-passage times in stochastic systems [46]. The origin of these different distributions was studied with respect to the parameters of the system. For parameter set θ_1, the central attractor is narrow and the two decided attractors are close by in terms of protein numbers (see Fig. 4B). Through fluctuations in protein numbers, the system is quickly driven out of the central attractor. Once it has left the central attractor, only few synthesis reactions are required to reach one or the other decided attractor. Therefore, the transition time between attractors is negligible, leading to an overall exponential first-passage time distribution whose parameters depend on the

high rate of escape from the central attractor. For parameter set θ_2, the central attractor is wider, and therefore more time passes until the system leaves the central attractor by chance. Additionally, significant time is required to bridge the distance to the decided attractors, which is larger than in parameter set θ_1 (note the log scale in Fig. 4B). The small rate of escape leads to the long tail of the first-passage time distribution of parameter set θ_2 in Fig. 4C, and the time spent moving between attractors causes a shift of the distribution to the right. The distribution in this case resembles a γ distribution, which is characteristic of random walks over long distances with a bias in one direction [46]. The first-passage time distributions observed in Fig. 4C may be interpreted in the context of differentiation probability. The exponentially distributed first-passage time corresponds to a time-independent differentiation probability $\lambda(g) = \lambda$, as the exponential distribution is memory-less. In this case, our simple branching model is a valid and useful abstraction of the system. However, if the first-passage time follows a non-exponential distribution (as seen in Fig. 4C), the differentiation probability is time-dependent and the simple branching process cannot reflect this property. Our molecular model therefore can induce either time-dependent or -independent differentiation probabilities, as determined by the kinetic rates of the switch. It shows how these two very different scenarios of differentiation can be traced to the same molecular origin: the first-passage time.

Next we investigate more systematically how the time-dependence of λ relies on the choice of parameters of the molecular model. Note that we only consider symmetric systems, i.e. the synthesis, degradation, binding and unbinding rates for both transcription factors are the same ($\delta_G = \delta_M = \delta$, etc.). Figure 4D shows $\lambda(g)$ for varying degradation rates δ and $N^a = \alpha/\delta$, representing the protein levels of the dominant transcription factors in the decided attractor (see Doc. S1). All curves show similar characteristics. After a transient time, the curves grow almost linearly before asymptotically approaching an upper bound. The higher the protein levels, the later the onset of growth in the curves. This is intuitive, as higher protein levels imply larger distances between the attractors, which sets a minimal time before the system is able to reach the decided attractors. The probability to observe a decided system before this time is 0. The different asymptotics of the curves may be attributed to the degradation rate, which sets the time scale of the system. For a constant protein level N^a, a high degradation rate δ implies a high synthesis rate α as $N^a = \alpha/\delta$. This speeds up the dynamics of the whole system, increasing the proportion of simulations that escape from the central attractor per unit time, which is in fact $\lambda(g)$.

Note that we disregard the tree structure of the genealogies in our molecular model, and instead simulate single branches corresponding to the time series of a single cell and its ancestors. However, by calculating the proportion of cells reaching a decided attractor within the time window of one generation, we can calculate the differentiation probability λ as a function of the generation g, and thus establish a one-to-one correspondence between the first-passage time in the molecular model and the probability density $P(g)$ in the branching process.

By adjusting the parameters of the model, we can also simulate a biased differentiation towards the one or the other lineage. To test this, we systematically changed the binding strength of the granulocytic homodimer (τ_G^+/τ_G^-) while keeping the binding strength of the monocytic homodimer constant. All remaining parameters are kept symmetric. In Fig. 4E, we show how the probability p_G for the granulocyte lineage changes in response to varying binding strength. Intuitively, as both binding strengths are equal, the system has equal probability of differentiating towards the granulocytic or the monocytic attractor. Increasing the binding strength τ_G^+ leads to stronger repression of the monocytic factor. Similarly, a decrease of binding strength leads to a disadvantage for the granulocytic factor and the probability for monocytic commitment is increased. Interestingly, the response of the differentiation bias p_G to the binding strength is influenced by the amount of proteins in the decided states N. For low protein numbers, small differences in binding strength of the two transcription factors still result in a balanced decision towards either G or M ($p_G \sim 0.5$). For higher protein numbers, even small differences in binding strength will destroy the balance between the two factors, and the favored transcription factor will almost certainly prevail. This sensitive response is interesting in the light of lineage instruction: in cytokine medium, Granulocyte/macrophage colony-stimulating factors (G-CSF and M-CSF) are able to instruct differentiation with a high reliability [30].

Bayesian parameter inference identifies a scale separation of decay rates

Having shown how different first-passage times and probabilities of commitment towards one or the other lineage emerge from a simple model of a toggle switch, it is now possible to fit the model to observed quantities in order to estimate molecular rates in a proof-of-concept way.

Analytical expressions of the first-passage time distribution and the lineage probabilities for the toggle switch are hard to derive. Therefore, we have to resort to approximate Bayesian computing [47] to infer molecular parameters from the observed differentiation and commitment probabilities. Approximate Bayesian computing allows parameter inference even though no analytical expression for the likelihood of the data given the parameters is available. The method substitutes the likelihood evaluation by forward simulation of the model and comparison of the simulated data to the observed data using a distance function. Instead of maximizing the likelihood, approximate Bayesian computing searches for parameters that minimize the distance function, thereby best fitting the data.

We fit the model parameters θ to the data with respect to the distance function

$$d(\theta) = d\left(\left(p_G^\theta, \lambda^\theta\right), \left(p_G^{obs}, \lambda^{obs}\right) \right) = \frac{1}{2}\left(\frac{1}{7}\sum_{g=1}^{7} \left|\lambda^\theta(g) - \lambda^{obs}(g)\right| + \left|p_G^\theta - p_G^{obs}\right| \right) \quad (17)$$

where $\lambda^{obs}(g)$ is the observed differentiation probability in generation $g = 1 \ldots 7$ (depicted in Fig. 3B). Similarly, $\lambda^\theta(g)$ is the differentiation probability obtained from simulations with parameters θ. Finally, p_G^{obs}, is the observed probability of differentiating to granulocytes (see Fig. 2) and p_G^θ is its simulated counterpart. Note that we use p_G^{obs} as obtained from colony assay data under the assumption of time-independent λ. More rigorously, p_G^{obs} would have to be replaced by an estimate that accounts for time-dependent λ, which we neglect in this study. The first term on the right side of eqn (17) quantifies how closely the parameter θ can reproduce the observed differentiation probabilities, whereas the second term quantifies its match to the observed lineage bias.

We used a standard ABC algorithm based on sequential monte carlo (ABC SMC) implemented by a customized version of the ABC-SysBio toolkit [48] to fit the toggle switch model to the observed data. We made a quasi-steady state assumption for the dimerization and DNA–protein binding reactions (see Doc. S3), leading to six parameters to be estimated: two synthesis rates (α_G, α_M), two degradation rates (δ_G, δ_M) and two dissociation constants ($K_G = \tau_G^-/\tau_G^+$, $K_M = \tau_M^-/\tau_M^+$) obtained from the quasi-steady-state assumption. We assigned flat prior distributions, constraining the parameters to biologically relevant regimes (see Table S2 for a list of the bounds used).

We iterated 10 populations consisting of 200 parameter sets, where λ^θ and p_G^θ were estimated from 1000 repeated simulations for each parameter set θ. The last population contained only parameter sets with $d(\theta) < 0.1$, giving a good fit to the data already (see Fig. 5A). Afterwards, we calculated the probability density $P(g)$ from the simulated $\lambda(g)$, and found that, apparently, moderate deviations in $\lambda(g)$ result in considerable deviations in $P(g)$ (see Fig. 5B). Thus, we calculate the distance $d(P^\theta, P^{obs}) = \sum_{g=1}^{7} \left|P^\theta(g) - P^{obs}(g)\right|$ for each parameter set to quantify its goodness of fit within the obtained posterior distribution.

We found asymmetric protein degradation rates for the best fits to the experimental data (black dots in Fig. 5C): a high δ_G implies a low δ_M, and *vice versa*. In contrast to the protein degradation, synthesis rates appear mostly correlated (see Fig. 5D). Therefore, the best fits result in systems where the degradation rate of one transcription factor may be c times that of the other (e.g. $\delta_G = c\ \delta_M$). At the same time, the number of proteins separates in an inverse fashion ($N_G^a = 1/c\, N_M^a$). This separation of scales induces qualitatively different $\lambda(g)$ curves for the two transcription factors: with regard to the above example, the granulocytic transcription factor not only is the slope in $\lambda(g)$ decreased due to a lower δ_G, but also the onset of λ is shifted due to the higher N_G^a. Note that, as shown in Fig. 4D and detailed in Doc. S5, the degradation rate δ but not the synthesis rate α determines the dynamics of the system. The differentiation probability $\lambda(g)$, as observed in the time-lapse data, is thus best fitted by a system where one transcription factor decides quickly, while the other takes longer to execute the differentiation decision. Interestingly, both transcription factors can act as the slow or the fast species in the system, implying that the asymmetry in degradation rates is independent of the lineage bias, i.e. the preference for one cell fate ($p_G > p_M$) does not induce separation of scales in the differentiation dynamics, but only the shape of $\lambda(g)$.

Discussion

Here we have illustrated how to link a branching process model to the molecular model of a toggle switch based on colony assay data and cellular genealogies from time-lapse microscopy. We report a discrepancy between the branching process with time-independent λ and the generation dependent $\lambda(g)$ observed in trees from single cell time-lapse microscopy. We are not the first to state limitations of this simple branching process model [49]; however, we do outline possible extensions. First, the explicit time dependence of λ can be

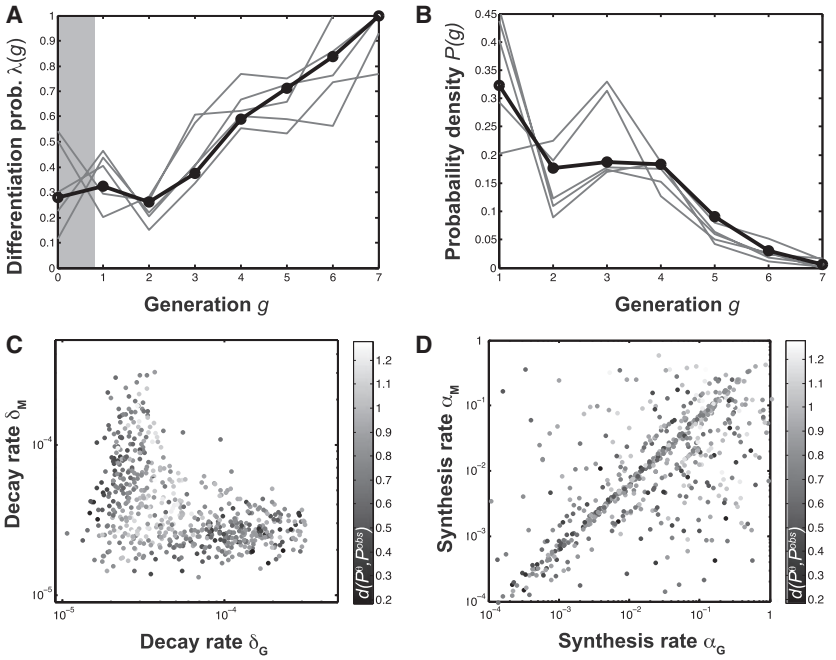

Fig. 5. Inference of model parameters using approximate Bayesian computing. (a) Fit of the best five parameter sets (gray) to experimentally observed $\lambda(g)$ (black). (B) The corresponding probability densities $P(g)$ show considerable deviations from the observed data due to their high sensitivity to small errors in the fit to λ. (C,D) Scatter plots of the degradation rate (C) and synthesis rate (D) in the last iteration of ABC sequential monte carlo algorithm. (population size increased from 200 to 600 by re-sampling). The distance $d(P^{\theta}, P^{obs})$ of each parameter set θ to the observed probability densities $P(g)$ as shown in (B) is indicated by shades of gray as defined on the right, where black corresponds to small distance. Interestingly, the best fits show a split of the decay rates of the two transcription factors, implying that differentiation towards one lineage is quick, while the other needs more time to reach the decided attractor.

incorporated into the branching process. However, within this framework, it is impossible to reconcile the lineage probabilities p_G and p_M inferred from the colony assay with the $\lambda(g)$ from the time-lapse experiments (data not shown). Second, a delay between loss of bi-potency and marker onset would shift onset times, allowing later onsets as observed in Fig. 3A. However, mere inclusion of a delay will not result in a generation-dependent $\lambda(g)$, but only in a shift of the constant differentiation probability λ. Moreover, a constant delay of, say, one generation is incompatible with cells differentiating in generation 0, as observed in the time-lapse data. Application of a more sophisticated methodology, such as the concept of hidden trees as described previously [50], appears promising to account for marker delay and possibly allows inference of the exact time of differentiation. Third, a heterogeneous differentiation status of the GMPs, induced by an impure sorted population of starting cells, may induce the observed time dependence of the differentiation probability. However, the effect of the contributing populations and their respective mixing is beyond the scope of the present study. Fourth, to account for the synchrony of marker onsets between sister and cousin cells, which are ignored in the molecular model, either spatial correlations via cell–cell signaling, medium change or marker delay should be incorporated in an extended version of the model. Finally, inclusion of cell death would alter the differentiation probability inferred from the colony assay data. For the permissive culture conditions used in the experiments [30], the exclusion of cell death appears to be an appropriate assumption, as the combination of interleukins 3 and 6 and SCF leads to strong proliferation and only limited cell death.

Most importantly, we showed how the observed time-dependent differentiation probability can emerge from a molecular toggle switch model for appropriate parameter sets. From our simple parameter fitting approach, we conclude that one can extract valuable information about molecular parameters from macroscopic observables such as $P(g)$ and p_G derived from genealogies only, without observing the time courses of the toggle switch system directly. This suggest that the parameters of more realistic but more complex models are still identifiable if one combines genealogies and time-lapse fluorescence microscopy. In a possible extension, the mRNA stage of gene expression may be

included, as this changes the overall dynamics of the toggle switch by introducing multiple progenitor states [26]. In such a model, progenitor cells consist of two distinct sub-populations, each inclined towards one lineage. However, the prediction derived from the simple toggle switch considered here is the split into a quick and a slow differentiating lineage induced by clearly differing decay rates of the antagonistic transcription factors.

In summary, we show that, while the bulk data from a colony assay provides indications of the differentiation dynamics, only time-lapse microscopy followed by cell tracking can reliably reveal dynamic features at the single-cell level: generation times, marker onsets, and subsequently the generation-dependent differentiation probability. In contrast to the expectations from the branching process, loss of bi-potency of GMPs (as detected by marker onset) appears later and in a more synchronized fashion. In a future analysis, additional features of the time-lapse microscopy data, such as cell–cell contact or morphological changes, may be taken into account. Finally, using a data set with lineage annotations, it may be possible to infer more molecular parameters and test the prediction of our model: a difference in differentiation dynamics of the two lineages.

Acknowledgements

We thank two unknown reviewers for helpful comment on the manuscript, and Max Endele and Oliver Hilsenbeck (both Helmholtz Zentrum München) for technical support. This work was supported by the Helmholtz Alliance on Systems Biology (project 'CoReNe'), the European Research Council (starting grant 'LatentCauses'), and the Deutsche Forschungsgemeinschaft (SPP 1356 'Pluripotency and Cellular Reprogramming').

References

1 Orkin SH & Zon LI (2008) Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644.
2 Dahl R (2009) Development of macrophages and granulocytes. In *Molecular Basis of Hematopoiesis* (Wickrema A & Kee B, eds), pp. 127–149. Springer, Berlin.
3 Viswanathan S & Zandstra P (2003) Towards predictive models of stem cell fate. *Cytotechnology* **41**, 75–92.
4 Whichard ZL, Sarkar CA, Kimmel M & Corey SJ (2010) Hematopoiesis and its disorders: a systems biology approach. *Blood* **115**, 2339–2347.
5 Till JE, McCulloch EA & Siminovitch L (1964) A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc Natl Acad Sci USA* **51**, 29–36.
6 Nordon RE, Ko KH, Odell R & Schroeder T (2011) Multi-type branching models to describe cell differentiation programs. *J Theor Biol* **277**, 7–18.
7 Wellard C, Markham J, Hawkins ED & Hodgkin PD (2010) The effect of correlations on the population dynamics of lymphocytes. *J Theor Biol* **264**, 443–449.
8 Glauche I, Cross M, Loeffer M & Roeder I (2007) Lineage specification of hematopoietic stem cells: mathematical modeling and biological implications. *Stem Cells* **25**, 1791–1799.
9 Colijn C & Mackey M (2007) Bifurcation and bistability in a model of hematopoietic regulation. *SIAM J Appl Dyn Syst* **6**, 378–394.
10 Kirouac DC, Madlambayan GJ, Yu M, Sykes EA, Ito C & Zandstra PW (2009) Cell–cell interaction networks regulate blood stem and progenitor cell fate. *Mol Syst Biol* **5**, 293.
11 Kirouac DC, Ito C, Csaszar E, Roch A, Yu M, Sykes EA, Bader GD & Zandstra PW (2010) Dynamic interaction networks in a hierarchically organized tissue. *Mol Syst Biol* **6**, 417.
12 Krumsiek J, Marr C, Schroeder T & Theis FJ (2011) Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS ONE* **6**, e22649.
13 Wang J, Zhang K, Xu L & Wang E (2011) Quantifying the Waddington landscape and biological paths for development and differentiation. *Proc Natl Acad Sci USA* **108**, 8257–8262.
14 Bokes P, King JR & Loose M (2009) A bistable genetic switch which does not require high cooperativity at the promoter: a two-timescale model for the PU.1–GATA-1 interaction. *Math Med Biol* **26**, 117–132.
15 Chickarmane V, Enver T & Peterson C (2009) Computational modeling of the hematopoietic erythroid–myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol* **5**, e1000268.
16 Duff C, Smith-Miles K, Lopes L & Tian T (2012) Mathematical modelling of stem cell differentiation: the PU.1–GATA-1 interaction. *J Math Biol* **64**, 449–468.
17 Huang S, Guo YP, May G & Enver T (2007) Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol* **305**, 695–713.
18 Roeder I & Glauche I (2006) Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. *J Theor Biol* **241**, 852–865.
19 Cherry JL & Adler FR (2000) How to make a biological switch. *J Theor Biol* **203**, 117–133.
20 Bialek W (2001) Stability and noise in biochemical switches. In *Advances in Neural Information Processing*

Systems 13: Proceedings of the 2000 Conference (Leen TK, Dietterich TG and Tresp V, eds), pp. 103. MIT Press, Cambridge, MA.
21 Warren P & Ten Wolde P (2004) Enhancement of the stability of genetic switches by overlapping upstream regulatory domains. *Phys Rev Lett* **92**, 128101.
22 Lipshtat A, Loinger A, Balaban NQ & Biham O (2006) Genetic toggle switch without cooperative binding. *Phys Rev Lett* **96**, 188101.
23 Walczak A, Sasai M & Wolynes P (2005) Self-consistent proteomic field theory of stochastic gene switches. *Biophys J* **88**, 828–850.
24 Fritz G, Buchler N, Hwa T & Gerland U (2007) Designing sequential transcription logic: a simple genetic circuit for conditional memory. *Syst Synth Biol* **1**, 89–98.
25 Barzel B & Biham O (2008) Calculation of switching times in the genetic toggle switch and other bistable systems. *Phys Rev E Stat Nonlin Soft Matter Phys* **78**, 041919.
26 Strasser M, Theis FJ & Marr C (2012) Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophys J* **102**, 19–29.
27 Orkin SH & Zon LI (2008) SnapShot: hematopoiesis. *Cell* **132**, 712.
28 Harris T (1963) *The Theory of Branching Processes*. Springer, Berlin.
29 Watson H & Galton F (1875) On the probability of the extinction of families. *J Anthropol Inst Great Britain Ireland* **4**, 138–144.
30 Rieger MA, Hoppe PS, Smejkal BM, Eitelhuber AC & Schroeder T (2009) Hematopoietic cytokines can instruct lineage choice. *Science* **325**, 217–218.
31 Clopper C & Pearson E (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.
32 Eilken HM, Nishikawa SI & Schroeder T (2009) Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature* **457**, 896–900.
33 Faust N, Varas F, Kelly LM, Heck S & Graf T (2000) Insertion of enhanced green fluorescent protein into the lysozyme gene creates mice with green fluorescent granulocytes and macrophages. *Blood* **96**, 719–726.
34 Lee ET & Go OT (1997) Survival analysis in public health research. *Annu Rev Public Health* **18**, 105–134.
35 Laslo P, Spooner CJ, Warmflash A, Lancki DW, Lee HJ, Sciammas R, Gantner BN, Dinner AR & Singh H (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* **126**, 755–766.
36 Iwasaki H, Somoza C, Shigematsu H, Duprez EA, Iwasaki-Arai J, Mizuno SI, Arinobu Y, Geary K, Zhang P, Dayaram T *et al.* (2005) Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood* **106**, 1590–1600.
37 Dakic A, Metcalf D, Di Rago L, Mifsud S, Wu L & Nutt SL (2005) PU.1 regulates the commitment of adult hematopoietic progenitors and restricts granulopoiesis. *J Exp Med* **201**, 1487–1502.
38 Dahl R, Walsh JC, Lancki D, Laslo P, Iyer SR, Singh H & Simon MC (2003) Regulation of macrophage and neutrophil cell fates by the PU.1:C/EBPα ratio and granulocyte colony-stimulating factor. *Nat Immunol* **4**, 1029–1036.
39 Laslo P, Pongubala JMR, Lancki DW & Singh H (2008) Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin Immunol* **20**, 228–235.
40 Shahrezaei V & Swain PS (2008) Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* **105**, 17256–17261.
41 Loinger A, Lipshtat A, Balaban NQ & Biham O (2007) Stochastic simulations of genetic switch systems. *Phys Rev E Stat Nonlin Soft Matter Phys* **75**, 021904.
42 Van Kampen NG (1992) *Stochastic Processes in Physics and Chemistry*. North-Holland Press, Amsterdam.
43 Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81**, 2340–2361.
44 Waddington CH (1956) *Principles of Embryology*. Allen & Unwin, London.
45 Walczak A, Onuchic J & Wolynes P (2005) Absolute rate theories of epigenetic stability. *Proc Natl Acad Sci USA* **102**, 18926–18931.
46 Bel G, Munsky B & Nemenman I (2010) The simplicity of completion time distributions for common complex biochemical processes. *Phys Biol* **7**, 016003.
47 Toni T, Welch D, Strelkowa N, Ipsen A & Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* **6**, 187–202.
48 Liepe J, Barnes C, Cule E, Erguler K, Kirk P, Toni T & Stumpf M (2010) ABC-SysBio – approximate Bayesian computation in Python with GPU support. *Bioinformatics* **26**, 1797–1799.
49 Bjerknes M (1986) A test of the stochastic theory of stem cell differentiation. *Biophys J* **49**, 1223–1227.
50 Olariu V, Coca D, Billings SA, Tonge P, Gokhale P, Andrews PW & Kadirkamanathan V (2009) Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages. *Bioinformatics* **25**, 2824–2830.

Supporting information

The following supplementary material is available:
Doc. S1. Definition of the attractors of the system.
Doc. S2. Calculation of the first-passage time from the simulations.

Doc. S3. Quasi-steady state approximation used in the approximate Bayesian computing.
Doc. S4. Parameters used in the simulation.
Doc. S5. Effect of degradation rate on the time scale of attractor transitions.
Table S1. Parameter sets for the molecular model used in Fig. 4B,C.
Table S2. Bounds of the uniform priors.

This supplementary material can be found in the online version of this article.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.