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 \( \ell \) cells. At any time, the probability that the configuration \( s_{j\ell} \) is observed in a sample of \( S \) cells can be written

\[
P(\{s\}) = S! \prod_{\ell=1}^{N_d} \prod_{j=1}^{y_\ell} \left( \frac{\ell}{N_d} \right)^{s_{j\ell}} \delta_{s_{j\ell}},
\]  

(A1)

where \( X = \sum_{\ell=1}^{N_d} \sum_{j=1}^{y_\ell} s_{j\ell} \), the factor \( N_d^{-s_{j\ell}} S! \) represents the probability that \( s_{j\ell} \) cells were drawn from the \( N_d \) within a sample of \( S \) cells, and \( \binom{y_\ell}{s_{j\ell}} \) 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

\[
m_k = \sum_{n=1}^{N_d} \sum_{m=1}^{y_n} \delta_{k,x_{mn}} .
\]  

(A2)

The expected value of this quantity is

\[
\langle m_k(t) \rangle = \sum_{\{s\}} P(\{s\}) \sum_{\ell=0}^{N_d} \sum_{j=1}^{y_\ell} \delta_{s_{j\ell}} .
\]  

(A3)

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

\[
G(z, t; S) = \sum_{\{s\}} P(\{s\}) \sum_{\ell=0}^{N_d} \sum_{j=1}^{y_\ell} z^{s_{j\ell}} = \frac{\partial}{\partial \beta} \sum_{\{s\}} P(\{s\}) \exp \left[ \beta \sum_{\ell=0}^{N_d} \sum_{j=1}^{y_\ell} s_{j\ell} \right] \bigg|_{\beta=0} .
\]  

(A4)

After using the Fourier representation of the Kronecker \( \delta \) function in Eq. A1, \( 2\pi \delta_n = \int_0^{2\pi} e^{iqn} dq \), we can further reduce the generating function to

\[
G(z, t; S) = \frac{\partial}{\partial \beta} \int_0^{2\pi} dq \frac{2\pi}{2\pi} e^{-iq0} \prod_{\ell=0}^{N_d} \prod_{j=1}^{y_\ell} \left( \frac{t}{N_d} \right)^{s_{j\ell}} \exp \left[ \beta z \right] \bigg|_{\beta=0} = \sum_{\ell=0}^{N_d} \sum_{j=1}^{y_j} \frac{y_j!}{2\pi} \int_0^{2\pi} dq e^{-iq0} \sigma(\ell, z) \prod_{i=1}^{y_j} \sigma(i, 1) \right]^{\frac{y_j}{N_d}},
\]  

(A5)

where

\[
\sigma(\ell, z) = \frac{\ell}{N_d} e^{i\ell q}. \quad (A6)
\]

Note that \( \sigma(\ell, z) \equiv \frac{1 + e^{i\ell q/N_d}}{N_d} \), and that \( \prod_{i=1}^{y_j} \sigma(i, 1) \right]^{\frac{y_j}{N_d}} \approx \left( 1 + e^{i\ell q/N_d} \right)^{N_d} \). Since \( N_d \gg S \gg 1 \), and \( N_d \sim 10^9 \sim 10^{10} \), we can take the large \( N_d \) limit before the large \( S \) limit to find \( 1 + e^{i\ell q/N_d} \right]^{N_d} \approx \exp \left[ e^{i\ell q} \right] \), \( \sigma(\ell, z) \approx \exp \left[ e^{i\ell q} \right] \), and
\[ G(z, t; S) \approx \sum_{\ell=0}^{\infty} y_{\ell}(t) S! \int_{0}^{2\pi} \frac{dq}{2\pi} \exp\left[A_{\ell}(z) e^{iq}\right] e^{-iqS}, \] (A7)

where \( A_{\ell}(z) = 1 + \ell(z - 1)/N_4 \). Note that the integral is simply Euler’s integral for \( 1/(S + 1) \). Namely, we find

\[
\int_{0}^{2\pi} \frac{dq}{2\pi} \exp\left[A_{\ell}(z) e^{iq}\right] e^{-iqS} = \frac{iA_{\ell}^{S}(z)}{2\pi} \int_{\mathbb{C}} d\xi \xi^{-(S+1)}
\]

\[
= A_{\ell}^{S}(z) \Gamma(S + 1).
\] (A8)

Since \( A_{\ell}^{S} = (1 + \ell(z - 1)/N_4)^S = (1 + (S/N_4)\ell(z - 1)/S))^S \approx e^{\ell S(z - 1)/N_4} \) for \( S \gg 1 \), we find, for \( \epsilon \equiv S/N_4 \ll 1 \),

\[
G(z, t; S) = \sum_{\ell=0}^{N_4} y_{\ell}(t) A_{\ell}^{S}(z) \approx \sum_{\ell=0}^{\infty} y_{\ell}(t) e^{\ell S(z - 1)},
\] (A9)

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

\[
F(q, t) \equiv \sum_{j=0}^{q} \sum_{n=0}^{\infty} e^{-nt} \frac{(n\epsilon)^j}{j!} y_{n}(t).
\] (A10)

The expected clone size distribution is thus defined as

\[
\langle m \rangle(t) = F(k, t) - F(k - 1, t) = \sum_{\ell=0}^{\infty} e^{-\ell \epsilon} \frac{(\ell \epsilon)^k}{k!} y_{\ell}(t).
\] (A11)

In general, further development of \( F(k, t) \) and \( \langle m \rangle(t) \) requires numerical solution of \( c_{\ell}(t) \) and \( y_{\ell}(t) = \sum_{n=0}^{\infty} y_{n}(t)^{k\ell}(t) \). The time-dependence of \( F(q, t) \) is further complicated by the time-dependence of \( \epsilon(t) = S/N_4(t) \), 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

\[
\langle m_k m_{k'} \rangle = \sum_{\langle s \rangle} P(\{s\}) \sum_{\ell=0}^{N_4} \sum_{j=1}^{y_{\ell}} \delta(s_{\ell} - k) \sum_{j'=0}^{N_4} \sum_{k'=1}^{y_{j'}} \delta(s_{j'} - k').
\] (A12)

This calculation requires evaluation of the two-dimensional generating function

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

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

\[
G(z, z', t) = \sum_{\ell,k} A_{\ell}^{S}(z) y_{\ell}(t) y_{\ell}(t) - \sum_{\ell} B_{\ell}^{S}(z, z') y_{\ell}(t) + \sum_{\ell} C_{\ell}^{S}(z z') y_{\ell}(t),
\] (A14)

where
where size distribution

Before this solution can be effectively used in Eq. A10, an explicit expression for the steady-state progenitor clone

explicitly indicating that the relative fluctuations in the measured number of clones of size

The total steady-state progenitor population is given by the solution to Eq. 5 to find

The diagonal variance is simply

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

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

\[
\sqrt{\frac{\langle m_k^2(t) \rangle - \langle m_k(t) \rangle^2}{\langle m_k(t) \rangle}} \lesssim \frac{1}{\sqrt{\langle m_k(t) \rangle}}
\]  

(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

\[
y_n = \sum_{k=0}^{\infty} y_n^{(k)} = \sum_{k=0}^{\infty} \left(\frac{uk}{n!}\right)^n e^{-uk} c_k, \quad w = \frac{(1 + \eta)\omega}{\mu_d}.
\]  

(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 \(\alpha C + [r(N_p) - \mu]N_p = 0\). The population-limited growth rate is given by Eq. 3 and the steady-state progenitor cell population is explicitly

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

(A19)

where \(\mu = \mu_\nu + \eta\omega\). After using Eq. A19 in Eq. 3, we find explicitly
\[ r(N_p) = \frac{\alpha}{p} \frac{U + C}{K} + \frac{p}{1 + \left( \frac{\alpha}{p} \frac{U + C}{K} + \frac{p}{1} \right)^2 + \frac{4\mu}{p} \frac{U + C}{K}} < \mu. \]  

(A20)

The total differentiated cell population found from \( \dot{N}_d = (1 + \eta)\omega N_p - \mu d N_d \approx 0 \) is \( N_d = (1 + \eta)\omega N_p / \mu d \equiv \omega N_p. \) Upon using these expressions in the steady-state limit of Eq. 1, we obtain

\[ c_{k\geq 1} = \frac{\alpha C}{k!} \frac{(1 - \bar{r})^a}{(a + k)^a} \prod_{\ell=1}^{k} (a + \ell), \]

\[ c_0 = C - \sum_{k=1}^{\infty} c_k = C(1 - \bar{r})^a, \]

(A21)

where

\[ a \equiv \frac{\alpha}{r} \equiv \frac{\alpha}{\mu r} \text{ and } \bar{r} \equiv \frac{r(N_p)}{\mu}. \]

(A22)

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

\[ C_p = \sum_{k=1}^{\infty} c_k = C \left[ 1 - (1 - \bar{r})^a \right], \]

\[ C_d = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} y_n^{(k)} = \sum_{k=1}^{\infty} (1 - e^{-w k}) c_k \]

\[ = C \left[ 1 - \left( \frac{1 - \bar{r}}{1 - \bar{r} e^{-w}} \right)^a \right], \]

(A23)

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

\[ C_s = \sum_{j=1}^{\infty} \langle m_j \rangle = \sum_{j=1}^{\infty} \sum_{\ell=1}^{\infty} e^{-\ell \varepsilon} \left( \frac{\ell \varepsilon}{j} \right)^j y_{\ell j} \]

\[ = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} (1 - e^{-\ell \varepsilon}) \left( \frac{w k}{\ell} \right)^{\ell - w k} c_k \]

\[ = \sum_{k=1}^{\infty} (1 - e^{-w k - \varepsilon}) c_k \]

\[ = C \left[ 1 - \left( \frac{1 - \bar{r}}{1 - \bar{r} e^{-w (1 - e^{-\varepsilon})}} \right)^a \right], \]

(A24)

where we have explicitly used Eq. A18 for \( y_{\ell j} \) and Eq. A21 for \( c_k \). This result can be further reduced in two limits

\[ C_s \approx \begin{cases} 
C \left[ 1 - \left( \frac{1}{\bar{r} \bar{w} e} \right)^a \left( \frac{1 - \bar{r}}{2 - \bar{r}} \right)^a \right], & \varepsilon w \gg \left( \frac{1 - \bar{r}}{\bar{r}} \right) \\
\alpha C \frac{\varepsilon w \bar{r}}{1 - \bar{r}}, & \varepsilon w \ll \left( \frac{1 - \bar{r}}{\bar{r}} \right). 
\end{cases} \]

(A25)

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

\[ C \gtrsim C_p \gtrsim C_d > C_s, \]

(A26)

with significant loss of clones due to sampling \( (C_d \gg C_s) \) only in the second case of Eq. A25 describing sample sizes \( S \ll \mu d N_d (1 - \bar{r}) / (\omega \bar{r}). \) Note that for clone sizes \( k \gg a, c_k \) in Eq. A21 can be approximated by
\[ c_{k \geq a} \approx A k^{a-1} e^{-k} \left[ 1 - \frac{a(1-a)}{2k} + O \left( \frac{1}{k^2} \right) \right], \]  

(\text{A27})

where

\[ A \equiv \frac{a^C}{\Gamma(a+1)} (1 - \bar{r})^a. \]  

(\text{A28})

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

\[ F(q) \equiv \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{(j\varepsilon)^j}{j!} \frac{(wk)^k}{k!} e^{-w} e^{-w k} c_k \]

\[ \equiv \sum_{k=0}^{\infty} \frac{(q+1, \varepsilon w k)}{\Gamma(q+1)} e^{-w k} c_k \]

\[ \approx \sum_{k=0}^{\infty} \frac{\Gamma(q+1, \varepsilon w k)}{\Gamma(q+1)} c_k, \]  

(\text{A29})

where a steepest descents approximation in \( \epsilon \ll 1 \) was used to derive the final approximation. The expected cumulative clone size frequency \( Q(q) \) is obtained by subtracting off the unrepresented clones \( (m_0) \equiv F(0) \) and normalizing by the total expected number of clones \( C^* = F(S) - F(0) \).

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 \( \varepsilon w \), 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

\[ F(q) \approx \sum_{k=0}^{\infty} \frac{\Gamma(q+1, \varepsilon w k)}{\Gamma(q+1)} c_k \approx c_0 + \int_{1}^{(q+1)/(\varepsilon w)} c_0 dk, \]  

(\text{A30})

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

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

(\text{A31})

This approximation shows that our distributions depend most strongly on only \( a = \alpha/\tau \) 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 \( A/\ln^a(1/\bar{r}) \approx a C \). Since \( (m_0) = F(q) - F(q-1) \), an approximate form useful for estimating the clone size distribution is

\[ (m_0) \approx a C [\Gamma(a, q/R) - \Gamma(a, (q+1)/R)]. \]  

(\text{A32})

Within physiologically-relevant regimes, our data can be well-fitted to \( (m_0) \) 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, \( \alpha/p \). Once the best fit parameters \( a^* \) and \( R^* \) have been estimated from fitting clonal frequency distributions, \( U + C \) and \( \alpha/p \) 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 \( \varepsilon w/R^* = (S/N_p)(\omega/\mu_\Delta)/R^* = S/(R^* N_p) \ll 1 \) and \( S/(R^* K) \ll 1 \). By using the definition of \( R \), we find \( \bar{r} = e^{-S/(R^* N_p)} \approx 1 - S/(R^* N_p) \). Since \( \bar{r} = (\Delta + 1)K/(N_p + K) \) also, these two independent expressions for \( \bar{r} \) furnish a quadratic equation for \( N_p \) in terms of \( R^* \). After comparing the positive root of this equation to the definition of the steady-state progenitor population \( N_p \) (Eq. A19), we find

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

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

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

(A34)

Note that to lowest order, $\alpha/p$ 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

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

(A35)

which is independent of $K$ and $\Delta$. 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 $\epsilon w \ll 1$. Our model allows one to make predictions on the number of expected clones in each pool, $C_p, C_t$, and the measured $C_p$ (Eqs. A23-A24), as well as expected clone size distributions (Eq. A32) as functions of sampling fraction $\epsilon$, turnover rate $\omega$, effective differentiation rate $a$, and effective growth rate $\bar{r}$. However, from the functional forms of $C_p, C_t, C_s$, 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 $\alpha$.

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 only on 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 $\alpha$. The master equation for the probability $p_k(t)$ of a single clone containing $k$ progenitor cells is

$$\frac{dp_k(t)}{dt} = \alpha [p_{k-1} - p_k] + r \left[ (k-1)p_{k-1} - lp_k \right] + \mu \left[ (k+1)p_{k+1} - kp_k \right],$$

(A36)

where in our application, $\mu \equiv \mu_0 + \eta \omega$. If the growth rate $r$ is assumed constant and independent of $k$, an analytic expression in terms of the corresponding generating function $\phi(z, t) \equiv \sum_{k=0}^\infty p_k z^k$ [45, 56]:

$$\phi(z, t) = \left( \frac{1 - \bar{r}}{1 - rz} - \frac{1 - \bar{r}}{1 - rz} e^{-\eta(1-\bar{r})t} \right)^a.$$

(A37)

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

$$\sum_{k=0}^\infty c_k z^k \approx C \phi(z, t)$$

(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

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

(A39)

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

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

(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 \( K_s \). If we denote \( x_k \) as the number of clones in the stem cell niche that is represented by exactly \( k \) stem cells,

\[
\frac{dx_1}{dt} = -(r_s + \mu_s)x_1 + 2\mu_s x_2 \\
\frac{dx_k}{dt} = r_s [(k - 1)x_{k-1} - kx_k] + \mu_s [(k + 1)x_{k+1} - kx_k],
\]

where the effective growth rate \( r_s \) is defined by the carrying capacity \( K_s \), and the total number of stem cells \( N_s \), labeled and unlabeled, in the stem cell compartment:

\[
r_s(N_s) = p_s \left(1 - \frac{N_s}{K_s}\right),
\]

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

\[
N_s(t) = U_s(t) + \sum_{k=1}^{\infty} kx_k(t).
\]

Upon summing Eqs. A41 and the equation for unlabeled cells, \( U_s = r_s U_s - \mu_s U_s \), we find that the total population decouples and obeys

\[
\frac{dN_s}{dt} = r_s(N_s)N_s - \mu_s N_s,
\]

which can be solved exactly, allowing one to find \( r_s \) 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 \( \mu_s = 0 \) and define a new time variable \( d\tau = r_s(t)dt \). Equations A41 for \( x_k(\tau) \) now have constant coefficients and can be solved by using the initial conditions \( x_1(\tau = 0) = C_s, x_{k>1}(0) = 0 \), and the Laplace transforms,

\[
\tilde{s}\tilde{x}_1 - C_s = -\tilde{x}_1, \quad \tilde{s}\tilde{x}_k = (k - 1)\tilde{x}_{k-1} - k\tilde{x}_k.
\]

The solution

\[
\tilde{x}(s) = \frac{C_s}{s + 1}, \quad \tilde{x}_k(s) = \frac{(k - 1)C_s}{\prod_{j=1}^{k-1}(s + j)}.
\]

can be inverted to yield

\[
x_1(\tau) = C_s e^{-\tau}, \quad x_k(\tau) = C_s (e^\tau - 1)^{k-1} e^{-k\tau}.
\]

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

\[
\tau(t) = \int_0^t r_s(t') dt' = p_s t - \ln(1 + C_s(e^{p_s t} - 1)/K_s).
\]

In the steady-state limit, \( t \to \infty \) corresponds to \( \tau \to \ln(K_s/C_s) + O(e^{-p_s t}) \). In this limit, Eq. A46 yields

\[
x_k(t \to \infty) = C_s f_s (1 - f_s)^{k-1}, \quad f_s \equiv \frac{C_s}{K_s}.
\]
The clone numbers in the progenitor cell population are modeled using

\[
\begin{align*}
\frac{dc_k^{(j)}}{dt} &= \alpha j \frac{x_j}{C_s} (c_0 - c_k^{(j)}) - (r + \mu_p) c_k^{(j)} + 2\mu_p c_k^{(j)}_1, \\
\frac{dc_k^{(j)}}{dt} &= \alpha j \frac{x_j}{C_s} (c_k^{(j)}_2 - c_k^{(j)}),
\end{align*}
\]

where \( c_0(t) = C_s - \sum_{j=1}^{\infty} c_k^{(j)}(t) \) 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

\[
\frac{d}{dt} \left( \sum_{k=1}^{\infty} c_k^{(j)}(t) \right) = A_j c_0(t) - \mu_p c_k^{(j)},
\]

and

\[
\frac{d}{dt} \left( \sum_{k=1}^{\infty} k c_k^{(j)}(t) \right) = A_j \sum_{k=1}^{\infty} c_k^{(j)} + (r_p(N_p) - \mu_p) \sum_{k=1}^{\infty} k c_k^{(j)}(t),
\]

where \( A_j(t) = \alpha j x_j(t) / C_s \). By further adding \( \dot{U}_p = \alpha U_p + (r_p - \mu_p) U_p \) to Eq. A51, we find

\[
\frac{dN_p}{dt} = \alpha U_p + \frac{\alpha}{C_s} \sum_{j,k=1}^{\infty} j x_j(t) c_k^{(j)}(t) + (r_p(N_p) - \mu_p) N_p(t).
\]

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

\[
c_0^{(j)} = \frac{A_j}{\mu_p} c_0 = j \frac{\alpha}{\mu_p} f_s(1 - f_s)^{-1} c_0,
\]

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

\[
c_k^{(j)} = \frac{c_0^{(j)}}{k!} \prod_{\ell=1}^{k} [a_j + (\ell - 1)],
\]

where

\[
a_j = \frac{2x_j}{C_s} \frac{\alpha}{\tau} = j f_s(1 - f_s)^{1-1} \frac{\alpha}{\tau}.
\]

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

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

and

\[
c_k^{(j)} = \frac{C_s}{Z k!} \prod_{\ell=1}^{k} [a_j + (\ell - 1)],
\]
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, \( f_s \to 1^- \), \( Z \to (1 - \tilde{r})^{-\alpha} \), 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

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

(A58)

when \( a_j \ll \ln k \). From Eq. A55, we know that \( a_j \) is strictly bounded above by \( \alpha / r \) and is typically \( \lesssim 0.5 \alpha / r \) for \( f_s \ll 0.5 \). 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

\[
c_k \approx \frac{e^p k}{Z} \sum_{j=1}^{\infty} \frac{k^{a_j}}{\Gamma(a_j)}
\]

(A59)

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

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

(A60)

where \( a \equiv \alpha / r \) and \( \gamma \approx 0.5772 \) is Euler’s constant. In order to find an explicit expression for \( Z \), note that \( \delta \equiv 1 - \tilde{r} \) is typically very small, on the order of \( 1 / K_s \). Therefore \( Z = 1 + \sum_{j=1}^{\infty} [\delta^{-a_j} - 1] \) 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

\[
Z \approx \frac{\sqrt{2 \pi e (1 - f_s)} \exp \left( \frac{a f_s}{(2 - f_s)(2 - f_s) \ln(1 - f_s)} \right)}{\sqrt{\alpha f_s \ln(1 - f_s) \ln(1 - f_s)}}
\]

(A61)

By using this approximation for \( Z \) 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,

\[
c_k \approx a K_s (1 - \tilde{r})^{\alpha / r} \frac{a}{2 \pi e} \ln \left( \frac{1}{1 - \tilde{r}} \right) \frac{p^k}{k}.
\]

(A62)

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

\[
Z \approx 1 - \frac{a}{f_s} \ln(1 - \tilde{r}),
\]

(A63)

and

\[
c_k \approx \left[ 1 + \frac{a(2 - f_s)(2 - f_s)(\gamma + \ln k)}{(2 - f_s)^2} \right] \frac{a K_s}{1 - a \ln(1 - \tilde{r})/f_s} \frac{p^k}{k}.
\]

(A64)

Thus, in the limit where the transplanted number of clones \( C_k \ll K_s \) 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:

$$a \equiv \frac{\alpha}{r}, \quad \bar{r} \equiv \frac{r}{\mu}, \quad \text{and} \quad \varepsilon w \equiv \frac{S}{N_d} \times \frac{\omega}{\mu d}. \quad (A65)$$

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:

$$P(q_i|a, \bar{r}, \varepsilon w) \approx \langle m_{q_i}(a, \bar{r}, \varepsilon w) \rangle = F(q_i|a, \bar{r}, \varepsilon w) - F(q_{i-1}|a, \bar{r}, \varepsilon w).$$

The likelihood of the parameters given the detected clone sizes $\{q_1, \ldots, q_n\}$ is then given by:

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

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 $U + C \sim 10^3 - 10^5$. 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. 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 2RC03. 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.
For completeness, we also calculate a rough goodness-of-fit metric. We do this by calculating the "diagonal" curvatures of the likelihood function $\frac{\partial^2 L}{\partial R^2}$ and $\frac{\partial^2 L}{\partial a^2}$ evaluated at the maximum $(R^*, a^*)$. Upon defining

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

a goodness-of-fit can be measured through the distribution of the values of the Fano factors $\sigma_a/a^*$ and $\sigma_R/R^*$ obtained by fitting each clone size distribution at each time point. The distributions of the logarithm of $\sigma_a/a^*$ and $\sigma_R/R^*$ (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**

