PLOS | COMPUTATIONAL BIOLOGY

Modeling large fluctuations of thousands of clones during hematopoiesis: The role of stem cell self-renewal and bursty progenitor dynamics in rhesus macaque

Song Xu[1], **Sanggu Kim**[2], **Irvin S. Y. Chen**[3], **Tom Chou**[1,4*]

1 Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, California, United States of America, **2** Department of Veterinary Biosciences, The Ohio State University, Columbus, Ohio, United States of America, **3** UCLA AIDS Institute and Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, California, United
States of America, **4** Department of Mathematics, University of California, Los Angeles, Los Angeles, California, United States of America

* tomchou@ucla.edu

Citation: Xu S, Kim S, Chen ISY, Chou T (2018) Modeling large fluctuations of thousands of clones during hematopoiesis: The role of stem cell selfrenewal and bursty progenitor dynamics in rhesus macaque. PLoS Comput Biol 14(10): e1006489. https://doi.org/10.1371/journal.pcbi.1006489

Editor: Thomas Ho¨fer, German Cancer Research Center, GERMANY

Received: March 23, 2017

Accepted: September 5, 2018

Published: October 18, 2018

DataAvailabilityStatement: Data analyzed have been previously published in Cell Stem Cell, volume 14, no. 4, pp 473-485, 2014 (https://doi.org/10.1016/j.stem.2013.12.012). No figures, plots, or tables in this manuscript, or any parts of them, have been previously published. Figures, plots, and tables have all been generated de novo from published raw data, specifically for this manuscript.

Funding: This study was supported by grants from the National Institutes of Health (www.nih.gov<http://www.nih.gov>, R56HL126544,

RESEARCH ARTICLE

Abstract

In a recent clone-tracking experiment, millions of uniquely tagged hematopoietic stem cells (HSCs) and progenitor cells were autologously transplanted into rhesus macaques and peripheral blood containing thousands of tags were sampled and sequenced over 14 years to quantify the abundance of hundreds to thousands of tags or “clones.” Two major puzzles of the data have been observed: consistent differences and massive temporal fluctuations of clone populations. The large sample-to-sample variability can lead clones to occasionally go “extinct” but “resurrect” themselves in subsequent samples. Although heterogeneity in HSC differentiation rates, potentially due to tagging, and random sampling of the animals’ blood and cellular demographic stochasticity might be invoked to explain these features, we show that random sampling cannot explain the magnitude of the temporal fluctuations. Moreover, we show through simpler *neutral* mechanistic and statistical models of hematopoiesis of tagged cells that a broad distribution in clone sizes can arise from stochastic HSC self-renewal instead of tag-induced heterogeneity. The very large clone population fluctuations that often lead to extinctions and resurrections can be naturally explained by a generation-limited proliferation constraint on the progenitor cells. This constraint leads to bursty cell population dynamics underlying the large temporal fluctuations. We analyzed experimental clone abundance data using a new statistic that counts clonal disappearances and provided least-squares estimates of two key model parameters in our model, the total HSC differentiation rate and the maximum number of progenitor-cell divisions.

Author summary

Hematopoiesis of virally tagged cells in rhesus macaques is analyzed in the context of a mechanistic and statistical model. We find that the clone size distribution and the temporal variability in the abundance of each clone (viral tag) in peripheral blood are consistent with (i) stochastic HSC self-renewal during bone marrow repair, (ii) clonal aging that restricts the number of generations of progenitor cells, and (iii) infrequent and smallsize samples. By fitting data, we infer two key parameters that control the level of fluctuations of clone sizes in our model: the total HSC differentiation rate and the maximum proliferation capacity of progenitor cells. Our analysis provides insight into the mechanisms of hematopoiesis and a framework to guide future multiclone barcoding/lineage tracking measurements.

R01AI110297, R01HL125030, U19AI117941 to ISYC and R00HL116234 to SK), the UCLA Center for Aids Research (http://aidsinstitute.ucla.edu/cfar, AI028697 to ISYC), the UCLA AIDS Institute (http://aidsinstitute.ucla.edu/), the James B. Pendleton Charitable Trust, the McCarthy Family Foundation (http://mccarthyfamilyfdn.org/), the Army Research Office (http://www.arl.army.mil W911NF-14-10472 to TC), and the National Science Foundation, (www.nsf.gov<http://www.nsf.gov>, DMS1516675 and DMS-1814364 to TC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competinginterests: Dr. Irvin S. Y. Chen has a financial interest in CSL Behring and Calimmune Inc. No funding was provided by these companies to support this work.

Introduction

Hematopoiesis is a process by which hematopoietic stem cells (HSCs) produce all the mature blood in an animal through a series of proliferating and differentiating divisions [1]. Maintenance of balanced hematopoietic output is critical for an organism's survival and determines its response to disease and clinical procedures such as bone marrow transplantation [2–5]. How the relatively small HSC population generates more than 10^{11} cells of multiple types daily over an organism's lifetime has yet to be fully understood. HSCs are defined primarily by their function but are often quiescent [6]. *In vivo*, it is hard to track the dynamics of individual HSCs, while HSCs *in vitro* do not typically proliferate or differentiate as efficiently. Therefore, the dynamics of HSCs can be inferred only from analyses of populations of progenitors and differentiated blood cells [7] and it is useful to investigate HSC dynamics through mathematical modeling and simulations [8–10].

While most studies model population-level HSC behavior [5, 11, 12], certain aspects of HSCs, such as individual-level heterogeneity in repopulation and differentiation dynamics, have to be studied on a single-cell or clonal level [13]. Single HSC transplant mouse data [14] and clonal tracking of HSCs [15, 16] in mice have shed some light on repopulation dynamics under homeostasis and after bone marrow transplantation [5, 17, 18]. However, murine studies usually involve only one or a few clones. How each individual HSC contributes to the blood production process over long times in much larger human and non-human primates is less clear and more difficult to study. Also, unlike in mice, there is no way to isolate and mark HSC populations in human [19].

Recently, results of a long-term clonal tracking of hematopoiesis in normal-state rhesus macaques has been made available [13, 20]. The experiment extracted and uniquely "labelled" hematopoietic stem and progenitor cells (HSPCs) from four rhesus macaques with viral tags that also carry an enhanced green fluorescent protein gene. After autologous transplantation, if any of the tagged HSPCs divide and differentiate, its progeny will inherit their unique tags and ultimately appear in the peripheral blood. Blood samples were drawn every few months over 4 − 14 years (depending on the animal) and the sampled cells were counted and sequenced. Of the *$10^6 - 10^7$ unique HSPC tags transplanted, *$10^2 - 10^3$ clones were detected in the sampled peripheral blood. In the original paper describing the clonal tracking experiment, Kim *et al*. [13] observed "A small fraction (4 − 10%) of tagged clones predominately contribute to a large fraction (25 − 71%) of total blood repopulation." They described the fluctuations of tags that appeared in each sample as "waves of clones", but did not address why some clones can disappear at certain times and reappear in a latter sample.

In this study, we seek to better understand the observed clone size distributions and the large temporal variability in clonal populations. To address these observations, we ask: Is heterogeneity in HSCs necessary for peripheral blood clone size heterogeneity, or can a neutral model explain clone size differences? Are clones that disappear and reappear from sample to sample simply missed by random blood sampling, or do other mechanisms of temporal variability need to be invoked?

Unlike other previous models that describe the evolution of lineages of different cell types and their regulation [8–10, 21], we will consider simpler neutral models that describe the dynamics of specifically granulocyte populations carrying different tags. Of central interest is the competition among the thousands of clones under a neutral environment that gives rise to fluctuations, extinctions, and resurrections in individual clone populations. Even when considering only one cell type, realistic mathematical models may need to include complex multilevel biochemical feedback mechanisms of regulation [8, 22–27]. Many mechanisms may contribute to temporal fluctuations, including extrinsic noise and heterogeneity of HSCs, progenitors, or mature granulocytes. Large time gaps between samplings (5 − 11 months) and small sample sizes also add to the uncertainty of the underlying dynamics. Trying to infer all possible mechanisms and associated parameters from the experimental data would essentially be an overfitting problem. In order to feasibly compare with experimental data, our modeling philosophy will be to recapitulate these complexities into simple, effective models and infer parameters that subsume some of these regulatory effects. This approach and level of modeling are similar to those taken by *e.g*., Yang, Sun, and Komarova [28, 29].

After careful consideration of a number of key physiological mechanisms, we hypothesize that stochastic HSC self-renewal, generation-limited progenitor cell proliferation, and smallsize sampling frequency statistics provide the simplest reasonable explanation for the observed clonal size variability and large temporal fluctuations. HSCs that are generated from self-renewal of the founder population share the same tag as their founder HSC. Thus, during intense self-renewal after myeloablative treatment and HSPC transplantation, each originally transplanted HSCs begets a clonal HSC subpopulation. Subsequently, heterogeneous clone sizes are stochastically generated even though each tag was initially represented by only a single cell. These expanded HSC clones then go on to repopulate the clones in the progenitor and mature blood population, which are also distinguishable by their corresponding tags.

Relative to HSCs, progenitor cells have limited proliferative potential that can explain the apparent extinctions of clones in blood samples. This limited proliferation potential can be thought of as an “aging” process. Different types of aging, including organism aging [23, 30, 31], replicative senescence of stem cells [32], and generation-dependent birth and death rates, have been summarized by Edelstein *et al*. [33]. Here, the clonal “aging” mechanism we invoke imposes a limit to the number of generations that can descend from each newly created (from HSC differentiation) “zeroth generation” progenitor cell. Possible sources of such a limit include differentiation-induced loss of division potential [34] and telomere shortening (as in the Hayflick limit) [35–37]. Mathematically, genealogical aging can be described by tracking cell populations within each generation. After a certain number of generations, progenitor cells of the final generation stop proliferating and can only differentiate into circulating mature cells or die.

In the following sections, we first present the mathematical equations and corresponding solutions (whenever possible) of a model that incorporates the above processes. We then develop a new statistical measure that tracks the numbers of absences of clones across the samples. Measured clone abundances of animal RQ5427 are statistically analyzed within our mechanistic model to infer estimates for key model parameters. The data and corresponding statistical analyses for animals 2RC003 and RQ3570 are also provided in the Results section.

Materials and methods

Below, we describe available clonal abundance data, mechanistic models, and a statistical model we will use for parameter inference.

Clone abundance data

In the experiments of Kim *et al*. [13], cells in samples of peripheral blood were sequenced and counted to extract ^$S^{þ}$ðt_jÞ, the total number of EGFP+ tagged cells in sample 1 j J taken at time t_j. After PCR amplification and sequencing, ^f_iðt_jÞ, the relative abundance of the i^{th} tag among all sampled, tagged cells is also quantified. The "^" notation will henceforth indicate experimentally measured quantities.

Within mature peripheral blood, lymphocytes such as T cells and B cells proliferate or transform in response to unpredictable but clone-specific immune signals [38]. They also vary greatly in their lifespans, ranging from days in the case of regular T and B cells to years in the case of memory B cells. On the other hand, mature granulocytes do not proliferate in peripheral blood and have relatively shorter life spans [7]. Granulocyte dynamics can thus be analyzed with fewer confounding factors [11]. Thus, in this paper, we restrict our analysis to granulocyte repopulation and extract all variables, including ^$S^{þ}$ðt_jÞ and ^f_iðt_jÞ described above, that are associated exclusively with granulocyte populations.

In Fig 1(a), we plot the total numbers of sampled granulocytes from one of the macaques, RQ5427. The subpopulation of EGFP+ granulocytes and the subset of EGFP+ granulocytes that were extracted for PCR amplification and analysis are also plotted. Data for two other animals, 2RC003 and RQ3570, are qualitatively similar. Blood samples from a fourth animal, 95E132, were not separated in to granulocyte and peripheral blood mononuclear cells (PBMCs) before sequencing. Thus, clonal abundances for granulocytes are not available from 95E132. There are only three animals for which we can analyze clonal abundances of granulocytes. For more specifics on the data, see supplemental files of the original experimental paper [13]. As shown in Fig 1(b), not only are the clone abundances ^f_iðt_jÞ heterogeneous,

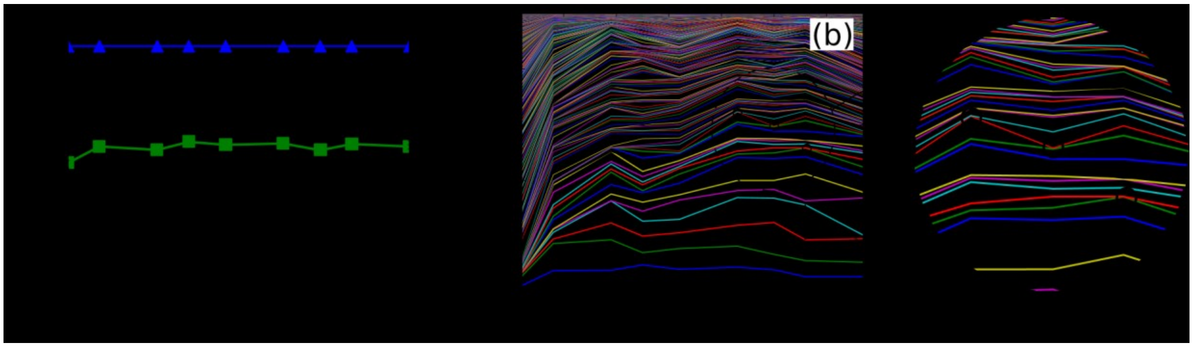


Fig 1. Blood sample data from animal RQ5427 [13]. (a) The total numbers of sampled granulocytes (blue triangles), EGFP+ granulocytes (green squares), and the subset of EGFP+ granulocytes that were properly tagged and quantifiable were extracted for PCR amplification and analysis (black circles). This last population defined by ^Sþðt_jÞ is used to normalize clone cell counts. We excluded the first sample at month 2 in our subsequent analysis so, for example, the sample at month 56 is labeled the 7th sample. There were 536 clones detected at least once across the eight samples taken over 67 months comprising an average fraction 0.052 of all granulocytes. The abundances of granulocyte clones are shown in (b). The relative abundance $\hat{f}_i$ðt_jÞ of granulocytes from the i^{th} clone measured at month t_j is indicated by the vertical distances between two adjacent curves. The relative abundances of individual clones feature large fluctuations over time. “Extinctions” followed by subsequent “resurrections,” were constantly seen in certain clones as indicated by the black circles in (b) and in the inset (c).

https://doi.org/10.1371/journal.pcbi.1006489.g001

but individual clone abundances vary across samples taken at different times. The variation is so large that many clones can go extinct and reappear from one sample to another, as shown in Fig 1(c). Since large numbers of progenitor and mature cells are involved in blood production, the observed clone size fluctuations cannot arise from intrinsic demographic stochasticity of progenitor- and mature-cell birth and death. Moreover, we will show later in the Results section that random sampling alone cannot explain the observed clonal variances and mechanisms that involve other sources of variation are required.

Nomenclature and lumped mechanistic model

Fig 2 depicts our neutral model of hematopoiesis which is composed of five successive stages, or compartments, describing the initial single-cell tagged HSC clonal populations immediately after transplantation (Compartment **0**), the heterogeneous HSC clonal populations after a short period of intense self-renewal (Compartment **1**), the transit-amplifying progenitor cell compartment (Compartment **2**), the peripheral blood pool (Compartment **3**), and the sampled peripheral blood (Compartment **4**), respectively. Each distinct color or shape in Fig 2 represents a distinct clone of cells with the same tag.

In each compartment, relevant parameters include (using Compartment **1** as example): the total cell count $H(t)$, the untagged cell count $H^{-}(t)$, the tagged cell count $H^{+}(t)$, the total

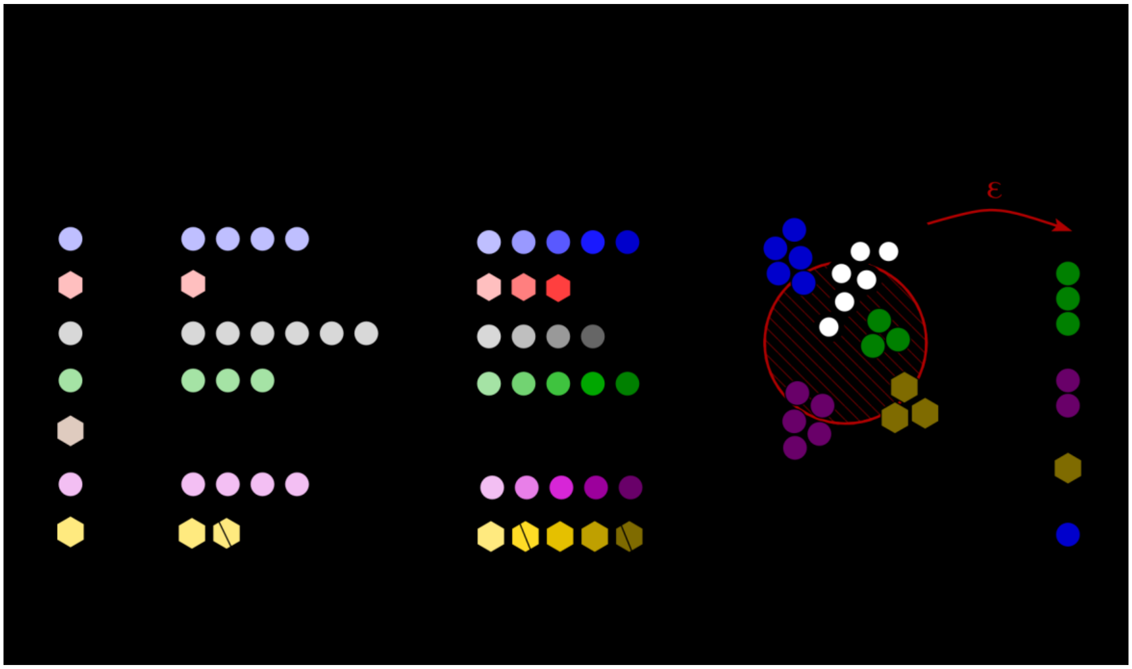


Fig 2. Schematic of a neutral multi-stage or multi-compartment hematopoiesis model. BM and PB refer to bone marrow and peripheral blood, respectively. Cells of the same clone have the same color. White circles represent untagged cells which were not counted in the analysis. Stages **0**, **1**, and **2** describe cell dynamics that occur mainly in the bone marrow. Stage **1** describes HSC clones ($C_h = 6$ in this example) after self-renewal that starts shortly after transplantation with rate r_h. After self-renewal, the relatively stable

HSC population ($H^+ = 20$ in this example) shifts its emphasis to differentiation (with per-cell differentiation rate α). Larger clones in Stage **1** (*e.g.*, the circular blue clone, $h_{\text{blue}} = 4$) will have a larger total differentiation rate αh_{blue} while smaller clones (*e.g.*, the red hexagonal clone, $h_{\text{red}} = 1$) will have smaller αh_{red}. The processes of progenitor-cell proliferation (with rate r_n) and maturation (with rate ω) in Compartments **2** and **3** are considered deterministic because of the large numbers of cells involved. The darker-colored symbols correspond to cells of later generations. For illustration, the maximum number of progenitor-cell generations allowed is taken to be $L = 4$. Compartment **4** represents a small sampled fraction ($\varepsilon(t_j)$ $2.8 \times 10^{-5} - 2 \times 10^{-4}$) of Compartment **3**, the entire peripheral blood of the animal. In the example pictured above, $C_s = 4$. Such small samples can lead to considerable sampling noise but is not the key driver of sample-to-sample variability.

https://doi.org/10.1371/journal.pcbi.1006489.g002

number of tagged clones $C_h(t)$, and the number $h_i(t)$ of HSCs carrying the i^{th} tag. These quantities are related through P$^{C}_{i¼^{h}1}$ $h_i$ðtÞ ¼ $H^þ$ðtÞ HðtÞ H ðtÞ.

In the progenitor pool, the total number of cells and the number with tag i are denoted $N(t)$ and $n_i(t)$, respectively. Further resolving these progenitor populations into those of the ℓ^{th} generation, we define $N^{(\ell)}(t)$ and $n^{ð}{}_{i}{}^{Þ}$ðtÞ. In the mature granulocyte pool, the total granulocyte population and that with tag i are labelled $M(t)$ and $m_i(t)$. In the sampled blood compartment, we use $S(t_j)$, $S^+(t_j)$, $s_i(t_j)$, and $C_s(t_j)$ to denote, at time t_j, the total number of sampled cells, the number of tagged sampled cells, the total number of tagged cells of clone i, and the total number of clones in the sample, respectively. In Compartment **4**, we further define $f_i(t_j) = s_i(t_j)/S^+(t_j)$ to denote the relative abundance of the i^{th} clone among all tagged clones.

By lumping together all clones (tagged and untagged) in each compartment, we can readily model the dynamics of total populations in each pool. After myeloablative treatment, the number of BM cells, including HSCs, is severely reduced. Repopulation of autogolously transplanted HSCs occurs quickly via self-renewal until their total number $H(t)$ reaches a steadystate. The repopulation of the *entire* HSC population and the subsequent entire progenitor and mature cell populations may be described via simple deterministic mass-action growth laws

_____ ¼ ðrhðHðtÞÞ mhÞHðtÞ; ð1Þ
dt

(
aHðtÞ ðrnð0Þ þ mðn0ÞÞNð0ÞðtÞ;
dN
_______ ð'Þ ðtÞ nð' 1Þ ð' 1ÞðtÞ ðr nð'Þ þ ð2Þ
mðn'ÞÞNð'ÞðtÞ; ' ¼ 0;
¼ 2r N
1 ' L
dt
2rnðL 1ÞNðL 1ÞðtÞ ðo þ mðnLÞÞNðLÞðtÞ; ' ¼ L; 1;

dMðtÞ ðLÞ ð3Þ
dHðtÞ

_____ ¼ oN ðtÞ mmMðtÞ: dt

HSC self-renewal is a regulated process involving signaling and feedback [22–24, 39, 40] and r_h may be a complicated function of many factors; however, we will subsume this complexity into a simple population-dependent logistic growth law $r_h(H(t))\ p_h(1 - H(t)/K_h)$ and assume a constant death rate μ_h. Alternatively, other studies have employed Hill-type growth functions [12, 28].

We assume the per cell HSC differentiation rate α is independent of the tag and that differentiation is predominantly an asymmetric process by which an HSC divides into one identical HSC and one progenitor cell that commits to differentiation into granulocytes. An initial generation-zero progenitor cell further proliferates with rate $r_n^{ð0Þ}$, contributing to the overall progenitor-cell population. Subsequent generation-ℓ progenitors, with population $N^{(\ell)}$, proliferate with rate $r_n^{ð'Þ}$ until a maximum number of generations L is reached. By keeping track of the generation index ℓ of any progenitor cell, we limit the proliferation potential associated with an HSC differentiation event by requiring that any progenitor cell of the final L^{th} generation to terminally differentiate into peripheral blood cells with rate ω or to die with rate $m^{ð_n^{Lþ}}$. For simplicity, we neglect any other source of regulation and assume α, $m^{ð}{}_{n}{}^{'þ}$ ¼ m_n, $r_n{}^{ð'þ}$ ¼ r_n and ω are all unregulated constants.

Our model analysis and data fitting will be performed using clone abundances sampled a few months after transplantation under the assumption that granulopoiesis in the animals has reached steady-state [4] after initial intensive HSC self-renewal. Steady-state solutions of Eqs (1), (2) and (3) are defined by H_{ss}, $N_{ss}{}^{ð'þ}$, and M_{ss}. The first constraint our model provides relates these steady-state total populations through

" $_L$#

M_{ss} —— —— ———— ——— —— o $N_{ssðLþ}$ ¼

mo_m ðoaþ$Hm_{ssðnLþ}$Þ r_n 2þr_nm_n $Am_{ssm}b$; ð4Þ ¼

m_m

where we have defined

L

o 2r

A_{ss} aH_{ss}; and b o——— ——— þ $m_{ðnLþ}$ r_n þ$_n m_n$

ð5Þ

as the total rate of HSC differentiation and the average number of granulocytes generated per HSC differentiation, respectively. These constraints also hold for the virally tagged, EGFP + subset (about 5% − 10%) of HSCs, *e.g.*, $M_{ss}{}^{þ}$ ¼ $A^{þ}{}_{ss}b$=m_m and $A_{ss}{}^{þ}$ ¼ $aH_{ss}{}^{þ}$. Since $M_{ss}{}^{þ}$ is inferred from the experiment, Eq (4) places a constraint between the total differentiation rate of labeled HSCs $A^{þ}{}_{ss}$ ¼ $aH_{ss}{}^{þ}$ and the typical per-differentiation amplification number β. This steady-state constraint will eventually be combined with statistics of the fluctuating clone abundances data to infer estimates for the underlying model parameters.

Clone-resolved mechanistic model

Although the lumped model above provides important constraints among the steady-state populations within each compartment, the clone-tracking experiment keeps track of the populations of sampled granulocytes that arise from "founder" HSCs that carry the same tag. Thus, we need to resolve the lumped model into the clonal subpopulations described by h_i, $n_i{}^{ð'þ}$, and m_i.

Even though the total HSC populations $H(t)$ and $H^{\pm}(t)$ are large, the total number of clones C_h 1 in compartment **1** is also large, and the number of cells with any tag (the size of any clone) can be small. The population of cells with any specific tag i is thus subject to large demographic fluctuations. Thus, we model the stochastic population of HSCs of any tag using a master equation for $P(h, t)$, the probability that at time t the number of HSCs of any clone is h: dPðh; tÞ

——— ¼ m_hðh þ 1ÞPðh þ 1; tÞ þ ðh 1Þr_hðHÞPðh 1; tÞ ½m_h þ r_hðHÞhPðh; tÞ: ð6Þ d*t*

Recall that immediately after transplantation, each HSC carries a distinct tag before selfrenewal ($h_i(0) = 1$) leading to the initial condition *Pðh;*0Þ ¼ **1**ðh*;*1Þ, where the indicator function **1**ðx*;*yÞ ¼ 1 if and only if $x = y$. Because $h = 0$ is an absorbing boundary, clones start to disappear at long times resulting in a decrease in the total number $C_h(t)$ of HSC clones. Before this "coarsening" process significantly depletes the entire population, each clone constitutes a small subpopulation among all EGFP+ cells, $h(t)H(t)$, and the stochastic dynamics of the population h of any clone can be approximated by the solution to Eq (6) with the logistic selfrenewal rate $r_h(H)$ $p_h(1 − H/K_h)$ replaced by $r_h(t) = p_h(1 − H(t)/K_h)$. Hence, evolution of each HSC clone follows a generalized birth-death process with time-dependent birth rate and constant death rate. We show in Appendix A in S1 Appendix that for H 1 the solution to Eq (6) can be written in the form [41]

Pðh; tÞ ¼ ð1   *Pð0; tÞÞð1 lðtÞÞlðtÞ*h* 1*;*   ð7Þ

where 0 $\lambda(t) < 1$ depends on $r_h(t)$ and μ_h. Here, $\lambda(t)$ determines "broadness" (level of clone size heterogeneity) of the clone size distribution. For the relevant initial condition of unique tags at $t = 0$, $\lambda(0) = 0$ and λ(*t* ! 1)!1. When $\lambda(t)$ is small, the distribution is weighted towards small h. For $\lambda(t) = 0$, *Pðh; tÞ* ¼ **1**ðh*;*1Þ which was the limit used in Goyal *et al*. [4] to assume no HSC self-renewal after transplantation. In the limit λ(*t*)!1, the distribution becomes flat and a clone is equally likely to be of any size 1 h H.

To further resolve the progenitor population into cells with distinct tags, we define $n^{(\ell)}(t)$ as the number of generation-ℓ progenitor cells carrying any one of the viral tags. The total number of progenitor cells with a specific tag is *nðtÞ* P*L*‘¼0 *n*ð‘Þ*ðtÞ*. Since the sizes h_i of individual clones may be small, differentiation of HSCs within each clone may be rare. However, since the size of each tagged progenitor clone quickly becomes large ($n(t)$ 1), we model the dynamics of $n^{(\ell)}(t)$ using deterministic mass-action growth laws:

(

ð‘ÞðtÞ ¼Poisson2*r n*ð‘ 1Þðð*ath*ÞðtÞÞð*r*n þð*rm*n nþÞ*nm*ð‘nÞÞð*nt*Þð0;ÞðtÞ*;*   ‘1¼0‘*; L*

1*;*   ð8Þ

d*n*

d*t*

n

2*r*n*n*ðL 1ÞðtÞ   ð*o* þ *m*ðnLÞÞ*n*ðLÞðtÞ*;*   ‘ ¼ *L:*

Our model is neutral (all clones have the same birth, death, and maturation rates), so these equations are identical to Eq (2). However, since creation of the zeroth-generation subpopulation $n^{(0)}(t)$ derives only from differentiation of HSCs of the corresponding clone, which has a relatively small population $h(t)$, we invoke a Poisson process with rate $\alpha h(t)$ to describe stochastic "injection" events associated with asymmetric differentiation of HSCs of said clone.

Each discrete differentiation event leads to a temporal burst in $n^{(\ell)}(t)$.

Finally, the dynamics of the population $m(t)$ of any granulocyte clone in the peripheral blood are described by an equation analogous to Eq (3):

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \omega n^{(L)}(t) - \mu_m m(t), \tag{9}$$

where we have assumed that only the generation-L progenitor cells undergo terminal differentiation with rate ω. An alternative model allows progenitor cells of earlier generations ($\ell < L$) to also differentiate and circulate but does not give rise to qualitatively different results (See Appendix B in S1 Appendix).

To study the dynamics of the burst in $n_b^{(0)}(t)$ immediately following a *single, isolated* asymmetric HSC differentiation event at $t = 0$, we set the initial condition $n_b^{(0)}(0) = 1$; $n_b^{(\ell)}(0) = 0$ $(1 \leq \ell \leq L)$, remove the Poisson ($\alpha h(t)$) term in Eq (8) and find,

$$n_b^{(\ell)}(t) = \begin{cases} \dfrac{(2r_n t)^{\ell}}{\ell!} e^{-(r_n+\mu_n)t}, & 0 \leq \ell \leq L-1; \\ 2r_n \displaystyle\int_0^t n_b^{(L-1)}(t')e^{-\omega(t-t')}dt', & \ell = L. \end{cases} \tag{10}$$

Bounded analytic solutions to $n_b^{(L)}(t)$ involving the lower incomplete gamma function can be found. Upon using the solution $n_b^{(L)}(t)$ in Eq (9) the mature blood population within a clone associated with a single HSC clone differentiation even is described by

$$m_b(t) = \omega \int_0^t n_b^{(L)}(t')e^{-\mu_m(t-t')}dt'. \tag{11}$$

The populations associated with a single HSC differentiation event, $n_b^{(\ell)}(t)$ and $m_b(t)$, are plotted below in Fig 3. of the Results section. Then, the total number $m_i(t)$ of mature granulocytes with the i^{th} tag at time t is obtained by summing up all $m_b(t - \tau_k)$ bursts initiated by HSC differentiations at separate times $\tau_k \leq t$ with the i^{th} tag.

Besides the burst dynamics described above, the data shown in Fig 1(a) are subject to the effects of small sampling size, uncertainty, and bias induced by experimental processing such

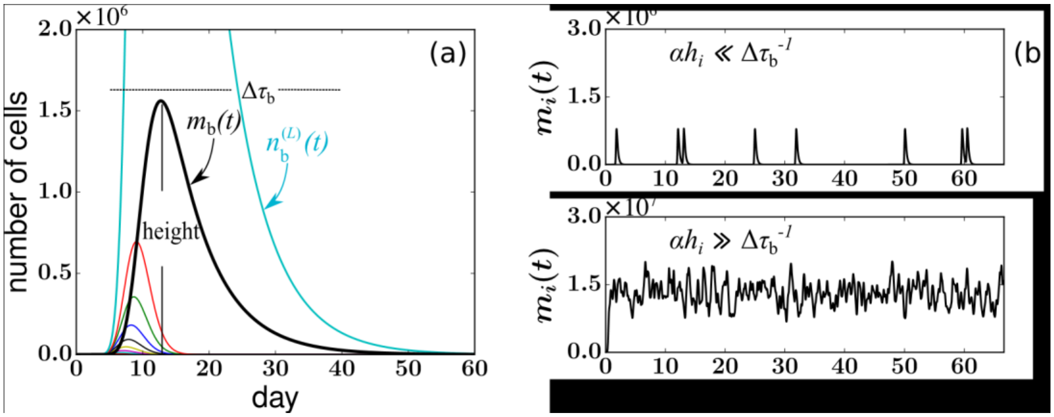


Fig 3. (a) A burst of cells is triggered by a single HSC differentiation event at time $t = 0$. A plot of representative solutions to Eqs (10) and (11) for $r_n = 2.5$, $L = 24$, m_n ¼ $m^{ð}{}_n{}^{Lþ}$ ¼ 0, $\mu_m = 1$, $A^{þ}_{ss}$ ¼ 14:7, and $\omega = 0.16$. Curves of different colors represent $n_b{}^{ð\ell þ}$ðtÞ, the progenitor cell population within each generation $\ell = 0, 1, 2, \ldots, L$, and $m_b(t)$, the number of mature granulocytes associated with the differentiation burst. All populations rise and fall. (b) Realizations of peripheral blood (PB) populations in a single clone arising from multiple successive differentiation events. The fluctuating populations are generated by adding together $m_b(t)$ associated with each differentiation event. Time series resulting from small ($h_i/H^* = 0.0003$) and large ($h_i/H^* = 0.03$) HSC clones are shown. Small clones are characterized by separated bursts of cells, after which the clone vanishes for a relatively long period of time. The number of mature peripheral blood cells of large clones reaches a relatively constant level and almost never vanishes.

https://doi.org/10.1371/journal.pcbi.1006489.g003

as PCR amplification, and data filtering. In this experimental system, PCR generates a smaller uncertainty than blood sampling so we focus on the statistics of random sampling. Each blood sample drawn from monkey RQ5427 contains about 10μg of genomic DNA [13]. After PCR

amplification, deep sequencing, and data filtering, the total number ^$S^{þ}$ðt_jÞ of quantifiable

tags corresponds to *5 × $10^3 - 3 \times 10^4$ tagged cells. The sample ratio is defined by εðt_jÞ

^$S^{þ}$ðt_jÞ=M^ $^{þ}{}_{ss}$ ¼ 3 10 ^{5} 2 10 ^{4} where M^ $_{ss}{}^{þ}$ 1:6 10^8 is the estimated total number of

tagged granuloctyes in the peripheral blood. The number of sampled cells with the i^{th} tag from

the j^{th} sample then approximately follows a Binomial distribution B $S^{þ}$ðt_jÞ; $^{m}{}_{\overline{M}}{}^{\overline{ð}}{}_{\overline{ssþ}}{}^{t_jþ}$

Bðm_iðt_jÞ;εðt_jÞÞ in our model. To quantitatively explore the feature of apparent extinctions of clones from a sample, we calculate the probability that no peripheral blood cell from clone i is found in a sample of size

! !

þ þ

ðt Þ M

$S_{þ}$ðt_jÞ≪ $M_{ssþ}$: Pðf_iðt_jÞ ¼ 0jm_iðt_jÞÞ ¼ M_{ss} $_{þ}$ð$m t_j$Þ$_i$ $_j$ = $S_{þ}$ð$_{ss}t_j$Þ exp m______$_i$ð$t M_j$Þ$S_{ssþþ}$ðt_jÞ .

Thus, if

S

m_iðt_jÞ < ε^{1}¼ M^ $^{þ}_{ss}$=^$S^{þ}$ðt_jÞ 2 10^4 the i^{th} clone is likely to be missed in the sample. The value ε^{-1} is also used to threshold the population $m_b(t)$ to define the measurable duration $\Delta\tau_b$ of a burst (as indicated in Fig 3(a)).

Parameter values

Parameters determined by the experimental procedure or estimated directly from the experiments include the weight of the animal, the sampling times t_j, the EGFP+ ratio, and the total number of tagged cells detected in each sample ^$S^{þ}$ðt_jÞ. Since the tagged granulocyte population M^ ^{þ}ðt_jÞ does not fluctuate much across samples, we use its average for M^ $_{ss}^{þ}$, and the relevant experimental parameters for each animal become y_{exp} ¼ fM^ $^{þ}_{ss}$;^$S^{þ}_i$ ðt_jÞ; t_jg. These will also be used as inputs to our models.

Table 1. Summary of parameters, including their biological interpretation, ranges of values, and references. All rate parameters are quoted in units of per day. Other parameters are chosen to be within their corresponding reported ranges from the referenced literature. How variations in parameter values affect our analysis will be described in the subsequent sections.

Parameter	Interpretation	Values & References
HSC pool (Compartment **1**)		
H_{ss}	total number of HSCs at steady state	$1.1 \times 10^4 - 1.1 \times 10^6$ [4, 11, 12]
α	per-cell HSC differentiation rate	$5.6 \times 10^{-4} - 0.02$ [4, 11, 12]
μ_h	HSC death rate	$10^{-3} - 0.1$ [12, 34]
Transit-Amplifying Progenitor pool (Compartment **2**)		
r_n	growth rate of progenitor cell	2 − 3 [12]
μ_n	death rate of progenitor cell (generation $\ell < L$)	0 [12, 34]
$m_{ðnLÞ}$	death rate of progenitor cell (generation $\ell = L$)	0 − 0.27 [12, 34]
ω	maturation rate of generation-L cells	0.15 − 0.17 [43, 44]
L	maximum generation of progenitor cells	15 − 21 [12, 34]
Peripheral Blood pool (Compartment **3**)		
M_{ss}	total number of peripheral blood granulocytes at steady state	$(2.5 - 5) \times 10^9$ [13, 42]
μ_m	death rate of peripheral blood granulocytes	0.2 − 2 [34, 44, 45]

https://doi.org/10.1371/journal.pcbi.1006489.t001

Our multi-stage model also contains many other intrinsic parameters, including y_{model} ¼ fl; C_h;a; r_n;m_n;$m^{ð}_{n}$^{LÞ}; L;o;m_mg. We first found parameter values that have been reliably independently measured. Some parameters were measured in human clinical studies rather than in rhesus macaques but can nonetheless serve as reasonable approximations for nonhuman primates due to multiple physiological similarities [42]. These estimates can certainly be improved once direct measurements on rhesus macaques become available. Model parameters, their estimates, and the associated references are given in Table 1 below.

Model properties and implementation

Using parameter estimates, we summarize the dynamical properties of our model and describe how the key model ingredients including stability of HSC clone distributions and subsequent "bursty" clone dynamics that follow differentiation can qualitatively generate the observed clone-size variances.

Slow homeostatic birth-death of HSCs—The first important feature to note is the slow homeostatic birth-death of HSCs. After the bone marrow is quickly repopulated, $r_{\mathrm{h}}(H(t)) - \mu_{\mathrm{h}}$ 0, and stochastic self-renewal slows down. Because $h = 0$ is an absorbing state, the size distribution of the clones may still slowly evolve and coarsen due to stochastic dynamics, leading to the slow successive extinction of smaller clones. The typical timescale for overall changes in h can be estimated by approximating $r_{\mathrm{h}}(H_{\mathrm{ss}})$ μ_{h} [46] and considering the mean

time $T(h)$ of extinction of a clone initially at size $h \ll H_{\mathrm{ss}}$. The standard result given in Gardiner [47] and also derived in Appendix C in S1 Appendix is TðhÞ $_{m}\frac{h}{}_{\mathrm{h}}$ 1 þ ln [redacted] $_{h}$ months (for $\mu_{\mathrm{h}} = 10^{-2}$, $H_{\mathrm{ss}} = 10^{4}$, $h = 10^{1}$; see Table 1 for applicable values). Since this timescale is larger than the time of the experiment (67 months for monkey RQ5427), mean HSC clone sizes do not change dramatically during the experiment, consistent with the stable number of clones observed in the samples shown in Fig 1(b). Thus, as a first approximation, we will use a static configuration $\{h_i\}$ drawn from $P(h)$ to describe how, through differentiation, HSC clones feed the progenitor pool.

Fast clonal aging of progenitors—In contrast to slow HSC coarsening, progenitor cells proliferate "transiently." In Fig 3(a) we plot a single population burst of progenitor and mature granulocytes, given by Eqs (10) and (11) and using the parameter values listed in Table 1. The characteristic duration, or "width" $\Delta\tau_{\mathrm{b}}$ associated with each temporal burst of cells is defined as the length of time during which the number $m_{\mathrm{b}}(t)$ is above the detection threshold within a sample of peripheral blood: ε ¹ ¼ M^ ᵖ$_{\mathrm{ss}}$=^Sᵖ 2 10⁴.

According to Eq (11), the burst width and height depend nonlinearly on the parameters L, r_{n}, μ_{n}, μ_{m}, and ω in their physiological ranges (see Table 1). The characteristic width of a burst scales as $\Delta\tau_{\mathrm{b}}$ * $L/r_{\mathrm{n}} + 1/\omega + 1/\mu_{\mathrm{m}}$. This estimate is derived by considering the L rounds of progenitor cell division, each of which takes time * $1/r_{\mathrm{n}}$. Terminal-generation progenitors then require time *$1/\omega$ to mature, after which mature granulocytes live for time * $1/\mu_{\mathrm{m}}$. In total, the expected life span of * $L/r_{\mathrm{n}} + 1/\omega + 1/\mu_{\mathrm{m}}$ approximates the timescale of a HSC-differentiation-induced burst of cells fated to be granulocytes. Using realistic parameter values, the typical detectable burst duration $\Delta\tau_{\mathrm{b}}$ * 1 − 2 months is much shorter than the typical sampling gaps $\Delta t_j = 5 - 11$ months.

With this "burst" picture in mind, we now show how fluctuations of sampled clone sizes can be explained. Small-h (where the clone-wise HSC differentiation rate $ah_i \ll$ $_{D}\frac{1}{}_{t\mathrm{b}}$) clones rarely appear in blood samples. Their appearance also depends on whether sampling is frequent and sensitive enough to catch the burst of cells after rare HSC differentiation events. On the other hand, large-h (ah_i $_{D}\frac{1}{}_{t\mathrm{b}}$) clones differentiate frequently and consistently appear in the peripheral blood. Their populations in blood samples are less sensitive to the frequency of taking samples. Fig 3(b) shows two multi-burst realizations of peripheral-blood populations $m_i(t)$ of clone i corresponding to a small clone and a large clone. The 2000-day trajectories were simulated by fixing h_i and stochastically initiating the progenitor proliferation process. Population bursts described by Eq (11) were added after each differentiation event distributed according to Poisson(αh_i). Using simulations, we confirm that the statistics of clone extinctions and resurrections are more sensitive to the overall clonal differentiation rate αh_i than to the precise shape of a mature cell burst, allowing a reduction in the number of effective parameters (Appendix D in S1 Appendix).

We can further pare down the number of remaining parameters by finding common dependences in the model and defining an effective maximum generation number. We can rewrite Eq (5) as b 2^{L_e}, where

$$L_e ¼ L \quad L\log_2 \frac{r^n þ m^n}{r_n} \quad \log_2 \frac{o þ m^{ðnLÞ}}{o} \qquad ð12Þ$$

is an *effective* (and noninteger) maximum generation parameter. Later in Appendix D in S1 Appendix, we show that uncertainties of the model structure, alternative mechanisms, and parameter values can be subsumed into L_e. Henceforth, in our quantitative data analysis, we will set the unmeasurable parameters $m_n ¼ m^{ð}{}_n{}^{LÞ} ¼ 0$ and subsume their uncertainties into an effective maximum generation L_e. Finally, we will invoke Eq (4) to find the constraint

$$A_{þss} b ¼ A_{þss} 2_{Le} ¼ M_{ssþ} m_m{:} \qquad ð13Þ$$

Since we can estimate $M_{ss}{}^{þ}$ of the animals in the experiment and the death rate of mature granulocytes μ_m has been reliably measured in the literature, Eq (13) provides a relationship between the total steady-state differentiation rate $A^{þ}{}_{ss}$ and the maximum number of progenitor generations L_e.

After assigning values to parameters using Table 1 (setting $\mu_n = 0$, $\omega = 0.16$ and $\mu_m = 1$), subsuming parameters into L_e (setting $m^{ð}{}_n{}^{LÞ} ¼ 0$), describing the configuration $\{h_i\}$ through the distribution shape factor λ and the total number of HSC clones C_h (setting the HSC death rate $\mu_h = 0$), and applying the constraint $A^{þ}{}_{ss} 2^{L_e} ¼ M\hat{}\ {}^{þ}{}_{ss} m_m$, we are left with four effective model parameters $\vartheta_{\text{model}} = \{\lambda, C_h, r_n, L_e\}$. Here we have included r_n in the key model parameters since it is not reliably measured and the cell burst width is sensitive to r_n. Once L_e is inferred, Eq (13) can be used to find $A^{þ}{}_{ss} ¼ 2\ {}^{L_e} M\hat{}\ {}^{þ}{}_{ss} m_m$.

Statistical model

The total number of tags observed across all samples (obtained by summing up the observed numbers of *unique* tags over J samples) can be used as a lower bound on C_h. Even though estimates for animal RQ5427 give $C_h * 550 - 1100$, uncertainties in the HSC self-renewal rate parameters p_h, K_h, and the initial HSC population $H(0)$ make λ and $P(h, t)$ difficult to quantify. Even if $P(h, t)$ were known, it is unlikely that the drawn $\{h_i\}$ would accurately represent those in the monkey, especially when λ 1 and $P(h)$ becomes extremely broad (the variance of $P(h)$ approaches infinity). Thus, we are motivated to find a statistical measure of the data that is insensitive to the exact configuration of $\{h_i\}$. The goal is to study the statistical correlations between various features of *only* the outputs, which should be insensitive to the input configuration $\{h_i\}$ but still encode information about the differentiation dynamics.

Two such features commonly used to fit simulated $f_i(t_j)$ to measured $\hat{f}_i(t_j)$ are the mean $y_i = \frac{1}{J}\sum_{j=1}^{J} f_i(t_j)$ and the variance $s_i^2 = \frac{1}{J}\sum_{j=1}^{J}(f_i(t_j) - y_i)^2$. However, the small number of measurement time points J and the frequent disappearance of clones motivated us to propose an even more convenient statistic that is based on

$$z_i = \sum_j \mathbf{1}(f_i(t_j), 0), \qquad (14)$$

the number of absences across all samples of a clone rather than on σ_i. Here, the indicator function $\mathbf{1}(x; x') = 1$ when $x = x'$ and $\mathbf{1}(x; x') = 0$ otherwise. In Appendix E in S1 Appendix, we illustrate alternatives such as data fitting based on σ_i and on an autocorrelation function but also describe the statistical insights gained from using statistics of z_i.

The level of correlation between the observed number $\hat{z}_i$ of absences of clone i and its average abundance $\hat{y}_i$ is measured by the average of $\hat{y}_i$ conditioned on $\hat{z}_i$ (dashed curve). In Fig 4, the distribution of the values of $\hat{y}_i$ at each $\hat{z}_i$ is clearly shown. To combine the correlated stochastic quantities z_i and y_i into a useful objective function, we take the expectation of y_i over only those clones that have a specific number $z_i = z$ absences across the time samples:

$$Y_z = \frac{\sum_i y_i \mathbf{1}(z_i; z)}{\sum_i \mathbf{1}(z_i; z)}. \qquad (15)$$

The normalizing denominator $\sum_i \mathbf{1}(z_i; z)$ is simply the number of clones with exactly z absences. In case no simulated or data-derived trajectories $f_i(t_j)$ exhibit exactly z absences, we set $Y_z = 0$ or $\hat{Y}_z = 0$. We then determine $Y_z(\vartheta_{\mathrm{model}})$ from simulating our model and $\hat{Y}_z$ from experiment and use the mean squared error (MSE) between the two as the objective function:

$$\mathrm{MSE}(\vartheta_{\mathrm{model}}) = \sum_{z=1}^{J-1} [Y_z(\vartheta_{\mathrm{model}}) - \hat{Y}_z]^2; \qquad (16)$$

where $\vartheta_{\mathrm{model}} = \{\lambda, C_h, r_n, L_e\}$. Y_0 is excluded from the MSE calculation because the y_i values of

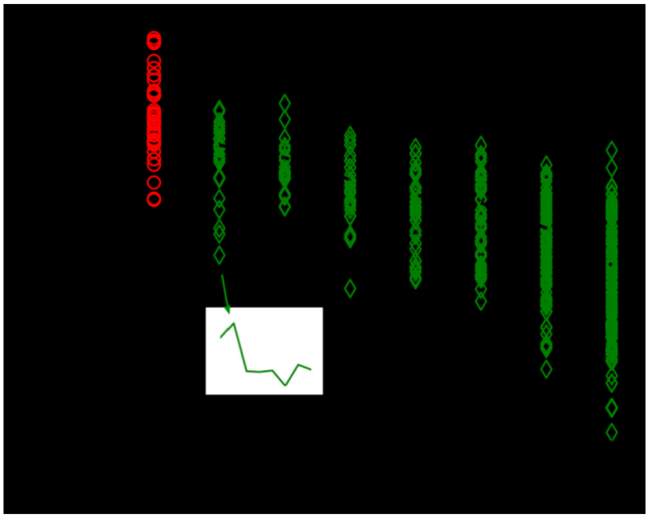


Fig 4. Scatterplot of clone trajectories of animal RQ5427 displayed in terms of ln ^*y* *i*, the log mean abundance of clone *i*, and ^*z* *i*, the number of samples in which clone *i* is undetected. The trajectory of each clone *i* is represented by a symbol located at a coordinate determined by its value of ln ^y_i and ^z_i. A trajectory of a clone that exhibits one absence within months 8 − 67 is shown in the inset. The first sample at month 2 is excluded because only long-term repopulating clones are considered. Clones that are absent in all eight samples are also excluded, so the largest number of absences considered for animal RQ5427 is 7. The dashed black line denotes ln $Y^{\wedge}_z$, where $Y^{\wedge}_z$ is the average of ^y_i calculated over *i* within each bin of *z* as shown in Eq (15). When later analyzing $Y^{\wedge}_z$, $Y^{\wedge}_0$ (red circles) is not included.

https://doi.org/10.1371/journal.pcbi.1006489.g004

clones that have $z_i = 0$ are not constrained by the burstiness of the model and Y_0 can be sensitive to the underlying configuration $\{h_i\}$ (see the Discussion and Appendix E in S1 Appendix).

We are now in a position to compare results of our model with experimental data. The general approach will be to choose a set of parameters, simulate the forward model (including sampling) to generate clone abundances $\{f_i(t_j)\}$, number of absences z_i, and ultimately $Y_z(\vartheta_{\mathrm{model}})$, which is then compared to data-derived $Y^{\wedge}_z$. By minimizing Eq (16) with respect to $\vartheta_{\mathrm{model}}$, we obtain the least square estimates (LSE) of $\vartheta_{\mathrm{model}}$. A schematic of our workflow is shown in Fig 5. We describe the details of the simulation of our model in Appendix F in S1 Appendix.

Results

By implementing the protocol outlined in Fig 5, we find a number of results including leastsquares-estimates (LSE) of the parameters, their sensitivity to other model features, validation of the mechanistic model, and robustness of our statistical methods to missing data and clone sampling thresholds. Our analyses allow us to effectively compare the results from the three different animals.

MSE function and estimates of L_e and A^{b}_{ss} for animal RQ5427

We first fix the HSC distribution shape parameter λ = 0.99 and the total number of HSC clones C_h = 500; this choice will be justified in the next subsection. The MSE objective function can now be plotted as a function of the proliferation rate r_n 2 [0.01, 10] and proliferation potential L_e 2 [19, 28] of progenitor cells in their respective biologically relevant ranges. Even after specifying $\vartheta_{model} = \{λ = 0.99, C_h = 500, r_n, L_e\}$, there is still uncertainty in the simulated values of $Y_z = \{Y_1, Y_2, ..., Y_7\}$ due to the uncertainty in the drawn configuration of HSC clone sizes $\{h_i\}$, the intrinsic stochastic mechanisms of the model (Poissonian HSC differentiation events), and random peripheral blood sampling. Therefore, we performed 200 simulations for each set of $\{r_n, L_e\}$, producing 200 sets of Y_z. The means of Y_z are used to construct the mean of MSE(λ = 0.99, C_h = 500, r_n, L_e), plotted in Fig 6.

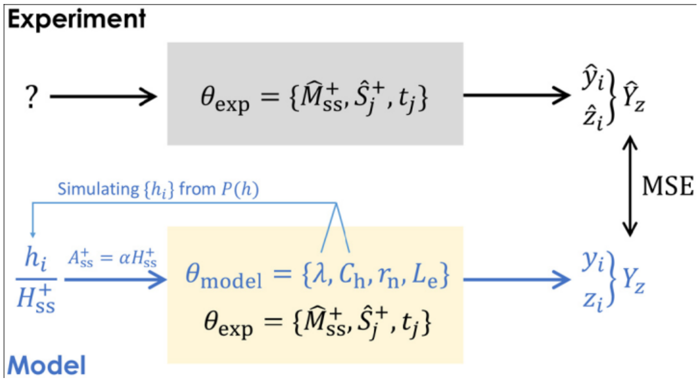

Experiment
?
$\theta_{\text{exp}} = \{\hat{M}^+_{ss}, \hat{S}^+_j, t_j\}$
$\hat{y}_i$
$\hat{z}_i$
$\hat{Y}_z$
MSE
Simulating $\{h_i\}$ from $P(h)$
$\frac{h_i}{H^+_{ss}}$
$A^+_{ss} = \alpha H^+_{ss}$
$\theta_{\text{model}} = \{\lambda, C_h, r_n, L_e\}$
$\theta_{\text{exp}} = \{\hat{M}^+_{ss}, \hat{S}^+_j, t_j\}$
y_i
z_i
Y_z
Model

Fig 5. Workflow for comparing parameter-dependent simulated data with measured clone abundances. The first step is drawing a configuration $\{h_i\}$, which is experimentally unmeasurable, from the HSC clone distribution $P(h)$. To define $P(h)$ requires an initial estimate of λ and C_h. Using known experimental parameters ϑ_{exp} and choosing r_n, L_e 2 ϑ_{model}, we compute the theoretical quantities y_i and z_i by simulating the multi-compartment mechanistic model and the peripheral-blood sampling. The corresponding ^y_i and ^z_i are extracted from data, and the theoretical $Y_z(\vartheta_{model})$ and the experimental Y^ $_z$ are compared through the MSE defined in Eq (16). The MSE is then minimized to find least squares estimates for ϑ_{model}.

https://doi.org/10.1371/journal.pcbi.1006489.g005

In the reported progenitor growth rate range of $r_n = 2 - 3$ (Table 1), the MSE function is quite insensitive to L_e. To interpret this observation, note that r_n does not affect the absolute value of β according to Eq (13), but it affects the typical time * $L/r_n + 1/\omega$ it takes for a generation 0 progenitor cell to form a mature granulocyte. When $r_n < \mu_m$, the proliferation of progenitors cannot "catch up" with the loss of granulocytes, resulting in a quickly vanishing burst in the granulocyte population $m_b(t)$ arising from a single-differentiation event $m_b(t)$. A larger L_e would be required to compensate. When r_n μ_m, the growth of any clone is much quicker

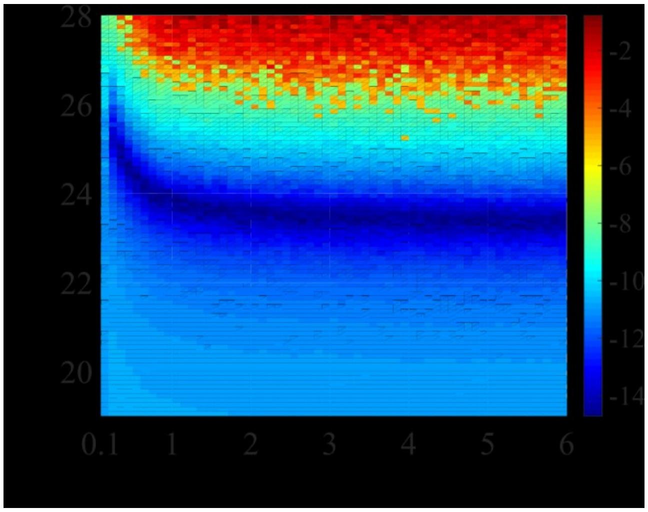


Fig 6. Dependence of the mean MSE defined in Eq (16) on r_n and L_e. For visualization purposes, we took the natural logarithms of MSE values and plotted them as a function of L_e and r_n. Blue areas denotes smaller MSE values, thus better fitting. This energy surface was generated by averaging over 200 simulations using $C_h = 500$ and λ = 0.99.

https://doi.org/10.1371/journal.pcbi.1006489.g006

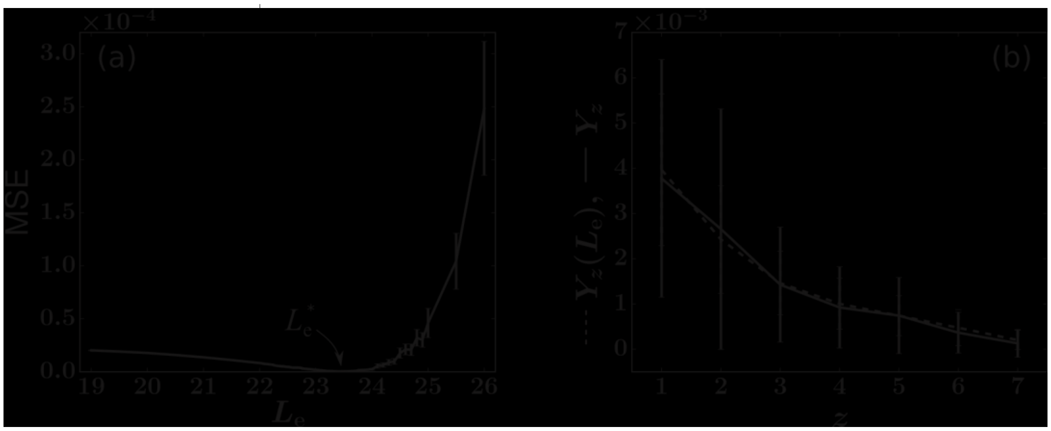


Fig 7. Finding the least squares estimate (LSE) L_e for animal RQ5427 by fitting the simulated Y_z to the experimental $\hat{Y}_z$. The values of (λ, C_h, r_n) are chosen to be (0.99, 500, 2.5). Simulations with $\{h_i\}$ set to f^y_igH_{ss}Þ instead of drawing from $P(h)$ generate similar results. (a) The LSE is L_e ¼ 23:4. Averages and standard deviations (error bars) of the 200 MSEs are plotted. (b) Comparisons between the experimental (solid) $\hat{Y}_z$ and simulated (dashed) Y_z with fixed L_e ¼ 23:4. The error bars are determined by considering the standard deviation of the average abundances (y_i or ^y_i) of all clones exhibiting z absences.

https://doi.org/10.1371/journal.pcbi.1006489.g007

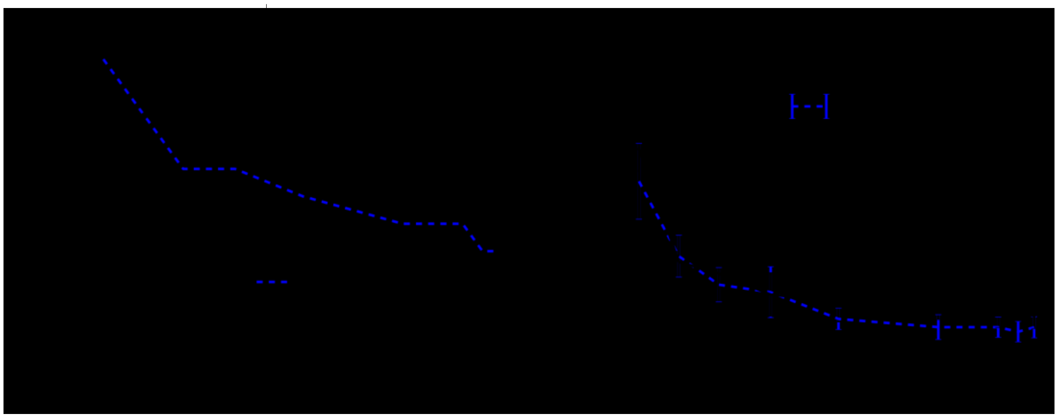
than its loss, so the burst size is relatively stable and L_e is not very sensitive to r_n. Thus, the MSE objective function is fairly insensitive to r_n in its biologically meaningful value range.

We then fix the progenitor proliferation rate r_n = 2.5 and plot the mean MSE(λ = 0.99, C_h = 500, r_n = 2.5, L_e) in Fig 7(a), which indicates a clear minimum at L_e ¼ 23:4 0:12. The error bars denote the standard deviation of MSEs obtained from the 200 simulations at different values of L_e and show that the variability is negligible for the purpose of determining the minimum. Upon applying the steady-state granulocyte balance constraint in Eq (13), we obtain a total HSC differentiation rate ð$A^þ_{ss}$Þ ¼ 14:7.

If we approximate m_n;$m^ð_n$^{LÞ} ¼ 0, L_e L. Substituting LSE values L_e ¼ 23:4 for L into the model for the peripheral blood bursts (the analytic solutions to $n^{(L)}(t)$ and $m_b(t)$ in Eqs (10) and (11)) yields a single burst duration of $\Delta\tau_b$ 32 days, consistent with our assumption $\Delta\tau_b \ll \Delta t_j$ = 5 − 11 months. Note that even though L is interpreted as an integer in Eq (8), analytic solutions of Eqs (10) and (11), $n^ð_b$^{Þ}ðtÞ and $m_b(t)$, depend on L in a continuous manner, interpolating the behavior to arbitrary values of L. Fig 7(b) shows how *one* simulation of Y_zðL_e ¼ 23:4Þ fits the experimentally measured $\hat{Y}_z$. Here, each error bar denotes the standard deviation across all mean abundances y_i (or ^y_i) within each value of z absences.

Insensitivity of analysis to HSC configurations

In Fig 8, we demonstrate the weak dependence of our least-squares estimate to λ, the parameter controlling the shape of the probability distribution of HSC clone sizes $P(h, t)$. For



each λ, we sample a fixed number (C_h = 500) of HSC clones from the theoretical distribution $P(h, t)$, fix r_n = 2.5, and let L_e vary between 19 and 28. The averages of the 200 simulated MSEs at each value of L_e are compared and the L_e that corresponds to the minimal average MSE is selected. The selected L_e as a function of λ is plotted in Fig 8(a). Fig 8(b) shows the averages and standard deviations of MSE ðL_eÞ at each value of λ. We then repeat the simulations with C_h = 1000. These results together show that L_e is insensitive to the distribution of h_i. This insensitivity might be understood by noticing that the quantity Y_z is defined as the *mean* of the values of y_i that are associated with z absences (dashed curve in Fig 4) and is not necessarily sensitive to

Fig 8. The LSE L_e is insensitive to the geometric distribution factor λ > 0 and to C_h 1. This implies that for a wide range of values of λ and C_h the LSEs are insensitive to the HSC configuration {h_i}. (a) L_e s found at each value of λ. (b) Averages and standard deviations (error bars) of MSE ðL_eÞ as a function of λ. The LSE and MSE(L_e) values associated with self-consistently using fh_ig=Hþ ¼ f^y $_i$g from experimental data are marked by arrows and "exp."

https://doi.org/10.1371/journal.pcbi.1006489.g008

how these values are distributed (vertically distributed markers at each value of z in Fig 4). Instead, Y_z incorporates the intrinsic relationship between a clone's mean abundance y_i and its number of absences z_i, averaged over all clones. It thus also encodes how heterogeneity in the HSC clone populations is translated into the burstiness seen in the sampled clone abundances $f_i(t_j)$. Although it is generally impossible to recover the exact {h_i} configuration, we find the HSC self-renewal-induced geometric distribution described by Eq (7) generally generates better fits to the sampled data when λ is large ($\gtrsim$ 0.5), suggesting significant heterogeneity in values of h_i.

Comparison of variability from simple sampling and best-fit model

We can check how our LSE result performs against the null hypothesis that clone size variations arise only from random sampling. An estimate of sampling-induced variability can be obtained by assuming a specific number of peripheral blood granulocytes of tag i and randomly drawing an experimentally determined fraction $\varepsilon(t_j)$ of peripheral blood cells. This is repeated J times from a constant peripheral pool $\{m_i\}$. Each draw results in $s_i(t_j)$ cells of clone i in the simulated sample. Normalizing by $S^+(t_j)$, the total number of tagged cells in the sample, we obtain simulated $f_i(t_j)$ from which we extract the mean abundance y_i and its standard devia-

qffi$\frac{1}{}$P$_J$ ðf_iðt_jÞ

y_iÞ$_2$ for each clone i. The simulated quantities ln y_i and σ_i associated

tion s_i ¼ $_J$ $_{j¼1}$

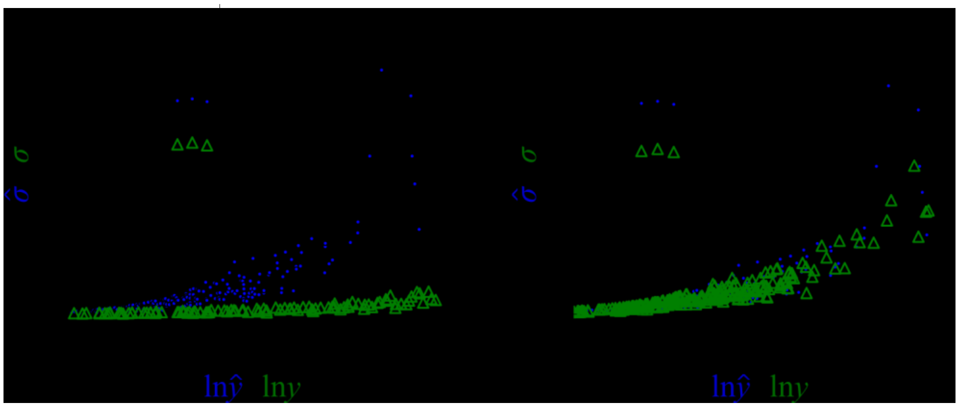
with each clone i are indicated by the green triangles in Fig 9(a). The corresponding values ln$\hat{y}_i$ and $\hat{s}_i$ derived from the data shown in Fig 1(b) are indicated by the blue dots. This simple heuristic test shows that the experimental fluctuations in clone abundances are significantly larger than those generated from random sampling alone and that additional mechanisms are responsible for the fluctuation of clone abundances in peripheral blood. Using LSE parameter values, Fig 9(b) shows the fluctuations in clone abundances obtained from random sampling of fluctuating mature clones simulated from our model. Here, the variability is a convolution of the fluctuations arising from intrinsic burstiness and from random sampling. The total variability fits those of the experimental data well except for several large-sized outlier clones.

Fig 9. (a) A plot of the standard deviation $\hat{s}_i$ *vs*. the log of the mean $\hat{y}_i$, extracted from abundance data (blue dots). For comparison, clonal tags distributed within the peripheral blood cells were randomly sampled (with the same sampling fraction $\varepsilon(t_j)$ at times t_j as in the experiment). The analogous quantity σ_i shown by the green triangles indicates a much lower standard deviation for a given value of ln y_i. This simple test implies that the clonal variability across time cannot be explained by random sampling. (b) The same test is performed after applying our model with the LSE parameter $L_e = 23.4$ (and the average of parameters listed in Table 1).

https://doi.org/10.1371/journal.pcbi.1006489.g009

Robustness of L_e to sampling frequency and threshold

We checked the robustness of our inference by leaving out time points from the experiment. Recall that the experimental data matrix for animal RQ5427 contains 536 rows, each representing a clone, and 8 columns, each representing a time point measured by month. By using only the first j = 8, 7, ...1 time points of data (leaving out 8 − j time points), seven additional simulation studies to find L_e were performed. As shown in Fig. G1 in Appendix G of S1 Appendix, reduction in the number of time samples flattens the MSE but preserves its minimum near L_e 23:4 23:6 provided at least 2-3 samples are used. We have also excluded intermediate samples to mimic larger sampling gaps Δt_j and found similar results.



Next, we examined the effects of sample thresholding on our parameter inference. By eliminating clones whose average abundances are under a certain threshold, we will observe fewer clones in the large-*z* bins depicted in Fig 4. Since larger clones with fewer absences contribute most to the MSE, our results will not be affected as long as the threshold is not too large. Provided we apply the same threshold to both the simulated and experimental data, there should not be systematic bias in our results. The MSEs generated using different thresholds are plotted in Fig. G2 in Appendix G of S1 Appendix and show that the inferred value L_{e} 23*:*4 remains essentially unchanged provided the threshold level is low enough to retain approximately at least 40% (about 200) of the clones (see Fig. G2(a-f) in Appendix G). With fewer clones retained (< 200), the LSE of L_e shifts only modestly to L_{e} 24*:*3. Thus, we conclude that our inference of L_{e} is robust to increases in sampling threshold as along as a reasonable number of clones ($\gtrsim$ 200) are counted.

Data analysis and fitting for animals 2RC003 and RQ3570

The data from the three different monkeys vary in their numbers of tagged clones transplanted and the lengths of the experiments. For animal RQ5427/2RC003/RQ3570, there are 536/1371/ 442 clones that are detected at least once within 67/103/38 months. The fraction of cells in all tracked clones in animal RQ5427/2RC003/RQ3570 was approximated by the average fraction of cells that were EGFP+ marked over time, around 0.052/0.049/0.086 (the ratios between

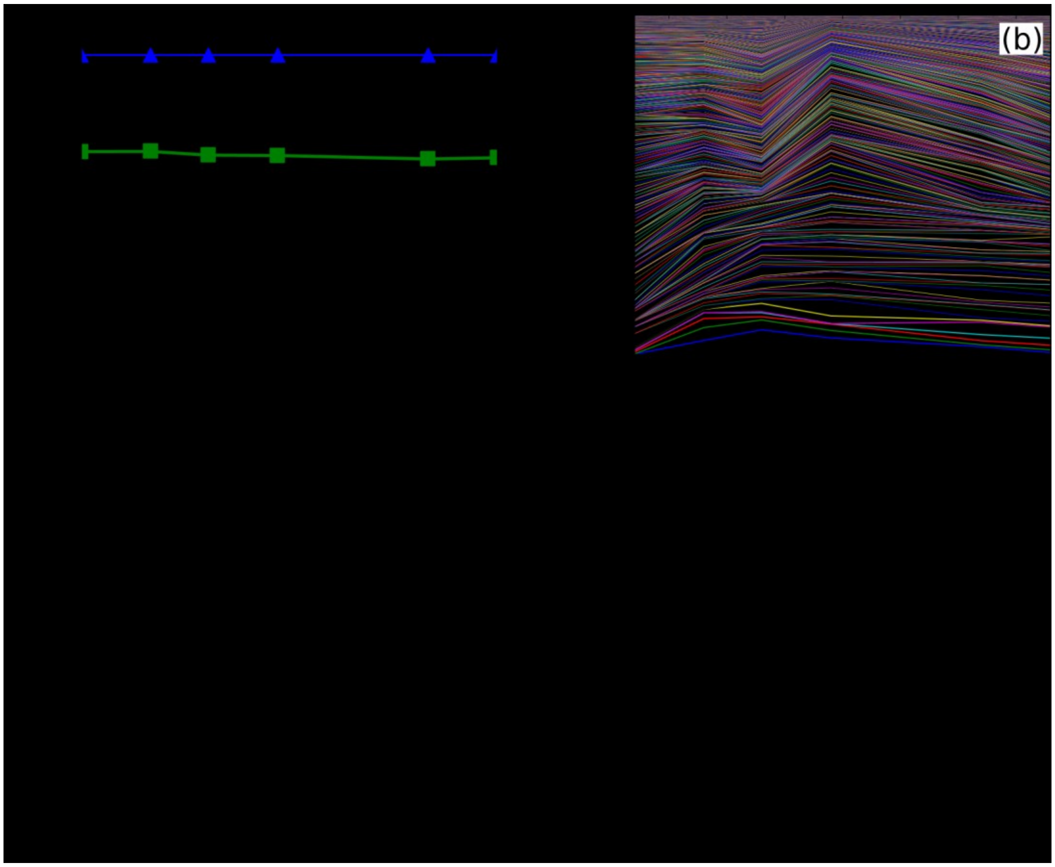


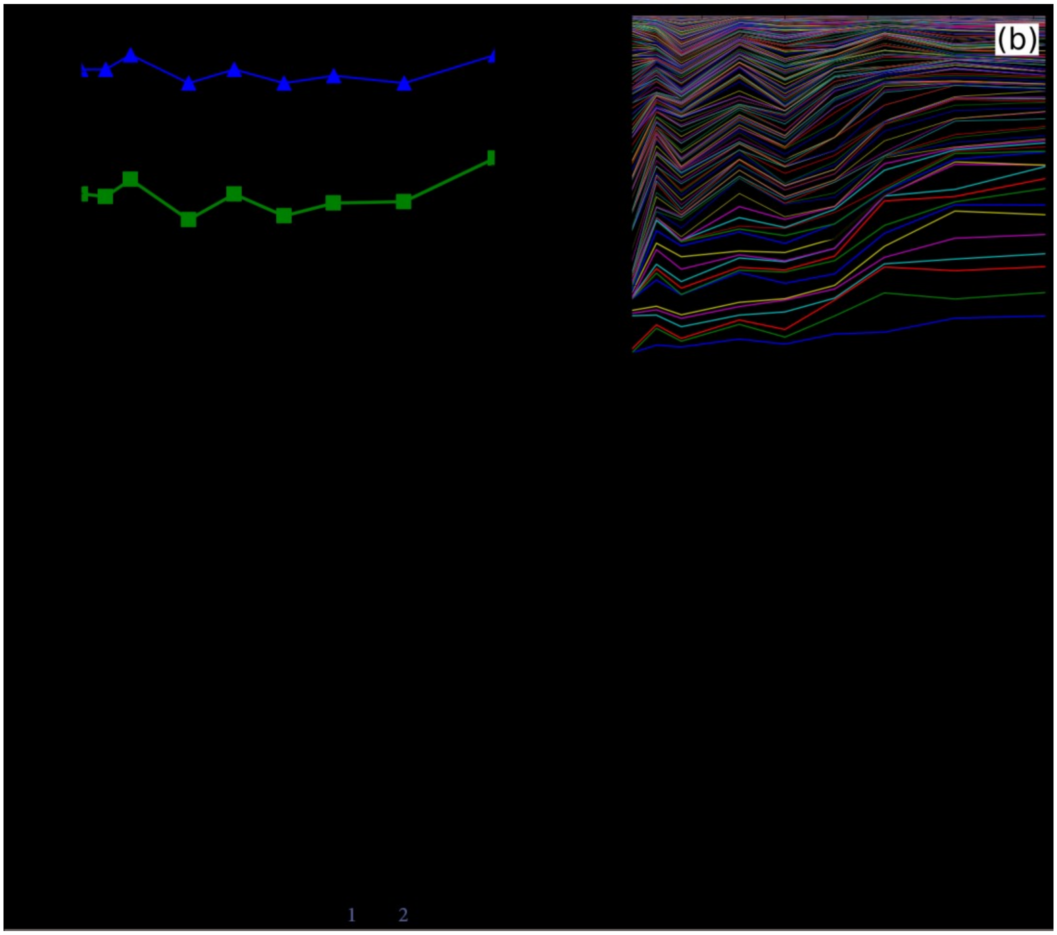
Fig 10. (a-b) Experimental data for animal 2RC003. (c) Difference between experimental $\hat{Y}_z$ and simulated $Y_z(L_e)$ as a function of L_e. The values of h_is are set to be equal to Hþ$\hat{y}_i$, and the model was simulated 200 times at each value of L_e. Other parameters are taken from Tables 1 and 2. The LSE L_e ¼ 25:0 and ðAþ$_{ss}$Þ ¼ 6:7. (d) Comparison of the optimal Y_z to the experimental $\hat{Y}_z$.

https://doi.org/10.1371/journal.pcbi.1006489.g010

green square and blue triangle markers in Figs 1(a), 10(a) and 11(a)), respectively. Figs 10 and 11 also show the clone abundances, the MSE functions, and the statistics of $Y(z)$.

Despite differences among the animals and the large variability in the estimated values of α and H_{ss} individually reported in the literature [4, 11, 12], the estimates of ðA_{ss}þÞ and L_e are rather similar across the three animals. For animal 2RC003, the optimal estimates are L_e 25:0, while for animal RQ3570, L_e ¼ 24:0. The corresponding estimates for A, after considering the constraint Eq (13) and the EGFP+ ratios in Table 2, are 282.7, 136.7, and 224.4.

We also compared how the simulated LSE Y_zðL_eÞ fits the experimental $\hat{Y}_z$ for all three animals. Note that for each specific z, the value of Y_z is the conditional mean of the values of y_i for which each clone i exhibits exactly z absences. Even though for any specific z, the distribution of the corresponding y_is is unknown, their mean Y_z should follow a normal distribution according to the central limit theorem. We use a one-sample t-test to compare $Y^\wedge{}_z$ against the mean of the Y_zðL_eÞ s generated from 10000 simulations using the optimal L_e ¼ L_e. For animal RQ5427, we actually performed seven one-sample t-tests on the $Y_z = \{Y_1, Y_2, ..., Y_7\}$ to find the seven p-values {0.69, 0.53, 0.58, 0.17, 0.68, 0.01, 3 × 10^{-5}}. Except for the last two p-values (corresponding to the bins $z = 6$ and $z = 7$), all other bins easily pass the one-sample t-test at a significance level of 0.05. Clones with $z = 6, 7$ are much smaller and more severely corrupted by noise, such as that induced during PCR amplification, and thus provide less reliable information.



https://doi.org/10.1371/journal.pcbi.1006489.g011

Comparisons of the test results among the three animals, together with comparisons among Figs 7(b), 10(d) and 11(d), show qualities of fit ordered according to RQ3570 < RQ5427 < 2RC003. This sequence of fitting qualities is consistent with the increasing experimental times

Table 2. Summary of specific parameter values for monkeys 2RC003 and RQ3570 derived from experimental measurements [13] or obtained by calculations (L_e and ð$A^{þ}_{ss}$Þ).

Parameter	Reference range or LSE value		
	RQ5427	2RC003	RQ3570
C^ $_s$	536	442	1371
ð$A_{þss}$Þ	14.7	6.7	19.3
A_{ss}	282.7	136.7	224.4
L_e	23.4	25.0	24.0
M_{ss}	3.2×10^9	4.6×10^9	3.8×10^9
$S^+(t_j)$	$(5.0 - 30) \times 10^3$	$(2.1 - 8.6) \times 10^3$	$(7.0 - 10.8) \times 10^3$
EGFP+ ratio	0.052	0.049	0.086
$\varepsilon(t_j)$	$(2.8 - 20) \times 10^{-5}$	$(1.2 - 4.2) \times 10^{-5}$	$(2.4 - 3.0) \times 10^{-5}$
Δt_j	150 − 330	180 − 660	150 − 260

https://doi.org/10.1371/journal.pcbi.1006489.t002

RQ3570 > RQ5427 > 2RC003, suggesting that age-associated changes of stem cell clone sizes cannot be fully neglected (which we did by fixing $\{h_i\}$) [48]. As is evident from Fig 10(a), several clones start to dominate after month 64; this coarsening phenomenon is not evident in the data of the other two animals. Animal RQ3570 was sacrificed at month 38, so no obvious coarsening is observed and no clones strongly dominate (see Fig 11). A summary of the parameters and fitting results for all animals is given in Table 2.

Discussion

In this study, we analyzed a decade-long clonal tracking experiment in rhesus macaques and developed mechanistic and statistical models that helped us understand two salient features of clone abundance data: the heterogeneous (nonuniform) distribution of clone sizes and the temporal fluctuation of clone sizes. Below, we further discuss the implications of our results, the structure of our mechanistic model, and the potential effects of including additional biological processes.

Comparison to previous studies

The long-term clonal tracking data we analyzed were generated from a huge number of initially tagged HSPCs ($C_h(0)$ * $10^6 - 10^7$) [13], a large number of observed clones (C_s * $10^2 - 10^3$), small numbers of sequenced cells that carry tags (^$S^þ$ðt_jÞ $10^3 10^4$), and infrequent sampling ($\Delta t_j > 5$ months). These features present significant challenges to the modeling and analysis over previous studies that mostly focused on one or a few clones [5, 15, 17, 18].

In a previous analysis, Goyal et al. [4] aggregated the clone abundance data across *all* mature cell types and studied the distribution of the *number* of clones of specific size. At each time point, they ordered the clones according to their sizes. Thus, the ordering can change across samples as some clones expand while others diminish. They found that the cumulative clone-number distribution (defined as the number of clones of a specific size or less) of the size-ordered clones becomes stationary as soon as a few months after transplantation. They proposed a neutral birth-death description of progenitor cells and fitted the *expected* value of clone counts in each sample by assuming h_i 18*iðPðh; t*Þ ¼ **1**ðh*;*1ÞÞ and tuning parameters in the downstream progenitor and mature-cell compartments. By focusing on aggregate clone counts, this study could not distinguish the dynamics of individual clones, nor could it predict the persistence of clone sizes over time. Since individual clone sizes (h_i, n_i, m_i, s_i of the same tag *i*) were not tracked, mechanisms driving the dynamics, and in particular, the variability and fluctuations of *individual* clone sizes that drive disappearances and reappearances, remain unresolved [4].

In our model, heterogeneity of clone sizes is explicitly generated by stochastic HSC selfrenewal of cells of each tag, and extinctions and resurrections arise from a generation-limited progenitor proliferation assumption. We infer model parameters as listed in Table 2. Combining the results with previous experimental and theoretical estimates of H_{ss} 1.1 × 10^4 − 2.2 × 10^4 [4, 49] results in $\alpha = 0.0045 - 0.027$, slightly larger than, but still consistent with, the estimates $\alpha = 0.0013 - 0.009$ by Shepherd *et al*. [11]. Previous studies that modeled total peripheral blood population estimated α 0.022 and H_{ss} 1.1 × 10^6/kg for dogs and α 0.044 and H_{ss} 1.1 × 10^6/kg for humans [12]. These estimates yield a value of αH_{ss} about 10^2 – 10^3 times greater than ours, which is nonetheless consistent with our steady-state constraint Eq (13) because they assumed a much smaller L 15 − 18 for dog and 16 − 21 for human. This difference in the estimates of L may be partially attributed to the transplant conditions under which the rhesus macaque experiments were performed [13]. Alternative model assumptions and differing values of other parameters may also contribute to this difference. For example, the extremely large value of H_{ss} 10^7 used in [34] will naturally decrease their estimate for L_e relative to that of our analysis.

Model structure, sensitivity to parameters, and cellular heterogeneity

Uncertainties in values of parameters such as μ_h, p_h, K_h, and other factors that tune the symmetric-asymmetric modes of HSC differentiation or involve HSC activation processes [50] will impart uncertainty in determining $P(h)$ and $\{h_i\}$. We have assumed $P(h)$ satisfies a master equation and depends on only two effective parameters λ and C_h. However, we have demonstrated that the statistical properties of Y_z are quite insensitive to the upstream configuration $\{h_i\}$ and hence to λ and C_h for a wide range of their values (see Fig 8). In other words, very little information in $\{h_i\}$ is retained in the sampled abundances $\hat{f}$ ðt_jÞ after HSCs differentiate and trigger random bursty peripheral blood-cell population dynamics.

Another feature we have ignored in our neutral model is cellular heterogeneity such as tagdependent differentiation, proliferation, and death rates. Cellular heterogeneity in HSC differentiation rates could be described by different α_i for each clone *i*, and the total differentiation

P^{C_h}

rate would be $A^{þ}_{ss}$ ¼ $_{i¼1} a_i h_i$. Differences in α_i can be subsumed into a modified configuration $\{h_i\}$ which, as we have seen, does not strongly influence our parameter estimation based on the Y_z statistics. Thus, given the available data and how information is lost along the stages of hematopoiesis and sampling, the present quasi-steady-state analyses cannot resolve heterogeneity across HSC clones.

We have not investigated how cellular heterogeneity in progenitor and mature cells would affect our results, but clone-dependences in their birth and death rates could affect sizes and durations of population bursts and quantitatively affect our analysis. However, unless the statistics of inter-burst times are highly variable across clones, we do not expect cellular heterogeneity to qualitatively affect our conclusions.

Changing downstream parameters such as μ_m or invoking alternative mechanisms of terminal differentiation (see Appendix B in S1 Appendix) can affect the shape of clonal bursts. We show in Appendix D in S1 Appendix that these effects can be subsumed into the effective maximum progenitor generation L_e. We have performed additional simulations to confirm that changing $\mu_m = 2$ will not influence the fitting of $A^{þ}_{ss}$ but increases L_e by one. In other words, inference of ð$A^{þ}_{ss}$Þ is robust against many upstream and downstream parameters, indicating that the intrinsic clone size fluctuations observed in the experimental data strongly constrain the total rate of HSC differentiation. On the other hand, uncovering the actual maximal generation L from L_e is possible only when uncertainties in these other parameters are resolved.

Clonal stability *vs* clonal succession

Our model reduction was based on the separation of timescales of the slow HSC dynamics and the fast clonal aging dynamics. Since HSC clone sizes vary extremely slowly for primates (Oð10^2Þ months), we ignored the homeostatic births/deaths of HSCs when fitting the temporal clonal variations. This is partially justified by visual inspection of Figs 1(b), 10(b) and 11(b) that show no significant variations of large clones' abundances is observed before 60 months. Instead, the random intermittent HSC differentiation events induce relatively short (Oð1Þ months) bursts of granulopoietic progeny that contribute strongly to temporal fluctuations of clone sizes. Such behavior are consistent to the "clonal stability" hypothesis [51–53], which assumes that a fixed group of HSCs randomly contributes to an organism's blood production at all times.

The alternative hypothesis of "clonal succession" [16, 54, 55] assumes that different groups of HSCs are sequentially recruited to the blood production at different times. This hypothesis would be consistent with our model only under a different set of parameters where HSCs selfrenew/die at a rate comparable to that of $\Delta\tau_b$, the duration of a granulocyte burst. For example, murine HSC turnover rates μ_h are hypothesized to be 10-fold higher than those in primates while the clonal aging dynamics (and its timescale $\Delta\tau_b$) are relatively conserved across species [56]. According to our result in Appendix C in S1 Appendix, such a 10-fold increase in HSC death rate would lead to a 10-fold increase in HSC clone extinction rate, bringing the lifespans of HSC clones closer to the (progenitor) clonal aging timescale $\Delta\tau_b$. This interpretation is consistent with the fact that hematopoiesis in large primates has been described in terms of "clonal stability" while hematopoiesis in mice has been described in terms of "clonal succession" [16, 51–55]. We thus predict that with even longer tracking (> 100 months), the "clonal succession" mechanism could be significant in primates also.

Summary and future directions

In summary, we have built mechanistic and statistical models that enable the quantitative analysis of noisy and infrequent clonal tracking data. We focused on the huge temporal variability observed in the sampled clone abundances and defined a robust statistical measure Y_z of sample-to-sample clone size variability through the number of clonal disappearances. Of course, there is a nearly endless list of details such cellular heterogeneity and more complex biology that we did not include, but given the noisy data, we propose and quantify the simplest explanation for the observed heterogeneous clone abundances and the temporal "extinctions and resurrections". The key ingredients in our mechanistic model are HSC self-renewal (quantified by the effective parameter λ), intermittent HSC differentiation (quantified by the parameter $A^{þ}_{ss}$), and an effective maximum progenitor generation (quantified by the effective parameter L_e). Although we cannot fully resolve λ from data, the obvious mismatch between experiment and our model when λ is small shows that a certain level of HSC clone-size heterogeneity (larger λ 1) is necessary to match the sampled data. Similarly, we cannot fully resolve α and $H_{ss}^{þ}$, but their product, the total tagged HSC differentiation rate $A^{þ}_{ss}$ ¼ $\alpha H_{ss}^{þ}$, is one of the key parameters constrained by our modeling. By minimizing an objective function of Y_z over effective model parameters, we found LSE values L_e ¼ 23   25 and ð$A^{þ}_{ss}$Þ ¼ 100   300 for the three rhesus macaques. These quantities could not be inferred from the total, more static cell populations. These results also imply that true dynamical changes in $A^{þ}_{ss}$ and L_e could be masked by the intrinsically bursty dynamics of each clone but provide a framework for future study into extrinsic perturbations.

Our analysis provides insight into the variables and experimental conditions to which parameter inference is most sensitive, possibly guiding the design of future experiments. The approach and models can also be readily extended to quantify white blood cells of other types. For example, the mechanistic model can be directly applied to monocytes since they also have relatively simple dynamics and do not proliferate in the periphery [57]. Peripheral lymphocytes, however, would require additional experimental information because their populations are more sensitive to the state of the animal and can homeostatically proliferate [38].

Supporting information

S1 Appendix.

(PDF)

Acknowledgments

The authors thank S. K. Lyons for help editing.

Author Contributions

Conceptualization: Song Xu, Tom Chou.

Data curation: Sanggu Kim, Irvin S. Y. Chen.

Formal analysis: Song Xu, Tom Chou.

Funding acquisition: Irvin S. Y. Chen, Tom Chou.

Investigation: Song Xu, Tom Chou.

Methodology: Song Xu.

Project administration: Tom Chou.

Resources: Irvin S. Y. Chen.

Writing – original draft: Song Xu, Tom Chou.

Writing – review & editing: Song Xu, Sanggu Kim, Tom Chou.

References

1. Abkowitz JL, Catlin SN, Guttorp P. Evidence that hematopoiesis may be a stochastic process in vivo. Nature Medicine. 1996; 2(2):190–197. https://doi.org/10.1038/nm0296-190 PMID: 8574964

2. Mendelson A, Frenette PS. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. Nature Medicine. 2014; 20(8):833. https://doi.org/10.1038/nm.3647 PMID: 25100529

3. Stiehl T, Ho A, Marciniak-Czochra A. The impact of CD34+ cell dose on engraftment after SCTs: personalized estimates based on mathematical modeling. Bone marrow transplantation. 2014; 49(1):30. https://doi.org/10.1038/bmt.2013.138 PMID: 24056742

4. Goyal S, Kim S, Chen IS, Chou T. Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Biology. 2015; 13(1):85. https://doi.org/10.1186/ s12915-015-0191-8 PMID: 26486451

5. Busch K, Klapproth K, Barile M, Flossdorf M, Holland-Letz T, Schlenner SM, et al. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature. 2015; 518(7540):542–546. https:// doi.org/10.1038/nature14242 PMID: 25686605

6. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2010; 2(6):640–653. https://doi.org/10.1002/wsbm.86 PMID: 20890962

7. Bystrykh LV, Verovskaya E, Zwart E, Broekhuis M, de Haan G. Counting stem cells: methodological constraints. Nature Methods. 2012; 9(6):567–574. https://doi.org/10.1038/nmeth.2043 PMID: 22669654

8. Sze´kely T, Burrage K, Mangel M, Bonsall M. Stochastic dynamics of interacting haematopoietic stem cell niche lineages. PLoS Computational Biology. 2014; 10:e1003794. https://doi.org/10.1371/journal. pcbi.1003794 PMID: 25188267

9. Stiehl T, Marciniak-Czochra A. Characterization of stem cells using mathematical models of multistage cell lineages. Mathematical and Computer Modelling. 2011; 53:1505–1517. https://doi.org/10.1016/j. mcm.2010.03.057

10. Ho¨fer T, Rodewald H. Output without input: the lifelong productivity of hematopoietic stem cells. Current Opinion in Cell Biology. 2016; 43:69–77. https://doi.org/10.1016/j.ceb.2016.08.003 PMID: 27620508

11. Shepherd BE, Kiem HP, Lansdorp PM, Dunbar CE, Aubert G, LaRochelle A, et al. Hematopoietic stemcell behavior in nonhuman primates. Blood. 2007; 110(6):1806–1813. https://doi.org/10.1182/blood2007-02-075382 PMID: 17526860

12. Zhuge C, Lei J, Mackey MC. Neutrophil dynamics in response to chemotherapy and G-CSF. Journal of Theoretical Biology. 2012; 293:111–120. https://doi.org/10.1016/j.jtbi.2011.10.017 PMID: 22037060

13. Kim S, Kim N, Presson A, Metzger M, Bonifacino A, Sehl M, et al. Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. Cell Stem Cell. 2014; 14(4): 473–485. https://doi.org/10.1016/j.stem.2013.12.012 PMID: 24702996

14. Sieburg HB, Rezner BD, Muller-Sieburg CE. Predicting clonal self-renewal and extinction of hematopoietic stem cells. Proceedings of the National Academy of Sciences. 2011; 108(11):4370–4375. https:// doi.org/10.1073/pnas.1011414108

15. Copley MR, Beer PA, Eaves CJ. Hematopoietic stem cell heterogeneity takes center stage. Cell Stem Cell. 2012; 10(6):690–697. https://doi.org/10.1016/j.stem.2012.05.006 PMID: 22704509

16. Sun J, Ramos A, Chapman B, Johnnidis JB, Le L, Ho YJ, et al. Clonal dynamics of native haematopoiesis. Nature. 2014; 514(7522):322–327. https://doi.org/10.1038/nature13824 PMID: 25296256

17. Muller-Sieburg CE, Sieburg HB, Bernitz JM, Cattarossi G. Stem cell heterogeneity: implications for aging and regenerative medicine. Blood. 2012; 119(17):3900–3907. https://doi.org/10.1182/blood2011-12-376749 PMID: 22408258

18. Verovskaya E, Broekhuis MJ, Zwart E, Ritsema M, van Os R, de Haan G, et al. Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. Blood. 2013; 122(4):523–532. https://doi.org/10.1182/blood-2013-01-481135 PMID: 23719303

19. Doulatov S, Notta F, Laurenti E, Dick JE. Hematopoiesis: A human perspective. Cell Stem Cell. 2012;
10(2):120–136. https://doi.org/10.1016/j.stem.2012.01.006 PMID: 22305562

20. Kim S, Kim N, Presson AP, An DS, Mao SH, Bonifacino AC, et al. High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones. Journal of Virology. 2010; 84(22):11771–11780.
https://doi.org/10.1128/JVI.01355-10 PMID: 20844053

21. Ho¨fer T, Busch K, Klapproth K, Rodewald H. Fate mapping and quantitation of hematopoiesis in vivo.
Annual Review of Immunology. 2016; 34(1):449–478. https://doi.org/10.1146/annurev-immunol032414-112019 PMID: 27168243

22. Crauste F, Pujo-Menjouet L, Ge´nieys S, Molina C, O G. Mathematical model of hematopoiesis dynamics with growth factor-dependent apoptosis and proliferation regulations. Journal of Theoretical Biology. 2008; 250:322–338.

23. Adimy M, Crauste F. Mathematical model of hematopoiesis dynamics with growth factor-dependent apoptosis and proliferation regulations. Mathematical and Computer Modelling. 2009; 49:2128–2137. https://doi.org/10.1016/j.mcm.2008.07.014

24. Hoyem M, Maloy F, Jakobsen P, Brandsdal B. Stem cell regulation: Implications when differentiated cells regulate symmetric stem cell division. Journal of Theoretical Biology. 2015; 380:203–219. https:// doi.org/10.1016/j.jtbi.2015.05.009 PMID: 25997796

25. Østby I, Rusten LS, Kvalheim G, Grøttum P. A mathematical model for reconstitution of granulopoiesis after high dose chemotherapy with autologous stem cell transplantation. Journal of Mathematical Biology. 2003; 47(2):101–136. https://doi.org/10.1007/s00285-003-0198-6 PMID: 12883857

26. Marciniak-Czochra A, Stiehl T, Ho AD, Ja¨ger W, Wagner W. Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation. Stem Cells and Development. 2009; 18(3):377–386. https://doi.org/10.1089/scd.2008.0143 PMID: 18752377

27. Manesso E, Teles J, Bryder D, Peterson C. Dynamical modelling of haematopoiesis: an integrated view over the system in homeostasis and under perturbation. Journal of the Royal Society Interface. 2013;
10(80):20120817. https://doi.org/10.1098/rsif.2012.0817

28. Sun Z, Komarova N. Stochastic modeling of stem-cell dynamics with control. Mathematical Biosciences. 2012; 240:231–240. https://doi.org/10.1016/j.mbs.2012.08.004 PMID: 22960597

29. Yang J, Sun Z, Komarova N. Analysis of stochastic stem cell models with control. Mathematical Biosciences. 2015; 266:93–107. https://doi.org/10.1016/j.mbs.2015.06.001 PMID: 26073965

30. Greenman CD, Chou T. Kinetic theory of age-structured stochastic birth-death processes. Physical Review E. 2016; 93:012112. https://doi.org/10.1103/PhysRevE.93.012112 PMID: 26871029

31. Chou T, Greenman CD. A hierarchical kinetic theory of birth, death and fission in age-structured interacting populations. Journal of Statistical Physics. 2016; 164:49–76. https://doi.org/10.1007/s10955016-1524-x PMID: 27335505

32. Marciniak-Czochra A, Stiehl T, Wagner W. Modeling of replicative senescence in hematopoietic development. Aging (Albany NY). 2009; 1(8):723. https://doi.org/10.18632/aging.100072

33. Edelstein-Keshet L, Israel A, Lansdorp P. Modelling perspectives on aging: Can mathematics help us stay young? Journal of Theoretical Biology. 2001; 213(4):509–525. https://doi.org/10.1006/jtbi.2001. 2429 PMID: 11742522

34. Bernard S, Be´lair J, Mackey MC. Oscillations in cyclical neutropenia: new evidence based on mathematical modeling. Journal of Theoretical Biology. 2003; 223(3):283–298. https://doi.org/10.1016/ S0022-5193(03)00090-0 PMID: 12850449

35. Rufer N, Bru¨mmendorf TH, Kolvraa S, Bischoff C, Christensen K, Wadsworth L, et al. Telomere fluorescence measurements in granulocytes and T lymphocyte subsets point to a high turnover of hematopoietic stem cells and memory T cells in early childhood. The Journal of Experimental Medicine. 1999;
190(2):157–168. https://doi.org/10.1084/jem.190.2.157 PMID: 10432279

36. Hodes RJ. Telomere length, aging, and somatic cell turnover. The Journal of Experimental Medicine.
1999; 190(2):153–156. https://doi.org/10.1084/jem.190.2.153 PMID: 10432278

37. Miller R. Telomere diminution as a cause of immune failure in old age: an unfashionable demurral. Biochemical Society Transactions. 2000; 28(2):241–245. https://doi.org/10.1042/bst0280241 PMID: 10816135

38. De Boer RJ, Perelson AS. Quantifying T lymphocyte turnover. Journal of Theoretical Biology. 2013; 327:45–87. https://doi.org/10.1016/j.jtbi.2012.12.025 PMID: 23313150

39. Muller-Sieburg C, Cho R, Thoman M, Adkins B, Sieburg H. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. Blood. 2002; 100(4):1302–1309. PMID: 12149211

40. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. Systems Biology and Medicine. 2010; 2(6):640–653. https://doi.org/10.1002/wsbm.86 PMID: 20890962

41. Kendall DG. On the generalized "birth-and-death" process. The Annals of Mathematical Statistics. 1948; p. 1–15. https://doi.org/10.1214/aoms/1177730285

42. Chen Y, Qin S, Ding Y, Wei L, Zhang J, Li H, et al. Reference values of clinical chemistry and hematology parameters in rhesus monkeys (Macaca mulatta). Xenotransplantation. 2009; 16(6):496–501. https://doi.org/10.1111/j.1399-3089.2009.00554.x PMID: 20042049

43. Dancey JT, Deubelbeiss KA, Harker LA, Finch CA. Neutrophil kinetics in man. Journal of Clinical Investigation. 1976; 58(3):705. https://doi.org/10.1172/JCI108517 PMID: 956397

44. Lahoz-Beneytez J, Elemans M, Zhang Y, Ahmed R, Salam A, Block M, et al. Human neutrophil kinetics:
modeling of stable isotope labeling data supports short blood neutrophil half-lives. Blood. 2016; 127(26):3431–3438. https://doi.org/10.1182/blood-2016-03-700336 PMID: 27136946

45. Pillay J, den Braber I, Vrisekoop N, Kwast LM, de Boer RJ, Borghans JA, et al. In vivo labeling with

2 H_2O reveals a human neutrophil lifespan of 5.4 days. Blood. 2010; 116(4):625–627. https://doi.org/10.
1182/blood-2010-01-259028 PMID: 20410504

46. Parsons TL, Quince C, Plotkin JB. Absorption and fixation times for neutral and quasi-neutral populations with density dependence. Theoretical Population Biology. 2008; 74(4):302–310. https://doi.org/10. 1016/j.tpb.2008.09.001 PMID: 18835288

47. Gardiner CW. Handbook of Stochastic Methods: For physics, chemistry, and natural sciences. Springer, Berlin; 1985.

48. Yu KR, Espinoza DA, Wu C, Truitt L, Shin TH, Chen S, et al. The impact of aging on primate hematopoiesis as interrogated by clonal tracking. Blood. 2018; 131:1195–1205. https://doi.org/10.1182/blood2017-08-802033 PMID: 29295845

49. Abkowitz JL, Catlin SN, McCallie MT, Guttorp P. Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. Blood. 2002; 100(7):2665–2667. https://doi.org/10.1182/blood2002-03-0822 PMID: 12239184

50. Wilson A, Laurenti E, Oser G, van der Wath RC, Blanco-Bose W, Jaworski M, et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. Cell. 2008;
135(6):1118–1129. https://doi.org/10.1016/j.cell.2008.10.048 PMID: 19062086

51. Abkowitz JL, Persik MT, Shelton GH, Ott RL, Kiklevich JV, Catlin SN, et al. Behavior of hematopoietic stem cells in a large animal. Proceedings of the National Academy of Sciences. 1995; 92(6):
2031–2035. https://doi.org/10.1073/pnas.92.6.2031

52. Prchal JT, Prchal JF, Belickova M, Chen S, Guan Y, Gartland GL, et al. Clonal stability of blood cell lineages indicated by X-chromosomal transcriptional polymorphism. Journal of Experimental Medicine.
1996; 183(2):561–567. https://doi.org/10.1084/jem.183.2.561 PMID: 8627167

53. McKenzie JL, Gan OI, Doedens M, Wang JC, Dick JE. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nature Immunology. 2006; 7(11):1225–1233. https://doi.org/10.1038/ni1393 PMID: 17013390

54. Jordan CT, Lemischka IR. Clonal and systemic analysis of long-term hematopoiesis in the mouse. Genes & Development. 1990; 4(2):220–232. https://doi.org/10.1101/gad.4.2.220

55. Drize NJ, Keller JR, Chertkov JL. Local clonal analysis of the hematopoietic system shows that multiple small short-living clones maintain life-long hematopoiesis in reconstituted mice. Blood. 1996; 88(8): 2927–2938. PMID: 8874189

56. Catlin SN, Busque L, Gale RE, Guttorp P, Abkowitz JL. The replication rate of human hematopoietic stem cells in vivo. Blood. 2011; 117(17):4460–4466. https://doi.org/10.1182/blood-2010-08-303537 PMID: 21343613

57. Auffray C, Sieweke MH, Geissmann F. Blood monocytes: development, heterogeneity, and relationship with dendritic cells. Annual Review of Immunology. 2009; 27:669–692. https://doi.org/10.1146/annurev.
immunol.021908.132557 PMID: 19132917

S1 Appendix: Mathematical Appendices for “Modeling large fluctuations of thousands of clones during hematopoiesis: the role of stem cell self-renewal and bursty progenitor dynamics in rhesus macaque”

Song Xu[1], Sanggu Kim[2], Irvin S. Y. Chen[3], Tom Chou[4,*]

1 Center for Biomedical Informatics Research, Dept. of Medicine, Stanford University, Stanford, CA 94305, USA
2 Dept. of Veterinary Biosciences, Ohio State Univ., Columbus, OH 41320, USA
3 UCLA AIDS Institute and Dept. of Microbiology, Immunology, and Molecular Genetics, UCLA, Los Angeles, CA 90095, USA
4 Dept. of Mathematics, UCLA, Los Angeles, CA 90095-1555, USA

A Stochastic evolution of HSC clone sizes

To solve Eq. (6) in the main text for $P(h,t)$, we transform the equation using the probability generating function [redacted]. We have also neglected the subscript i because our model is “neutral” and $P(h,t)$ can describe the size of any HSC clone i. If the HSC self-renewal rate is approximated as $r_{\mathrm{h}}(H(t)) \equiv r_{\mathrm{h}}(t)$, the solution for $Q(s,t)$ takes on the following form [1]:

[redacted], (A1) [redacted] and . (A2)

where [redacted]
Note that for $h \geq 1$,

[redacted] and [redacted]. (A3)

These solutions obey the initial condition $P(h,0) = 1(h,1)$ and as $t \to\infty$, $\psi(t) \to \psi(\infty) \in (0,1)$, $\varphi \to\infty$, and $P(h,t) \to 0$. For [redacted] and $P(0,t \to\infty) \to 1$, indicating eventual extinction at long times [1, 2].

Using forms given in Eq. (A3), since both φ and ψ are independent of h, we can define

[redacted]. (A4)

[redacted]. (A5)

Thus, the probability distribution $P(h,t)$ can be written as

B Alternative model of progenitor aging

An alternative model to the one we have analyzed allows younger-generation progenitor cells ($\ell < L$) to differentiate into peripheral blood. Since each generation can differentiate with rate ω, the progenitor cell dynamics are slightly modified from those in our main model:

Poisson($\alpha h(t)$) − [redacted] ($r_{\mathrm{n}} + \mu_{\mathrm{n}} + \omega$)$n^{(0)}(t)$, $\ell = 0$, [redacted] , (B1)

Moreover, the dynamics of the mature peripheral blood population obey

$$\frac{m(t)}{\mathrm{d}t} = \sum_{\ell=0}^{L} \omega n^{(\ell)}(t) - \mu_{\mathrm{m}} m(t) \quad \mathrm{d} \quad . \tag{B2}$$

The solution to Eqs. (B1) and (B2) following a single differentiation event is

[redacted] , (B3)

These results can be applied to the model and analyzed and simulated using the same procedures as described in the main text. However, certain parameters have to be re-interpreted. For example, using the same value of $\omega = 0.16$ will significantly increase the effective death rate for progenitor cells of each generation. Fortunately, as we will show later, this alternative mechanism should not affect our main conclusion as the parameter-fitting results are not sensitive to the exact shape of cell bursts.

C Mean extinction time for a clone

As a function of the initial number h of HSCs in a clone, the mean extinction time (MET) $T(h)$ under the steady-state approximation $r_{\mathrm{h}} = \mu_{\mathrm{h}}$ obeys [3, 4]

$$[T(h+1) - T(h)]\mu_{\mathrm{h}}h - [T(h) - T(h-1)]\mu_{\mathrm{h}}h = -1. \tag{C1}$$

with an absorbing boundary condition $T(0) = 0$. By iterating Eq. (C1), we find

[redacted] , (C2) [redacted] (C3)

[redacted] , (C4)

[redacted] (C5)

which can be again iterated to obtain

To solve for $T(1)$, we invoke a reflecting boundary condition $T(H_{\mathrm{ss}}) - T(H_{\mathrm{ss}} - 1) = 1/(\mu_{\mathrm{h}}H_{\mathrm{ss}})$ [5], where

to find

Upon using Eq. (C5) in Eq. (C3), we find

[redacted] $\equiv T_{\mathrm{discrete}}(h),$ (C6)

which is the MET for a discrete system.

We can also approximate $T(h)$ by considering h as a continuous variable, and replace the summations in Eq. (C6) by integrations to find a simpler, more insightful approximation to $T(h)$:

$$h \quad {}_{H} \quad \mathrm{d}\ell \quad 1 \quad {}_{h\,1} \quad {}_{k}\mathrm{d}\ell$$

$$T_{\text{continous}}(h) = \mathrm{Z}\ {}_{\text{ss}} -\!\mathrm{Z}\!-\ \mathrm{d}k\mathrm{Z}\ \_\mu_{\text{h}}\ {}_{\ell=1}\ell\ \mu_{\text{h}}\ {}_{k=1}\ {}_{\ell=1}\ \ell$$

$$= h\ln H_{\text{ss}} - (h-1)\ln(h-1) + h - 2 \qquad \text{(C7)}$$

$$\mu_{\text{h}}$$

[redacted],

where we have used $\mathrm{R}^{x}(1/x')\mathrm{d}x' = \ln x$ and $\mathrm{R}^{x}\ln x'\mathrm{d}x' = x\ln x - x$. The continuous approximation to the MET matches the exact result quite well (relative error . 5%) for all values of h.

D Effective parameters and symmetric HSC differentiation

There are differing reports on the measured death rates for circulating granulocytes. We have used the most recently reported value $\mu_{\text{m}} = 1$ per day for humans. The effect of changing the value of $\mu_{\text{m}} \to \mu_{\text{m}}'$ on our analysis is a reinterpretation of L_{e}. By rewriting Eq. (13) as [redacted], we rearrange the expression to [redacted] and find [redacted]$_{\text{e}}$ [redacted]). For example, [redacted]= 2 would lead to [redacted]+ 1, where one additional round of progenitor doubling compensates for the doubled loss rate of mature granulocytes. One may argue that the change in μ_{m} can also be compensated for by doubling A^{+}_{ss}, which would have a different effect on the burstiness of the model compared to doubling L_{e}. However, when re-fitting the data with [redacted]= 2 or 0.2, we observed that $(A^{+}_{\text{ss}})^{*}$ did not change much, with most of the effect of modifying μ_{m} absorbed by changes in L_{e}^{*}.

Similarly, uncertainties in other parameters can also be subsumed into L_{e}. For example, setting [redacted]$\omega > 0$ implies that only half of the generation-L progenitors contribute to the peripheral blood. For a model with [redacted] = 0 to generate an equivalent effect, we can halve the number of mature cells by using an effective maximum generation parameter [redacted] 1. This indicates that the intrinsic clone size fluctuations demonstrated in the experimental data strongly constrain A^{+}_{ss}.

Another possible modification of our mechanistic model is to allow for the possibility of symmetric HSC differentiation. The effect of symmetric differentiation can again be subsumed into the parameter L_{e} without qualitatively affecting our analysis. Assume a proportion $0 \leq q \leq 1$ of HSC differentiations are symmetric, producing on average $1+q$ generation-0 progenitor cells. After L_{e} rounds of proliferation, the $1+q$ generation0 progenitors produce on average $(1+q)\times 2^{L_{\text{e}}}$ mature cells. This is equivalent to an exclusively asymmetric differentiation model ($q = 0$) with [redacted]+ 1). We also expect symmetric differentiation to slightly increase the speed of coarsening since each HSC differentiation is also accompanied by the HSC's death and clones represented by a single HSC would disappear under symmetric differentiation. However, given the small rate α of HSC differentiation, the large number C_{h} of clones, and the insensitivity of our results to the distribution h_i, the data cannot quantitatively resolve the symmetric-asymmetric modes of HSC differentiation.

E Alternative objective functions and statistical insights

We developed our data analysis based on the statistics of the quantity y_i, the time averaged relative clone sizes for those clones exhibiting z absences across their longitudinal samples. While reasonable parameter estimates were obtained from fitting to data, we also considered alternative objective functions. Specifically, we looked at the standard deviation [redacted] quantifying the temporal fluctuations of the relative sizes of each clone i. The way we construct an alternative objective function is similar to the way we constructed Y_z. Recall for Y_z, we calculated the average abundance across only those clones with the same $z_i = z$ absences across time. However, unlike z_i which takes a finite set of discrete values $\{1, 2, ..., J-1\}$, σ_i is a continuous variable so we have to artificially bin their values. Instead, we bin clones with similar y_i and study the average of their associated σ_i's. Since the distribution y_i is non-linear with a long tail, we evaluated $\ln y_i$ to obtain the near-linear distribution shown in Fig. E1(a), sorted $\ln y_i$ into equal-width bins, and calculated the average of the associated σ_is. Dividing the values of $\ln y_i$ into bins labeled by k, we compute

$$U_k = \frac{\sum_i \sigma_i 1(\text{clone}\, i \in \text{bin}\, k)}{\sum_i 1(\text{clone}\, i \in \text{bin}\, k)} \tag{E1}$$

in analogy with the definition of Y_z. The objective function can be straightforwardly defined as

$$\text{MSE}_\sigma(\theta_{\text{model}}) = \sum_k (U_k(\theta_{\text{model}}) - \hat{U}_k)^2. \tag{E2}$$

It is also unclear how to set upper and lower bounds on the range of y_i for comparison (in contrast to the natural bound on $1 \le z \le J-1$) because an unconstrained set of clones will be sensitive to the underlying h_i distribution (an undesirable property). In Fig. E1(b), we fit the data from animal RQ5427 using MSE_σ and find [redacted]4, consistent with our previous estimate using Y_z.

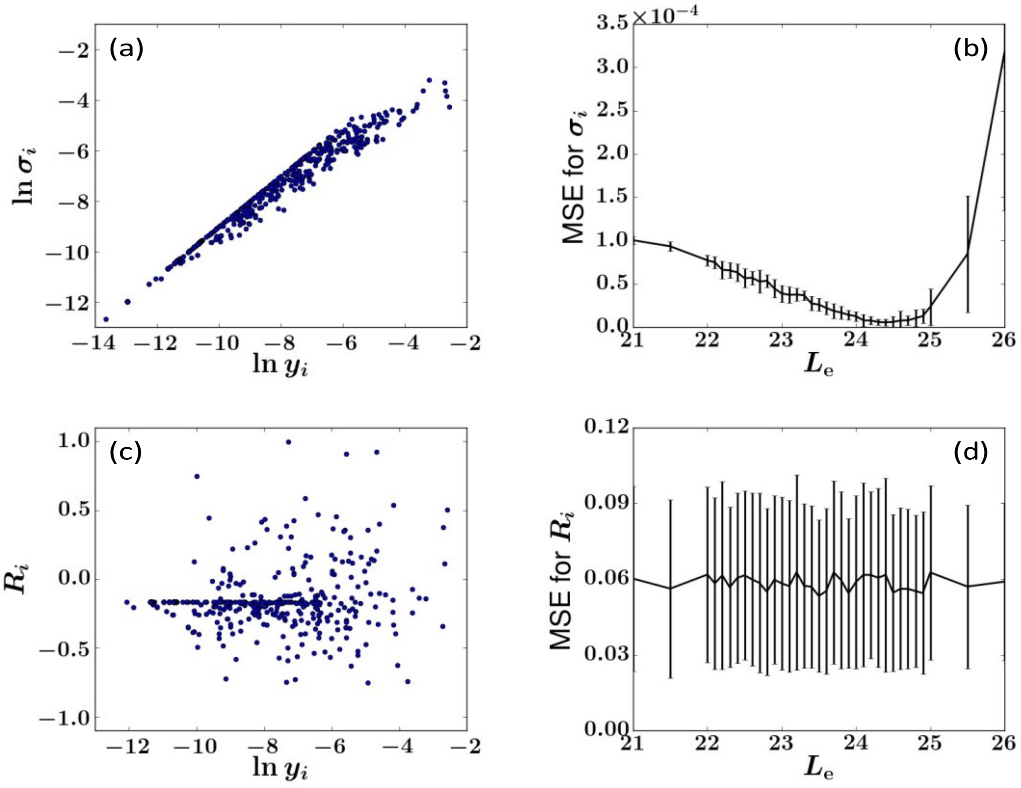


Figure E1: Statistics of the two alternative fluctuation measures and their fitting results. Each dot represents a clone. (a) Log standard deviation plotted against log average abundances. Clones are near-linearly distributed in the log average abundance space. (b) Objective function MSE$_\sigma$ *vs.* L_e. Clones of similar y_i are binned, and their averaged σ_i were used to compute U_k. (c) Autocorrelations R_i *vs.* log of average abundances u_i. There is no clear pattern in the distribution of R_is. (d) MSE$_R$ *vs.* L_e. This objective function cannot resolve the LSE L_e^*.

While it is also possible to choose σ_i as a measure of clone population fluctuations, we list several advantages of $\hat{}z_i$ over σ_i for the current dataset. Note that the number of disappearances z_i of each individual clone is defined on a finite set of integers (unlike the continuously measured σ_i), making it easier to bin clones with the same z values. Different clones i will exhibit different time-averaged abundances y_i but may have the same value of z_i. As shown in Fig. 4 in the main text, the larger $\hat{}z_i$ is, the smaller the corresponding ln $\hat{}y_i$ tends to be. The robust correlation between z_i and y_i encodes the level of fluctuations for a clone of certain size. For a given y_i, the larger z_i, the "burstier" the dynamics, implying a smaller number of tagged HSC differentiations per unit time (a smaller A^+_ss).

Another advantage of using z_i statistics emerges when fitting model results to the pattern of the measured data in Fig. 4 in the main text. Average sizes y_i (and the underlying h_i) associated with clones having $1 \leq z \leq 7$ all contain at least one absence. This constraint naturally controls the upper and lower bounds of h_i in a particular z bin ($1 \leq z \leq 7$), based on the burstiness of the model. Exact knowledge of the configuration $\{h_i\}$ is not required for fitting these y_i data.

Thus, dividing clones into z bins provides us with a natural way to exclude unconstrained clone sizes. In other words, the theoretical values of y_i (and the underlying h_i) associated with bin $z_i = 0$ can be arbitrarily and unreasonably large, and such a possibility should be excluded. Similarly, all y_i below a threshold size generate $z_i = J$ (clones that never appeared in the sampled blood) and do not provide any statistical power. This advantage of using z_i can also be confirmed by visual inspection of Fig. 9(b) in the main text. Several very large clones do not follow the general statistical pattern and show extremely large variances. Without manually filtering out these clones, our fitting in Fig. 1(b) results in a larger $L_\mathrm{e}^* = 24.4$ than the $L_\mathrm{e}^* = 23.4$ obtained in the main text using Y_z statistics. Finally, another option for comparing model with data is to use correlation functions. In this approach, the sampling gap Δt_j varies between 5 and 11 months, so the usual autocorrelation function with equal time gaps cannot be rigorously defined. We use the one-sample-gap autocorrelation function

[redacted]) (E3)

and bin values of lny_i in analogy to Eq. (E1) to define

P

P

and construct an autocorrelation-based objective function

$$\mathrm{MSE}_R(\theta_\mathrm{model}) = \mathrm{X}(W_k(\theta_\mathrm{model}) - \hat{W}_k)2. \qquad \text{(E5)}$$

$$W_k = \frac{\quad {}_i\, R_i 1(\mathrm{clone}\, i \in \mathrm{bin}\, k)}{1(\mathrm{clone}\, i \in \mathrm{bin}\, k)_i} \qquad \text{(E4)}$$

k

Since the inter-sample intervals Δt_j are larger than a typical burst size $\Delta\tau_\mathrm{b} \approx 32$ days, cells in different samples likely originate from different HSC differentiation events. Thus, the fluctuations of clone sizes are uncorrelated from sample to sample, as shown in Fig. E1(c). Randomly distributed between -1 and 1, the values of R_i are centered about the line[redacted], corresponding to the majority of clones that have $z_i = J-1$ (only 1 non-zero sample). Data fitting using R_i and MSE$_R$ is ill-conditioned and cannot resolve L_e^*, as shown in Fig. E1(d).

F Simulation of the forward model

To generate predictions, we first choose values of $\theta_{\text{model}} = \{\lambda, C_h, r_n, L_e\}$ and simulate our model, including sampling, to find $s_i(t_j)$. To simulate each realization of our model we

1. Specify the static HSC clone size distribution $P(h)$ by choosing the pair (λ, C_h) and draw $\{h_i\}$ from the geometric distribution C_h times using the Python package np.random.geometric. Normalize to construct the configuration[redacted]. Alternatively, we can also use the data $\hat{y}_i$ to approximate the configuration $\{h_i\}/H_{ss}^+$.

2. Fix all parameters θ_{model}, construct the total clone i differentiation rate [redacted] for each clone i. Generate realizations of sets of HSC differentiation event times [redacted]for each clone i based on the rate $\alpha h_i = A_{ss}^+ h_i / H_{ss}^+$.
3. Evaluate Eqs. (10) and (11) in the main text. Sum up the peripheral blood bursts initiated by eachdifferentiation event of each clone i to find[redacted]
4. Sample a fraction [redacted] of the total peripheral cell count $M^+(t_j) = \mathrm{P}_i\, m_i(t_j)$. Here, $\hat{S}^+(t_j)$, $\hat{M}^+(t_j)$, and the times t_j are defined by the experiment. We used the Python package numpy.random.binomial. The cell counts of each clone are $s_i(t_j)$. Use the simulated total tagged cell counts in the samples $S^+(t_j) = \mathrm{P}_i\, s_i(t_j)$ to normalize[redacted]). Up to this point, we have generated a data matrix $f_i(t_j)$ of size $C_h \times J$.

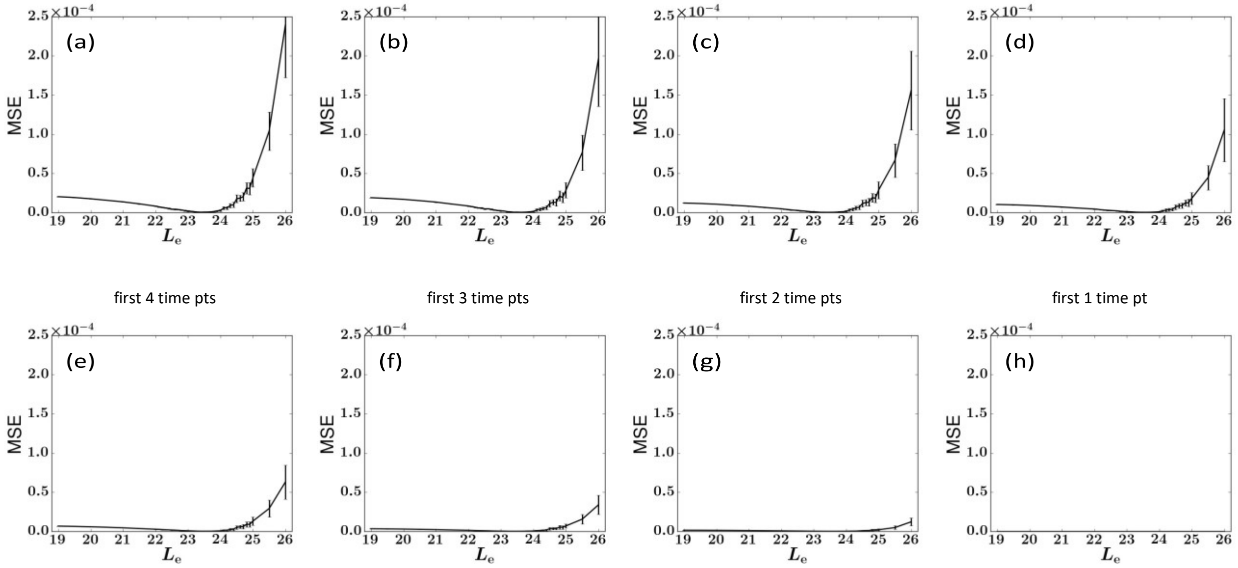
5. Increment L_e within the desired interval and repeat steps 2-4 200 times. For each value of L_e, the 200 simulations generate 200 $f_i(t_j)$ matrices. These repeats are to ensure that the noise induced from drawing values of h_i from $P(h)$ and sampling $s_i(t_j)$ from $m_i(t_j)$ do not significantly corrupt our parameter estimation.

The simulated, model-derived configurations $f_i(t_j)$ are then compared with experimentally measured values $f_i(t_j)$. The parameter L_e that minimizes the mean-squared error will be chosen as the least-squares estimate L_e^*.

G Robustness to samping frequency and threshold

The robustness of our inference of L_e^* to sampling frequency is demonstrated for animal RQ5427 by excluding some time samples. In Figs. G1(a-h), we plot the MSE function by including only the first $j = (8,7,...,1)$ time samples of the data. In this data set (animal RQ5427), the MSE remains meaningful, and the reconstruction of L_e^* is unchanged as long as at least four or five time samples are used. This conclusion is independent of which sampling time points are excluded. Since the system is well-approximated by a statistical steady state, the key determinant for robust inference is the *number* of samples included in the analysis.

all 8 time pts first 7 time pts first 6 time pts first 5 time pts



first 4 time pts

first 3 time pts

first 2 time pts

first 1 time pt

Figure G1: Simulated MSEs with λ = 0.99,C_h = 500,r_m = 2.5 for different numbers of time samples. From (a-h), only the first j = (8,7,...,1) time samples are used to fit the model. Provided at least two time samples are used, the reconstruction of L_e^* ≈ 23.4 − 23.6 remains fairly robust.

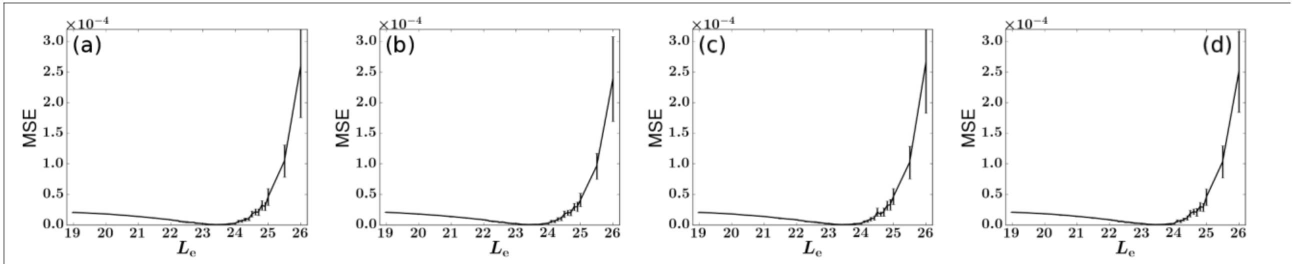
Robustness to a larger threshold of clone sizes is also demonstrated by eliminating clones whose average abundances are under a certain threshold in both the experimental and simulated data. In Figs. G2(a-h), we plot the MSE corresponding to the clone frequency thresholds $1.16 \times 10^{-5}, 2.03^{-5}, 3.41 \times 10^{-5}, 8.84 \times 10^{-5}, 1.66\times10^{-4}, 3.30\times10^{-4}, 6.78\times10^{-4}, 1.46\times10^{-3}$, respectively. Using these thresholds, the numbers of clones retained in the analysis are 482, 428, 375, 322, 268, 215, 159, and 107, corresponding to 90%, 80%, 70%, 60%, 50%, 40%, 30%, and 20% of the 536 total number of clones detected in animal RQ5427. Figs. G2 show that as long as & 200 clones are included (a-f), the MSE yields a clear LSE [redacted]6. Only at very high thresholds, where only 20-30% of the clones are retained, does the minimum of the MSE shift to slightly higher values [redacted]3 as shown in Figs. G2(g-h), respectively. Thus, we conclude that the inference of L_e^* from the data is fairly insensitive to sampling threshold provided a reasonable number of clones (typically & 200) are included in the analysis.

top 90% clones

top 80% clones

top 70% clones

top 60% clones



top 50% clones

top 40% clones

top 30% clones

top 20% clones

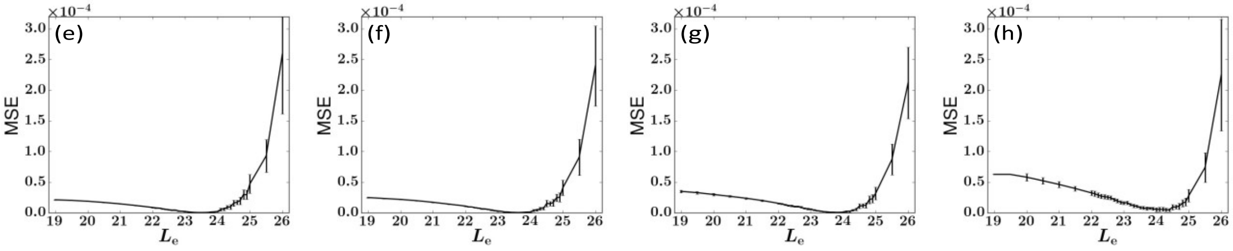


Figure G2: MSEs for animal RQ5427 using successively higher clone detection thresholds. Unique reconstruction of L_e^* is robust (a-f) even if only 40-50% of the clones are counted. (g-h) At even higher thresholds, the LSE for L_e^* increases only very slightly.

References

[1] Wang M. Nonhomogeneous Birth-death Processes, M. S. Thesis: California State Polytechnic University, Pomona. M. S. Thesis: California State Polytechnic University, Pomona; 2005.

[2] Yang J, Sun Z, Komarova N. Analysis of stochastic stem cell models with control. Mathematical Biosciences. 2015;266:93–107.

[3] Gardiner CW. Handbook of Stochastic Methods: For physics, chemistry, and natural sciences. Springer, Berlin; 1985.

[4] Allen L. An Introduction to Stochastic Processes with Applications to Biology. Taylor and Francis; 2010.

[5] Doering CR, Sargsyan KV, Sander LM. Extinction Times for Birth-Death Processes: Exact Results, Continuum Asymptotics, and the Failure of the Fokker–Planck Approximation. Multiscale Modeling & Simulation. 2005;3(2):283–299.