Data S1: Supplementary analysis of neutral evolution. Related to figure 4 and *STAR* methods.

Our framework of HSC proliferation allows for the explicit calculation of the number and frequencies of neutral variants in the HSC pool. First we analyze the Moran process describing the turnover phase of HSC proliferation after birth. We derive the frequency distribution of neutral variants arising after birth and the expected number of these variants in each HSC. Then we analyze a Yule process describing the growth phase of HSC establishment before birth and assess the fate of neutral variants arising before birth. Finally we combine our results from both analyses to calculate the overall frequency distribution of neutral variants in an adult patient. We evaluate the sensitivity of this result to the HSC proliferation rate across a patient's lifespan.

During the growth phase, a single cell (the zygote) grows to N cells by repeated division according to a Yule process (Otto and Day, 2007). Then, beginning at patient age 0, the critical phase sees the population of N cells evolve according to a continuous-time Moran process with proliferation rate $b(t)$. Throughout both phases, we track the dynamics of neutral variants in individual cells. At each division, the two daughter cells inherit all parental variants and independently may receive an random number of additional variants that is Poisson distributed with mean equal to the mutation rate u. Variants adhere to the infinite sites assumption (Nowak 2006, Wakeley 2008), such that a variant arising in a particular cell will be present in all its descendants but in no other cells.

Moran transition framework

We describe the trajectories of variant frequencies in terms of the transition probabilities of the Moran process. To define these transition probabilities, we construct the transition rate matrix Q of a standard Moran process with unit per-cell division rate by writing its entries $Q = (q_{i,j})_{i,j=0}^{N}$as

$$q_{i,j} = \begin{cases} i\frac{N-i}{N-1}, & j = i-1 \text{ or } j = i+1; \\ -2i\frac{N-i}{N-1}, & j = i; \\ 0, & \text{otherwise.} \end{cases}$$

We then obtain the corresponding transition rate probabilities p from the matrix exponential of Q, so that $(p_{i,j}(t))_{i,j=0}^{n} = e^{tQ}$. For a time-homogeneous Moran process, $p_{i,j}(t)$ is the probability that i given cells at time 0 have exactly j descendants at time t. We note that an equivalent expression for $p_{i,j}(t)$ can be given in terms of the Hahn polynomials (Karlin and McGregor, 1962). Since our model has an inhomogenous division rate $b(t)$, each of the $p_{i,j}(t)$ need to be rescaled in time: the probability that i given cells at time s have exactly j descendants at time $t > s$ is given by

$$P_{i,j}(s,t) = p_{i,j}\left(\int_s^t b(\tau)d\tau\right). \tag{9}$$

Site frequency spectrum

We define the function $\eta(j,t)$ to be the number of variants present in exactly j HSCs at time t; this is the *site frequency spectrum*. The following result gives the expected site frequency spectrum.

Proposition 1.1. For $j \in \{1,..,N\}$ and $t \geq 0$,

$$\mathbb{E}[\eta(j,t)] = \sum_{i=1}^{N} \mathbb{E}[\eta(i,0)] P_{i,j}(0,t) + 2Nu \int_0^{\int_0^t b(s)ds} p_{1,j}(s)ds.$$

The initial condition $\eta(i,0)$ at birth is discussed in Section "*Variant frequencies at birth.*"

Proof of Proposition 1.1. Let $\mathcal{M}$ denote the set of possible variants. Each cell is assigned a subset of $\mathcal{M}$ which denotes the variants carried by the cell. Denote the initial cells by $1,..,N$, and write $\mathcal{M}_r \subset \mathcal{M}$ for the variants carried by cell r. For a variant $z \in \mathcal{M}$, write

$$I_z = \{r \in \{1,..,N\} : z \in \mathcal{M}_r\}$$

for the subset of intitial cells which carry variant z, and write $X_r(t)$ for the number of descendants of cell r at time t. Then the number of variants that are both present at birth and present in exactly j cells at time t is

$$\eta_0(j,t) = \sum_{z\in\mathcal{M}} 1_{\{\sum_{r\in I_z} X_r(t)=j\}}.$$

Conditioning on the initial variants and applying linearity of expectation,

$$\mathbb{E}[\eta_0(j,t)|(\mathcal{M}_r)] = \sum_{z\in\mathcal{M}} \mathbb{P}\left[\sum_{r\in I_z} X_r(t) = j\right] = \sum_{z\in\mathcal{M}} P_{|I_z|,j}(0,t) = \sum_{i=1}^{N} \eta(i,0) P_{i,j}(0,t).$$

Then, taking the unconditional expectation, we obtain

$$\mathbb{E}[\eta_0(j,t)] = \sum_{i=1}^{N} \mathbb{E}[\eta(i,0)] P_{i,j}(0,t). \tag{10}$$

Next we count the variants arising after birth. Denote the two daughter cells of the kth cell division by $(k,1)$ and $(k,2)$, for $k \in \mathbb{N}$. Write $M_{k,l}$ for the number of new variants which arise at the birth of cell (k,l), and write $X_{k,l}(t)$ for the number of descendants of cell (k,l) time t after its birth. Let T_k be the time of the kth cell division, and let $D(t) = |\{k \in \mathbb{N} : T_k < t\}|$ be the number of cell divisions before time t. Then the number of variants that arise after birth and are present in exactly j cells at time t is

$$\eta_+(j,t) = \sum_{k=1}^{D(t)} \sum_{l=1}^{2} M_{k,l} 1_{\{X_{k,l}(t-T_k)=j\}}.$$

Evaluating the expectation conditioned on cell division times, this means that

$$\begin{aligned}\mathbb{E}\left[\eta_+(j,t)|D(t)=d,(T_k)_{k=1}^d=(t_k)_{k=1}^d\right] =& \sum_{k=1}^{d}\sum_{l=1}^{2} \mathbb{E}[M_{k,l}]\mathbb{P}[X_{k,l}(t-t_k)=j] \\ =& 2U\sum_{k=1}^{d} P_{1,j}(t_k,t).\end{aligned} \tag{11}$$

Observe that the division times come from a Poisson process on $[0,\infty)$ with intensity $Nb(\cdot)$. Hence $D(t)$ is Poisson distributed with mean $N\int_0^t b(s)ds$, and when conditioned on $D(t)=d$, the unordered times are independent and identically distributed on the interval $[0,t]$ with density proportional to $b(\cdot)$. Integrating over the distribution of cell division times, Eq. (11) becomes

$$\mathbb{E}[\eta_+(j,t)] = 2Nu\int_0^{\int_0^t b(s)ds} p_{1,j}(s)ds. \tag{12}$$

Finally, we sum Eqs. (10) and (12) to account for both variants arising before and after birth, $\eta(j,t) = \eta_0(j,t) + \eta_+(j,t)$, giving the result. □

Expected number of variants in each cell

For a given site frequency spectrum $\eta(j,t)$, the mean number of variants per cell at time t is

$$H = \frac{1}{N}\sum_{j=1}^{N} j\eta(j,t).$$

Averaging over all possible site frequency spectra, we obtain its expected value as follows.

Proposition 2.2. The expected number of variants per cell at time t is

$$\mathbb{E}[H] = \frac{1}{N}\sum_{j=1}^{N} j\mathbb{E}[\eta(j,0)] + 2u\int_0^t b(s)ds.$$

Proof. Using the result from Proposition 1 and interchanging expectations, summations, and integrals, we evaluate the expected number of variants per cell,

$$\mathbb{E}[H] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\eta(i,0)]\sum_{j=1}^{N} jP_{i,j}(0,t) + 2u\int_0^{\int_0^t b(s)ds}\sum_{j=1}^{N} jp_{1,j}(s)ds.$$

As $\sum_{j=1}^{N} jp_{i,j}(s)$ is just the expected number of descendants of i cells, we know that this must be equal to i in the neutral Moran process. □

Remark 2.3. The second term of the above result for the expected number of variants per cell can alternatively be derived by counting cell divisions. Choose a cell at time t and trace its lineage back to time 0. The expected number of cell divisions in this lineage is $2\int_0^t b(s)ds$. Multiply this by the expected number of variants per cell division, u, to obtain the second term.

Remark 2.4. The rate at which variants accumulate per cell is

$$\frac{d}{dt}\mathbb{E}[H] = 2ub(t),$$

which is the product of the proliferation rate and the expected number of variants per cell division.

Variant frequencies at birth

To generate the initial distribution of variant frequencies $\eta(j,0)$, we suppose that the population grows from one cell to $\bar{N}$ cells as a Yule process, and then N of the $\bar{N}$ cells are randomly selected as the HSC pool at birth. During this Yule process, as in the Moran process, each new cell receives a Poisson number of new variants with mean u. The next result gives the expected site frequency spectrum at birth, which initializes the Moran process as described in the previous section.

Proposition 3.5. For all variant frequencies $j \in \{1,..,N\}$,

$$\mathbb{E}[\eta(j,0)] = \frac{2\bar{N}u}{\binom{\bar{N}}{N}}\sum_{i=j}^{\bar{N}-N+j}\frac{\binom{i}{j}\binom{\bar{N}-i}{N-j}}{i(i+1)}.$$

Proof. In the Yule process, write $\psi_k(j)$ for the number of variants present in exactly j cells when the total number of cells reaches k, for $1 \leq j < k \leq \bar{N}$. Note that $\mathbb{E}[\psi_2(1)] = 2u$, which is the expected number of variants to arrive at the first cell division. We now prove by induction that

$$\mathbb{E}[\psi_k(j)] = \frac{2ku}{j(j+1)}. \tag{13}$$

Fix k and suppose that Eq. (13) is true for all $j \leq k-1$. Give each of the variants present when there are k cells an individual name by letting $[1,j],..,[\psi_k(j),j]$ be the variants present in exactly j cells. Then write $C_{i,j}$ for the number of cells which contain variant $[i,j]$ after the next cell division (that is the cell division which takes the total number of cells from k to $k+1$). For $j \geq 2$,

$$\psi_{k+1}(j) = \sum_{i=1}^{\psi_k(j-1)} 1_{\{C_{i,j-1}=j\}} + \sum_{i=1}^{\psi_k(j)} 1_{\{C_{i,j}=j\}}.$$

Applying the expectation operator to each term in this equation,

$$\mathbb{E}[\psi_{k+1}(j)] = \mathbb{E}[\psi_k(j-1)]\mathbb{P}[C_{i,j-1}=j] + \mathbb{E}[\psi_k(j)]\mathbb{P}[C_{i,j}=j]. \tag{14}$$

The distribution of $C_{i,j}$ is given by the probability mass function

$$\mathbb{P}[C_{i,j}=r] = \begin{cases} 1-\frac{j}{k}, & r=j; \\ \frac{j}{k}, & r=j+1. \end{cases}$$

Substituting in the distribution of $C_{i,j}$ and recalling the inductive hypothesis, Eq. (14) becomes

$$\mathbb{E}[\psi_{k+1}(j)] = \frac{2u}{j} + \frac{2u}{j}\,\frac{k-j}{j+1} = \frac{2u}{j}\,\frac{k+1}{j+1}$$

as required. It only remains to check the case $j=1$, which requires consideration of new variants arising at the kth cell division. Write M for the number of new variants at the kth cell division, which is Poisson distributed with mean $2u$. Then

$$\psi_{k+1}(1) = M + \sum_{i=1}^{\psi_k(1)} 1_{\{C_{i,1}=1\}}.$$

Evaluating the expectation of each term as before,

$$\begin{aligned} \mathbb{E}[\psi_{k+1}(1)] &= 2u + \mathbb{E}[\psi_k(1)]\mathbb{P}[C_{i,1}=1] \\ &= 2u + ku(k-1)/k \\ &= u(k+1). \end{aligned}$$

Therefore (13) holds for $1 \leq j < k \leq \bar{N}$. Finally we need to select N cells from the final $\bar{N}$ cells in the Yule process. To see the frequencies of variants after this selection process, denote the variants present in exactly j cells as before: $[1,j],..,[\psi_N(j),j]$ denote the variants present in exactly j cells out of the $\bar{N}$ cells. Write $F_{i,j}$ for the number of cells at the beginning of the Moran process that carry variant $[i,j]$, such that

$$\eta(k,0) = \sum_{j=1}^{\bar{N}-1} \sum_{i=1}^{\psi_{\bar{N}}(j)} 1_{\{F_{i,j}=k\}}.$$

Now taking the expectation of each term and applying Eq. (13),

$$\mathbb{E}[\eta(k,0)] = \sum_{j=1}^{\bar{N}-1} \mathbb{E}[\psi_{\bar{N}}(j)]\mathbb{P}[F_{i,j}=k] = \sum_{j=1}^{\bar{N}-1} \frac{2\bar{N}u}{j(j+1)}\mathbb{P}[F_{i,j}=k].$$

Finally, we observe that $F_{i,j}$ follows a hypergeometric distribution,

$$\mathbb{P}[F_{i,j}=k] = \frac{\binom{j}{k}\binom{\bar{N}-j}{N-k}}{\binom{\bar{N}}{N}},$$

for $\max(0, N-\bar{N}+j) \leq k \leq \min(j, N)$, which completes the derivation. □

Total burden of neutral variants in each cell

The expected site frequency spectrum at time 0, as given by Proposition 3.5, can be substituted into Propositions 1.1 and 2.2 to give the expected site frequency spectrum and mean number of variants per cell at time t. We note that H, the expected number of total variants per cell, is

$$\mathbb{E}[H] = \mathbb{E}\Big[\frac{1}{N}\sum_{j=1}^{N} j\eta(j,t)\Big] = 2u\sum_{i=2}^{\bar{N}}\frac{1}{i} + 2u\int_0^t b(s)ds. \tag{15}$$

where the sum is one less than the $\bar{N}^{\text{th}}$ harmonic number. Since $\bar{N}$ is large (we consider $\bar{N}=10^{13}$ cells), this sum almost exactly equal to $\gamma-1+\ln\bar{N}$, where $\gamma \approx 0.557$ denotes the Euler-Mascheroni constant. The error in this approximation is less than $(2\bar{N})^{-1}$.

An alternative derivation of this result follows by counting the expected number of cell divisions in a randomly chosen cell's lineage. During the Yule process, a cell division that takes the number of cells from k to $k+1$ involves 2 out of the $k+1$ cells. Hence the cell division has a $2/(k+1)$ probability of occuring along a particular lineage, accounting for the harmonic sum in this result.

This result highlights the central role of the HSC proliferation rate in the accumulation of neutral variants. If the proliferation rate $b(t)$ of a patient is higher than usual at any point in time, this will lead to an accumulation of additional variants per cell in the second term of this result. Specifically, if a patient experiences a β-fold increase in their otherwise constant proliferation rate, with onset age T and with $\beta > 1$, then at some later age t they will have an excess of $2u(\beta-1)(t-T)$ neutral variants relative to a patient without this increase.