frontiers | Frontiers in Systems Biology

TYPE Original Research
PUBLISHED 09 February 2023
DOI 10.3389/fsysb.2023.893366

OPEN ACCESS

EDITED BY
Jayajit Das,
The Ohio State University, United States

REVIEWED BY
Steven M. Abel,
The University of Tennessee, United States
Wenjie Sun,
Institute Curie, France

*CORRESPONDENCE
Maria R. D'Orsogna,
dorsogna@csun.edu
Min Tang,
tangmin@sjtu.edu.cn
Tom Chou,
tomchou@ucla.edu

SPECIALTY SECTION
This article was submitted to Integrative Systems Immunology, a section of the journal Frontiers in Systems Biology

RECEIVED 10 March 2022
ACCEPTED 09 January 2023
PUBLISHED 09 February 2023

CITATION
Pan Y, D'Orsogna MR, Tang M, Stiehl T and Chou T (2023), Clonal abundance patterns in hematopoiesis: Mathematical modeling and parameter estimation.
Front. Syst. Biol. 3:893366.
doi: 10.3389/fsysb.2023.893366

Clonal abundance patterns in hematopoiesis: Mathematical modeling and parameter estimation

Yunbei Pan[1,2], Maria R. D'Orsogna[1,2], Min Tang[3]*, Thomas Stiehl[4] and Tom Chou[1,5]*

[1]Department of Computational Medicine, UCLA, Los Angeles, CA, United States, [2]Department of Mathematics, California State University at Northridge, Los Angeles, CA, United States, [3]Institute of Natural Sciences, Shanghai Jiaotong University, Shanghai, China, [4]Institute of Computational Biomedicine, RWTH Aachen University, Aachen, Germany, [5]Department of Mathematics, UCLA, Los Angeles, CA, United States

Hematopoiesis has been studied *via* stem cell labeling using barcodes, viral integration sites (VISs), or *in situ* methods. Subsequent proliferation and differentiation preserve the tag identity, thus defining a clone of mature cells across multiple cell type or lineages. By tracking the population of clones, measured within samples taken at discrete time points, we infer physiological parameters associated with a hybrid stochastic-deterministic mathematical model of hematopoiesis. We analyze clone population data from Koelle et al. (Koelle et al., 2017) and compare the states of clones (mean and variance of their abundances) and the state-space density of clones with the corresponding quantities predicted from our model. Comparing our model to the tagged granulocyte populations, we find parameters (stem cell carrying capacity, stem cell differentiation rates, and the proliferative potential of progenitor cells, and sample sizes) that provide reasonable fits in three out of four animals. Even though some observed features cannot be quantitatively reproduced by our model, our analyses provides insight into how model parameters influence the underlying mechanisms in hematopoiesis. We discuss additional mechanisms not incorporated in our model.

KEYWORDS

stem cells, hematopoiesis, barcodes, clonal tracking, differentiation

Introduction

Hematopoiesis, the process by which hematopoietic stem cells (HSCs) generate all mature blood cells in an animal through proliferation and differentiation plays a crucial role in an organism's immune response and maintaining overall homeostasis. Estimates of the number of actively cycling HSC range from 50000–200000 in humans (Lee-Six et al., 2018) and approximately 5,000 in mice (Busch et al., 2015; Mayle et al., 2015). It is well-known that these small numbers of hematopoietic stem cells can generate 10^{10}–10^{12} cells of multiple cell types daily, over an organism's lifetime (Fliedner, 2002; Doulatov et al., 2012). Understanding the mechanisms of hematopoiesis can help guide clinical treatment, especially those related to bone marrow transplantation and in the context of blood cancers (Mendelson and Frenette, 2014; Busch et al., 2015; Goyal et al., 2015).

HSCs are often quiescent (Seita and Weissman, 2010), making them hard to track *in vivo* and difficult to control *in vitro*. Thus, the HSC dynamics *in vivo* can only be straightforwardly interrogated through analysis of populations of more downstream progenitors and differentiated blood cells (Bystrykh et al., 2012). One way to quantitatively probe the

hematopoiesis process is the labeling of multipotent HSCs by tagging their genomes. The tags can take the form of viral integration sites or barcodes (Grosselin et al., 2013; Kim et al., 2014; Wu et al., 2014; Biasco et al., 2016; Koelle et al., 2017). In a typical *in vivo* experiment, $CD34^+$ stem cells are extracted, tagged, and then autologously transplanted back into the animal, typically a mouse or a rhesus macaque. $CD34^+$ cells contain HSCs as well as early hematopoietic stem and progenitor cells (HSPCs), with a wide range of estimated relative proportions (Corso et al., 2005; Attar, 2014; Parmentier et al., 2020). The downstream progenitor and mature cells that derive from proliferation of HSCs of a particular tag will form a clone of cells that share the same tag. Clonal tracking of cell tags is thus a powerful tool for interrogating the differentiation process during hematopoiesis (Lyne et al., 2018; Challen and Goodell, 2020; Cordes et al., 2021). For example, the abundances of the different tags that appear in the different types of mature cells can shed light on the branching structure of differentiation and on proliferation dynamics, particularly when coupled with mathematical models and/or simulations (Stiehl and Marciniak-Czochra, 2011; Sun and Komarova, 2012; Székely et al., 2014; Höfer and Rodewald, 2016; Xu J. et al., 2018).

Clonal tracking in mice (Copley et al., 2012; Sun et al., 2014) has revealed the timescales of repopulation dynamics under homeostasis and after bone marrow transplantation (Muller-Sieburg et al., 2012; Verovskaya et al., 2013; Busch et al., 2015), but typically involves very few clones that cover only a small fraction of the HSC population. To transplant many HSC clones in order to see patterns of how clones are distributed during hematopoiesis requires experiments on animals larger than mice.

Transplant experiments in rhesus macaque on the other hand allow for hundreds or thousands of clones to be engrafted into an organism that exhibits population levels and timescales closer to those in humans. One experiment in rhesus macaque involved tracking HSC clonal dynamics of lentivirus-tagged HSCs and early progenitor cells (HSPCs), and following hematopoiesis over a time period comparable to the animal's life-span (Kim et al., 2010, 2014). Here, $CD34^+$ HSPCs from the bone marrow, which include various progenitor cells, were marked *via* the integration of a lentivirus vector with an accompanying green fluorescent protein (GFP) tag at random viral integration sites (VISs). After sublethal myeloablative irradiation to eliminate a substantial number of cells in the bone marrow, the tagged HSPCs were autologously transplanted. If these cells divide and differentiate after transplantation, their progeny will inherit the unique VISs. Sampling and sequencing of these mature cells indicates which ones are descendants of a founder HSC. Data collected from four macaques over 14 years were analyzed showing how bias towards the lymphoid or myeloid differentiation branches changes over time. More detailed analyses were also performed in order to connect clonal patterns during hematopoiesis with a mathematical model that describes how self-renewal, differentiation, and subsampling of a multiclone population affects clone abundances and their fluctuations across time (Goyal et al., 2015; Xu S. et al., 2018). By fitting a simple mechanistic model to abundances of hundreds to thousands of clones, random initial differentiation events that each led to a subsequent burst of mature cells was proposed as a mechanism to explain observed population fluctuations. The number of generations L that progenitor cells traverse along a differentiation pathway (lineage) before terminal differentiation was also estimated to be $L \sim 24$ for the granulocyte lineage (Xu S. et al., 2018). To obtain this result, a mean-field model for HSC self-renewal was developed and applied to experimental data on granulocytes, using only the mean and variance of clone populations in the data fitting.

In this paper, we improve on the model used in (Xu S. et al., 2018) by developing a framework that can explain population transients and that can predict the density of the number of clones with respect to mean clone sizes. Instead of analyzing VIS data from (Kim et al., 2010, 2014), we consider the barcode data from (Wu et al., 2014; Koelle et al., 2017). In the latter experimental studies, replication-incompetent HIV-derived lentiviral barcoding vectors were used to tag HSCs that were transplanted into four rhesus macaques. The barcode consists of a six base-pair library identification and a 35 base pair high-diversity cellular barcode. As with the VIS experiments, barcoded cells were reinfused in the animals after myeloablative total-body irradiation. Purified samples of blood cells were then subject to low-cycle PCR amplification with the two primers bracketing the barcode. This barcoding approach provides more precise quantification relative to other clonal tracking protocols such as VIS (Kim et al., 2014) and transposon tagging (Sun et al., 2014) approaches. Thus, we will analyze the barcoding data *via* a mathematical model with the goal of more accurately estimating physiological parameters such as HSC carrying capacity, progenitor cell division rates, and progenitor cell proliferative potential for the granulocyte cell lineage. Although clonal structure of mature cells of different lineages, such as T, B, monocytes, and NK cells, were quantified in (Wu et al., 2014; Koelle et al., 2017), lymphocyte maturation is more complex, involving additional intermediate steps and subsequent immune signaling and mature cell proliferation. Thus, we focus on the simpler and abundant mature granulocyte population (Bystrykh et al., 2012).

In the following Materials and Methods section, we briefly describe the raw data and present the mathematical model. In the Data Analysis and Results section, we describe how measured clone data is compared to predicted clone abundances and show that minimization of the difference leads to reasonable estimates of parameter estimates. Finally, in the Discussion and Conclusions, we provide qualitative insight into how model parameters affect the predicted clonal patterns and discuss further improvements and potential new modeling directions.

Materials and methods

In this section, we describe information extracted from the granulocyte abundance data in (Koelle et al., 2017) and the mathematical model used to describe this data. The experimental parameters associated with the experiments are listed in Table 1, which lists the number of cells (tagged and untagged) transplanted, the barcode library size used, and the total number of different barcodes observed across all samples of all lineages for each animal. These values inform us on the typical magnitude of experimental parameters to which our subsequent model must conform. In Table 2, we list parameters used in our mathematical model as determined either from experimental data or through estimates.

TABLE 1 Transplant parameters. The initial transplant populations for the four animals ZH33, ZG66, ZH19, and ZJ31. The total library size for the cell preparation was in the range $C_L = 53319 - 109085$. The total number of cells injected was $H = 2.3 \times 10^7 - 4.8 \times 10^7$, of which $H^* = 8.0 \times 10^6 - 1.67 \times 10^7$ were barcoded (corresponding to 23%–35% GFP+ labeling). Across all peripheral blood samples and cell lineages, the total number of barcodes detected in each animal was in the range $\hat{C}_s = 21450 - 62354$, *i.e.*, roughly half of injected HSC barcodes were detected in the peripheral blood samples. Among granulocytes, the total sampled richness (across all time points) ranged from 2660 – 32363.

Variable / Animal	Library size C_L	Injected cells H	Injected GFP+ H^*	Total C_s	C_s (grans)
ZH33	63469	3.2×10^7	1.11×10^7	25325	9221
ZG66	53613	4.8×10^7	1.67×10^7	21450	2660
ZH19	53319	4.8×10^7	1.1×10^7	31929	10964
ZJ31	109085	2.3×10^7	8.0×10^6	62354	32363

TABLE 2 Overview of variables and parameters. Parameters and variables and their estimated values if known. Some values need to be calculated from our model and are denoted "calc.," while others need to be self-consistently estimated. For example, from GFP tagging, the fraction of tagged HSCs is approximately $\sum_{i=1}^{C_h} h_i(0) / \sum_{i=0}^{C_h} h_i(0) \approx 15 - 35\%$ but can slowly vary in time. Values relating to sampled cell populations are derived from animal ZH33 in the experiment of (Koelle et al., 2017). HSC proliferation and death rates have been estimated in (Shepherd et al., 2007) and (Catlin et al., 2011). Numbers specific to granulocytes are indicated as such.

Variables/Parameters	Definition	Value
t_j, $j = 1, \ldots, J$	Sampling time points	~month
$\hat{s}_i(t_j)$, $\quad i = 1, \ldots, C_s$	No. of cells with tag i in sample drawn at t_j (data)	$\sim 0 - 10^4$
$\hat{s}_i = \frac{1}{J}\sum_{i=1}^{C_s} \hat{s}_i(t_j)$	Mean no. of cells with tag i in sample (data)	$\sim 0 - 10^4$
$\hat{S}(t_j) = \sum_{i=1} \hat{s}_i(t_j)$	Total no. of tagged granulocytes in each sample at t_j (data)	$\sim 2 \times 10^6$
$\hat{C}_s(t_j > 2$ months)	Total no. of clones (richness) in sample j	~ 1000 (grans)
$\hat{C}_s^{>2}$	Total richness across all $t > t_2$	2,335 (grans)
h_0	Untagged HSCs in bone marrow (model)	unknown, $\sim 10^4$
h_i, $i = 1, \ldots, C_h$	HSCs with barcode i in BM (model)	1–1,000
$n_i^{(\ell)}(t)$	No. of ℓ^{th}-generation progenitor cells with tag/barcode i (model)	calc
$m_i(t)$	No. mature cells with tag i (model)	calc
$s_i(t)$	No. of cells with tag i in sample (model)	calc
K	HSC niche carrying capacity (model)	inferred, 10^4–10^5
$C_h(t)$	Total no. of engrafted clones (model)	unknown, $\sim 10^4$
$C_s(t)$	Total no. of clones sampled at t (model)	simulated
$r_h(0)$	Intrinsic HSC self-renewal rate	$\gg$ 0.01/day
μ_h	HSC death rate	$<$ 0.01/day
α	HSC differentiation rate	10^{-3} – 0.02/day, inferred
r_n	Progenitor cell division rate	1–5/day
μ_n	Progenitor cell death rate	unknown, ~ 0/day
ω	Progenitor cell terminal differentiation rate	unknown, $< r_n$
L	Proliferative potential of progenitor cells	inferred, $L^* = 22$ (grans)
μ_m	Mature cell death rate	0.185/day (grans)
η	Average sample fraction	$\sim 10^{-5} - 10^{-4}$
$\eta(t_j)$	Fraction of sample j	$\sim 10^{-5} - 10^{-4}$

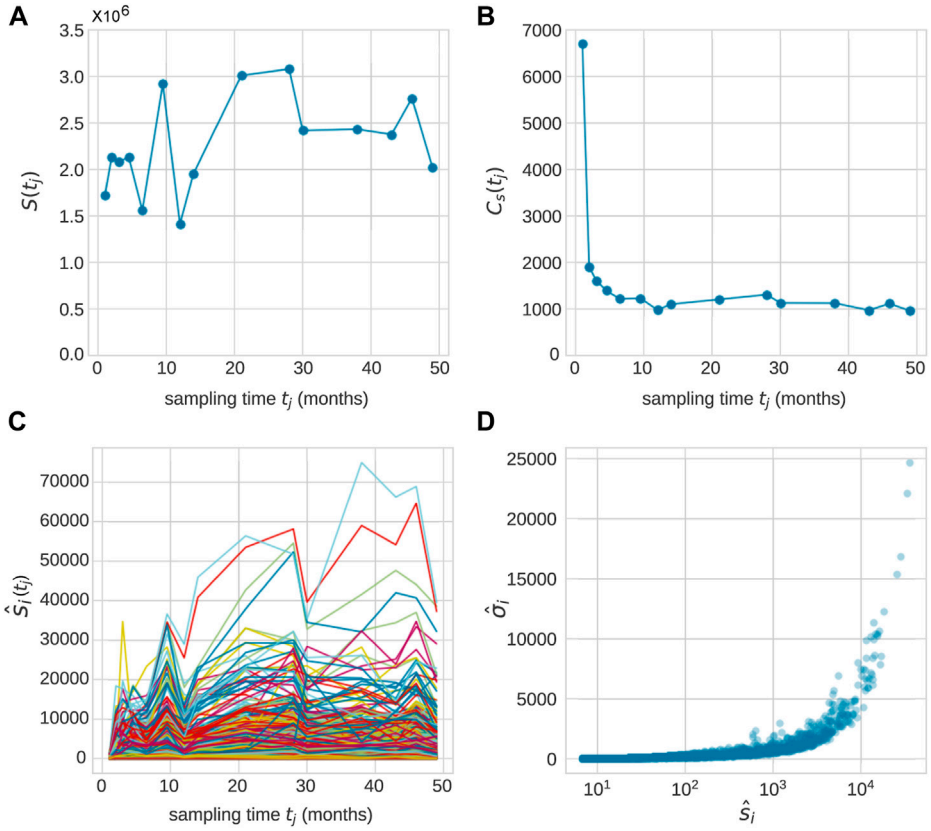

FIGURE 1
After transplantation, peripheral blood samples were taken across J time points t_j, $j = 1, \ldots,$ **(J)**. Typically, measurements were taken over 20–49 months and $J = 10 - 15$. **(A)** The total population $\hat{S}(t_j)$ of granulocytes sampled from animal ZH33 (Koelle et al., 2017) at times t_j = (1, 2, 3, 4.5, 6.5, 9.5, 12, 14, 21, 28, 30, 38, 43, 46, 49) months. **(B)** The total richness in each sample, $\hat{C}_s(t_j)$. The richness at the first two time samples are large (as we shall see, due to transplantation of barcoded progenitor cells). After the first two time points, where the richness will arise from the richness of the transplanted HSCs, the typical richness at each time point $\hat{C}_s(t_j \geq 2) \approx 1000$, while the richness across all $J - 2$ time points (for $t_j > 2$ months) is $\hat{C}_s^{>2} = 2335$. Across *all J time points*, 9,221 unique granulocyte clones were detected (out of a total of 25325 across *all cell types*). The individual clone abundances in the sampled granulocyte population are shown in **(C)** where the abundances of clone i in a sample taken at time t_j are denoted by $\hat{s}_i(t_j)$. The mean and standard deviation $\hat{\sigma}_i$ of the abundances of all clones across all sampling times are calculated using Eq. 1 and scatter-plotted in **(D)**. Each point represents one of the 2,335 detected granulocyte clones.

Measured quantities

First, we consider the observed data associated with each animal (Koelle et al., 2017), as shown in Figure 1. Granulocytes in blood samples drawn from each animal at times points t_j, $j = 1, 2, \ldots, J$ are sequenced and clonal (barcode) abundances tabulated. The total abundance (number of mature cells of a given cell type), $\hat{S}(t_j)$, and the richness $\hat{C}_s(t_j)$ (the total number of different barcodes detected in each sample) are also recorded and plotted in Figures 1A, B. In this study, $\hat{S}(t_j)$ denotes the total measured number of granulocytes (barcoded and unbarcoded) in a sample taken at time t_j. The fluctuations of $\hat{S}(t_j)$ and $\hat{C}_s(t_j)$ across t_j may arise from varying sampled sizes across time points and/or fluctuations in the state of the animal.

An example of the abundances of each clone within the granulocyte population from Koelle et al. (Koelle et al., 2017) is shown in Figure 1C. In these experiments, tagged stem cells are transplanted back into a rhesus macaque at $t = 0$ so that initially each clone consists of a single cell. A series of $1 \leq j \leq J$ samples are taken at time t_j after implementation, yielding a set of mature cells. We denote the abundance of clone i (among granulocyte cells) in the sample taken at time t_j after transplantation as $\hat{s}_i(t_j)$. The J measurements allow each clone of a particular mature cell type i to be characterized by a mean $\hat{s}_i$ and variance $\hat{\sigma}_i^2$ defined by

$$\begin{aligned} \hat{s}_i &= \frac{1}{J}\sum_{j=1}^{J}\hat{s}_i(t_j) \\ \hat{\sigma}_i^2 &= \frac{1}{J}\sum_{j=1}^{J}\left(\hat{s}_i(t_j) - \hat{s}_i\right)^2. \end{aligned} \tag{1}$$

Note that the total measured population of any cell type $\hat{S}(t_j) = \sum_{i=1}^{C_s}\hat{s}_i(t_j)$. A scatter plot of $\hat{\sigma}_i$ versus $\hat{s}_i$ for all clones detected in a sample of granulocytes is shown in Figure 1D.

For each clone at $(\hat{s}_i, \hat{\sigma}_i)$ we can evaluate the local density $\hat{\rho}$, the number of clones within some size window. This density can be viewed as the concentration of data points shown in Figure 1D as a function of mean clone size $\hat{s}$, and will be constructed using kernel density estimation (Rosenblatt, 1956; Parzen, 1962) of the data points in $(\hat{s}, \hat{\sigma})$ space. The unknown density function $\hat{\rho}$ is obtained by concatenating isotropic Gaussian kernel functions about each point and using an optimal, common bandwidth parameter, typically chosen as the value that minimizes the mean integrated squared error, or Kernel Density Estimation (KDE) (Silverman, 1986). The reconstructed density function can be thought of as a probability that a

random clone arises in the volume $(s, s + \mathrm{d}s) \times (\sigma, \sigma + \mathrm{d}\sigma)$. For each clone at $(\hat{s}_i, \hat{\sigma}_i)$ we can evaluate the local density $\hat{\rho}$.

In the remainder of this paper we will develop a mathematical model that we can simulate to generate total populations, clonal populations, and their associated attributes (s_i, σ_i, ρ_i). Note that while there are many existing mathematical models of hematopoiesis (Colijn and Mackey, 2005; Peixoto et al., 2011; De Souza and Humphries, 2019), they describe only time variations in total populations, rather than that of lower-population, individual clones. We will tune parameters of the model so that its predictions provide a reasonable match to the aforementioned measurements, paying particular attention to clone abundances and clone size variability.

Mathematical modeling

Our mathematical model incorporates known and accepted features of hematopoiesis. Three main cell compartments are considered: hematopoietic stem cells (HSCs), transit amplifying progenitor cells, and peripheral mature cells. Although the stem cell population in bone marrow is large and can be described using a deterministic model, the initial populations within each clone are small and require a discrete stochastic description. We will then assume that a small sample of mature cells is drawn from the animals at times t_j, $j = 1, 2, \ldots, J$ and sequenced.

We first describe the initial conditions including the number of HSCs and HSPCs injected into each animal. As listed in Table 1, H is the total number of HSCs injected into each animal, among which H_0 are untagged and $H_i \ll H_0$ contain barcode $1 \le i \le C_H$ (and are GFP+). The total tagged HSC population is $H^\star \equiv \sum_{i=1}^{C_H} H_i$ so that $H = H_0 + H^\star$. The richness $C_H \lesssim C_L$ is the number of barcodes transferred into the animal, which is comparable to the richness of the barcode library C_L used in each experiment. Since $H_0 \sim 10^7 \gg H_i$, we will consider the probability distribution of only the tagged populations, which is described by the multinomial

$$P(\mathbf{H}) = (H^\star)! \prod_{i=1}^{C_H} \left(\frac{1}{C_H}\right)^{H_i} \frac{1}{H_i!}, \tag{2}$$

where $\mathbf{H} \equiv (H_1, H_2, \ldots, H_{C_H})$ and $H^\star$ is the total number of GFP+ (barcoded) cells. Specifically, for animal ZH33 studied in (Koelle et al., 2017)) $H \approx 3 \times 10^7$, $H^\star/H \approx 0.35$, $\sum_{i=1}^{C_H} H_i \approx 1.1 \times 10^7$, $C_H \lesssim C_L \approx 6 \times 10^5$. Thus, the typical $H_i \approx H^\star/C_H \sim 180$.

A certain fraction η_0 of the H HSCs home into the bone marrow, successfully engraft, and subsequently actively self-renewal and/or differentiate. Engrafted HSC populations are defined by $\mathbf{h}(0) = (h_1(0), h_2(0), \ldots, h_{C_h(0)}(0))$, where the richness of engrafted HSCs in the bone marrow is $C_\mathrm{h} \lesssim C_H$. Transplantation efficiencies are typically single-digit percentages (Abbuehl et al., 2017; Radtke et al., 2020) and transplanted CD34$^+$ cells contain significant numbers of progenitor cells. Thus, the fraction $\eta_0 \ll 1$; if η_0 is sufficiently small (approximately $\lesssim 1/180$), then we can safely assume that the initial clone populations in bone marrow are represented by very few cells. For simplicity, we approximate $h_i(0) \approx 1$. Even if $\eta_0 \not\lesssim 1/180$, most barcodes will be represented by very few cells. We have verified that an initial condition in our model that allows for, say, some $h_i(0) = 2, 3$ does not qualitatively affect the mature cell populations.

The random selection of cells into the bone marrow can be thought of as a sampling (without replacement) process. Including the untagged population, the probability distribution of engrafted cells resulting from the injected tagged cell population $\mathbf{H} = (H_1, H_2, \ldots H_{C_H})$ is given by

$$P(\mathbf{h}(0)|\mathbf{H}) = \frac{1}{\binom{H^\star}{h^\star(0)}} \prod_{i=1}^{C_H} \binom{H_i}{h_i(0)}, \tag{3}$$

where $H^\star$ and $h^\star(0) = \sum_{i=1}^{C_H} h_i(0)$ are the total initial numbers of barcoded injected cells and engrafted barcoded HSCs, summed over all clones. Note that the number of untagged transplanted cells $h_0(0) \gg 1$ is large so that we can approximate it by its deterministic value $h_0(0) \approx \eta_0 H_0$.

To extract the overall probability of initial condition $\mathbf{h}(0)$, we average Eq. 3 over the prior $P(H_1, \ldots, H_{C_H})$ and find

$$\begin{aligned} P(\mathbf{h}(0)) &= \sum_{\mathbf{H}} P(\mathbf{h}(0)|\mathbf{H}) P(\mathbf{H}) \\ &= h^\star(0)! \prod_{i=1}^{C_H} \left(\frac{1}{C_H}\right)^{h_i(0)} \frac{1}{h_i(0)!}, \quad h^\star(0) \equiv \sum_{i=1}^{C_H} h_i(0). \end{aligned} \tag{4}$$

Besides the initial condition $h_i(t = 0) = 1$, the initial number of untagged HSCs $h_0(t = 0)$ is related to the transplantation efficiency and is generally unknown. Barcodes are associated with a GFP tag and the initial fraction of sampled cells that are GFP+ is $\sim 35\%$. Since we assume a neutral model, it is reasonable to assume that the fraction of injected tagged cells $H^\star/H$ is equivalent to the fraction of tagged cells in the engrafted population $\sum_{i=1}^{C_\mathrm{h}(0)} h_i(0) / \sum_{i=0}^{C_\mathrm{h}(0)} h_i(0) \approx 0.35$ (although this ratio slowly decreases *via* extinction). The precise richness of HSC population in stem cell niche, $C_\mathrm{h}(t) < C_H$ is also unknown, but except for fluctuations, has a lower bound of $\hat{C}_\mathrm{s}$, the total number of unique clones detected across all samples across all cell types. Thus, we take $h(0) = \sum_{i=0}^{C_\mathrm{h}(0)} h_i(0) = h_0(0) + C_\mathrm{h}(0) \approx C_\mathrm{h}(0)/0.35$.

Self-renewal, death, and differentiation into progenitor cells all contribute to the stochastic dynamics of h_i. Although the *total* HSC population in the niche $h(t) \approx \sum_{i=0}^{C_\mathrm{h}(t)} h_i(t)$ is large and can be approximated deterministically, the HSC population of each clone $h_i(t)$ may be small and must be treated stochastically. Under our neutral assumption, the intrinsic self-renewal rate r_h of HSCs does not depend on the barcode identity i. Since HSCs reside in niches in the bone marrow that place limits on growth, we assume the HSC proliferation rate follows a linearly decreasing form defined by a carrying capacity and the engrafted HSC population $h(t)$

$$\begin{aligned} r_\mathrm{h}(h(t)) &= r_\mathrm{h}(0)\left(1 - \frac{h(t)}{K}\right), \\ h(t) &\equiv \sum_{i=0}^{C_\mathrm{h}} h_i(t), \end{aligned} \tag{5}$$

where $r_\mathrm{h}(0)$ is the intrinsic proliferation rate of a single, isolated HSC. Note that the untagged HSCs are included through h_0. Finally, we assume that HSCs die at rate μ_h and differentiate at rate α and that these rates, like the growth rate in Eq. 5, do not depend on barcode identity.

As shown in the Supplementary Material, the richness $C_\mathrm{h}(t)$ may progressively decrease from random HSC death and extinction and can be estimated by solving for the stochastic birth-death process (neglecting outflux from differentiation) and using generating functions to find

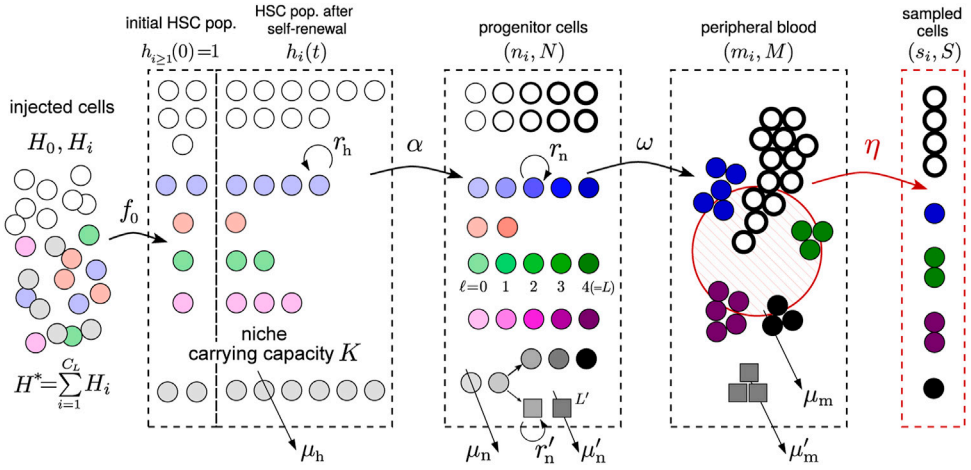

FIGURE 2
Schematic of the hybrid stochastic-deterministic model. Tagged (barcoded) stem cells are transplanted into an animal initially with one cell ($h_i(t = 0) \approx 1$) per clone. These cells, together with the untagged ones ($h_0(t = 0) \gg 1$) then undergo self-renewal and death, at rates $r_h\,(h\,(t) = \sum_{i=0} h_i\,(t))$ and μ_h, respectively, in the bone marrow. HSCs in all clones are also assumed to undergo asymmetric differentiation with rate α, forming a zeroth-generation progenitor cells. The population of ℓ^{th}-generation (or stage) progenitor cells, denoted $n_i^{(\ell)}$, further symmetrically differentiate with each division, up to a maximum of $\ell = L$ generations. The final-generation cell in clone i with population $n_i^{(L)}$ can then undergo terminal differentiation at rate ω to form mature, circulating peripheral blood cells. Mature cells at population m_i are then randomly sampled (with sampling fraction η and generating a sample population s_i) and sequenced. We wish to infer some of the parameters of the model by comparing the predicted means, standard deviations, and clone number densities with those from data (Figure 1). Lineage differentiation is schematically shown as a splitting of the grey clone between generations $\ell = 1$ and $\ell = 2$, where a new cell type (squares) branches off. The division and death rates of progenitor cells in this new lineage, r_n' and μ_n', may be different, as may the maximum number of generations L'. The mature cells turn over with rate a μ_m that may depend on lineage (but not clone identity within each lineage). In this paper, we assume that the lineages diverge at the zeroth-generation progenitor cell and analyze the model after the first differentiation step (rate α) independently for different cell types (in this paper, granulocytes).

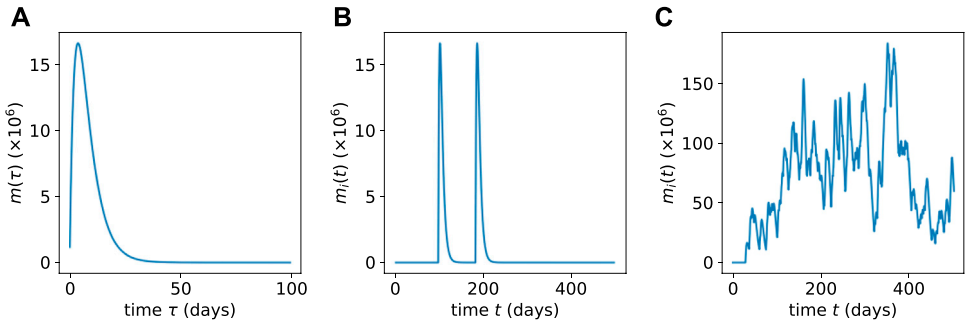

FIGURE 3
(A) The population $m_i(t)$ of mature cells resulting from a single HSC differentiation event as obtained from Eq. 12. **(B)** Multiple concatenated bursts from a low-population HSC clone showing well-separated intermittent pulses obtained *via* Eq. 13. **(C)** When the HSC population of a clone is large, the resulting mature cell population bursts merge together and exhibit lower relative variability.

$$\mathbb{E}[C_h\,(t)] \approx \frac{C_h\,(0)}{\psi\,(t) + \phi\,(t)}, \tag{6}$$

where

$$\psi\,(t) \equiv e^{-\int_0^t [r_h\,(t') - \mu_h]\,dt'}, \quad \phi\,(t) \equiv \int_0^t r_h\,(t')\psi\,(t')dt'. \tag{7}$$

In this expression, $r_h(t)$ is approximated by $r_h\,(\bar{h}_0\,(t) + \bar{h}^*\,(t))$, where $\bar{h}_0 + \bar{h}^*\,(t)$ is given by the explicit solution to the deterministic birth-death process with carrying capacity K. Thus, given $C_h(0)$, $r_h(0)$, K, and μ_h determine how the expected richness decreases. Henceforth we treat $C_h(t)$ in our model as the expected value $\mathbb{E}[C_h\,(t)]$ derived from our stochastic birth-death model, *i.e.*, we use $C_h(0)/(\psi(t) + \phi(t))$ as the model for $C_h(t)$.

We will also simulate the stochastic birth-death process for HSCs (see Supplementary Material for details), with the differentiation rate α that allows HSCs to differentiate into the progenitor/transit amplifying cell compartment (see Figure 2). Progenitor cells are further distinguished by their generation ℓ. Thus, ℓ not only measures generation number but also an effective differentiation state. Each HSC differentiation event leads to an $\ell = 0$ progenitor cell. Since the number of HSCs of any one clone is small, the initial differentiation events from clone i follow a Poisson process with rate αh_i. The

populations of the subsequent generations of progenitor cells in clone i are denoted by $n_i^{(\ell)}$. Once the $\ell = 0$ generation cells are generated, the number of progenitor cells quickly expand, so their dynamics will be described by a deterministic model as developed in (Xu et al., 2018)

$$\frac{\mathrm{d}n_i^{(\ell)}(t)}{\mathrm{d}t} = \begin{cases} \text{Poisson}\left(\alpha h_i(t)\right) - \left(r_{\mathrm{n}}^{(0)} + \mu_{\mathrm{n}}^{(0)}\right) n_i^{(0)}(t) & \ell = 0, \\ 2r_{\mathrm{n}}^{(\ell-1)} n_i^{(\ell-1)}(t) - \left(r_{\mathrm{n}}^{(\ell)} + \mu_{\mathrm{n}}^{(\ell)}\right) n_i^{(\ell)}(t) & 1 \le \ell \le L-1, \\ 2r_{\mathrm{n}}^{(L-1)} n_i^{(L-1)}(t) - \left(\omega + \mu_{\mathrm{n}}^{(L)}\right) n_i^{(L)}(t) & \ell = L, \end{cases} \tag{8}$$

where Poisson$(\alpha h_i(t))$ is the time-inhomogeneous point Poisson process describing HSC differentiation events. In other words, after a differentiation event at time t_1, the probability density of the time Δt to the next differentiation event is given by $\alpha h_i(t_1 + \Delta t) \exp[-\alpha \int_0^{\Delta t} h_i(t_1 + s)\mathrm{d}s]$. We will use the values of $h_i(t)$ from our stochastic simulations and sample from this inter-event time density to simulate realizations of differentiation events.

In Eq. 8, $r_{\mathrm{n}}^{(\ell)}$ and $\mu_{\mathrm{n}}^{(\ell)}$ represent the proliferation and death rates of generation-ℓ progenitor cells, respectively, and ω is the terminal differentiation rate into mature blood. Each division can also be thought of symmetric differentiation producing successively more differentiated progenitor cells. To model the finite proliferative potential of progenitor cells, we set the maximum number of generations to $\ell = L$, after which the L^{th} generation cell can only terminally differentiate to mature blood.

Consider a single isolated differentiation event of an HSC at $t = 0$ belonging to a particular clone. The resulting progenitor cell population after this event is described by Eq. 8 without the Poisson$(\alpha h_i(t))$ term but with an initial condition corresponding to a single $\ell = 0$ cell:

$$n_i^{(\ell)}(0) = \begin{cases} 1 & \ell = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The subsequent populations at time t form a temporal "burst" of cells that are described by $n_i^{(\ell)}(t)$ which is the solution of Eq. 8 without the Poisson$(\alpha h_i(t))$ term but using the initial condition in Eq. 9. If we assume that all progenitor generations carry the same division and death rates, $r_{\mathrm{n}}^{(\ell)} = r_{\mathrm{n}}$ for $0 \le \ell \le L-1$ and $\mu_{\mathrm{n}}^{(\ell)} = \mu_{\mathrm{n}}$ for $0 \le \ell \le L$, we can find an analytical solution associated with a single isolated burst as

$$n_i^{(L)}(t) = \frac{e^{-(\omega+\mu_{\mathrm{n}})t}}{(L-1)!}\left(\frac{2r_{\mathrm{n}}}{r_{\mathrm{n}} - \omega}\right)^L \int_0^{(r_{\mathrm{n}}-\omega)t} z^{L-1} e^{-z}\mathrm{d}z. \tag{10}$$

We can evaluate all populations $n_i^{(\ell)}(t)$ for $\ell < L$ by solving Eq. 8 and using Eq. 10, as detailed in the Supplementary Material.

If we assume that mature cells do not appreciably proliferate[1], the mature cell population in clone i obeys

$$\frac{\mathrm{d}m_i(t)}{\mathrm{d}t} = \omega n_i^{(L)}(t) - \mu_{\mathrm{m}} m_i(t), \tag{11}$$

where μ_{m} is the lineage-dependent turnover rate of mature cells. Using the solution to $n_i^{(L)}(t)$, we solve Eq. 11 to find (see Supplementary Material)

$$m_i(t) = \omega \int_0^t n_i^{(L)}(t') e^{-\mu_{\mathrm{m}}(t-t')}\mathrm{d}t'. \tag{12}$$

The mature cell population burst (of a specific clone) arising from a single, isolated differentiation event is plotted in Figure 3A. Note that the expression for a mature cell burst given in Eq. 12 is derived from the specific initial condition Eq. 9; however, some low-ℓ progenitors are also initially transplanted (see below). Thus, $m_i(t)$ will in general depend on the initial numbers of $n^{(\ell>0)}(0)$.

Since Eq. 8 are linear, populations arising from a sequence of Poisson-distributed differentiation events can be constructed by adding those derived from single events occurring at times T_k. In this case, the resulting mature cell population at time t is given by

$$m_i(t) = \sum_{k=1}^{k_{\max}} m_i(t - T_k), \quad T_{k_{\max}+1} > t > T_{k_{\max}}, \tag{13}$$

where $m_i(t - T_k)$ is the solution given in Eq. 12 (for the specific initial condition in Eq. 9). Two different sequences of bursts are shown in Figures 3B, C. In Figure 3B, we consider a clone with few HSCs such that $\alpha h_i(t) \ll \mu_{\mathrm{m}}$. This limit gives rise to differentiation events that occur rarely over the lifespan of mature cells, as depicted Figure 3B. In Figure 3C, we plot a sequence of more frequent mature cell population bursts that arise for more frequent differentiation events $\alpha h_i(t) \gg \mu_{\mathrm{m}}$ from HSCs that are in higher population clones.

Recall that the calculations involving Eqs 8–12 are performed for each clone i, resulting in a series of time-dependent expressions for $m_i(t)$ as per Eq. 12. These single event responses are then summed according to Eq. 13 to arrive at the total, time-dependent population of mature cells of a specific type and carrying the same barcode. These predicted whole-organism populations depend on the model parameters $\{C_{\mathrm{h}}(0), \alpha, K, r_{\mathrm{n}}, \mu_{\mathrm{h}}, \mu_{\mathrm{n}}, \mu_{\mathrm{m}}, L, \omega\}$. Since growth of transit amplifying progenitor cells is fast, we will henceforth assume $r_{\mathrm{n}} \gg \mu_{\mathrm{n}} \approx 0$. All other rates are given in units of per day.

Since the clone abundances are derived from sequencing cells in a small fraction η (for rhesus macaque, $\eta \sim 10^{-5} - 10^{-4}$) of the animal's blood taken at times t_j, we also need expressions for mature cell populations within a sample. Given the small sample sizes η, low population clones in the mature cell pool can easily be missed. For a given population m_i in the whole animal, the probability that s_i cells are captured in the sample is given by (Chao and Lin, 2012; Xu et al., 2020)

$$\begin{aligned} \mathbb{P}\left(\{s_i\}|\{m_i\}, S, M\right) &= \frac{1}{\binom{M}{S}} \prod_{i=0}^{C_{\mathrm{h}}} \binom{m_i}{s_i} \\ &\approx \prod_{i=0} \binom{m_i}{s_i} \eta^{s_i} (1-\eta)^{m_i - s_i}, \end{aligned} \tag{14}$$

where, for a given cell type (for example granulocytes), $S(t) = \sum_{i=0} s_i(t)$ is the total number of sampled cells (including untagged ones), $M(t) = \sum_{i=0} m_i(t)$ is the total number of circulating mature cells (including untagged ones), and the sampling fraction is $\eta = S/M$ (which we first assume is the same at each t_j).

After computing m_i at time t_j using Eq. 12, we take the nearest integer value and use it for m_i in Eq. 14. We then draw a single value $s_i(t_j)$ from the binomial distribution, assuming η is given. Finally, we simulate our model to generate trajectories of $M(t)$ and then determine the tagged sampled fraction

1 Certain lineages such as T cells can intermittently proliferate, e.g., upon antigen activation, but we neglect this and use small effective death rates μ_{m} for such cell types.

$$S(t) \approx \frac{H^{\star}}{H}\eta M(t). \tag{15}$$

Equations 5–15 represent a hybrid stochastic-deterministic model since the self-renewal process of a small number of HSCs in each clone and the final sampling step (Eq. 14) are modeled as discrete stochastic processes, while proliferation and differentiation of higher population progenitor and mature cells are treated deterministically *via* Eqs 8–12. The values $s_i(t_j)$ obtained from the model in Eqs 8–14 are used to generate the predicted clone population mean and standard deviations, s_i and σ_i, according to Eq. 1. Both s_i, σ_i are then used to determine the density of points ρ_i. These three values for each clone are then compared to their corresponding values constructed from data $\hat{s}_i(t_j)$, as we detail in the next section. These predictions, along with the predicted richness $C_s(t)$ and total sampled granulocyte population $S(t)$ provide the basis for comparing with measured data and parameter inference.

Data analysis and results

Data presentation

The published experimental data in (Koelle et al., 2017) provide data for four rhesus macaques ZH33, ZG66, ZH19, and ZJ31 over a period of 49, 42, 36 and 20 months, respectively. Samples were taken after autologous transplantation with myeloablative conditioning at times t_j, $1 \le j \le J$. Possible sampled cells are of five types: T cells, B cells, monocytes, granulocytes and NK cells. Barcoded myeloid (granulocytes and monocytes) and B-cells reach a noisy equilibrium after approximately 1 month, whereas for T-cells the time frame is longer, between 5 to 17 months. Furthermore, since granulocytes comprise a majority of white blood cells, we apply our mathematical and statistical model to the granulocyte lineage. At each sampling time t_j, the experimental data from (Koelle et al., 2017) reveals how many cells are sampled from each clone and what type each cell is. Across sampling time points, these sampled populations contain information on the overall abundance, how these abundances fluctuate in time, and the density of the number of clones detected as a function of abundance and abundance variability. Since ZH33 has the longest follow-up period, we use data from this macaque to compare experimental data with our mathematical predictions.

Matching model to data

Validation of our model will rely on matching predictions with available data in the form of $\hat{C}_s(t_j)$, $\hat{S}(t_j)$, and $(\hat{s}_i, \hat{\sigma}_i, \hat{\rho}(s,\sigma))$. Since the model contains many parameters and the data is noisy and "sparse," the model will likely overfit. Therefore, we carry out the parameter estimation by hand in stages, imposing limit on parameter values that are physiologically feasible.

First, we compare $C_h(t)$ (Eq. 6) with the richness $\hat{C}_s(t_j)$ shown in Figure 1A to provide a constraint among μ_h, K, $C_h(0)$, and $r_h(0)$. We assume that the $C_h(0)$ associated with granulocytes is slightly greater than the total richness across all samples after the first two, $C_h(0) \gtrsim \hat{C}_s^{>2}$. This is equivalent to assuming that granulocyte richness after about 2 months arises solely from barcoded HSCs. The cumulative post-two-month richnesses $\hat{C}_s^{>2}$ for animals ZH33, ZG66, ZH19, and ZJ31 are 2,335, 2007, 4,007, and 30732, respectively. Although the sample specific $\hat{C}_s(t_j)$ quickly decreases for $t > t_1$, our model prediction for $C_h(t)$ follows Eq. 6 and decays more slowly. By estimating $C_h(0)$ and using Eq. 6, we generate the predicted $C_s(t)$ and $S(t)$ by simulating our full model and comparing them to $\hat{C}_s(t_j)$ and $\hat{S}(t_j)$. This allows us to further constrain the parameters $C_h(0)$, $r_h(0)$, K, and μ_h. Note that no analytic formula exists for $C_s(t)$.

We first set $K \approx 100C_h(0)$ as the niche carrying capacity since smaller or larger values of K cannot provide the correct average clone sizes or approximately matching values of C_s. This comparison allowed us to obtain rough constraints and approximations to some parameters, particularly μ_h, $r_h(0)$, and $C_h(0)$. Discrete sets of values that are consistent with $\hat{C}_s(t_j)$ were selected and further pruned by using the remaining data.

The sampled richness $\hat{C}_s(t_j)$ in ZH33 exhibits a sharp decrease after the first sample time, without a corresponding collapse in the sampled abundances of granulocytes $\hat{S}(t_j)$. Our model explains this phenomenon by the initial condition; namely, the initially transplanted population of $CD34^+$ cells contains some partially differentiated HSPCs. Progenitor cells of barcode i (described in our model by the populations $n_i^{(\ell)}$) are initially transplanted so that some $n_i^{(\ell)}(t=0) > 0$ particularly for small ℓ (cells with a low degree of differentiation).

As shown in Figure 4, a fraction of the initial clones are HSPCs. Once these HSPC clones generate a burst of mature cells, they disappear from the animal without being renewed since there are no corresponding HSCs carrying the same barcode. Thus, the HSPC contribution to the overall sampled richness $\hat{C}_s(t_j)$ largely disappears after about 2 months. However, the total mature granulocyte population $\hat{S}(t_j)$ does not suffer a decline since HSPCs lost due to terminal differentiation are replaced by HSC differentiation. The subsequently sampled mature cell richness then reflects the richness $C_h(0)$ of the initially transplanted HSCs. We propose this partial HSPC transplantation as a mechanism for the observed rapid decrease in $\hat{C}_s$ observed in some animals. The shape of $\hat{\rho}(s)$ (see Figure 7D) can inform our estimate of the initial progenitor population $n^{(\ell)}(t=0)$. Maxima in $\hat{\rho}(s)$ can be accounted for by offspring of initially transplanted progenitor cells of different stages ℓ, with $n^{(\ell)}(0)$ generating smaller clones for larger ℓ (fewer remaining generations to expand).

Next, we consider the small clones, predominantly arising at short times, and find their average value at the first sample taken at t_1. These small clones also yield the highest density values $\hat{\rho}(s)$, but mostly disappear at longer times. Therefore, we assume they predominantly arise from initial progenitor cells. We can then generate the prediction $m(t=t_1)$ from our model assuming an initial condition $n^{\ell}(0)$ (and assuming no HSC contribution by setting $\alpha = 0$). This approximation provides a constraint on the deterministic progenitor cell parameters r_n, L, ω, η. In these experiments, the typical $\eta \sim 10^{-5}$, so we find that $L = 22$, and collect a set of feasible values for r_n, ω, and η that provide a good starting point for estimating the other parameters in the model. Note that r_n, L, ω, η can largely compensate each other at this level of comparison. In other words, sets of different ranges of values of one parameter will fit equally well provided some other parameters are also correspondingly adjusted.

Next, we least-squares minimize $\hat{S}(t_j) - S(t_j)$, where the model prediction $S(t)$ is given by Eq. 15. This further helps fix α. Once a set of parameters that allow for a reasonable match of $C_s(t_j)$ and $S(t_j)$ have

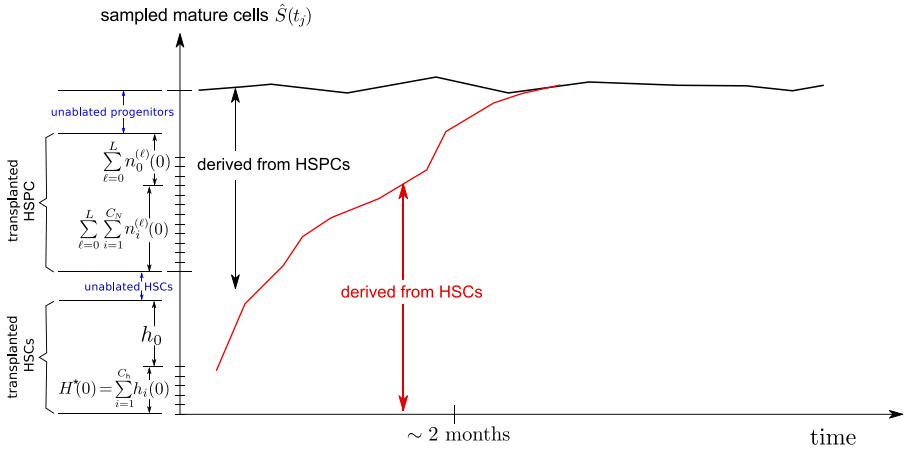

FIGURE 4
Schematic of a mixed HSC/HSPC initial condition. The transplanted CD34$^+$ cells contain HSCs and some progenitor cells (HSPCs), which are exhausted after about 2 months. The remaining richness arises mainly from that of the long term HSCs $C_h(0)$ which then slowly decreases as certain HSC clones become extinct.

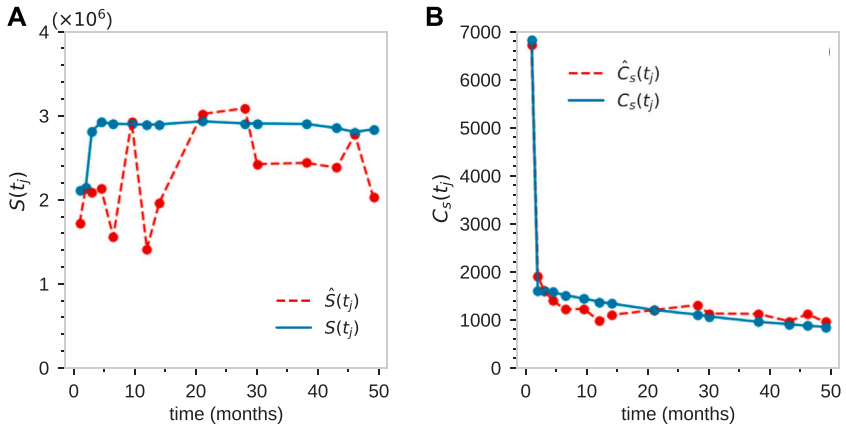

FIGURE 5
For animal ZH33, coarse fitting to the total population of tagged granulocytes $S(t_j)$ **(A)**, and the total number of clones at each time point $C_s(t_j)$ **(B)**. This initial rough matching was achieved using parameter values $L = 22$, $r_n \approx 2$, $\omega \approx 0.2$, and sampling size $\eta = 10^{-5}$.

been defined, as shown in Figures 5A, B, we then refine the fitting to the mean–standard deviation $(\hat{s}_i, \hat{\sigma}_i)$ scatter plot shown in Figure 1D.

To compare our modeling results to the remaining experimental data $(\hat{s}_i, \hat{\sigma}_i$, and $\hat{\rho}_i)$ and to fine-tune the estimates of the other parameters, the most natural intuition would be to compute the Euclidean distance (or any other relevant distance metric) between model predictions and experimental datasets and tune parameters so as to minimize this distance. However, since a number of clones go extinct in our model, the data size (for $\hat{s}_i, \hat{\sigma}_i$, and $\hat{\rho}_i$) varies in time.

Thus, before comparing predicted clone size distributions to measured results, we first cluster the data according to the values of $\hat{s}_i$ and $\hat{\sigma}_i$. Recall that $\hat{\rho}(s, \sigma) \approx \hat{\rho}(s)$ (since σ is highly correlated with s) is still determined from KDE using the raw, unclustered data $(\hat{s}_i, \hat{\sigma}_i)$. Clustering is performed using k-means to partition the data into multiple regions (in $\hat{s}$- and $\hat{\sigma}$-space) such that the Euclidean distance between a point and the center of its cluster is smaller than its distance to all other cluster centers. The goal is not to cluster the $(\hat{s}_i, \hat{\sigma}_i)$ points according to any real feature, but to simply reduce the dimensionality of the problem and to control the number of effective data points before applying least squares comparisons. Although there are no obvious features in the $(\hat{s}_i, \hat{\sigma}_i)$ data, k-means clustering of the distribution of points does yield an optimal number of clusters $k^\star$ *via* the "elbow" method where the curvature of the sum of square errors (distortion score) is maximal (Yuan and Yang, 2019). After implementing k-means clustering using the Python *yellowbrick* package, we find that the optimal number of clusters is typically $k^\star \approx 50 \pm 3$ depending on the initial randomization and partitioning process. Subsequent results, however, are insensitive to the precise numbers of clusters used as long as $k^\star \approx 50$.

Figure 6A compares $(\hat{s}_i, \hat{\sigma}_i)$ from experiments and from our hybrid multicompartment model. Figure 6B shows the distortion score (blue) and the convergence time (green) as a function of the number of clusters. The optimal number of clusters $k^\star = 48$ arises at the elbow of the distortion score curve. Figure 6C shows the clustered

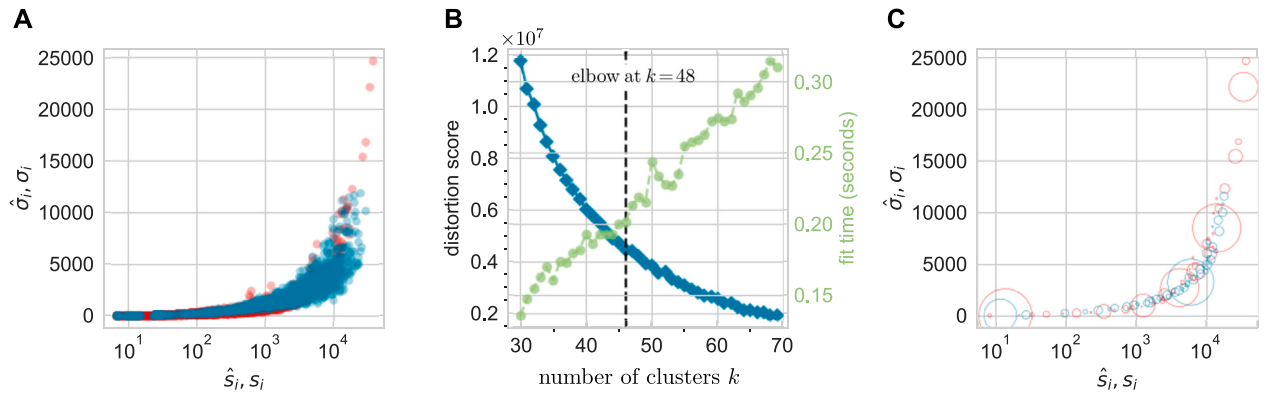

FIGURE 6
A total of 48 clusters are identified *via* k-means clustering of the data from animal ZH33. **(A)** The unclustered $(\hat{s}_i, \hat{\sigma}_i)$ data and model predictions for an arbitrary set of parameters. **(B)** The squared errors in k-means clustering or distortion score (blue) and convergence time (green) as a function of the number of clusters. The optimal value from the elbow signalling diminishing returns was found in this case to be $k^* = 48$. **(C)** The location of the cluster centers plotted on the $(\hat{s}, \hat{\sigma})$ plane. The radius of each circle indicates the fraction of all clones w_k that are associated with cluster k and is set to $2000 w_k$.

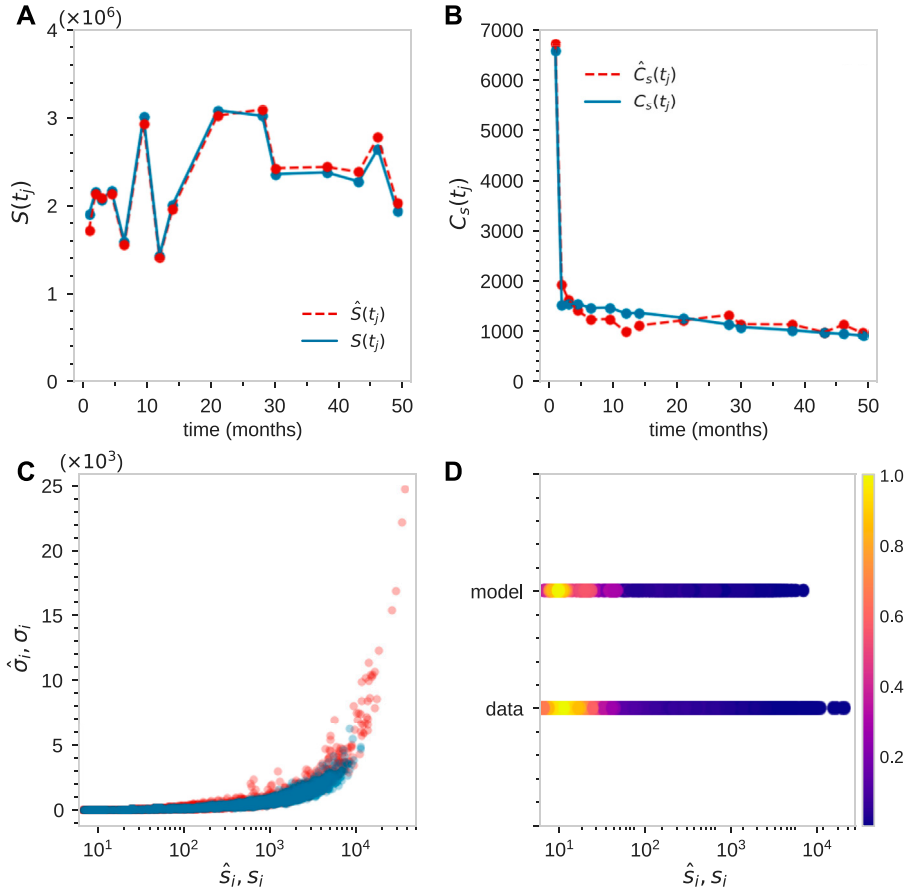

FIGURE 7
Fitting of tagged granulocyte populations for animal ZH33 after time-dependent adjustment of sampling size $\eta(t_j)$. For this animal t_j = [1, 2, 3, 4.5, 6.5, 9.5, 12, 14, 21, 28, 30, 38, 43, 46, 49] months. **(A)** By initially best-matching $S(t_j)$, we find $\eta(t_j)$ = [1, 2.03, 1.07, 0.94, 0.68, 1.25, 0.61, 0.83, 1.29, 1.29, 1.01, 1.01, 0.99, 1.14, 0.83]×10^{-5}. **(B–D)** By searching parameter space to find a good match to $\hat{C}_s(t_j)$, $\hat{s}_i$, $\hat{\sigma}_i$, and $\hat{\rho}$ (which is shown with its maximum nnormalized to one), which is plotted with its maximum normalized to unity. We find parameters [in units of/day] that fit the data are $\mu_h = 0.02$, $r_h(0) = 0.08$, $C_h(0) = 2500$, $C_n(\ell = 0) = 800$, $C_n(\ell = 1) = 1600$, $C_n(\ell = 2) = 3200$, $K = 2.5\times10^5$, $\alpha = 0.016$, $r_n = 2$, $L = 22$, $\mu_n = 0$, $\mu_m = 0.185$ and $\omega = 0.2$.

data (for animal ZH33) against the clustered model predictions. The radius of each circle, w_k, $k = 1, \ldots, 53$ denotes the fraction of data points (fraction of the total number of observed clones) assigned to cluster k and is thus a coarse-grained representation of the local data density.

Thus, we have clustered measured data according to $P = \{p_1, \ldots, p_{k^*}\} \equiv \{(\hat{s}_1, \hat{\sigma}_1, \hat{w}_1), \ldots, (\hat{s}_{k^*}, \hat{\sigma}_{k^*}, \hat{w}_{k^*})\}$, where $(\hat{s}_k, \hat{\sigma}_k)$ denotes the center values of cluster k and $\hat{w}_k$ is the area of the k^{th} cluster. For a fixed set of all model parameters, we generate predictions $Q = \{q_1, \ldots, q_{\ell^*}\} \equiv \{(s_1, \sigma_1, w_1), \ldots, (s_{\ell^*}, \sigma_{\ell^*}, w_{\ell^*})\}$ (in this expression and in the following, ℓ, ℓ^* denote matrix indices and not progenitor cell generations). The optimal number of clusters derived from our hybrid stochastic-deterministic model, ℓ^* will in general be different from the number of clusters k^* derived from data, but is typically also about $\ell^* \approx 50$.

In order to compare the clustered data P to the model prediction Q, which are matrices with different numbers of rows, we use the Wasserstein metric, or Earth mover's distance (EMD) (Villani, 2009; Kolouri et al., 2017) to define a distance between them. Let $d_{k,\ell}$ denote the distance between cluster p_k and cluster q_ℓ so that the matrix a $\mathbf{d} = \{d_{k,\ell}\}$ catalogues all possible cluster-cluster distances. We aim to compute a flow map $\mathbf{f} = \{f_{k,\ell}\}$ that yields the minimal distance between clusters $\mathbf{p} = \{p_k\}$ and clusters $\mathbf{q} = \{q_\ell\}$ by finding

$$\min_{\mathbf{f}} \sum_{k=1}^{k^*} \sum_{\ell=1}^{\ell^*} f_{k,\ell} d_{k,\ell} \tag{16}$$

Subject to the constraints

$$f_{k,\ell} \geq 0, \qquad \sum_{\ell=1}^{k^*} f_{k,\ell} \leq w_k, \qquad \sum_{k=1}^{\ell^*} f_{k,\ell} \leq w_\ell, \tag{17}$$

For all $1 \leq k \leq k^*$ and $1 \leq \ell \leq \ell^*$ and

$$\sum_{k=1}^{k^*} \sum_{\ell=1}^{\ell^*} f_{k,\ell} = \sum_{k=1}^{k^*} w_k = \sum_{\ell=1}^{\ell^*} w_\ell. \tag{18}$$

After finding the optimal flow $\mathbf{f}$, we evaluate the EMD as

$$\text{EMD}(P, Q) = \frac{\sum_{k=1}^{k^*} \sum_{\ell=1}^{\ell} f_{k,\ell} d_{k,\ell}}{\sum_{k'=1}^{k^*} \sum_{\ell'=1}^{\ell^*} f_{k',\ell'}}. \tag{19}$$

Model parameters are varied until our model-derived predictions best match the clustered data by minimizing the EMD. In this paper, we consider only the granulocyte lineage since it is the most abundant and reliably measured with minimal complex dynamics and regulation.

Finally, note the fluctuations in $S(t_j)$ which are too large to be captured by the intrinsic stochasticity in our model, as indicated in Figure 5B. These "unknown" fluctuations can arise from a number of processes, including variable sampling fractions $\eta(t_j)$ at each time point t_j and fluctuating animal state due to infections, stress, inflammation, etc. These may influence total mature cell populations month to month. Some of these effects can be effectively accounted for by adjusting the mean sample fraction $\eta = 10^{-5}$ by an amount $\Delta\eta(t_j)$ at each time point. Using these values of $\Delta\eta(t_j)$ to match the data $S(t_j)$, and then readjusting the parameters, we find good comparison between the experimental measurements and our model, as shown in Figures 7A–C.

To further show consistency, we then plotted the predicted density and compare it with the data-derived (using KDE) density $\hat{\rho}$ in Figure 7D. A few large, highly variable clones remain not well reproduced by our model and are discussed in the next section. The parameter estimation procedure was applied to the three other animals in (Koelle et al., 2017), showing reasonable, consistent matching (see Supplementary Material).

Discussion and conclusion

In this paper, we analyzed data from stem cell transplantation experiments in rhesus macaque (Koelle et al., 2017) in which barcoded stem and progenitor cells (HSPCs) were autologously transplanted after myeloablative conditioning. Typically, of the ~ 30 million cells transplanted only ~ 15–35% are tagged and then only a fraction homes to functional bone marrow niches. Nonetheless, typically hundreds or thousands of barcodes are detected in peripheral blood samples. The counts of circulating mature cells derived from each clone fluctuate from sample to sample; these fluctuations are significantly larger than those expected from random small samples (Xu et al., 2018) and thus arise from intrinsic stochasticity (Abkowitz et al., 1996) during hematopoiesis and/or physiological changes in the animals over the months that samples were being drawn.

In order to explain clone size variability, we extend a mathematical model first presented in (Xu et al., 2018). Our hybrid stochastic-deterministic model delineates all the clone populations and assumes regulation of the stem cell proliferation rate through a carrying capacity K, a finite differentiation potential of progenitor cells, terminal differentiation after a fixed number of divisions, and a final sampling step. Since the numbers of HSCs within each clone are typically small, we treated the self-renewal of HSCs within its niche stochastically by a coupled (through the carrying capacity) discrete birth-death process for each clone. In (Xu et al., 2018), the coupling in the stochastic HSC birth-death process was treated using a mean field approximation, which leads to a smaller clone size variability, everything else equal. Random times of asymmetric differentiation of each HSC into the first-stage progenitor cell is described by a rate α. After L differentiation steps, the L^{th}-stage progenitor cell terminally differentiates into a mature, circulating blood cell. The progenitor and mature cell pools are treated deterministically. We performed stochastic simulations using the Gillespie algorithm (Bortz et al., 1975; Gillespie, 2007) of the entire HSC pool and solve for the progenitor and mature cell populations numerically. Additional feedback mechanisms between the HSC and progenitor pools can be implemented (Klose et al., 2019), but in our case would require a more complex model incorporating two-way coupling between stochastic and deterministic dynamics.

Our model suggests that the sampled clone abundance variation arises primarily from random differentiation events by HSC clones that occur at rate α. Each differentiation event leads to a temporal burst of mature cells of the same clone. We hypothesize that these bursts of mature cell production lead to the variability in sampled clone populations. Other physiological factors related to animal state may still also play a role, but are not considered in our model. Another feature captured by our current model is the transient richness immediately after transplantation. This behavior is explained by our initial condition that contains short-term HSPCs that are within the initial $CD34^+$ pool. These HSPCs quickly differentiate but are not replenished at long times. This distributed initial condition may depend on the experimental protocol and may be indicative of the efficiency, and especially, the composition of the transplant.

Our model also allows us to explore how clone abundance predictions change with parameters. For example, we find that the range of larger clone sizes increase with increases in L, K, and α, in this order. All parameters affect the density of predicted data and cluster sizes. Mature cell death rates,

which vary across different cell types, also affect the predicted abundance variations, especially for larger values of μ_m.

We also note that while percentage of GFP+ cells within each mature cell type (lineage) fluctuated from sample to sample, they generally increased with animal age (after transplant) for all animals. T cells had the largest increase in their barcoded populations during the handful of months after transplantation. Since this timescale is much longer than the progenitor cell transients, this slow increase in tagged cell populations suggests that, assuming a neutral model (barcodes and barcode integration sites do not affect cell proliferation and death rates), (i) that a certain fraction of HSCs remained *in situ* after radiative ablation, and (ii) the $CD34^+$ barcoded and transplanted HSC population with an enriched GFP+ fraction were slowly activated. Scenario (i) means that some HSCs remained in their niche and continuously generated blood. The transplanted cells with barcoded HSCs (GFP+) can increase in importance if they slowly become more proliferative as they settle into the animal. Thus, the slow increase in GFP+ fraction in all five measured lineages suggest that transplanted cells may recover slowly from the transplantation procedure and increase their contribution to mature blood cell formation.

Finally, we note that there appears to be additional fluctuation mechanisms that are not accounted for in our model. In all animals, there appears to be a few very large clones with very high variability $\hat{\sigma}_i$. Within a stochastic HSC population, adjusting birth-death parameters that allow for larger clones and larger variances would suppress the richness to below what is observed. We have extensively explored feasible regimes of all parameters and conclude that allowing large clones that vary in abundance precludes agreement with other basic observables such as $\hat{S}(t_j)$ and $\hat{C}_s(t_j)$. Nonetheless, such unexplained features can be mechanistically informative and we discuss a number of reasonable factors that may account for them. First, the fluctuations in the measured total abundances of the different mature cell lineages did not correlate, implying that the specific set of sampling sizes $\eta(t_j)$ used to explain granulocyte population variations (as shown in Figure 7) cannot be used to analyze those of other mature cell lineages. The seemingly uncorrelated cross-lineage populations imply that time variations in animal state arise further downstream, affecting the development of individual lineages. If fluctuations occurred in stem or multipotent progenitor cells, they would affect multiple cell lineages in similar ways and lead to inter-lineage population correlations.

Animal ZJ31 (see Supplementary Material) appears to be uniquely different from the others in that it exhibited a much larger richness $\hat{C}_s(t_j)$ as well as a much smaller maximum clone size (which nonetheless had high variance $\hat{\sigma}_i$). For example, C_s at $t_j = 5$ months dips to a very low value, while the abundance $\hat{S}(t_j)$ seems to be at a local maximum. Thus, a small number of granulocyte clones expanded dramatically, potentially squeezing out the many smaller clones below sampling. At month seven, $\hat{S}$ is extremely low but $\hat{C}_s$ has recovered to its long term value, indicating that the previously large granulocyte clones were quickly cleared out. The results for ZJ31 may indicate a lower level of competitive exclusion, but also some other mechanism contributing to high variability. Therefore, the overall observation of high variability and the magnitudes of $\hat{C}_s$ indicates that other model features should be considered.

One assumption of our model that is likely an oversimplification is the neutrality of barcodes. Although different barcodes themselves may not influence cells, different VISs may. For example, aberrant self-renewal arises when using lentiviral vectors (Espinoza et al., 2019) and different VISs of HIV have been shown to affect cellular proliferation rates (*e.g.*, if the VIS is near an oncogene) Yeh et al. (2021). Besides the non-neutrality, we have also neglected stochastic or variable proliferative potential L and the time course of the HSC homing and engraftment into the bone marrow. A random but distributed L would allow a few randomly selected clones to expand further. We also expect that HSC migration and successful settlement into the bone marrow niche is a time-continuous process that provide a proliferation head start for a few early arriving clones. This would ultimately result in fewer clones C_s with some of them at higher populations $s_i(t)$. An instantaneous (more abrupt) HSC engraftment and a larger spread in L would be more consistent with animal ZJ31 than with the others.

Additional information can be extracted from the clone abundance data to further identify and interrogate such "opposing" behaviors. For example, we have only considered the average autocorrelation of the clone abundances (the variance) and have not constructed cross-correlations between cell types/lineages or correlations across time. Except for the initial period after transplantation, we have assumed a time-inhomogeneous process and have not considered explicit time-dependence such as aging (Muller-Sieburg et al., 2012; de Haan and Lazare, 2018). Physiological aging can be straightforwardly incorporated by *e.g.*, allowing for slow degradation of HSCs, changes in progenitor proliferative potential (Marciniak-Czochra et al., 2009), or changes in HSC niche carrying capacity $K(t)$. Mutations that arise with age may also *increase* HSC self-renewal (Challen and Goodell, 2020) which could be modeled by a $r_h(0)$ that increases with time. Thymic interruptions or involution with age (Lewkiewicz et al., 2019a,b) could also be modeled by assuming a decreasing maturation rate $\omega(t)$ when considering the T cell lineage.

While our current model contains a large number of parameters, it seems that a number of them are compensatory and control specific properties of the model predictions. For example, we found that α, L, ω, and η can compensate for each other and form an unknown effective parameter function $f(\alpha, L, \omega, \eta)$. This feature effectively reduces overfitting and might be better analyzed using machine learning methods. Incorporating the more realistic mechanisms discussed above would yield additional effective parameters allowing the model to more accurately reproduce the measured quantities; nonetheless, intermittent differentiation of HSCs remains the key proposed mechanism for understanding intersample clone abundance variations.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.1182/blood-2016-07-728691.

Author contributions

TC, YP, and MD'O developed and analyzed the model and wrote the manuscript. YP organized the data, performed numerical analyses, and data fitting. MT contributed to the analysis of the model and helped shape the stochastic simulation approach. TS provided insight on stem cell biology.

Funding

This work was supported by grants from the NIH through grant R01HL146552 (TC) and the Army Research Office through grant W911NF-18-1-0345 (MD'O) and DMS-1814090 (MD'O).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2023.893366/full#supplementary-material

References

Abbuehl, J. P., Tatarova, Z., Held, W., and Huelsken, J. (2017). Long-term engraftment of primary bone marrow stromal cells repairs niche damage and improves hematopoietic stem cell transplantation. *Cell. Stem Cell.* 21, 241–255. doi:10.1016/j.stem.2017.07.004

Abkowitz, J. L., Catlin, S. N., and Guttorp, P. (1996). Evidence that hematopoiesis may be a stochastic process *in vivo*. *Nat. Med.* 2, 190–197. doi:10.1038/nm0296-190

Attar, A. (2014). Changes in the cell surface markers during normal hematopoiesis: A guide to cell isolation. *Glob. J. Hematol. Blood Transfus.* 1, 20–28. doi:10.15379/2408-9877.2014.01.01.4

Biasco, L., Pellin, D., Scala, S., Dionisio, F., Basso-Ricci, L., Leonardelli, L., et al. (2016). *In vivo* tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell. Stem Cell.* 19, 107–119. doi:10.1016/j.stem.2016.04.016

Bortz, A. B., Kalos, M. H., and Lebowitz, J. L. (1975). A new algorithm for Monte Carlo simulation of ising spin systems. *J. Comput. Phys.* 17, 10–18. doi:10.1016/0021-9991(75)90060-1

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S. M., et al. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*. *Nature* 518, 542–546. doi:10.1038/nature14242

Bystrykh, L. V., Verovskaya, E., Zwart, E., Broekhuis, M., and de Haan, G. (2012). Counting stem cells: Methodological constraints. *Nat. Methods* 9, 567–574. doi:10.1038/nmeth.2043

Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P., and Abkowitz, J. L. (2011). The replication rate of human hematopoietic stem cells *in vivo*. *Blood* 117, 4460–4466. doi:10.1182/blood-2010-08-303537

Challen, G. A., and Goodell, M. A. (2020). Clonal hematopoiesis: Mechanisms driving dominance of stem cell clones. *Blood* 136, 1590–1598. doi:10.1182/blood.2020006510

Chao, A., and Lin, C. W. (2012). Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics* 68, 912–921. doi:10.1111/j.1541-0420.2011.01739.x

Colijn, C., and Mackey, M. C. (2005). A mathematical model of hematopoiesis: II. Cyclical neutropenia. *J. Theor. Biol.* 237, 133–146. doi:10.1016/j.jtbi.2005.03.034

Copley, M. R., Beer, P. A., and Eaves, C. J. (2012). Hematopoietic stem cell heterogeneity takes center stage. *Cell. Stem Cell.* 10, 690–697. doi:10.1016/j.stem.2012.05.006

Cordes, S., Wu, C., and Dunbar, C. E. (2021). Clonal tracking of haematopoietic cells: Insights and clinical implications. *J. Haematol.* 192, 819–831. doi:10.1111/bjh.17175

Corso, A., Varettoni, M., Mangiacavalli, S., Zappasodi, P., Klersy, C., Rusconi, C., et al. (2005). Bone marrow CD34+ cell count is predictive for adequate peripheral progenitor cell collection. *Leukemia Res.* 29, 159–163. doi:10.1016/j.leukres.2004.05.018

de Haan, G., and Lazare, S. S. (2018). Aging of hematopoietic stem cells. *Blood* 131, 479–487. doi:10.1182/blood-2017-06-746412

De Souza, D. C., and Humphries, A. R. (2019). Dynamics of a mathematical hematopoietic stem-cell population model. *SIAM J. Appl. Dyn. Syst.* 18, 808–852. doi:10.1137/18m1165086

Doulatov, S., Notta, F., Laurenti, E., and Dick, J. E. (2012). Hematopoiesis: A human perspective. *Cell. Stem Cell.* 10, 120–136. doi:10.1016/j.stem.2012.01.006

Espinoza, D. A., Fan, X., Yang, D., Cordes, S. F., Truitt, L. L., Calvo, K. R., et al. (2019). Aberrant clonal hematopoiesis following lentiviral vector transduction of HSPCs in a rhesus macaque. *Mol. Ther.* 27, 1074–1086. doi:10.1016/j.ymthe.2019.04.003

Fliedner, M. C. (2002). Research within the field of blood and marrow transplantation nursing: How can it contribute to higher quality of care? *Int. J. Hematol.* 76, 289–291. doi:10.1007/BF03165135

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81, 2340–2361. doi:10.1021/j100540a008

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* 58, 35–55. doi:10.1146/annurev.physchem.58.032806.104637

Goyal, S., Kim, S., Chen, I. S., and Chou, T. (2015). Mechanisms of blood homeostasis: Lineage tracking and a neutral model of cell populations in rhesus macaques. *BMC Biol.* 13, 85. doi:10.1186/s12915-015-0191-8

Grosselin, J., Sii-Felice, K., Payen, E., Chretien, S., Roux, D. T. L., and Leboulch, P. (2013). Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. *Stem Cells* 31, 2162–2171. doi:10.1002/stem.1383

Höfer, T., and Rodewald, H. (2016). Output without input: The lifelong productivity of hematopoietic stem cells. *Curr. Opin. Cell. Biol.* 43, 69–77. doi:10.1016/j.ceb.2016.08.003

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

Kim, S., Kim, N., Presson, A. P., An, D. S., Mao, S. H., Bonifacino, A. C., et al. (2010). High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones. *J. Virology* 84, 11771–11780. doi:10.1128/JVI.01355-10

Klose, M., Florian, M. C., Gerbaulet, A., Geiger, H., and Glauche, I. (2019). Hematopoietic stem cell dynamics are regulated by progenitor demand: Lessons from a quantitative modeling approach. *Stem Cells* 37, 948–957. doi:10.1002/stem.3005

Koelle, S. J., Espinoza, D. A., Wu, C., Xu, J., Lu, R., Li, B., et al. (2017). Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants. *Blood* 129, 1448–1457. doi:10.1182/blood-2016-07-728691

Kolouri, S., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* 34, 43–59. doi:10.1109/MSP.2017.2695801

Lee-Six, H., Friesgaard-Obro, N., Shepherd, S., Grossmann, S., Dawson, K., Belmonte, M., et al. (2018). Population dynamics of normal human blood inferred from somatic mutations. *Nature* 561, 473–478. doi:10.1038/s41586-018-0497-0

Lewkiewicz, S., Chuang, Y. L., and Chou, T. (2019a). A mathematical model of the effects of aging on naive T cell populations and diversity. *Bull. Math. Biol.* 81, 2783–2817. doi:10.1007/s11538-019-00630-z

Lewkiewicz, S., Chuang, Y. L., and Chou, T. (2019b). Dynamics of T cell receptor distributions following acute thymic atrophy and resumption. *Math. Biosci. Eng.* 17, 28–55. doi:10.3934/mbe.2020002

Lyne, A. M., Kent, D. G., Laurenti, E., Cornils, K., Glauche, I., and Perie, L. (2018). A track of the clones: New developments in cellular barcoding. *Exp. Hematol.* 68, 15–20. doi:10.1016/j.exphem.2018.11.005

Marciniak-Czochra, A., Stiehl, T., and Wagner, W. (2009). Modeling of replicative senescence in hematopoietic development. *Aging (Albany NY)* 1, 723–732. doi:10.18632/aging.100072

Mayle, A., Luo, M., Jeong, M., and Goodell, M. A. (2015). Flow cytometry analysis of murine hematopoietic stem cells. *Nature* 518, 542–546.

Mendelson, A., and Frenette, P. S. (2014). Hematopoietic stem cell niche maintenance during homeostasis and regeneration. *Nat. Med.* 20, 833–846. doi:10.1038/nm.3647

Muller-Sieburg, C. E., Sieburg, H. B., Bernitz, J. M., and Cattarossi, G. (2012). Stem cell heterogeneity: Implications for aging and regenerative medicine. *Blood* 119, 3900–3907. doi:10.1182/blood-2011-12-376749

Parmentier, S., Kramer, M., Weller, S., Schuler, U., Ordemann, R., Rall, G., et al. (2020). Reevaluation of reference values for bone marrow differential counts in 236 healthy bone marrow donors. *Ann. Hematol.* 99, 2723–2729. doi:10.1007/s00277-020-04255-4

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statistics* 33, 1065–1076. doi:10.1214/aoms/1177704472

Peixoto, D., Dingli, D., and Pacheco, J. M. (2011). Modelling hematopoiesis in health and disease. *Math. Comput. Model.* 53, 1546–1557. doi:10.1016/j.mcm.2010.04.013

Radtke, S., Colonna, L., Perez, A. M., Hoffman, M., Kean, L. S., and Kiem, H. P. (2020). Isolation of a highly purified HSC-enriched $CD34^{+}CD90^{+}CD45RA^{-}$ cell subset for allogeneic transplantation in the nonhuman primate large-animal model. *Transplant. Direct* 6, e579. doi:10.1097/TXD.0000000000001029

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statistics* 27, 832–837. doi:10.1214/aoms/1177728190

Seita, J., and Weissman, I. L. (2010). Hematopoietic stem cell: Self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 640–653. doi:10.1002/wsbm.86

Shepherd, B. E., Kiem, H. P., Lansdorp, P. M., Dunbar, C. E., Aubert, G., LaRochelle, A., et al. (2007). Hematopoietic stem-cell behavior in nonhuman primates. *Blood* 110, 1806–1813. doi:10.1182/blood-2007-02-075382

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Stiehl, T., and Marciniak-Czochra, A. (2011). Characterization of stem cells using mathematical models of multistage cell lineages. *Math. Comput. Model.* 53, 1505–1517. doi:10.1016/j.mcm.2010.03.057

Sun, J., Ramos, A., Chapman, B., Johnnidis, J. B., Le, L., Ho, Y. J., et al. (2014). Clonal dynamics of native haematopoiesis. *Nature* 514, 322–327. doi:10.1038/nature13824

Sun, Z., and Komarova, N. (2012). Stochastic modeling of stem-cell dynamics with control. *Math. Biosci.* 240, 231–240. doi:10.1016/j.mbs.2012.08.004

Székely, T., Burrage, K., Mangel, M., and Bonsall, M. (2014). Stochastic dynamics of interacting haematopoietic stem cell niche lineages. *PLoS Comput. Biol.* 10, e1003794. doi:10.1371/journal.pcbi.1003794

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

Villani, C. (2009). *Optimal transport, old and new*. Berlin Heidelberg: Springer-Verlag.

Wu, C., Li, B., Lu, R., Koelle, S. J., Yang, Y., Jares, A., et al. (2014). Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell. Stem Cell.* 14, 486–499. doi:10.1016/j.stem.2014.01.020

Xu, J., Wang, Y., Guttorp, P., and Abkowitz, J. L. (2018a). Visualizing hematopoiesis as a stochastic process. *Blood Adv.* 1, 2637–2645. doi:10.1182/bloodadvances.2018023705

Xu, S., Böttcher, L., and Chou, T. (2020). Diversity in biology: Definitions, quantification and models. *Phys. Biol.* 17, 031001. doi:10.1088/1478-3975/ab6754

Xu, S., Kim, S., Chen, I. S. Y., and Chou, T. (2018b). Modeling large fluctuations of thousands of clones during hematopoiesis: The role of stem cell self-renewal and bursty progenitor dynamics in rhesus macaque. *PLoS Comput. Biol.* 14, e1006489. doi:10.1371/journal.pcbi.1006489

Yeh, Y. H. J., Yang, K., Razmi, A., and Ho, Y. C. (2021). The clonal expansion dynamics of the HIV-1 reservoir: Mechanisms of integration site-dependent proliferation and HIV-1 persistence. *Viruses* 13, 1858. doi:10.3390/v13091858

Yuan, C., and Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J* 2, 226–235. doi:10.3390/j2020016