Goyal *et al. BMC Biology* (2015) 13:85
DOI 10.1186/s12915-015-0191-8

BMC Biology

RESEARCH ARTICLE Open Access

Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques

Sidhartha Goyal[1†], Sanggu Kim[2†], Irvin SY Chen[2,3] and Tom Chou[4*]

Abstract

Background: How a potentially diverse population of hematopoietic stem cells (HSCs) differentiates and proliferates to supply more than 10^{11} mature blood cells every day in humans remains a key biological question. We investigated this process by quantitatively analyzing the *clonal* structure of peripheral blood that is generated by a population of transplanted lentivirus-marked HSCs in myeloablated rhesus macaques. Each transplanted HSC generates a clonal lineage of cells in the peripheral blood that is then detected and quantified through deep sequencing of the viral vector integration sites (VIS) common within each lineage. This approach allowed us to observe, over a period of 4-12 years, hundreds of distinct clonal lineages.

Results: While the distinct clone sizes varied by three orders of magnitude, we found that collectively, they form a steady-state clone size-distribution with a distinctive shape. Steady-state solutions of our model show that the predicted clone size-distribution is sensitive to only two combinations of parameters. By fitting the measured clone size-distributions to our mechanistic model, we estimate both the effective HSC differentiation rate and the number of active HSCs.

Conclusions: Our concise mathematical model shows how slow HSC differentiation followed by fast progenitor growth can be responsible for the observed broad clone size-distribution. Although all cells are assumed to be statistically identical, analogous to a neutral theory for the different clone lineages, our mathematical approach captures the intrinsic variability in the times to HSC differentiation after transplantation.

Keywords: Hematopoiesis, Stem cell clones, Lineage tracking, Mathematical modeling

Background

Around 10^{11} new mature blood cells are generated in a human every day. Each mature blood cell comes from a unique hematopoietic stem cell (HSC). Each HSC, however, has tremendous proliferative potential and contributes a large number and variety of mature blood cells for a significant fraction of an animal's life. Traditionally, HSCs have been viewed as a homogeneous cell population, with each cell possessing equal and unlimited proliferative potential. In other words, the fate of each HSC (to differentiate or replicate) would be determined by its intrinsic stochastic activation and signals from its microenvironment [1, 2].

However, as first shown in Muller-Sieburg et al. [3], singly transplanted murine HSCs differ significantly in their long-term lineage (cell-type) output and in their proliferation and differentiation rates [4–7]. Similar findings have been found from examining human embryonic stem cells and HSCs in vitro [8, 9]. While cell-level knowledge of HSCs is essential, it does not immediately provide insight into the question of blood homeostasis at the animal level. More concretely, analysis of single-cell transplants does not apply to human bone marrow transplants, which involve millions of CD34-expressing primitive hematopoietic and committed progenitor cells. Polyclonal blood regeneration from such hematopoietic stem and progenitor cell (HSPC) pools is more complex

*Correspondence: tomchou@ucla.edu
†Equal contributors
[4]Departments of Biomathematics and Mathematics, UCLA, Los Angeles, USA
Full list of author information is available at the end of the article

and requires regulation at both the individual cell and system levels to achieve stable [10, 11] or dynamic [12] homeostasis.

To dissect how a population of HSPCs supplies blood, several high-throughput assay systems that can quantitatively track repopulation from an individual stem cell have been developed [6, 11, 13, 14]. In the experiment analyzed in this study, as outlined in Fig. 1, each individual CD34+ HSPC is distinctly labeled by the random incorporation of a lentiviral vector in the host genome before transplantation into an animal. All cells that result from proliferation and differentiation of a distinctly marked HSPC will carry identical markings defined by the location of the original viral vector integration site (VIS). By sampling nucleated blood cells and enumerating their unique VISs, one can quantify the cells that arise from a single HSPC marked with a viral vector. Such studies in humans [15] have revealed highly complex polyclonal repopulation that is supported by tens of thousands of different clones [15–18]; a clone is defined as a population of cells of the same lineage, identified here by a unique VIS. These lineages, or clones, can be distributed across all cell types that may be progeny of the originally transplanted HSC after it undergoes proliferation and differentiation. However, the number of cells of any VIS lineage across certain cell types may be different. By comparing abundances of lineages across blood cells of different types, for example, one may be able to determine the heterogeneity or bias of the HSC population or if HSCs often switch their output. This type of analysis remains particularly difficult in human studies since transplants are performed under diseased settings and followed for only 1 or 2 years.

We analyze here a systematic clone-tracking study that used a large number of HSPC clones in a transplant and competitive repopulation setting comparable to that used in humans [19]. In these nonhuman primate rhesus macaque experiments, lentiviral vector-marked clones were followed for up to a decade post-transplantation (equivalent to about 30 years in humans if extrapolated by average life span). All data are available in the supplementary information files of Kim et al. [19]. This long-term study allows one to distinguish clearly HSC clones from other short-term progenitor clones that were included in the initial pool of transplanted CD34+ cells. Hundreds to thousands of detected clones participated in repopulating the blood in a complex yet highly structured fashion. Preliminary examination of some of the clone populations suggests waves of repopulation with short-lived clones that first grow then vanish within the first 1 or 2 years, depending on the animal [19].

Subsequent waves of HSC clones appear to rise and fall sequentially over the next 4–12 years. This picture is consistent with recent observations in a transplant-free transposon-based tagging study in mice [20] and in human gene therapy [15, 16]. Therefore, the dynamics of a clonally tracked nonhuman primate HSPC repopulation provides rich data that can inform our understanding of regulation, stability, HSPC heterogeneity, and possibly HSPC aging in hematopoiesis.

Although the time-dependent data from clonal repopulation studies are rich in structure, in this study we focus on one specific aspect of the data: the number of clones that are of a certain abundance as described in Fig. 2. Rather than modeling the highly dynamic populations

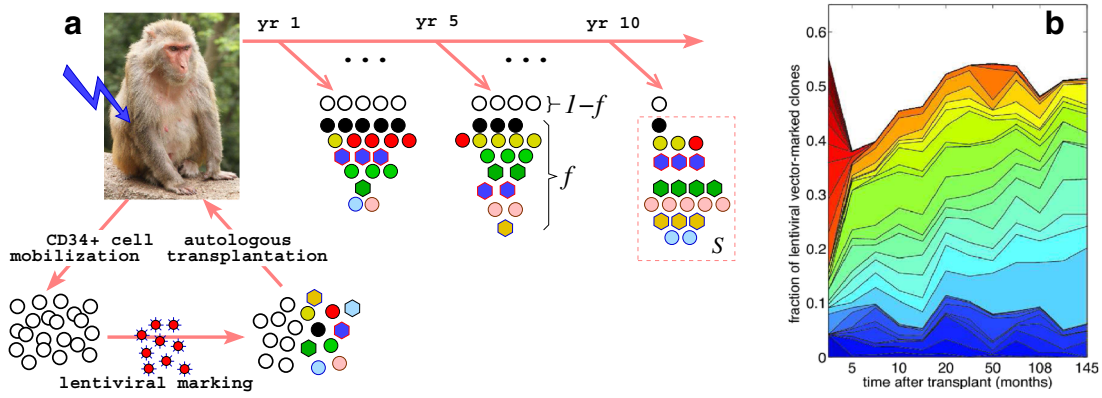

Fig. 1 Probing hematopoietic stem and progenitor cell (HSPC) biology through polyclonal analysis. **a** Mobilized CD34+ bone marrow cells from rhesus macaques are first marked individually with lentiviral vectors and transplanted back into the animal after nonlethal myeloablative irradiation [19]. Depending on the animal, 30–160 million CD34+ cells were transplanted, with a fraction ~0.07–0.3 of them being lentivirus-marked. The clonal contribution of vector-marked HSPCs is measured from blood samples periodically drawn over a dozen years [19]. An average fraction f ~0.03–0.1 of the sampled granulocytes and lymphocytes in the peripheral blood was found to be marked. This fraction is smaller than the fraction of marked CD34+ cells due probably to repopulation by surviving unmarked stem cells in the marrow after myeloablative conditioning. Within any post-transplant sample, $S = 1342–44{,}415$ (average 10,026) viral integration sites of the marked cells were sequenced (see [14, 19] for details). **b** The fraction of all sequenced VIS reads belonging to each clone is shown by the thickness of the slivers. Small clones are not explicitly shown

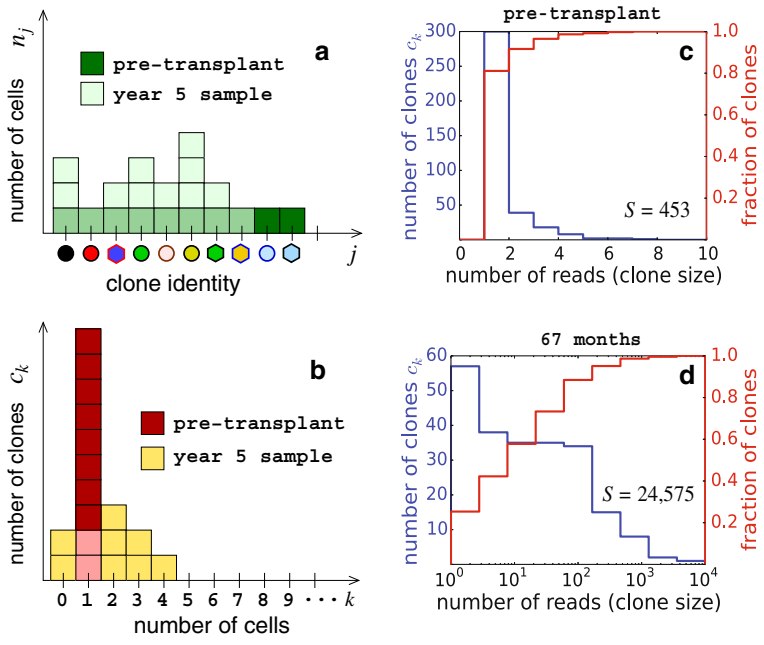

Fig. 2 Quantification of marked clones. **a** Assuming each transplanted stem cell is uniquely marked, the initial number of CD34+ cells representing each clone is one. **b** The pre-transplant clone size distribution is thus defined by the total number of transplanted CD34+ cells and is peaked at one cell. Post-transplant proliferation and differentiation of the HSC clones result in a significantly broader clone size distribution in the peripheral blood. The number of differentiated cells for each clone and the number of clones represented by exactly *k* cells, 5 years' post-transplantation (corresponding to Fig. 1a), are overlaid in (**a**) and (**b**) respectively. **c** Clone size distribution (*blue*) and the cumulative normalized clone size distribution (*red*) of the pre-transplant CD34+ population. **d** After transplantation, clone size distributions in the transit-amplifying (TA) and differentiated peripheral cell pools broaden significantly (with clones ranging over four decades in size) but reach a steady state. The corresponding cumulative normalized distribution is less steep

of each clone, our aim here is to develop first a more global understanding of how the total number of clones represented by specific numbers of cells arises within a mechanistically reasonable model of hematopoiesis. The size distributions of clones present in the blood sampled from different animals at different times are characterized by specific shapes, with the largest clones being a factor of 100–1000 times more abundant than the most rarely detected clones. Significantly, our analysis of renormalized data indicates that the clone size distribution (measuring the number of distinct lineages that are of a certain size) reaches a stationary state as soon as a few months after transplantation (see Fig. 4 below). To reconcile the observed stationarity of the clone size distributions with the large diversity of clonal contributions in the context of HSPC-mediated blood repopulation, we developed a mathematical model that treats three distinct cell populations: HSCs, transit-amplifying progenitor cells, and fully differentiated nucleated blood cells (Fig. 3). While multistage models for a detailed description of differentiation have been developed [21], we lump different stages of cell types within the transit-amplifying progenitor pool into one population, avoiding excess numbers of unmeasurable parameters. Another important feature of our model is the overall effect of feedback and regulation, which we incorporate via a population-dependent cell proliferation rate for progenitor cells.

The effective proliferation rate will be modeled using a Hill-type suppression that is defined by the limited space for progenitor cells in the bone marrow. Such a regulation term has been used in models of cyclic neutropenia [22] but has not been explicitly treated in models of clone propagation in hematopoiesis. Our mathematical model is described in greater detail in the next section and in Additional file 1.

Our model shows that both the large variability and the characteristic shape of the clone size distribution can result from a slow HSC-to-progenitor differentiation followed by a burst of progenitor growth, both of which are generic features of hematopoietic systems across different organisms. By assuming a homogeneous HSC population and fitting solutions of our model to available data, we show that randomness from stochastic activation and proliferation and a global carrying capacity are sufficient to describe the observed clonal structure. We estimate that only a few thousand HSCs may be actively contributing toward blood regeneration at any time. Our model can be readily generalized to include the role of

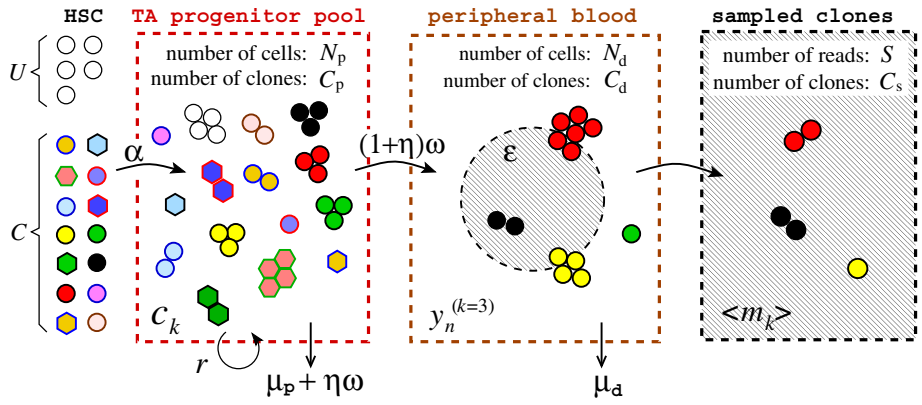

Fig. 3 Schematic of our mathematical model. Of the $\sim 10^6$–10^7 CD34+ cells in the animal immediately after transplantation, C active HSCs are distinctly labeled through lentiviral vector integration. U HSCs are unlabeled because they were not mobilized, escaped lentiviral marking, or survived ablation. All HSCs asymmetrically divide to produce progenitor cells, which in turn replicate with an effective carrying capacity-limited rate r. Transit-amplifying progenitor cells die with rate μ_p or terminally differentiate with rate ω. The terminal differentiation of the progenitor cells occurs symmetrically with probability η or asymmetrically with probability $1-\eta$. This results in a combined progenitor-cell removal rate $\mu = \mu_p + \eta\omega$. The differentiated cells outside the bone marrow are assumed not to be subject to direct regulation but undergo turnover with a rate μ_d. The mean total numbers of cells in the progenitor and differentiated populations are denoted N_p and N_d, respectively. Finally, a small fraction $\varepsilon \ll 1$ of differentiated cells is sampled, sequenced, and found to be marked. In this example, $S = \varepsilon N_d = 5$. Because some clones may be lost as cells successively progress from one pool to the next, the total number of clones in each pool must obey $C \geq C_p \geq C_d \geq C_s$. Analytic expressions for the expected total number of clones in each subsequent pool are derived in Additional file 1. *HSC* hematopoietic stem cell, *TA* transit-amplifying

heterogeneity and aging in the transplanted HSCs and provides a framework for quantitatively studying physiological perturbations and genetic modifications of the hematopoietic system.

Mathematical Model

Our mathematical model explicitly describes three subpopulations of cells: HSCs, transit-amplifying progenitor cells, and terminally differentiated blood cells (see Fig. 3). We will not distinguish between myeloid or lymphoid lineages but will use our model to analyze clone size distribution data for granulocytes and peripheral blood mononuclear cells independently. Our goal will be to describe how clonal lineages, started from distinguishable HSCs, propagate through the amplification and terminal differentiation processes.

Often clone populations are modeled directly by dynamical equations for $n_j(t)$, the number of cells of a particular clone j identified by its specific VIS [23]. Since all cells are identical except for their lentiviral marking, mean-field rate equations for $n_j(t)$ are identical for all j. Assuming identical initial conditions (one copy of each clone), the expected populations $n_j(t)$ would be identical across all clones j. This is a consequence of using identical growth and differentiation *rates* to describe the evolution of the *mean* number of cells of each clone.

Therefore, for cells in any specific pool, rather than deriving equations for the mean number n_j of *cells* of each distinct clone j (Fig. 2a), we perform a hodograph transformation [24] and formulate the problem in terms of the number of *clones* that are represented by k cells, $c_k = \sum_j \delta_{k,n_j}$ (see Fig. 2b), where the Kronecker δ function $\delta_{k,n_j} = 1$ only when $k = n_j$ and is 0 otherwise. This counting scheme is commonly used in the study of cluster dynamics in nucleation [25] and in other related models describing the dynamics of distributions of cell populations. By tracking the number of clones of different sizes, the intrinsic stochasticity in the *times* of cell division (especially that of the first differentiation event) and the subsequent variability in the clone abundances are quantified. Figure 2a, b qualitatively illustrates n_j and c_k, pre-transplant and after 5 years, corresponding to the scenario depicted in Fig. 1a. Cells in each of the three pools are depicted in Fig. 3, with different clones grouped according to the number of cells representing each clone.

The first pool (the progenitor-cell pool) is fed by HSCs through differentiation. Regulation of HSC differentiation fate is known to be important for efficient repopulation [26, 27] and control [28] and the balance between asymmetric and symmetric differentiation of HSCs has been studied at the microscopic and stochastic levels [29–32]. However, since HSCs have life spans comparable to that of an animal, we reasoned that the total number of HSCs changes only very slowly after the initial few-month transient after transplant. For simplicity, we will assume, consistent with estimates from measurements [33], that HSCs divide only asymmetrically. Therefore, upon differentiation, each HSC produces one partially differentiated progenitor cell and one replacement HSC. How symmetric HSC division might affect the resulting clone sizes is

discussed in Additional file 1 through a specific model of HSC renewal in a finite-sized HSC niche. We find that the incorporation of symmetric division has only a small quantitative effect on the clone size distribution that we measure and ultimately analyze.

Next, consider the progenitor-cell pool. From Fig. 3, we can count the number of clones c_k represented by exactly k cells. For example, the black, red, green, and yellow clones are each represented by three cells, so $c_3 = 4$. Each progenitor cell can further differentiate with rate ω into a terminally differentiated cell. If progenitor cells undergo symmetric differentiation with probability η and asymmetric differentiation with probability $1-\eta$, the effective rate of differentiation is $2\eta\omega + (1-\eta)\omega = (1+\eta)\omega$. In turn, fully differentiated blood cells (not all shown in Fig. 3) are cleared from the peripheral pool at rate μ_{d}, providing a turnover mechanism. Finally, each measurement is a small-volume sample drawn from the peripheral blood pool, as shown in the final panel in Fig. 3.

Note that the transplanted CD34+ cells contain both true HSCs and progenitor cells. However, we assume that at long times, specific clones derived from progenitor cells die out and that only HSCs contribute to long-lived clones. Since we measure the number of clones of a certain size rather than the dynamics of individual clone numbers, transplanted progenitor cells should not dramatically affect the steady-state clone size distribution. Therefore, we will ignore transplanted progenitor cells and assume that after transplantation, effectively only U unlabeled HSCs and C labeled (lentivirus-marked) HSCs are present in the bone marrow and actively asymmetrically differentiating (Fig. 3). Mass-action equations for the expected number of clones c_k of size k are derived from considering simple birth and death processes with immigration (HSC differentiation):

$$\frac{\mathrm{d}c_k}{\mathrm{d}t} = \underbrace{\alpha\left[c_{k-1}-c_k\right]}_{\text{HSC differentiation}} + \underbrace{r\left[(k-1)c_{k-1}-kc_k\right]}_{\text{progenitor birth}} + \underbrace{\mu\left[(k+1)c_{k+1}-kc_k\right]}_{\text{progenitor death}}, \tag{1}$$

where $k = 1, 2, \ldots, C$ and $c_0(t) \equiv C - \sum_{k=1}^{\infty} c_k(t)$ is the number of clones that are not represented in the progenitor pool. Since C is large, and the number of clones that are of size comparable to C is negligible, we will approximate $C \to \infty$ in our mathematical derivations. We have suppressed the time dependence of $c_k(t)$ for notational simplicity. The constant parameter α is the asymmetric differentiation rate of all HSCs, while r and μ are the proliferation and overall clearance rates of progenitor cells. In our model, HSC differentiation events that feed the progenitor pool are implicitly a rate-α Poisson process. The appreciable number of detectable clones (Fig. 1b) implies the initial number C of HSC clones is large enough that asymmetric differentiation of individual HSCs is uncorrelated. The alternative scenario of a few HSCs undergoing synchronized differentiation would not lead to appreciably different results since the resulting distribution c_k is more sensitive to the progenitor cells' *unsynchronized* replication and death than to the statistics of the immigration by HSC differentiation.

The final differentiation from progenitor cell to peripheral blood cell can occur through symmetric or asymmetric differentiation, with probabilities η and $1-\eta$, respectively. If parent progenitor cells are unaffected after asymmetric terminal differentiation (i.e., they die at the normal rate μ_{p}), the dynamics are feed-forward and the progenitor population is not influenced by terminal differentiation. Under symmetric differentiation, a net loss of one progenitor cell occurs. Thus, the overall progenitor-cell clearance rate can be decomposed as $\mu = \mu_{\mathrm{p}} + \eta\omega$. We retain the factor η in our equations for modeling pedagogy, although in the end it is subsumed in effective parameters and cannot be independently estimated from our data.

The first term in Eq. 1 corresponds to asymmetric differentiation of each of the C active clones, of which c_k are of those lineages with population k already represented in the progenitor pool. Differentiation of this subset of clones will add another cell to these specific lineages, reducing c_k. Similarly, differentiation of HSCs in lineages that are represented by $k-1$ progenitor cells adds cells to these lineages and increases c_k. Note that Eq. 1 are mean-field rate equations describing the evolution of the expected number of clones of size k. Nonetheless, they capture the intrinsic dispersion in lineage sizes that make up the clone size distribution. While all cells are assumed to be statistically identical, with equal rates α, p, and μ, Eq. 1 directly model the evolution of the *distribution* $c_k(t)$ that arises ultimately from the distribution of times for each HSC to differentiate or for the progenitor cells to replicate or die. Similar equations have been used to model the evolving distribution of virus capsid sizes [34].

Since the equations for $c_k(t)$ describe the evolution of a distribution, they are sometimes described as master equations for the underlying process [34, 35]. Here we note that the solution to Eq. 1, $c_k(t)$, is the *expected* distribution of clone sizes. Another level of stochasticity could be used to describe the evolution of a *probability distribution* $P_b(\boldsymbol{b};t) = P_b(b_0, b_1, \ldots, b_{N_{\mathrm{p}}}; t)$ *over the integer numbers* b_k. This density represents the *probability* that at time t, there are b_0 unrepresented lineages, b_1 lineages represented by one cell in the progenitor pool, b_2 lineages represented by two cells in the progenitor pool, and so on. Such a probability distribution would obey an N_{p}-dimensional master equation rather than a one-dimensional equation, like Eq. 1, and once known, can

be used to compute the mean $c_k(t) = \sum_{\mathbf{b}} b_k P(\mathbf{b};t)$. To consider the entire problem stochastically, the variability described by probability distribution P_b would have to be propagated forward to the differentiated cell pool as well. Given the modest number of measured data sets and the large numbers of lineages that are detectable in each, we did not attempt to use the data as samples of the distribution P_b and directly model the mean values c_k instead. Variability from both intrinsic stochasticity and sampling will be discussed in Additional file 1.

After defining $u(t)$ as the number of unlabeled cells in the progenitor pool, and $N_\mathrm{p}(t) = u(t)+\sum_{k=1}^{\infty} kc_k(t)$ as the total number of progenitor cells, we find $\dot{u} = (r-\mu)u+\alpha U$ and

$$\frac{\mathrm{d}N_\mathrm{p}(t)}{\mathrm{d}t} = \alpha\,(U+C) + (r-\mu)\,N_\mathrm{p}(t). \tag{2}$$

Without regulation, the total population $N_\mathrm{p}(t\to\infty)$ will either reach $N_\mathrm{p} \approx \alpha(U+C)/(\mu-r)$ for $\mu > r$ or will exponentially grow without bound for $r > \mu$. Complex regulation terms have been employed in deterministic models of differentiation [28] and in stochastic models of myeloid/lymphoid population balance [36]. For the purpose of estimating macroscopic clone sizes, we assume regulation of cell replication and/or spatial constraints in the bone marrow can be modeled by a simple effective Hill-type growth law [22, 37]:

$$r = r(N_\mathrm{p}) \equiv \frac{pK}{N_\mathrm{p}+K} \tag{3}$$

where p is the intrinsic replication rate of an isolated progenitor cell. We assume that progenitor cells at low density have an overall positive growth rate $p > \mu$. The parameter K is the progenitor-cell population in the bone marrow that corresponds to the half-maximum of the effective growth rate. It can also be interpreted as a limit to the bone marrow size that regulates progenitor-cell proliferation to a value determined by K, p, and μ and is analogous to the carrying capacity in logistic models of growth [38]. For simplicity, we will denote K as the carrying capacity in Eq. 3 as well. Although our data analysis is insensitive to the precise form of regulation used, we chose the Hill-type growth suppression because it avoids negative growth rates that confuse physiological interpretation. An order-of-magnitude estimate of the bone marrow size (or carrying capacity) in the rhesus macaque is $K \sim 10^9$. Ultimately, we are interested in how a limited progenitor pool influences the overall clone size distribution, and a simple, single-parameter (K) approximation to the progenitor-cell growth constraint is sufficient.

Upon substituting the growth law $r(N_\mathrm{p})$ described by Eq. 3 into Eq. 2, the total progenitor-cell population $N_\mathrm{p}(t\to\infty)$ at long times is explicitly shown in Additional file 1: Eq. A19 to approach a finite value that depends strongly on K. Progenitor cells then differentiate to supply peripheral blood at rate $(1+\eta)\omega$ so that the total number of differentiated blood cells obeys

$$\frac{\mathrm{d}N_\mathrm{d}(t)}{\mathrm{d}t} = (1+\eta)\omega N_\mathrm{p} - \mu_\mathrm{d} N_\mathrm{d}. \tag{4}$$

At steady state, the combined peripheral nucleated blood population is estimated to be $N_\mathrm{d} \sim 10^9$–10^{10} [39], setting an estimate of $N_\mathrm{d}/N_\mathrm{p} \approx (1+\eta)\omega/\mu_\mathrm{d} \sim 1$–10. Moreover, as we shall see, the relevant factor in our steady-state analysis will be the numerical *value* of the effective growth rate r, rather than its functional form. Therefore, the chosen form for regulation will not play a role in the mathematical results in this paper except to define parameters (such as K) explicitly in the regulation function itself.

To distinguish and quantify the clonal structure within the peripheral blood pool, we define $y_n^{(k)}$ to be the number of clones that are represented by exactly n cells in the differentiated pool *and* k cells in the progenitor pool. For example, in the peripheral blood pool shown in Fig. 3, $y_1^{(3)} = y_2^{(3)} = y_4^{(3)} = y_6^{(3)} = 1$. This counting of clones across both the progenitor and peripheral blood pools is necessary to balance progenitor-cell differentiation rates with peripheral blood turnover rates. The evolution equations for $y_n^{(k)}$ can be expressed as

$$\frac{\mathrm{d}y_n^{(k)}}{\mathrm{d}t} = (1+\eta)\omega k\left(y_{n-1}^{(k)} - y_n^{(k)}\right) + (n+1)\mu_\mathrm{d} y_{n+1}^{(k)} - n\mu_d y_n^{(k)}, \tag{5}$$

where $y_0^{(k)} \equiv c_k - \sum_{n=1}^{\infty} y_n^{(k)}$ represents the number of progenitor clones of size k that have not yet contributed to peripheral blood. The transfer of clones from the progenitor population to the differentiated pool arises through $y_0^{(k)}$ and is simply a statement that the number of clones in the peripheral blood can increase only by differentiation of a progenitor cell whose lineage has not yet populated the peripheral pool. The first two terms on the right-hand side of Eq. 5 represent immigration of clones represented by $n-1$ and n differentiated cells *conditioned upon* immigration from only those specific clones represented by k cells in the progenitor pool. The overall rate of addition of clones from the progenitor pool is thus $(1+\eta)\omega k$, in which the frequency of terminal differentiation is weighted by the stochastic division factor $(1+\eta)$. By using the Hill-type growth term $r(N_\mathrm{p})$ from Eq. 3, Eq. 1 can be solved to find $c_k(t)$, which in turn can be used in Eq. 5 to find $y_n^{(k)}(t)$. The number of clones in the peripheral blood represented by exactly n differentiated cells is thus $y_n(t) = \sum_{k=1}^{\infty} y_n^{(k)}(t)$.

As we mentioned, Eqs. 1 and 5 describe the evolution of the expected clone size distribution. Since each measurement represents one realization of the distributions $c_k(t)$ and $y_n(t)$, the validity of Eqs. 1 and 5 relies on a sufficiently

large C such that the marked HSCs generate enough lineages and cells to allow the subsequent peripheral blood clone size distribution to be sampled adequately. In other words, measurement-to-measurement variability described by e.g., $\langle c_k(t)c_{k'}(t)\rangle - \langle c_k(t)\rangle\langle c_{k'}(t)\rangle$ is assumed negligible (see Additional file 1). Our modeling approach would not be applicable to studying single HSC transplant studies [4–6] unless the measured clone sizes from multiple experiments are aggregated into a distribution.

Finally, to compare model results with animal blood data, we must consider the final step of sampling small aliquots of the differentiated blood. As derived in Additional file 1: Eq. A11, if S marked cells are drawn and sequenced successfully (from a total differentiated cell population N_d), the expected number of clones $\langle m_k(t)\rangle$ represented by k cells is given by

$$\begin{aligned}\langle m_k(t)\rangle &= F(q,t) - F(q-1,t)\\ &= \sum_{\ell=0}^{\infty} \mathrm{e}^{-\ell\varepsilon}\frac{(\ell\varepsilon)^k}{k!}y_\ell(t),\end{aligned} \tag{6}$$

where $\varepsilon \equiv S/N_\mathrm{d} \ll 1$ and $F(q,t) \equiv \sum_{k=0}^{q}\langle m_k(t)\rangle$ is the sampled, expected cumulative size distribution. Upon further normalization by the total number of detected clones in the sample, $C_\mathrm{s}(t) = F(S,t) - F(0,t)$, we define

$$Q(q,t) \equiv \frac{F(q,t) - F(0,t)}{F(S,t) - F(0,t)} \tag{7}$$

as the fraction of the total number of sampled clones that are represented by q or fewer cells. Since the data represented in terms of Q will be seen to be time-independent, explicit expressions for $c_k, y_n^{(k)}$, $\langle m_k\rangle$, and $Q(q)$ can be derived. Summarizing, the main features and assumptions used in our modeling include:

- A neutral-model framework [40] that directly describes the distribution of clone sizes in each of the three cell pools: progenitor cells, peripheral blood cells, and sampled blood cells. The cells in each pool are statistically identical.
- A constant asymmetric HSC differentiation rate α. The appreciable numbers of unsynchronized HSCs allow the assumption of Poisson-distributed differentiation times of the HSC population. The level of differentiation symmetry is found to have little effect on the steady-state clone size distribution (see Additional file 1). The symmetry of the terminal differentiation step is also irrelevant for understanding the available data.
- A simple one-parameter (K) growth regulation model that qualitatively describes the finite maximum size of the progenitor population in the bone marrow. Ultimately, the specific form for the regulation is unimportant since only the steady-state value of the growth parameter r affects the parameter fitting.

Using only these reasonable model features, we are able to compute clone size distributions and compare them with data. An explicit form for the expected steady-state clone size distribution $\langle m_k\rangle$ is given in Additional file 1: Eq. A32, and the parameters and variables used in our analysis are listed in Table 1.

Results and discussion

In this section, we describe how previously published data (the number of cells of each detected clone in a sample of the peripheral blood, which are available in the supplementary information files of Kim et al. [19]) are used to constrain parameter values in our model. We emphasize that our model is structurally different from models used to track lineages and clone size distributions in retinal and epithelial tissues [41, 42]. Rather than tracking only the lineages of stem cells (which are allowed to undergo asymmetric differentiation, symmetric differentiation, or symmetric replication), our model assumes a highly proliferative population constrained by a carrying capacity K and slowly fed at rate α by an asymmetrically dividing HSC pool of C fixed clones. We have also included terminal differentiation into peripheral blood and the effects of sampling on the expected clone size distribution. These ingredients yield a clone size distribution different from those previously derived [41, 42], as described in more detail below.

Stationarity in time

Clonal contributions of the initially transplanted HSC population have been measured over 4–12 years in four different animals. As depicted in Fig. 4a, populations of individual clones of peripheral blood mononuclear cells from animal RQ5427, as well as all other animals, show significant variation in their dynamics. Since cells of any detectable lineage will number in the millions, this variability in lineage size across time cannot be accounted for by the intrinsic stochasticity of progenitor-cell birth and death. Rather, these rises and falls of lineages likely arise from a complicated regulation of HSC differentiation and lineage aging. However, in our model and analysis, we do not keep track of lineage sizes n_i. Instead, define $Q(\nu)$ as the fraction of clones arising with relative frequency $\nu \equiv fq/S$ or less (here, q is the number of VIS reads of any particular clone in the sample, f is the fraction of all sampled cells that are marked, and S is the total number of sequencing reads of marked cells in a sample). Figure 4b shows data analyzed in this way and reveals that $Q(\nu)$ appears stationary in time.

The observed steady-state clone size distribution is broad, consistent with the mathematical model developed

Table 1 Model parameters and variables. Estimates of steady-state values are provided where available. We assume little prior knowledge on all but a few of the more established parameters. Nonetheless, our modeling and analysis place constraints on combinations of parameters, allowing us to fit data and provide estimates for steady-state values of $U + C \sim 10^3$–10^4 and $\alpha(N_p + K)/(pK) \sim 0.002$–$0.1$

Symbol	Parameter or variable	Estimate	Ref.
α	Single HSC asymmetric differentiation rate	∼0.1–0.3 per month	[46, 51]
p	Free progenitor-cell proliferation rate		
μ_p	Progenitor-cell death rate		
μ_d	Differentiated cell death rate	∼0.01–0.3 per day	[52, 53]
η	Symmetric differentiation probability		
ω	Terminal differentiation rate		
K	Progenitor-cell capacity	$\sim 10^9$	
U	Number of active unmarked HSCs		
C	Number of active viral-marked HSCs		
C_p	Number of progenitor clones		Additional file 1: Eq. A23
C_d	Number of differentiated clones		Additional file 1: Eq. A23
C_s	Number of sampled clones		Additional file 1: Eq. A24
S	Number of sequences read	$\sim 10^3$–10^4	[14]
c_k	Number of progenitor clones of size k		
u	Unlabeled progenitor-cell population		
N_p	Total progenitor-cell population		
N_d	Total differentiated blood population	$\sim 10^9$–10^{10}	
$y_n^{(k)}$	Number of differentiated clones of size n arising from progenitors of size k		

above. The handful of most populated clones constitutes up to 1–5 % of the entire differentiated blood population. These dominant clones are followed by a large number of clones with fewer cells. The smallest clones sampled in our experiment correspond to a single read $q = 1$, which yields a minimum measured frequency $\nu_{\min} = f/S$. A single read may comprise only 10^{-4}–10^{-3} % of all differentiated blood cells. Note that the cumulative distribution $Q(\nu)$ exhibits higher variability at small sizes simply because fewer clones lie below these smaller sizes.

Although engraftment occurs within a few weeks and total blood populations N_p and N_d (and often immune function) re-establish themselves within a few months after successful HSC transplant [43, 44], it is still

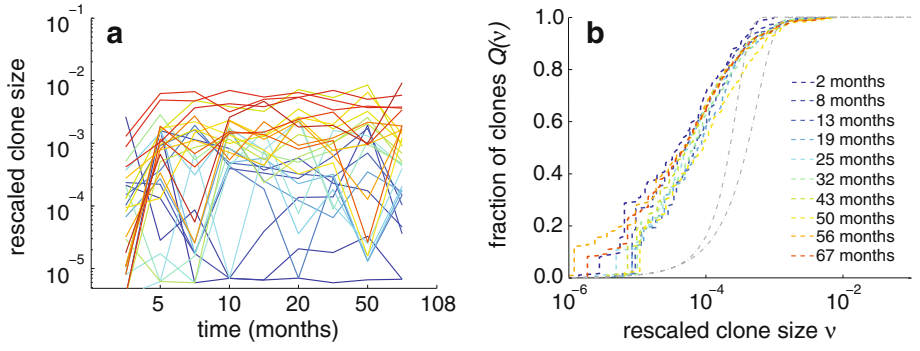

Fig. 4 Rescaled and renormalized data. **a** Individual clone populations (here, peripheral blood mononuclear cells of animal RQ5427) show significant fluctuations in time. For clarity, only clones that reach an appreciable frequency are plotted. **b** The corresponding normalized clone size distributions at each time point are rescaled by the sampled and marked fraction of blood, $\nu = q/S \times f$, where q is the number of reads of a particular clone within the sample. After an initial transient, the fraction of clones (*dashed curves*) as a function of relative size remains stable over many years. For comparison, the *dot-dashed gray curves* represent binomial distributions (with $S = 10^3$ and 10^4 and equivalent mean clone sizes) and underestimate low population clones

surprising that the clone size distribution is relatively static within each animal (see Additional file 1 for other animals). Given the observed stationarity, we will use the steady-state results of our mathematical model (explicitly derived in Additional file 1) for fitting data from each animal.

Implications and model predictions

By using the exact steady-state solution for c_k (Additional file 1: Eq. A21) in Additional file 1: Eq. A18, we can explicitly evaluate the expected clone size distribution $\langle m_k \rangle$ using Eq. 6, and the expected cumulative clone fraction $Q(q)$ using Eq. 7. In the steady state, the clone size distribution of progenitor cells can also be approximated by a gamma distribution with parameters $a \equiv \alpha/r$ and $\bar{r} \equiv r/\mu$: $c_k \sim \bar{r}^k k^{-1+a}$ (see Additional file 1: Eq. A27). In realistic steady-state scenarios near carrying capacity, $r = r(N_\mathrm{p}) \lesssim \mu$, as calculated explicitly in Additional file 1: Eq. A20. By defining $\bar{r} = r/\mu = 1 - \delta$, we find that δ is inversely proportional to the carrying capacity:

$$\delta \approx \frac{\alpha}{\mu}\frac{\mu}{p-\mu}\frac{U+C}{K} \ll 1. \tag{8}$$

The dependencies of $\langle m_q \rangle$ on δ and $a = \alpha/r$ are displayed in Fig. 5a, in which we have defined $w \equiv (1+\eta)\omega/\mu_\mathrm{d}$.

Although our equations form a mean-field model for the expected number of measured clones of any given size, randomness resulting from the stochastic differentiation times of individual HSCs (all with the same rate α) is taken into account.

This is shown in Additional file 1: Eqs. A36–A39, where we explicitly consider the fully stochastic population of a single progenitor clone that results from the differentiation of a single HSC. Since independent unsynchronized HSCs differentiate at times that are exponentially distributed (with rate α), we construct the expected clone size distribution from the birth–death–immigration process [45] to find a result equivalent to that derived from our original model (Eq. 1 and Additional file 1: Eq. A21). Thus, we conclude that if $a = \alpha/r$ is small, the shape of the expected clone size distribution is mainly determined at short times by the initial repopulation of the progenitor-cell pool.

Our model also suggests that the expected number of sampled clones relative to the number of active transplanted clones (see Additional file 1: Eq. A24) can be expressed as:

$$\begin{aligned}\frac{C_\mathrm{s}}{C} &\approx \left[1 - \left(\frac{\delta}{1-(1-\delta)e^{-\varepsilon w}}\right)^a\right] \\ &\approx \frac{\alpha}{r}\ln\left(\frac{\varepsilon w}{\delta}+1\right),\end{aligned} \tag{9}$$

where the last approximation is accurate for $\varepsilon w \ll 1$ and $C_\mathrm{s}/C \ll 1$. The clonal diversity one expects to measure in the peripheral blood sample is sensitive to the combination of biologically relevant parameters and rates δ and $a = \alpha/r$. Figure 5b shows the explicit dependence of the fraction of active clones on a and the combination of parameters defining δ, for $\varepsilon w = \varepsilon(1+\eta)\omega/\mu_\mathrm{d} = 5\times10^{-5}$.

Our analysis shows how scaled measurable quantities such as C_s/C and $C^{-1}\langle m_q \rangle$ depend on just a few combinations of experimental and biological parameters. This small domain of parameter sensitivity reduces the number of parameters that can be independently

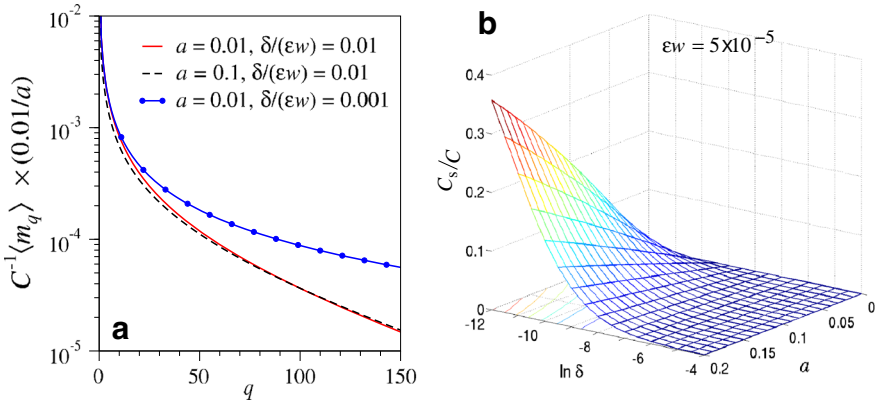

Fig. 5 Clone size distributions and total number of sampled clones. **a** Expected clone size distributions $C^{-1}\langle m_q \rangle$ derived from the approximation in Additional file 1: Eq. A32 are plotted for various a and $\delta/(\varepsilon w)$ [where $w \equiv (1+\eta)\omega/\mu_\mathrm{d}$]. The nearly coincident *solid* and *dashed curves* indicate that variations in a mainly scale the distribution by a multiplicative factor. In contrast, the combination $\delta/(\varepsilon w)$ controls the weighting at large clone sizes through the population cut-off imposed by the carrying capacity. Of the two controlling parameters, the steady-state clone size distribution is most sensitive to $R \cong \delta/(\varepsilon w)$. The dependence of data-fitting on these two parameters is derived in Additional file 1 and discussed in the next section. **b** For $\varepsilon w = 5 \times 10^{-5}$, the expected fraction C_s/C of active clones sampled as a function of $\ln\delta$ and α. The expected number of clones sampled increases with carrying capacity K, HSC differentiation rate $a = \alpha/r$, and the combined sampling and terminal differentiation rate εw

extracted from clone size distribution data. For example, the mode of terminal differentiation described by η clearly cannot be extracted from clonal tracking measurements. Similarly, models that are more detailed of the complex regulation processes would introduce additional parameters that are not resolved by these experiments. Nonetheless, we shall fit our data and known information contained in the experimental protocol to our model to estimate biologically relevant parameters, such as the total number of activated HSCs $U + C$, and thus indirectly C.

Model fitting

Our mathematical model for $\langle m_k \rangle$ (and $F(q)$ and $Q(q)$) includes numerous parameters associated with the processes of HSC differentiation, progenitor-cell amplification, progenitor-cell differentiation, peripheral blood turnover, and sampling. Data fitting is performed using clone size distributions derived separately from the read counts from both the left and right ends of each VIS (see [14] for details on sequencing). Even though we fit our data to $\langle m_k \rangle$ using three independent parameters, $a = \alpha/r$, $\bar{r} = r/\mu$, and εw, we found that within the relevant physiological regime, all clone distributions calculated from our model are most sensitive to just two combinations of parameters (see Additional file 1 for an explicit derivation):

$$a \equiv \frac{\alpha}{r} \quad \text{and} \quad R \equiv \frac{\varepsilon w}{\ln(1/\bar{r})} \approx \frac{\varepsilon w}{\delta} = \frac{(1+\eta)\omega S}{N_{\mathrm{d}}\mu_{\mathrm{d}}\delta}, \tag{10}$$

where the last approximation for R is valid when $1 - \bar{r} = \delta \ll 1$. While the fits are rather insensitive to εw this parameter can fortunately be approximated from estimates of S and the typical turnover rate of differentiated blood. Consequently, we find two maximum likelihood estimates (MLEs) for a and R at each time point. It is important to note that fitting our model to steady-state clone size distributions does not determine all of the physiological parameters arising in our equations. Rather, they provide only two constraints that allow one to relate their values.

For ease of presentation, henceforth we will show all data and comparisons with our model equations in terms of the fraction $Q(\nu)$ or $Q(q)$ (Figs. 4b and 6a, b). Figure 6a, b shows MLE fitting to the raw data $\langle m_k \rangle$ plotted in terms of the normalized but unrescaled data $Q(q)$ for two different peripheral blood cell types from two animals (RQ5427 and RQ3570). Data from all other animals are shown and fitted in Additional file 1, along with overall

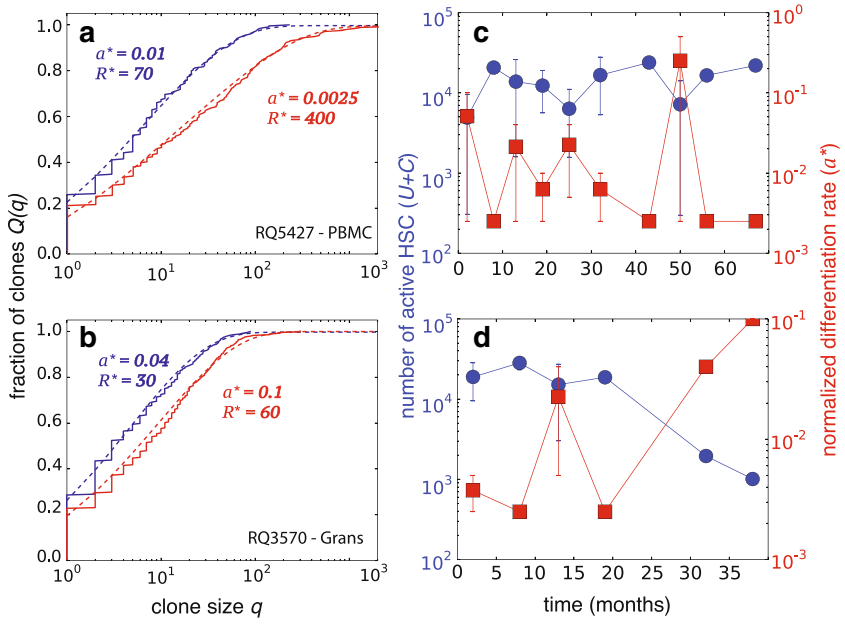

Fig. 6 Data fitting. **a** Fitting raw (not rescaled, as shown in Figure 4) clone size distribution data to $\langle m_k \rangle$ from Eq. 6 at two time points for animal RQ5427. The maximum likelihood estimates (MLEs) are ($a^* \approx 0.01$, $R^* \approx 70$) and ($a^* \approx 0.0025$, $R^* \approx 400$) for data taken at 32 (*blue*) and 67 (*red*) months post-transplant, respectively. Note that the MLE values for different samples vary primarily due to different values of S (and hence ε) used in each measurement. **b** For animal RQ3570, the clone fractions at 32 (*blue*) and 38 (*red*) months yield ($a^* \approx 0.04$, $R^* \approx 30$) and ($a^* \approx 0.1$, $R^* \approx 60$), respectively. For clarity, we show the data and fitted models in terms of $Q(q)$. **c** Estimated number of HSCs $U + C$ (*circles*) and normalized differentiation rate a (*squares*) for animal RQ5427. **d** $U + C$ and a for animal RQ3570. Note the temporal variability (but also long-term stability) in the estimated number of contributing HSCs. Additional details and fits for other animals are qualitatively similar and given in Additional file 1. *HSC* hematopoietic stem cell, *PBMC*, peripheral blood mononuclear cell Grans, granulocytes

goodness-of-fit metrics. Raw cell count data are given in Kim et al. [19].

HSC asymmetric differentiation rate

The MLE for $a = \alpha/r$, a^*, was typically in the range 10^{-2}–10^{-1}. Given realistic parameter values, this quantity mostly provides an estimate of the HSC relative differentiation rate $a^* \sim \alpha/(\mu_p + \eta\omega)$. The smallness of a^* indicates slow HSC differentiation relative to the progenitor turnover rate μ_p and the final differentiation rate $\eta\omega$, consistent with the dominant role of progenitor cells in populating the total blood tissue. Note that besides the intrinsic insensitivity to εw, the goodness-of-fit is also somewhat insensitive to small values of a^* due to the weak dependence of $c_k \sim 1/k^{1-a}$ on a (see Additional file 1). The normalized relative differentiation rates estimated from two animals are shown by the squares (right axis) in Fig. 6c, d.

Number of HSCs

The stability of blood repopulation kinetics is also reflected in the number of estimated HSCs that contribute to blood (shown in Fig. 6c, d). The total number of HSCs is estimated by expressing $U + C$ in terms of the effective parameters, R and a, which in turn are functions of microscopic parameters ($\alpha, p, \mu_p, \mu_d, w$, and K) that cannot be directly measured. In the limit of small sample size, $S \ll R^*K$, however, we find $U + C \approx S/(R^*a^*)$ (see Additional file 1), which can then be estimated using the MLEs a^* and R^* obtained by fitting the clone size distributions. The corresponding values of $U + C$ for two animals are shown by the circles (left axis) in Fig. 6c, d. Although variability in the MLEs exists, the fluctuations appear stationary over the course of the experiment for each animal (see Additional file 1).

Conclusions

Our clonal tracking analysis revealed that individual clones of HSCs contributed differently to the final differentiated blood pool in rhesus macaques, consistent with mouse and human data. Carefully replotting the raw data (clone sizes) in terms of the normalized, rescaled cumulative clone size distribution (the fraction of all detected clones that are of a certain size or less) shows that these distributions reach steady state a few months after transplantation. Our results carry important implications for stem cell biology. Maintaining homeostasis of the blood is a critical function for an organism. Following a myeloablative stem cell transplant, the hematopoietic system must repopulate rapidly to ensure the survival of the host. Not only do individual clones rise and fall temporally, as previously shown [19], but as any individual clone of a certain frequency declines, it is replaced by another of similar frequency. This exchange-correlated mechanism of clone replacement may provide a mechanism by which overall homeostasis of hematopoiesis is maintained long term, thus ensuring continued health of the blood system.

To understand these observed features and the underlying mechanisms of stem cell-mediated blood regeneration, we developed a simple neutral population model of the hematopoietic system that quantifies the dynamics of three subpopulations: HSCs, transit-amplifying progenitor cells, and fully differentiated nucleated blood cells. We also include the effects of global regulation by assuming a Hill-type growth rate for progenitor cells in the bone marrow but ignore cell-to-cell variation in differentiation and proliferation rates of all cells.

Even though we do not include possible HSC heterogeneity, variation in HSC activation, progenitor-cell regulation, HSC and progenitor-cell aging (progenitor bursting), niche- and signal molecule-mediated controls, or intrinsic genetic and epigenetic differences, solutions to our simple *homogeneous* HSC model are remarkably consistent with observed clone size distributions. As a first step, we focus on how the intrinsic stochasticity in just the cellular birth, death, and differentiation events gives rise to the progenitor clone size distribution.

To a large extent, the exponentially distributed first HSC differentiation times and the growth and turnover of the progenitor pool control the shape of the expected long-time clone size distribution. Upon constraining our model to the physiological regime relevant to the experiments, we find that the calculated shapes of the clone size distributions are sensitive to effectively only two composite parameters. The HSC differentiation rate α sets the scale of the expected clone size distribution but has little effect on the shape. Parameters, including carrying capacity K, active HSCs $U + C$, and birth and death rates p, ω, μ_p, μ_d, influence the shape of the expected clone size distribution $\langle m_q \rangle$ only through the combination R, and only at large clone sizes.

Our analysis allowed us to estimate other combinations of model parameters quantitatively. Using a MLE, we find values for the effective HSC differentiation rate $a^* \sim 10^{-2}$–10^{-1} and the number of HSCs that are contributing to blood within any given time frame $U + C \sim 10^3$–10^4. Since the portion of HSCs that contribute to blood may vary across their typical life span $L \sim 25$ years, the total number of HSCs can be estimated by $(U + C) \times L/\tau$, where $\tau \sim 1$ year [19]. Our estimate of a total count of $\sim 3 \times 10^4$–3×10^5 HSCs is about 30-fold higher than the estimate of Abkowitz et al. [33] but is consistent with Kim et al. [19]. Note that the ratio of C to the total number of initially transplanted CD34+ cells provides a measure of the overall potency of the transplant towards blood regeneration. In the extreme case in which one HSC is significantly more potent (through, e.g., a faster

differentiation rate), this ratio would be smaller. An example of this type of heterogeneity would be an HSC with one or more cancer-associated mutations, allowing it to out-compete other transplanted normal HSCs. Hence, our clonal studies and the associated mathematical analysis can provide a framework for characterizing normal clonal diversity as well as deviations from it, which may provide a metric for early detection of cancer and other related pathologies.

Several simplifying assumptions have been invoked in our analysis. Crucially, we assumed HSCs divided only asymmetrically and ignored instances of symmetric self-renewal or symmetric differentiation. The effects of symmetric HSC division can be quantified in the steady-state limit. In previous studies, the self-renewal rate for HSCs in primates is estimated as 4–9 months [46, 47], which is slightly longer than the short timescale (∼2–4 months) on which we observe stabilization of the clone size distribution. Therefore, if the HSC population slowly increases in time through occasional symmetric division, the clone size distribution in the peripheral blood will also shift over long times. The static nature of the clone distributions over many years suggests that size distributions are primarily governed by mechanisms operating at shorter timescales in the progenitor pool. For an HSC population (such as cancerous or precancerous stem cells [48]) that has already expanded through early replication, the initial clone size distribution within the HSC pool can be quantified by assuming an HSC pool with separate carrying capacity K_{HSC}. Such an assumption is consistent with other analyses of HSC renewal [49]. All our results can be used (with the replacement $C \rightarrow K_{\mathrm{HSC}}$) if the number of transplanted clones $C \geq K_{\mathrm{HSC}}$ because replication is suppressed in this limit. When $K_{\mathrm{HSC}} \gg C \gg 1$, replicative expansion generates a broader initial HSC clone size distribution (see Additional file 1). The resulting final peripheral blood clone size distribution can still be approximated by our result (Eq. 6) if the normalized differentiation rate $a \ll 1$, exhibiting the insensitivity of the differentiated clone size distribution to a broadened clone size distribution at the HSC level. However, if HSC differentiation is sufficiently fast ($a \not\ll 1$), the clonal distribution in the progenitor and differentiated pools may be modified.

To understand the temporal dynamics of clone size distributions, a more detailed numerical study of our full time-dependent neutral model is required. Such an analysis can be used to investigate the effects of rapid temporal changes in the HSC division mode [41]. Temporal models would also allow investigation into the evolution of HSC mutations and help unify concepts of clonal stability (as indicated by the stationarity of rescaled clone size distributions) with ideas of clonal succession [10, 11] or dynamic repetition [12] (as indicated by the temporal fluctuations in the estimated number $U + C$ of active HSCs). Predictions of the time-dependent behavior of clone size distributions will also prove useful in guiding future experiments in which the animals are physiologically perturbed via e.g., myeloablation, hypoxiation, and/or bleeding. In such experimental settings, regulation may also occur at the level of HSC differentiation (α) and a different mathematical model may be more appropriate.

We have not addressed the temporal fluctuations in *individual* clone abundances evident in our data (Fig. 4a) or in the wave-like behavior suggested by previous studies [19]. Since the numbers of detectable cells of each VIS lineage in the whole animal are large, we believe these fluctuations do not arise from intrinsic cellular stochasticity or sampling. Rather, they likely reflect slow timescale HSC transitions between quiescent and active states and/or HSC aging [50]. Finally, subpopulations of HSCs that have different intrinsic rates of proliferation, differentiation, or clearance could then be explicitly treated. As long as each subtype in a heterogeneous HSC or progenitor-cell population does not convert into another subtype, the overall aggregated clone size distribution $\langle m_k \rangle$ will preserve its shape. Although steady-state data are insufficient to provide resolution of cell heterogeneity, more resolved temporal data may allow one to resolve different parameters associated with different cell types. Such extensions will allow us to study the temporal dynamics of individual clones and clone populations in the context of cancer stem cells and will be the subject of future work.

Additional file

Additional file 1: Mathematical appendices and data fitting.
(PDF 327 kb)

Abbreviations

HSC: hematopoietic stem cell; HSPC: hematopoietic stem and progenitor cell; MLE: maximum likelihood estimate; VIS: viral vector integration site.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TC and SG designed and developed the mathematical modeling and data analysis. TC, SG, and SK wrote the manuscript. SK and IC participated in study design and data interpretation. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by grants from the National Institutes of Health (R01AI110297 and K99HL116234), the California Institute of Regenerative Medicine (TRX-01431), the University of California, Los Angeles, AIDS Institute/Center for AIDS Research (AI28697), the National Science Foundation (PHY11-25915 KITP/UCSB), and the Army Research Office (W911NF-14-1-0472). The authors also thank B Shraiman and RKP Zia for helpful discussions.

Author details

[1]Department of Physics, University of Toronto, St George Campus, Toronto, Canada. [2]Department of Microbiology, Immunology, and Molecular Genetics,

UCLA, Los Angeles, USA. [3]UCLA AIDS Institute and Department of Medicine, UCLA, Los Angeles, USA. [4]Departments of Biomathematics and Mathematics, UCLA, Los Angeles, USA.

Received: 9 June 2015 Accepted: 12 September 2015
Published online: 20 October 2015

References

1. Enver T, Heyworth CM, Dexter TM. Do stem cells play dice? Blood. 1998;92:348–52.
2. Hoang T. The origin of hematopoietic cell type diversity. Oncogene. 2004;23:7188–98.
3. Muller-Sieburg CE, Cho RH, Thoman M, Adkins B, Sieburg HB. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. Blood. 2002;100:1302–9.
4. Copley MR, Beer PA, Eaves CJ. Hematopoietic stem cell heterogeneity takes center stage. Cell Stem Cell. 2012;10:690–7.
5. Muller-Sieburg CE, Sieburg HB, Bernitz JM, Cattarossi G. Stem cell heterogeneity: implications for aging and regenerative medicine. Blood. 2012;119:3900–7.
6. Lu R, Neff NF, Quake SR, Weissman IL. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat Biotechnol. 2011;29:928–33.
7. Huang S. Non-genetic heterogeneity of cells in development: more than just noise. Development. 2009;136:3853–62.
8. Osafune K, Caron L, Borowiak M, Martinez RJ, Fitz-Gerald CS, Sato Y, et al. Marked differences in differentiation potential among human embryonic stem cell lines. Nat Biotechnol. 2008;26:313–15.
9. Pang WW, Price EA, Sahoo D, Beerman I, Maloney WJ, Rossi DJ, et al. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. Proc Natl Acad Sci USA. 2011;108:20012–17.
10. Harrison DE, Astle CM, Lerner C. Number and continuous proliferative pattern of transplanted primitive immunohematopoietic stem cells. Proc Natl Acad Sci USA. 1988;85:822–6.
11. Verovskaya E, Broekhuis MJC, 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;122:523–32.
12. Takizawa H, Regoes RR, Boddupalli CS, Bonhoeffer S, Manz MG. Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. J Exp Med. 2011;208:273–84.
13. Gerrits A, Dykstra B, Kalmykowa OJ, Klauke K, Verovskaya E, Broekhuis MJC, et al. Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood. 2010;115:2610–18.
14. Kim S, Kim N, Presson AP, An DS, Mao SH, Bonifacino AC, et al. High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones. J Virol. 2010;84:11771–80.
15. Biffi A, Montini E, Lorioli L, Cesani M, Fumagalli F, Plati T, et al. Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. Science. 2013;341:1233158.
16. Aiuti A, Biasco L, Scaramuzza S, Ferrua F, Cicalese MP, Baricordi C, et al. Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott–Aldrich syndrome. Science. 2013;341:1233151.
17. Cavazzana-Calvo M, Payen E, Negre O, Wang G, Hehir K, Fusil F, et al. Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. Nature. 2010;467:318–22.
18. Cartier N, Hacein-Bey-Abina S, Bartholomae CC, Veres G, Schmidt M, Kutschera I, et al. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. Science. 2009;326:818–23.
19. Kim S, Kim N, Presson AP, Metzger ME, Bonifacino AC, Sehl M, et al. Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. Cell Stem Cell. 2014;14:473–85.
20. Sun J, Ramos A, Chapman B, Johnnidis JB, Le L, Ho YJ, et al. Clonal dynamics of native haematopoiesis. Nature. 2014;514:322–7.
21. Loeffler M, Roeder I. Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models – a conceptual approach. Cells Tissue Organs. 2002;171:8–26.
22. Bernard S, Belair J, Mackey MC. Oscillations in cyclical neutropenia: new evidence based on mathematical modeling. J Theor Biol. 2003;223:283–98.
23. Dingli D, Pacheco JM. Modeling the architecture and dynamics of hematopoiesis. Wiley Interdiscip Rev Syst Biol Med. 2010;2:235–44.
24. Courant R. Differential and integral calculus. Vol. II. London: Blackie & Son; 1936.
25. D'Orsogna MR, Lakatos G, Chou T. Stochastic self-assembly of incommensurate clusters. J Chem Phys. 2012;136:084110.
26. Marciniak-Czochra A, Stiehl T, Ho AD, Jager W, Wagner W. Modeling of asymmetric cell division in hematopoietic stem cells – regulation of self-renewal is essential for efficient repopulation. Stem Cells Dev. 2009;18:377–85.
27. Kent DG, Li J, Tanna H, Fink J, Kirschner K, Pask DC, et al. Self-renewal of single mouse hematopoietic stem cells is reduced by JAK2V617F without compromising progenitor cell expansion. PLoS Biol. 2013;11:1001576.
28. Lander AD, Gokoffski KK, Wan FYM, Nie Q, Calof AL. Cell lineages and the logic of proliferative control. PLoS Biol. 2009;7:1000015.
29. Hoffmann M, Chang HH, Huang S, Ingber DE, Loeffler M, Galle J. Noise-driven stem cell and progenitor population dynamics. PLoS One. 2008;3:2922.
30. Roshan A, Jones PH, Greenman CD. Exact, time-independent estimation of clone size distributions in normal and mutated cells. J Roy Soc Interface. 2014;11:20140654.
31. McHale PT, Lander A. The protective role of symmetric stem cell division on the accumulation of heritable damage. PLoS Comput Biol. 2014;10:1003802.
32. Antal T, Krapivsky PL. Exact solution of a two-type branching process: Clone size distribution in cell division kinetics. J Stat Mech. 2010; P07028.
33. Abkowitz JL, Caitlin SN, McCallie MT, Guttorp P. Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. Blood. 2002;100:2665–7.
34. Morozov AY, Bruinsma R, Rudnick J. Assembly of viruses and the pseudo-law of mass action. J Chem Phys. 2009;131:155101.
35. Krapivsky PL, Ben-Naim E, Redner S. Statistical physics of irreversible processes. Cambridge, UK: Cambridge University Press; 2010.
36. Szkely TJr, Burrage K, Mangel M, Bonsall MB. Stochastic dynamics of interacting haematopoietic stem cell niche lineages. PLoS Comput Biol. 2014;10:1003794.
37. Mackay R. Unified hypothesis for the origin of aplastic anemia and periodic hematopoiesis. Blood. 1978;51:941–56.
38. Keshet-Edelstein L. Mathematical models in biology. New York, NY: SIAM; 2005.
39. Wolfensohn S, Lloyd M. Handbook of laboratory animal management and welfare, 3rd ed. Oxford: Blackwell Publishing; 2003.
40. Kimura M. Population genetics, molecular evolution, and the neutral theory: selected papers In: Takahata N, editor. Chicago, IL: University of Chicago Press; 1995.
41. He J, Zhang G, Almeida AD, Cayoutte M, Simons BD, Harris WA. How variable clones build an invariant retina. Neuron. 2012;75:786–98.
42. Blanpain C, Simons BD. Unravelling stem cell dynamics by lineage tracing. Nat Rev Mol Cell Biol. 2013;14:489–502.
43. Guillaume T, Rubenstein DB, Symann M. Immune reconstitution and immunotherapy after autologous hematopoietic stem cell transplantation. Blood. 1998;92:1471–90.
44. Tzannou I, Leen AM. Accelerating immune reconstitution after hematopoietic stem cell transplantation. Clin Transl Immunol. 2014;3:11.
45. Allen LJS. An introduction to stochastic processes with applications to biology. Upper Saddle, NJ: Pearson Prentice Hall; 2003.
46. Shepherd BE, Guttorp P, Lansdorp PM, Abkowitz JL. Estimating human hematopoietic stem cell kinetics using granulocyte telomere lengths. Exp Hematol. 2004;32:1040–50.
47. Shepherd BE, Kiem HP, Lansdorp PM, Dunbar CE, Aubert G, LaRochelle A, et al. Hematopoietic stem-cell behavior in nonhuman primates. Blood. 2007;110:1806–13.
48. Driessens G, Beck B, Caauwe A, Simons BD, Blanpain C. Defining the mode of tumour growth by clonal analysis. Nature. 2012;488:527–30.
49. Sieburg HB, Cattarossi G, Muller-Sieburg CE. Lifespan differences in hematopoietic stem cells are due to imperfect repair and unstable mean-reversion. PLoS Comput Biol. 2013;9:1003006.
50. Weiss GH. Equations for the age structure of growing populations. Bull Math Biophys. 1968;30:427–35.

51. Catlin SN, Busque L, Gale RE, Guttorp P, Abkowitz JL. The replication rate of human hematopoietic stem cells in vivo. Blood. 2011;117:4460–6.
52. DeBoer RJ, Mohri H, Ho DD, Perelson AS. Turnover rates of B cells, T cells, and NK cells in simian immunodeficiency virus-infected and uninfected rhesus macaques. J Immunol. 2003;170:2479–87.
53. Pillay J, den Braber I, Vrieskoop N, Kwast LM, de Boer RJ, Borghans JAM, et al. In vivo labeling with $^{2}H_{2}O$ reveals a human neutrophil lifespan of 5.4 days. Blood. 2010;116:625–7.

Additional Files: Mathematical Appendices and Data Fitting

Here, we provide additional details of our model, including explicit derivations of the sampled clone size distribution, clone size distributions in the steady-state limit, and the effective parameters that accurately describe our data. We also describe the maximum likelihood estimation used to estimate these parameters.

Derivation of sampled clone size distribution:

We first derive an expression for the expected clone size distribution $\langle m_k(t)\rangle$ in a sample of the differentiated blood, as given by Eq. 6. Define $s_{j\ell}$ to be the number of cells sampled from the j^{th} clone of those that are represented by ℓ cells. At any time, the probability that the configuration $s_{j\ell}$ is observed in a sample of S cells can be written

$$P(\{s\}) = S! \prod_{\ell=1}^{N_{\rm d}} \prod_{j=1}^{y_\ell} \binom{\ell}{s_{j\ell}} \frac{\delta_{X,S}}{N_{\rm d}^{s_{j\ell}}}, \tag{A1}$$

where $X \equiv \sum_{\ell=1}^{N_{\rm d}} \sum_{j=1}^{y_\ell} s_{j\ell}$, the factor $N_{\rm d}^{-s_{j\ell}} S!$ represents the probability that $s_{j\ell}$ cells were drawn from the $N_{\rm d}$ within a sample of S cells, and $\binom{\ell}{s_{j\ell}}$ is the number of ways of drawing $s_{j\ell}$ cells. Finally, the last Kronecker $\delta-$function forces the sum over all $s_{j\ell}$ to equal the total number of cells sampled and sequenced. In any particular sample, the number of clones with size k is exactly

$$m_k = \sum_{n=1}^{N_{\rm d}} \sum_{m=1}^{y_n} \delta_{k,s_{mn}}. \tag{A2}$$

The expected value of this quantity is

$$\langle m_k(t)\rangle = \sum_{\{s\}} P(\{s\}) \sum_{\ell=0}^{N_{\rm d}} \sum_{j=1}^{y_\ell} \delta_{k,s_{j\ell}}, \tag{A3}$$

which can be found by using the generating function $G(z,t) \equiv \sum_{k=0}^{\infty} \langle m_k(t)\rangle z^k$:

$$\begin{aligned} G(z,t;S) &= \sum_{\{s\}} P(\{s\}) \sum_{\ell=0}^{N_{\rm d}} \sum_{j=1}^{y_\ell} z^{s_{j\ell}} \\ &= \frac{\partial}{\partial \beta} \sum_{\{s\}} P(\{s\}) \exp\left[\beta \sum_{\ell=0}^{N_{\rm d}} \sum_{j=1}^{y_\ell} z^{s_{j\ell}}\right]\Big|_{\beta=0}. \end{aligned} \tag{A4}$$

After using the Fourier representation of the Kronecker $\delta-$function in Eq. A1, $2\pi\delta_n = \int_0^{2\pi} e^{iqn} \mathrm{d}q$, we can further reduce the generating function to

$$\begin{aligned} G(z,t;S) &= \frac{\partial}{\partial \beta} \int_0^{2\pi} \frac{\mathrm{d}q}{2\pi} S! e^{-iqS} \prod_{\ell=0}^{N_{\rm d}} \prod_{j=1}^{y_\ell} \left[\sum_{s=0}^{\ell} \binom{\ell}{s} \frac{e^{iqs}}{N_{\rm d}^s} e^{\beta z^s}\right]_{\beta=0} \\ &= \sum_{\ell=0} y_\ell S! \int_0^{2\pi} \frac{\mathrm{d}q}{2\pi} e^{-iqS} \frac{\sigma(\ell,z)}{\sigma(\ell,1)} \prod_{i=1} [\sigma(i,1)]^{y_i}, \end{aligned} \tag{A5}$$

where

$$\sigma(\ell,z) \equiv \sum_{s=0}^{\ell} \binom{\ell}{s} \frac{e^{iqs}}{N_{\rm d}^s} z^s. \tag{A6}$$

Note that $\sigma(\ell,z) \equiv (1 + z e^{iq}/N_{\rm d})^{\ell}$, and that $\prod_{i=1} [\sigma(i,1)]^{y_i} \approx (1 + e^{iq}/N_{\rm d})^{N_{\rm d}}$. Since $N_{\rm d} \gg S \gg 1$, and $N_{\rm d} \sim 10^9 - 10^{10}$, we can take the large $N_{\rm d}$ limit before the large S limit to find $(1 + e^{iq}/N_{\rm d})^{N_{\rm d}} \approx \exp\left[e^{iq}\right]$, $\sigma(\ell,z) \approx \exp[e^{iq} z\ell/N_{\rm d}]$, and

$$G(z,t;S) \approx \sum_{\ell=0}^{\infty} y_\ell S! \int_0^{2\pi} \frac{\mathrm{d}q}{2\pi} \exp\left[A_\ell(z)e^{iq}\right] e^{-iqS}, \tag{A7}$$

where $A_\ell(z) = 1 + \ell(z-1)/N_\mathrm{d}$. Note that the integral is simply Euler's integral for $1/\Gamma(S+1)$. Namely, we find

$$\begin{aligned}\int_0^{2\pi} \frac{\mathrm{d}q}{2\pi} \exp\left[A_\ell(z)e^{iq}\right] e^{-iqS} &= \frac{iA_\ell^S(z)}{2\pi} \int_\mathcal{C} \mathrm{d}\xi e^\xi \xi^{-(S+1)} \\ &= \frac{A_\ell^S(z)}{\Gamma(S+1)}.\end{aligned} \tag{A8}$$

Since $A_\ell^S = (1+\ell(z-1)/N_\mathrm{d})^S = (1+(S/N_\mathrm{d})\ell(z-1)/S)^S \approx e^{\ell S(z-1)/N_\mathrm{d}}$ for $S \gg 1$, we find, for $\varepsilon \equiv S/N_\mathrm{d} \ll 1$,

$$G(z,t;S) = \sum_{\ell=0}^{N_\mathrm{d}} y_\ell(t) A_\ell^S(z) \approx \sum_{\ell=0} y_\ell(t) e^{\ell\varepsilon(z-1)}, \tag{A9}$$

Next, we define the fraction of clones of size $1 \le q \le S$ or less. This distribution includes unrepresented or lost clones, and is defined as $F(q,t) \equiv \sum_{i=0}^q \langle m_i(t)\rangle$. By using Eq. A9 and the definition of $G(z,t)$, we find

$$F(q,t) \equiv \sum_{j=0}^{q} \sum_{n=0}^{\infty} e^{-n\varepsilon} \frac{(n\varepsilon)^j}{j!} y_n(t). \tag{A10}$$

The expected clone size distribution is thus defined as

$$\langle m_k(t)\rangle = F(k,t) - F(k-1,t) = \sum_{\ell=0}^{\infty} e^{-\ell\varepsilon} \frac{(\ell\varepsilon)^k}{k!} y_\ell(t). \tag{A11}$$

In general, further development of $F(k,t)$ and $\langle m_k(t)\rangle$ requires numerical solution of $c_k(t)$ and $y_\ell(t) = \sum_{k=0}^\infty y_\ell^{(k)}(t)$. The time-dependence of $F(q,t)$ is further complicated by the time-dependence of $\varepsilon(t) = S/N_\mathrm{d}(t)$, requiring the solution to Eq. 4.
The variability of m_k due to sampling can be also estimated by calculating $\langle m_k m_{k'}\rangle$, which we write as

$$\langle m_k m_{k'}\rangle = \sum_{\{s\}} P(\{s\}) \sum_{\ell=0}^{N_\mathrm{d}} \sum_{j=1}^{y_\ell} \delta(s_{j\ell} - k) \sum_{\ell'=0}^{N_\mathrm{d}} \sum_{j'=1}^{y_{\ell'}} \delta(s_{j'\ell'} - k'). \tag{A12}$$

This calculation requires evaluation of the two-dimensional generating function

$$G(z,z',t) = \sum_{k,k'} \langle m_k(t) m_{k'}(t)\rangle z^k z'^{k'}. \tag{A13}$$

After using Eq. A1 for $P(\{s\})$ in Eq. A12 and performing some algebra, we find

$$G(z,z',t) = \sum_{\ell,\ell'} A_{\ell,\ell'}^S(z,z') y_\ell(t) y_{\ell'}(t) - \sum_\ell B_\ell^S(z,z') y_\ell(t) + \sum_\ell C_\ell^S(zz') y_\ell(t), \tag{A14}$$

where

$$A_{\ell,\ell'}(z,z') = 1 + \frac{\varepsilon(z'-1)\ell'}{S} + \frac{\varepsilon(z-1)\ell}{S}$$

$$B_{\ell}(z,z') = 1 + \frac{\varepsilon(z'-1)\ell}{S} + \frac{\varepsilon(z-1)\ell}{S}$$

$$C_{\ell}(zz') = 1 + \frac{\varepsilon(zz'-1)\ell}{S}.$$

Using $(1+x/S)^S \approx e^x$, and expanding in powers of z and z', we find

$$\begin{aligned}\langle m_k(t) m_{k'}(t)\rangle = \quad & \left[\sum_{\ell} y_{\ell}(t) e^{-\varepsilon\ell}\frac{(\varepsilon\ell)^k}{k!}\right]\left[\sum_{\ell'} y_{\ell'}(t) e^{-\varepsilon\ell'}\frac{(\varepsilon\ell')^{k'}}{k'!}\right] \\ & - \sum_{\ell} y_{\ell}(t) e^{-2\varepsilon\ell}\frac{(\varepsilon\ell)^k}{k!}\frac{(\varepsilon\ell)^{k'}}{k'!} + \sum_{\ell} y_{\ell}(t) e^{-\varepsilon\ell}\frac{(\varepsilon\ell)^k}{k!}\delta_{k,k'}.\end{aligned} \tag{A15}$$

The diagonal variance is simply

$$\begin{aligned}\langle m_k^2(t)\rangle - \langle m_k(t)\rangle^2 &= \sum_{\ell} y_{\ell}(t) e^{-\varepsilon\ell}\frac{(\varepsilon\ell)^k}{k!}\left[1 - e^{-\varepsilon\ell}\frac{(\varepsilon\ell)^k}{k!}\right] \\ &= \langle m_k(t)\rangle - \sum_{\ell} y_{\ell}(t)\frac{e^{-2\varepsilon\ell}(\varepsilon\ell)^{2k}}{(k!)^2}.\end{aligned} \tag{A16}$$

The second term is much smaller than the first except for very small values of k. Therefore, the relative fluctuation in m_k due to sampling is

$$\frac{\sqrt{\langle m_k^2(t)\rangle - \langle m_k(t)\rangle^2}}{\langle m_k(t)\rangle} \lesssim \frac{1}{\sqrt{\langle m_k(t)\rangle}}, \tag{A17}$$

explicitly indicating that the relative fluctuations in the measured number of clones of size k decreases as the square-root of its expected value.

Steady-state solution:
As was discussed, the total peripheral blood population in the animals recovered quickly, usually within a few weeks after transplantation. Moreover, from our data, the overall qualitative shape of the clone size distribution also reaches steady-state only after a few months post-transplant, with no discernible systematic time-dependence. Therefore, we consider the steady-state solutions to our model (Eqs. 1 and 5). Henceforth, all quantities will be assumed to be those at steady-state. First, we can start from $n = 1$ and inductively solve for the steady-state form of Eqs. 5 to find

$$y_n = \sum_{k=0}^{\infty} y_n^{(k)} = \sum_{k=0}^{\infty}\frac{(wk)^n}{n!}e^{-wk}c_k,\ w = \frac{(1+\eta)\omega}{\mu_{\rm d}}. \tag{A18}$$

Before this solution can be effectively used in Eq. A10, an explicit expression for the steady-state progenitor clone size distribution c_k is needed.
The total steady-state progenitor population is given by the solution to $\alpha C + [r(N_{\rm p}) - \mu]N_{\rm p} = 0$. The population-limited growth rate is given by Eq. 3 and the steady-state progenitor cell population is explicitly

$$\frac{N_{\rm p}}{K} = \frac{1}{2}\left[\frac{\alpha}{\mu}\frac{U+C}{K} + \frac{p}{\mu} - 1 + \sqrt{\left(\frac{\alpha}{\mu}\frac{U+C}{K} + \frac{p}{\mu} - 1\right)^2 + \frac{4\alpha}{\mu}\frac{U+C}{K}}\right], \tag{A19}$$

where $\mu = \mu_{\rm p} + \eta\omega$. After using Eq. A19 in Eq. 3, we find explicitly

$$r(N_{\rm p}) = \frac{2p}{\frac{\alpha}{\mu}\frac{U+C}{K} + \frac{p}{\mu} + 1 + \sqrt{\left(\frac{\alpha}{\mu}\frac{U+C}{K} + \frac{p}{\mu} - 1\right)^2 + \frac{4\alpha}{\mu}\frac{U+C}{K}}} < \mu. \tag{A20}$$

The total differentiated cell population found from $\dot{N}_{\rm d} = (1+\eta)\omega N_{\rm p} - \mu_{\rm d} N_{\rm d} \approx 0$ is $N_{\rm d} = (1+\eta)\omega N_{\rm p}/\mu_{\rm d} \equiv w N_{\rm p}$. Upon using these expressions in the steady-state limit of Eq. 1, we obtain

$$\begin{aligned} c_{k\geq 1} &= \frac{aC}{k!}\frac{\bar{r}^k(1-\bar{r})^a}{(a+k)}\prod_{\ell=1}^{k}[a+\ell], \\ c_0 &= C - \sum_{k=1}^{\infty} c_k = C(1-\bar{r})^a, \end{aligned} \tag{A21}$$

where

$$a \equiv \frac{\alpha}{r} \equiv \frac{\alpha}{\mu\bar{r}} \quad \text{and} \quad \bar{r} \equiv \frac{r(N_{\rm p})}{\mu}. \tag{A22}$$

From these results, the total number of clones in each compartment can be explicitly found:

$$\begin{aligned} C_{\rm p} &= \sum_{k=1}^{\infty} c_k = C\left[1-(1-\bar{r})^a\right], \\ C_{\rm d} &= \sum_{n=1}^{\infty}\sum_{k=1}^{\infty} y_n^{(k)} = \sum_{k=1}^{\infty}(1-e^{-wk})c_k \\ &= C\left[1 - \left(\frac{1-\bar{r}}{1-\bar{r}e^{-w}}\right)^a\right]. \end{aligned} \tag{A23}$$

For the total expected number of clones observed in the sample,

$$\begin{aligned} C_{\rm s} &= \sum_{j=1}^{\infty}\langle m_j\rangle = \sum_{j=1}^{\infty}\sum_{\ell=1}^{\infty} e^{-\ell\varepsilon}\frac{(\ell\varepsilon)^j}{j!}y_\ell \\ &= \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty}(1-e^{-\ell\varepsilon})\frac{(wk)^\ell}{\ell!}e^{-wk}c_k \\ &= \sum_{k=1}^{\infty}(1-e^{-wke^{-\varepsilon}})c_k \\ &= C\left[1-\left(\frac{1-\bar{r}}{1-\bar{r}e^{-w(1-e^{-\varepsilon})}}\right)^a\right], \end{aligned} \tag{A24}$$

where we have explicitly used Eq. A18 for y_ℓ and Eq. A21 for c_k. This result can be further reduced in two limits

$$C_{\rm s} \approx \begin{cases} C\left[1-\left(\frac{1}{\bar{r}w\varepsilon}\right)^a\left(\frac{1-\bar{r}}{2-\bar{r}}\right)^a\right], & \varepsilon w \gg \frac{(1-\bar{r})}{\bar{r}} \\ aC\frac{\varepsilon w\bar{r}}{1-\bar{r}}, & \varepsilon w \ll \frac{(1-\bar{r})}{\bar{r}}. \end{cases} \tag{A25}$$

As expected, the total numbers of clones present in each pool follow the progression

$$C \gtrsim C_{\rm p} \gtrsim C_{\rm d} > C_{\rm s}, \tag{A26}$$

with significant loss of clones due to sampling ($C_{\rm d} \gg C_{\rm s}$) only in the second case of Eq. A25 describing sample sizes $S \ll \mu_{\rm d} N_{\rm d}(1-\bar{r})/(\omega\bar{r})$. Note that for clone sizes $k \gg a$, c_k in Eq. A21 can be approximated by

$$c_{k\gg a} \approx Ak^{a-1}\bar{r}^k\left[1-\frac{a(1-a)}{2k}+O\left(\frac{1}{k^2}\right)\right], \tag{A27}$$

where

$$A \equiv \frac{aC}{\Gamma(a+1)}(1-\bar{r})^a. \tag{A28}$$

Finally, in steady-state, using Eq. A18 in Eq. A10, we find the cumulative clone size distribution

$$\begin{aligned} F(q) &= \sum_{j=0}^{q}\sum_{\ell=0}^{\infty}\sum_{k=0}^{\infty}\frac{(\ell\varepsilon)^j}{j!}\frac{(wk)^\ell}{\ell!}e^{-\ell\varepsilon}e^{-wk}c_k \\ &\equiv \sum_{\ell=0}^{\infty}\sum_{k=0}^{\infty}\frac{\Gamma(q+1,\ell\varepsilon)}{\Gamma(q+1)}\frac{(wk)^\ell}{\ell!}e^{-wk}c_k \\ &\approx \sum_{k=0}^{\infty}\frac{\Gamma(q+1,\varepsilon wk)}{\Gamma(q+1)}c_k, \end{aligned} \tag{A29}$$

where a steepest descents approximation in $\varepsilon \ll 1$ was used to derive the final approximation. The expected cumulative clone size frequency $Q(q)$ is obtained by subtracting off the unrepresented clones $\langle m_0\rangle \equiv F(0)$ and normalizing by the total expected number of clones $C_{\rm s} = F(S) - F(0)$.
A number of numerical procedures can be used to evaluate $Q(q)$ using the final approximation in Eq. A29. For large values of q, and small εw, the ratio of $\Gamma-$functions is near unity for $k \lesssim (q+1)/(\varepsilon w)$, and quickly decreases to zero for larger k. One approach for numerically evaluating $F(q)$ is to explicitly separate small q and small k terms in the sum. For small k, the exact form of c_k should be used. For larger k, the asymptotic form (Eq. A27) can be used and the sum can be approximated as an integral. We find that even for small values $q \approx 5$, this approximation results in a relative error $< 1\%$. Even the crudest approximation of replacing the sum by the integral

$$F(q) \approx \sum_{k=0}^{\infty}\frac{\Gamma(q+1,\varepsilon wk)}{\Gamma(q+1)}c_k \approx c_0 + \int_1^{(q+1)/(\varepsilon w)} c_k \mathrm{d}k, \tag{A30}$$

and using the asymptotic form Eq. A27, provides a reasonable estimate of $F(q)$. This rough approximation also provides an informative analytical expression:

$$F(q) \approx c_0 + \frac{A}{\ln^a(1/\bar{r})}\left[\Gamma\left(a, \ln\left(\frac{1}{\bar{r}}\right)\right) - \Gamma\left(a, \frac{(q+1)}{\varepsilon w}\ln\left(\frac{1}{\bar{r}}\right)\right)\right]. \tag{A31}$$

This approximation shows that our distributions depend most strongly on only $a = \alpha/r$ and $R = (\varepsilon w)/\ln(1/\bar{r})$ (Eq. 10). Since $\bar{r} \approx 1 - \alpha(U+C)/(K(p-\mu))$ is only very slightly smaller than unity, $\ln(1/\bar{r})$ is a small positive number and $A/\ln^a(1/\bar{r}) \approx aC$. Since $\langle m_q\rangle = F(q) - F(q-1)$, an approximate form useful for estimating the clone size distribution is

$$\langle m_q\rangle \approx aC\left[\Gamma(a, q/R) - \Gamma(a, (q+1)/R)\right]. \tag{A32}$$

Within physiologically-relevant regimes, our data can be well-fitted to $\langle m_k\rangle$ by varying just a and R. The other physiological parameters, $K, p, \mu, U+C$, etc., are then related to each other through the most likely numerical values a^* and R^* found from fitting the data.
Consider the total number of active HSC cells, $U+C$, and the ratio of the rate of HSC differentiation to the rate of self-renewal of progenitor cells, α/p. Once the best fit parameters a^* and R^* have been estimated from fitting clonal frequency distributions, $U+C$ and α/p can be expressed in terms of K, S, and $\Delta \equiv p/\mu - 1$. Note that S, the number of sequencing reads detected in each sample, is an experimentally determined parameter.
To find these relationships, we first assume that $\varepsilon w/R^* = (S/N_{\rm d})(\omega/\mu_{\rm d})/R^* = S/(R^*N_{\rm p}) \ll 1$ and $S/(R^*K) \ll 1$. By using the definition of R, we find $\bar{r} = e^{-S/(R^*N_{\rm p})} \approx 1 - S/(R^*N_{\rm p})$. Since $\bar{r} = (\Delta+1)K/(N_{\rm p}+K)$ also, these two independent expressions for $\bar{r}$ furnish a quadratic equation for $N_{\rm p}$ in terms of R^*. After comparing the positive root of this equation to the definition of the steady-state progenitor population $N_{\rm p}$ (Eq. A19), we find

$$\left(\frac{\alpha}{p}\right)\left(\frac{U+C}{K}\right) \approx \frac{S}{(\Delta+1)R^*K}. \tag{A33}$$

Eq. A33 can be then used to find an expression for $N_{\rm p}$ that is independent of α/p and $U+C$. Using this form of $N_{\rm p}$ in the definition $a=\alpha/r=(\alpha/p)\,[N_{\rm p}/K+1]$, we find an explicit expression for the best-fit value $\alpha/p=a^*K/(N_{\rm p}+K)$. Further assuming that $S/(R^*K)\ll\Delta$, we find

$$\left(\frac{\alpha}{p}\right)\approx\frac{a^*}{\Delta+1}\left[1-\frac{1}{\Delta}\left(\frac{S}{R^*K}\right)+\ldots\right]. \tag{A34}$$

Note that to lowest order, α/p can be estimated from a^* and $\Delta=p/\mu-1$. Finally, substituting $(\alpha/p)\approx a^*/(\Delta+1)$ from solving Eq. A34 into Eq. A33, we find

$$U+C\approx\frac{S}{a^*R^*}, \tag{A35}$$

which is independent of K and Δ. Note that these parameters can be extracted out of the many parameters in the model because of the limiting values of $\bar{r}\lesssim 1$ and $\varepsilon w\ll 1$. Our model allows one to make predictions on the number of expected clones in each pool, $C_{\rm p}, C_{\rm d}$, and the measured $C_{\rm s}$ (Eqs. A23-A24), as well as expected clone size distributions (Eq. A32) as functions of sampling fraction ε, turnover rate w, effective differentiation rate a, and effective growth rate $\bar{r}$. However, from the functional forms of $C_{\rm p}, C_{\rm d}, C_{\rm s}$, and because $\bar{r}\approx 1$, the numerical determination of the number of clones in each pool is highly sensitive to high values of R and low values of a.

Expected clone size distributions from stochastic clones sizes:
Here, we explicitly show how the neutral assumption (identical transition rates and fitness for all clones) of our populations allows mean-field equations for the expected clone size distribution to be derived from considerations of the stochastic dynamics of an individual clone. Analysis of individual clones is more natural in settings where each clone can be easily isolated and imaged, such as in epidermal systems and geometries [42, 54, 55]. An important feature of our neutral model is that the steady-state clone size distribution depends on only the value of the effective growth rate at steady-state and not on the specific form of the regulation. In other words, the *relative* sizes of neutral clones are independent of the growth law common to all clones. Therefore, we first consider the corresponding birth-death process of a single isolated clone in the presence of constant immigration occurring at rate α. The master equation for the probability $p_k(t)$ of a single clone containing k progenitor cells is

$$\frac{{\rm d}p_k(t)}{{\rm d}t}=\alpha\left[p_{k-1}-p_k\right]+r\left[(k-1)p_{k-1}-kp_k\right]+\mu\left[(k+1)p_{k+1}-kp_k\right], \tag{A36}$$

where in our application, $\mu\equiv\mu_{\rm p}+\eta\omega$. If the growth rate r is assumed constant and independent of k, an analytic expression expressed in terms of the corresponding generating function $\phi(z,t)\equiv\sum_{k=0}^{\infty}p_kz^k$ [45, 56]:

$$\phi(z,t)=\left[\frac{(1-\bar{r})}{(1-\bar{r}z)-\bar{r}(1-z)e^{-\mu(1-\bar{r})t}}\right]^a. \tag{A37}$$

We now identify $c_k(t)$ with C times the probability that any independent clone is of size k. Thus,

$$\sum_{k=0}c_kz^k\simeq C\phi(z,t) \tag{A38}$$

and the variability in clone sizes arises from the variability of the times of differentiation of HSC cells to create progenitor cells of different lineages. In the $t\to\infty$ steady-state limit, we find

$$\phi(z,t\to\infty)=\left(\frac{1-\bar{r}}{1-\bar{r}z}\right)^a=(1-\bar{r})^a\sum_{k=0}^{\infty}\frac{\Gamma(k+a)}{\Gamma(a)k!}(\bar{r}z)^k, \tag{A39}$$

in which $\bar{r}=r/\mu$. Thus, the single stochastic clone construction of the expected clone size distribution yields

$$c_k=C(1-\bar{r})^a\frac{\Gamma(k+a)\bar{r}^k}{\Gamma(a)k!}\approx C(1-\bar{r})^a\frac{\bar{r}^k}{k^{1-a}}, \tag{A40}$$

which matches the result in Eq. A27. This derivation explicitly shows that the exponentially distributed initial differentiation times sets the progenitor cell clone size distribution c_k. This distribution is preserved even in the

mean-field setting of the hodograph-transformed model described by Eq. 1 and is independent of the specific form chosen for the growth law r.

HSC self-renewal:
Rather than assuming that HSCs differentiate only asymmetrically, leaving each unique HSC clone unchanged, we now consider the effects of symmetric HSC replication on the measured clone size distribution. We also assume a separate HSC niche with a corresponding carrying capacity $K_{\rm s}$. If we denote x_k as the number of clones in the stem cell niche that is represented by exactly k stem cells,

$$\begin{aligned}\frac{{\rm d}x_1}{{\rm d}t} &= -(r_{\rm s}+\mu_{\rm s})x_1 + 2\mu_{\rm s}x_2 \\ \frac{{\rm d}x_k}{{\rm d}t} &= r_{\rm s}\left[(k-1)x_{k-1} - kx_k\right] + \mu_{\rm s}\left[(k+1)x_{k+1} - kx_k\right],\end{aligned} \tag{A41}$$

where the effective growth rate $r_{\rm s}$ is defined by the carrying capacity $K_{\rm s}$, and the total number of stem cells $N_{\rm s}$, labeled and unlabeled, in the stem cell compartment:

$$r_{\rm s}(N_{\rm s}) = p_{\rm s}\left(1 - \frac{N_{\rm s}}{K_{\rm s}}\right). \tag{A42}$$

Here, we have used logistic growth for mathematical convenience and to simply illustrate the insensitivity of the final clone size distribution to the model of HSC differentiation. The total stem cell population is defined as $N_{\rm s}(t) = U_{\rm s}(t) + \sum_{k=1}^{\infty} kx_k(t)$. Upon summing Eqs. A41 and the equation for unlabeled cells, $\dot{U}_{\rm s} = r_{\rm s}U_{\rm s} - \mu_{\rm s}U_{\rm s}$, we find that the total population decouples and obeys

$$\frac{{\rm d}N_{\rm s}}{{\rm d}t} = r_{\rm s}(N_{\rm s})N_{\rm s} - \mu_{\rm s}N_{\rm s}, \tag{A43}$$

which can be solved exactly, allowing one to find $r_{\rm s}$ explicitly as a function of time. Eqs. A41 can then be solved numerically to find the stem cell clone frequencies in the stem cell compartment. To simplify the calculations and find a tractable solution, we will set $\mu_{\rm s} = 0$ and define a new time variable ${\rm d}\tau = r_{\rm s}(t){\rm d}t$. Equations A41 for $x_k(\tau)$ now have constant coefficients and can be solved by using the initial conditions $x_1(\tau = 0) = C_{\rm s}, x_{k>1}(0) = 0$, and the Laplace transforms,

$$s\tilde{x}_1 - C_{\rm s} = -\tilde{x}_1, \quad s\tilde{x}_k = (k-1)\tilde{x}_{k-1} - k\tilde{x}_k. \tag{A44}$$

The solution

$$\tilde{x}(s) = \frac{C_{\rm s}}{s+1}, \quad \tilde{x}_k(s) = \frac{(k-1)C_{\rm s}}{\prod_{j=1}^{k}(s+j)}, \tag{A45}$$

can be inverted to yield

$$x_1(\tau) = C_{\rm s}e^{-\tau}, \quad x_k(\tau) = C_{\rm s}(e^{\tau}-1)^{k-1}e^{-k\tau}. \tag{A46}$$

To transform back to $x_k(t)$, we need to invert

$$\tau(t) = \int_0^t r_{\rm s}(t'){\rm d}t' = p_{\rm s}t - \ln(1 + C_{\rm s}(e^{p_{\rm s}t}-1)/K_{\rm s}). \tag{A47}$$

In the steady-state limit, $t \to \infty$ corresponds to $\tau \to \ln(K_{\rm s}/C_{\rm s}) + O(e^{-p_{\rm s}t})$. In this limit, Eq. A46 yields

$$x_k(t\to\infty) = C_{\rm s}f_{\rm s}(1-f_{\rm s})^{k-1}, \quad f_{\rm s} \equiv \frac{C_{\rm s}}{K_{\rm s}}. \tag{A48}$$

The clone numbers in the progenitor cell population are modeled using

$$\begin{aligned}\frac{\mathrm{d}c_1^{(j)}}{\mathrm{d}t} &= \alpha j \frac{x_j}{C_\mathrm{s}}(c_0 - c_1^{(j)}) - (r+\mu_\mathrm{p})c_1^{(j)} + 2\mu_\mathrm{p}c_2^{(j)} \\ \frac{\mathrm{d}c_k^{(j)}}{\mathrm{d}t} &= \underbrace{\alpha j \frac{x_j}{C_\mathrm{s}}(c_{k-1}^{(j)} - c_k^{(j)})}_{\text{HSC asym. differentiation}} + \underbrace{r((k-1)c_{k-1}^{(j)} - kc_k^{(j)})}_{\text{progenitor birth}} + \underbrace{\mu((k+1)c_{k+1}^{(j)} - kc_k^{(j)})}_{\text{progenitor death}},\end{aligned} \tag{A49}$$

where $c_0(t) \equiv C_\mathrm{s} - \sum_{j,\ell=1}^{\infty} c_\ell^{(j)}(t)$ is the total number of clones that do not appear in the progenitor cell pool at time t. Upon summing all the above equations to find the zeroth and first moments of $c_k^{(j)}$, we find

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{k=1}^{\infty} c_k^{(j)}(t)\right) = A_j c_0(t) - \mu_\mathrm{p} c_1^{(j)}, \tag{A50}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{k=1}^{\infty} k c_k^{(j)}(t)\right) = A_j \sum_{k=1}^{\infty} c_k^{(j)} + (r_\mathrm{p}(N_\mathrm{p}) - \mu_\mathrm{p}) \sum_{k=1}^{\infty} k c_k^{(j)}(t), \tag{A51}$$

where $A_j(t) = \alpha j x_j(t)/C_\mathrm{s}$. By further adding $\dot{U}_\mathrm{p} = \alpha U_\mathrm{s} + (r_\mathrm{p} - \mu_\mathrm{p})U_\mathrm{p}$ to Eq. A51, we find

$$\frac{\mathrm{d}N_\mathrm{p}}{\mathrm{d}t} = \alpha U_\mathrm{s} + \frac{\alpha}{C_\mathrm{s}} \sum_{j,k=1}^{\infty} j x_j(t) c_k^{(j)}(t) + (r_\mathrm{p}(N_\mathrm{p}) - \mu_\mathrm{p})N_\mathrm{p}(t). \tag{A52}$$

If we assume steady-state in both the stem cell and progenitor cell populations, $A_j = j\alpha x_j(\infty)/C_\mathrm{s} = j\alpha f_\mathrm{s}(1-f_\mathrm{s})^{j-1}$ and Eq. A50 yields

$$c_1^{(j)} = \frac{A_j}{\mu_\mathrm{p}} c_0 = j\frac{\alpha}{\mu_\mathrm{p}} f_\mathrm{s}(1-f_\mathrm{s})^{j-1} c_0, \tag{A53}$$

which, when used in the steady-state limit of Eq. A51 yields

$$c_k^{(j)} = \frac{c_0 \bar{r}^k}{k!} \prod_{\ell=1}^{k} \left[a_j + (\ell - 1)\right], \tag{A54}$$

where

$$a_j = \frac{j x_j}{C_\mathrm{s}} \frac{\alpha}{r} = j f_\mathrm{s}(1-f_\mathrm{s})^{j-1} \frac{\alpha}{r}. \tag{A55}$$

The coefficient c_0 can now be self-consistently calculated by noting that $c_0 + \sum_{j,k=1}^{\infty} c_k^{(j)} = C_\mathrm{s}$. Upon double-summing Eq. A54, we find $c_0 = C_\mathrm{s}/\mathcal{Z}$, where

$$\mathcal{Z} \equiv 1 + \sum_{j=1}^{\infty} \left[(1-\bar{r})^{-a_j} - 1\right], \tag{A56}$$

and

$$c_k^{(j)} = \frac{C_\mathrm{s} \bar{r}^k}{\mathcal{Z} k!} \prod_{\ell=1}^{k} \left[a_j + (\ell - 1)\right], \tag{A57}$$

The clone numbers in the progenitor pool are thus $c_k = \sum_{j=1}^{\infty} c_k^{(j)}$. Note that when the initial transplantation fills the entire stem cell niche, $f_s \to 1^-$, $\mathcal{Z} \to (1-\bar{r})^{-a}$, and the only term in Eq. A54 that survives is $j = 1$, leading to our previous result as expected. For the general product in Eq. A54, we can approximate

$$\prod_{\ell=1}^{k} [a_j + (\ell - 1)] \approx (k-1)! \frac{k^{a_j}}{\Gamma(a_j)}, \tag{A58}$$

when $a_j \ll \ln k$. From Eq. A55, we know that a_j is strictly bounded above by α/r and is typically $\lesssim 0.5\alpha/r$ for $f_s < 0.5$. Since we expect $\alpha/r < 1$, the approximation in Eq. A58 is valid for essentially all values of $k \gtrsim 2$. In order to compute c_k, we perform the sum

$$c_k \approx \frac{C_s \bar{r}^k}{\mathcal{Z} k} \sum_{j=1}^{\infty} \frac{k^{a_j}}{\Gamma(a_j)}. \tag{A59}$$

By Taylor-expanding in small a_j first, we can further approximate the sum as

$$\sum_{j=1}^{\infty} \frac{k^{a_j}}{\Gamma(a_j)} \approx \frac{a}{f_s} + \frac{a^2}{f_s} \frac{(2 - f_s(2 - f_s))}{(2-f_s)^3} (\gamma + \ln k) + O(a^3) \tag{A60}$$

where $a \equiv \alpha/r$ and $\gamma \approx 0.5772$ is Euler's constant. In order to find an explicit expression for $\mathcal{Z}$, note that $\delta \equiv 1 - \bar{r}$ is typically very small, on the order of $1/K_p$. Therefore $\mathcal{Z} = 1 + \sum_{j=1}^{\infty} [\delta^{-a_j} - 1]$ can be approximated using $\delta^{-a_j} = \exp[-a_j \ln \delta] \gg 1$ and Laplace's method on the sum. The dominant term in the sum arises for $j^* \approx -1/\ln(1 - f_s)$. Approximating the sum by an integral over a Gaussian centered about j^*, we find

$$\mathcal{Z} \approx \frac{\sqrt{2\pi e r (1-f_s)} \exp\left[\left(\frac{\alpha f_s}{er}\right) \frac{\ln(1-\bar{r})}{(1-f_s)\ln(1-f_s)}\right]}{\sqrt{\alpha f_s \ln(1-\bar{r}) \ln(1-f_s)}}. \tag{A61}$$

By using this approximation for $\mathcal{Z}$ and Eq. A60 in Eq. A59, we can find the leading behavior of c_k in the small f_s and $a \ln k \ll 1$ limits,

$$c_k \approx a K_s (1-\bar{r})^{a/e} \sqrt{\frac{a}{2\pi e} \ln\left(\frac{1}{1-\bar{r}}\right)} \frac{\bar{r}^k}{k}. \tag{A62}$$

This approximation is not good for small a since it suppresses the largeness of the asymptotic parameter $-\ln(1-\bar{r})$. Instead, for small $|a \ln(1-\bar{r}) \ll 1$, we Taylor expand $\exp[-a_j \ln \delta] \approx 1 - a_j \ln \delta$ to find

$$\mathcal{Z} \approx 1 - \frac{a}{f_s} \ln(1-\bar{r}), \tag{A63}$$

and

$$c_k \approx \left[1 + \frac{a(2 - f_s(2-f_s))(\gamma + \ln k)}{(2-f_s)^3}\right] \frac{a K_s}{1 - a \ln(1-\bar{r})/f_s} \left(\frac{\bar{r}^k}{k}\right). \tag{A64}$$

Thus, in the limit where the transplanted number of clones $C_s \ll K_s$ is much less than the HSC carrying capacity, the marked HSCs greatly expand before reaching K_s; however, the resulting clone size distribution c_k remains qualitatively unchanged.

Maximum Likelihood Estimation:
We start with the counts of unique sequencing reads on the macaque genome. *i.e.* number of times the read was sequenced. We refer to each unique read as a "clone." Since sequencing of each end of a unique viral sequence is performed independently, we treat the two data sets as independent measurements at each time. The reads are then pooled according to which end of a read was sequenced. For more details of the sequencing and filtering of the reads, see Kim *et al.* [14].

Assume a sequencing run from a particular animal, at a particular time and at one of the sequencing ends, yields n unique clones with $\{q_1, \ldots, q_i, \ldots, q_n\}$ read counts. We calculate the likelihood of observing this data within our model given a particular set of parameter values. Our mathematical model contains three independent parameter combinations:

$$a \equiv \frac{\alpha}{r}, \quad \bar{r} \equiv \frac{r}{\mu}, \quad \text{and} \quad \varepsilon w \equiv \frac{S}{N_{\mathrm{d}}} \times \frac{\omega}{\mu_{\mathrm{d}}}. \tag{A65}$$

Since the distinguishable clones are otherwise physiologically identical, we associate the distribution of sizes of any particular clone with the expected value of clone size distribution:
$\mathcal{P}(q_i|a, \bar{r}, \varepsilon w) \approx \langle m_{q_i}(a, \bar{r}, \varepsilon w)\rangle = F(q|a, \bar{r}, \varepsilon w) - F(q-1|a, \bar{r}, \varepsilon w)$. The likelihood of the parameters given the detected clone sizes $\{q_1, \ldots, q_n\}$ is then given by:

$$\mathcal{L}(a, \bar{r}, \varepsilon w|\{q_1, \ldots, q_n\}) = \prod_{i=1}^{n} \mathcal{P}(q_i|a, \bar{r}, \varepsilon w). \tag{A66}$$

The most likely parameters are then estimated by numerically maximizing the likelihood over the parameters. However, as shown previously, the distribution of clone sizes depends most strongly on only a and R given by Eqs. 10.
The figures below show normalised and rescaled clone size distributions extracted from granulocyte or peripheral blood mononuclear cell (PBMC) subpopulations of blood from all animals in the original study. The MLE values of a^* and R^* all fall within regimes such that $U + C \sim 10^3 - 10^4$. The fluctuations in $U + C$ are predominantly due to changes in the fraction f at different time points. Such fluctuations are the result of internal dynamics not considered in our model and do not exhibit any discernible trend.

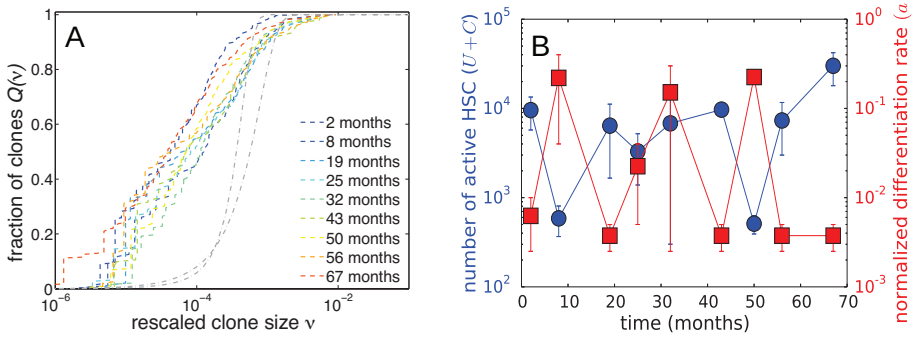

Fig. A1: (A) Rescaled clone size distributions from granulocytes in animal RQ5427. The stationarity in this subpopulation is evident. (B) The MLE fits for $U + C$ (blue circles) and a^* (red squares).

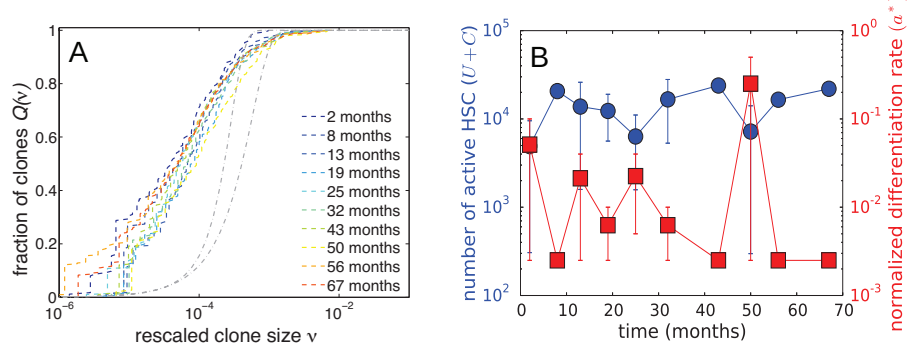

Fig. A2: (A) Rescaled PBMC clone size distributions from animal RQ5427. (B) MLE fits for $U + C$ and a^*.

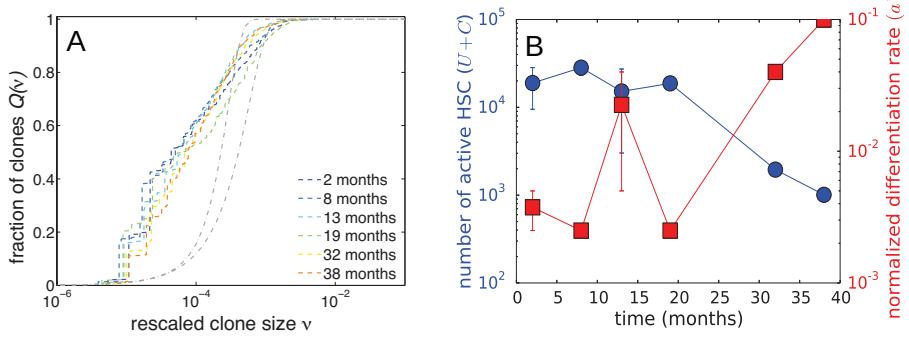

Fig. A3: (A) Rescaled granulocyte clone size distributions from animal RQ3570. (B) MLE fits for $U + C$ and a^*.

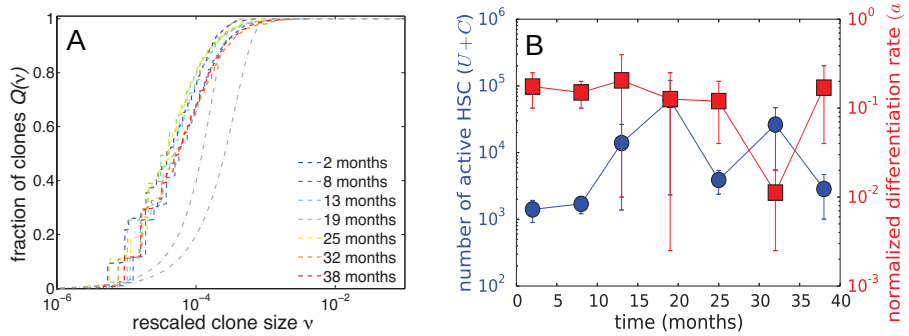

Fig. A4: (A) Rescaled PBMC clone size distributions from animal RQ3570. (B) MLE fits for $U + C$ and a^*.

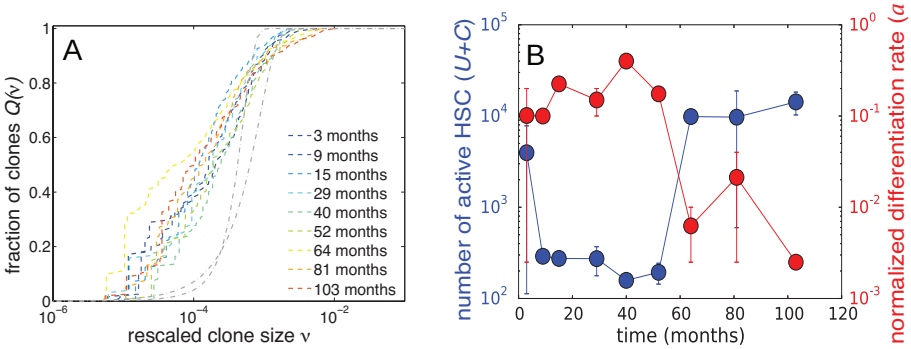

Fig. A5: (A) Rescaled clone size distributions from granulocytes in animal 2RC003. The stationarity in this subpopulation is evident. (B) The MLE fits for $U + C$ (blue circles) and a^* (red squares).

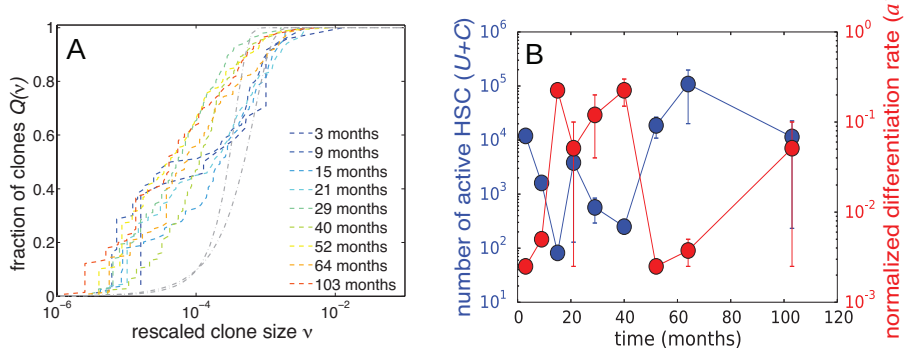

Fig. A6: (A) Rescaled PBMC clone size distributions from animal 2RC003. (B) MLE fits for $U + C$ and a^*.

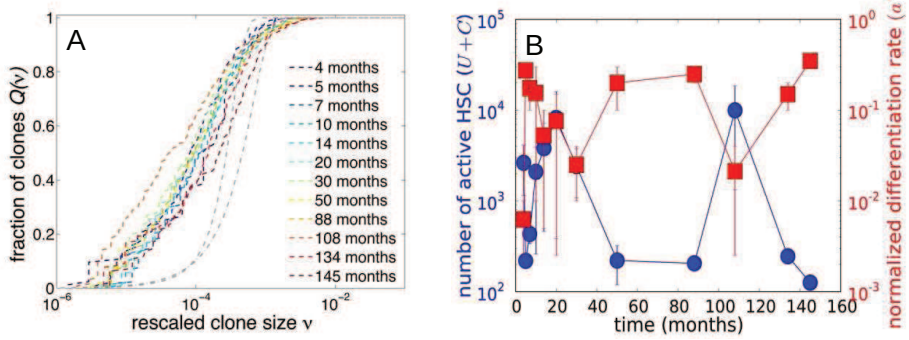

Fig. A7: For animal 95E132 clone sizes were measured only for within the total peripheral blood cells (PBC) population. (A) Rescaled clone size distributions from animal 95E132. (B) MLE fits for $U + C$ and a^* from the combined PBC clone size distributions.

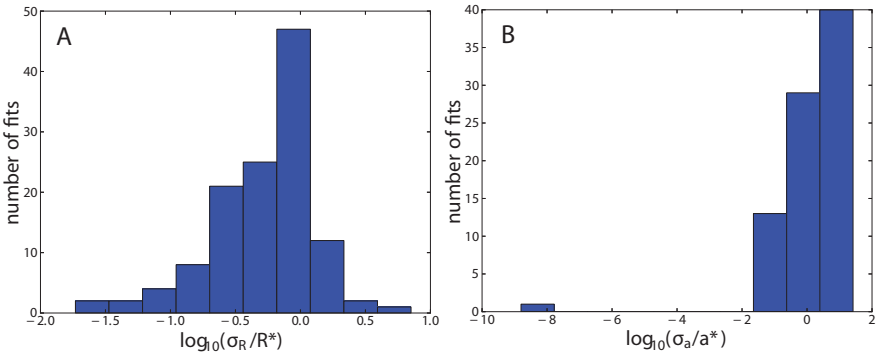

Fig. A8: (A) Distribution of $\log(\sigma_R/R^*)$ values across all data sets. (B) Distribution of $\log(\sigma_a/a^*)$ values.

For completeness, we also calculate a rough goodness-of-fit metric. We do this by calculating the "diagonal" curvatures of the likelihood function $\partial^2\mathcal{L}/\partial R^2$ and $\partial^2\mathcal{L}/\partial a^2$ evaluated at the maximum (R^*, a^*). Upon defining

$$\sigma_R = \left[\left(\frac{\partial^2\mathcal{L}}{\partial R^2}\right)_{R^*,a^*}\right]^{-1/2}, \quad \sigma_a = \left[\left(\frac{\partial^2\mathcal{L}}{\partial a^2}\right)_{R^*,a^*}\right]^{-1/2}, \tag{A67}$$

a goodness-of-fit can be measured through the distribution of the values of the Fano factors σ_a/a^* and σ_R/R^* obtained by fitting each clone size distribution at each time point. The distributions of the logarithm of σ_a/a^* and σ_R/R^* (sampled from fitting at all times points for all animals) are plotted below. We see that the fitting for R at most time points is reasonably good, but that some of the fits, particularly for the small values of a^*, are not particularly well-conditioned.

Additional file references

54. Klein AM, Doup DP, Jones PH, Simons BD. Mechanism of murine epidermal maintenance: cell division and the voter model. Phys Rev E. 2008;77:031907.

55. Klein AM, Nikolaidou-Neokosmidou V, Doup DP, Jones PH, Simons BD. Patterning as a signature of human epidermal stem cell regulation. J R Soc Interface. 2011;8:181524.

56. Chou T, Wang Y. Fixation times in differentiation and evolution in the presence of bottlenecks, deserts, and oases. J Theor Biol. 2015;372:6573.