Estimation of population size and cell life-span parameters of a stem cell population

1 The phylogeny of single cell genomes

We have a sample of cells obtained from a bone marrow aspirate, for each of which we have a single cell genome, obtained using whole genome sequencing. From these single cell genomes it was possible to construct a phylogenetic tree of cells. In population genetic studies we can estimate branch lengths and node ages (measured in years) from mutation counts on branches and estimated mutation rates. Here we know the age of our volunteer, and hence the time interval from our sample of stem cells to the root node of the phylogeny. We can estimate branch lengths and node ages from mutation counts alone, without the need for independent estimates of mutation rates.

In section 3 we describe a model of the stem cell population. This model belongs to a class of multi-type birth-death processes. The *types* alluded to here are the *genotypes* of the stem cells. More specifically, the genotype of a stem cell here refers to the set of mutations which have been identified in the genome of the cell. It is assumed that the birth rates and death rates of cells are the same for all genotypes. In population genetics terminology this is a *neutral* model. Under this model, the phylogeny of cells is a realisation of a coalescent process (Kingman, 1982).

In section 5 we explain how we can use the phylogeny of cells to estimate parameters of the coalescent process. Analysis using the method of Lan et al. (2015) supports a model where the population of stem cells (and their ancestor cells) underwent a rapid growth phase, from the first division of the zygote, to a time t_0 (measured in years) where the stem cell population attained a stable size N.

2 NGS read data from the peripheral blood sample

We have a sample of granulocytes isolated from a sample of peripheral blood. This sample has been split into L sub-samples of equal size. These sub-samples are what we refer to as *biological replicates*. The cells in each biological replicate are lysed, and NGS read data is obtained from the lysate. The biological replicate are kept entirely separate throughout these DNA extraction and sequencing steps.

A bait-set was designed to target a the genomic sites of a set of mutations (labelled 1, ..., m). Each of these m mutations has been placed on the phylogenetic tree of single cell genomes. The positions of these mutations on the phylogenetic tree determines the set of single cell genotypes which are allowed (assuming that each mutation has occurred only once). Only a subset of these genotypes may be represented in the sample of peripheral blood, and we do not know with certainty which of these genotypes are present. We know only which mutations are present in each biological replicate (sub-sample of granulocytes).

The array of read count data is denoted by $\boldsymbol{Y}$. This array contains elements $Y_{j,k,\ell}$, where j (= 1, ..., m) is the mutation, k (= 1, ..., L) is the biological replicate, and ℓ (= 0, 1) indicates the state (wild-type or mutated) of the reads. Thus $Y_{j,k,0}$ is the count of reads, in biological replicate k, which cover the site of mutation j but which do not report mutation j, while $Y_{j,k,1}$ is the count of reads which do report mutation j. The total read depth covering the site of mutation j is $n_{j,k} = Y_{j,k,0} + Y_{j,k,1}$. The array of read depth data is denoted by $\boldsymbol{n}$. The element $n_{j,k}$ in row j and column k of this array is the read depth at the site of mutation j, in biological replicate k.

3 The Moran model applied to a stem cell population

We assume that the population of stem cells remains stable in size during adulthood (see Figure 3), and furthermore, that the population of stem cells behaves to a close approximation like the Moran model of population genetics (Moran, 1958). The Moran model is a kind of birth-death process, in which the population consists of a constant number of individuals N. The original setting for the Moran model was the allele composition at a single haploid locus in an outcrossing population, undergoing *neutral* evolution (the allelic variation at the locus in question is not subject to natural selection). Subsequently, the Moran Model has been used as a model for the genotype composition of an asexual population (see for example Yu et al. (2010)). Here we are

applying the Moran model to the (non-recombining) genomes of an asexual population of stem cells. Here we assume that the genomic variation in the population of stem cells is not subject to natural selection. We apply the *neutral* Moran model, and we make use of the remarkably simple reverse-time genealogical process of the neutral Moran model.

The individuals in the population can be labelled using the integers in the set $\mathbb{N}(N) = \{1, 2, \ldots, N\}$. This labelling is arbitrary. Each individual (allele copy) $i \in \mathbb{N}(N)$ in the population at generation t has a *type* (genotype) $X_t(i)$. The sequence of types $X_t = \{X_t(1), X_t(2), \ldots, X_t(N)\}$ specifies the state of the population at generation t, and this state can change only at discrete generations. At each generation, one individual (chosen at random) dies, and another individual (again chosen at random) gives birth to a daughter, thus replacing the individual who was removed from the population.

In the original Moran model, an individual is chosen at random from the set $\mathbb{N}(N)$, to give birth, and then an individual is chosen at random from the same set to die. So it is possible for the same individual to be chosen to give birth and to die. This makes sense if, in each generation, the birth event precedes the death event. There is an alternative version of the Moran model (Gladstein (1978)) in which an individual is chosen at random from the set $\mathbb{N}(N)$ to die, and then an individual is chosen at random from the set of survivors to give birth. This version of the Moran model makes sense if, in each generation, the death event precedes the birth event. It is this modified version of the Moran model which we consider here.

In this modified version of the Moran model, at each generation t, one individual a, chosen at random $\mathbb{N}(N)$, dies (is removed from the population), and another individual b ($b \neq a$), chosen at random from the remaining $N - 1$ individuals) gives birth to a daughter who inherits the type of her mother. (We ignore for now the process of mutation.) This daughter individual acquires the label a of the individual which it has replaced.

So far we have given the conventional description of the Moran model, in which an individual is said to give birth to a daughter individual, while the mother individual

(which retains its original label) persists to the next generation. We now adopt an alternative description of these *birth* events, which is more natural when we are applying the Moran model to a population of cells. According to this description, in each generation, one cell a, chosen at random $\mathbb{N}(N)$, dies, and another individual b ($b \neq a$), chosen at random from the remaining $N-1$ cells) divides into two daughter cell. One of these daughter cells retains the label b from its mother, while the other daughter cell acquires the label a of the cell which has died. We can now refer to the *life-span* of an individual cell as the time from the birth of the cell (in a cell division) until the cell either dies or undergoes a cell division.

The sampling of pairs of individuals (to give birth and to die) occurs independently in each generation. From this it follows that the state of the population in the next generation depends only on its state in the current generation (and not on any preceding generations). In other words, the Moran model is a Markov chain.

From this description, it follows that for an individual cell chosen at random from the population, at each generation there are three mutually exclusive possibilities. The first possibility is that the cell dies, which occurs with probability $1/N$. The second possibility is that the cell divides to leave two daughter cells. This can only occur only if the chosen cell has not been removed from the population, and so has probability

$$\left(\frac{N-1}{N}\right)\frac{1}{N-1} = \frac{1}{N}.$$

The final possibility is that the cell persists for another generation. This outcome has probability $1 - p_{cell}$, where $p_{cell} = 2/N$.

Let W_{cell} denote the number of (Moran model) generations for which a cell persists, from the generation of its birth (in a cell division), to the generation in which it either dies or divides. Recall that the sampling of pairs of cells (once cell to die and one cell to divide) occurs independently in each generation. Therefore W_{cell} has the geometric distribution

$$\mathbb{P}(W_{cell} = t) = p_{cell}(1 - p_{cell})^{t-1},$$

for $t = 1, 2, \ldots$, with parameter $p_{cell} = 2/N$ and expectation

$$\mathbb{E}\left[W_{cell}\right] = \frac{1}{p_{cell}} = \frac{N}{2}.$$

If the duration of one generation of the Moran model is δ_M (in years), then the mean life-span of a cell δ_{cell}, is

$$\delta_{cell} = \mathbb{E}\left[W_{cell}\right] \delta_M = \frac{N}{2} \delta_M \tag{1}$$

years. The cell life-span parameter δ_{cell} is one of the quantities which is of direct interest to us.

The description of the Moran model as a forward-time Markov chain implies a remarkably simple reverse-time genealogical process. The genealogical process describes what happens to the lines of descent which trace back from a sample of n individuals (cells) sampled (without replacement) from the population.

In every generation of the Moran model, there is a pair of cells who are daughters of a single cell in the previous generation. We may refer to this pair of individuals as the *most recent sisters*. Recall the formula

$$\begin{aligned} h\left(k, k' \,|K, K', n\right) &= h\left(k, n-k \,|K, N-K, n\right) \\ &= \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \end{aligned}$$

for the probability that a sample of size n drawn without replacement from a population of size N, contains exactly k black balls, given that the population contains exactly K black balls and N - K white balls. (This is the hypergeometric distribution.) Therefore, the probability that both of these (most recent) sister individuals are present in a sample of size n drawn (without replacement) from a population (of size N) is

$$h\left(2, n-2 \,|2, N-2, n\right) = \frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n\left(n-1\right)}{N\left(N-1\right)}.$$

The labelling of the pair of chosen cells (one to die and one to divide), and hence the labelling of the pair of most recent sisters, occurs independently in each generation. Therefore the number of (Moran model) generations $W_M\left(n\right)$ for which the number of lines of descent tracing back from a sample (of size n), remains equal to n, has the geometric distribution

$$\mathbb{P}\left(W_M\left(n\right) = t\right) = p_M\left(n\right)\left(1 - p_M\left(n\right)\right)^{t-1},$$

for $t = 1, 2, \ldots$, with parameter

$$p_M\left(n\right) = \frac{n\left(n-1\right)}{N\left(N-1\right)},$$

and expectation

$$\mathbb{E}\left[W_M\left(n\right)\right] = \frac{1}{p_M\left(n\right)} = \frac{N\left(N-1\right)}{n\left(n-1\right)}.$$

The variable $W_M\left(n\right)$ is the waiting time to the first coalescent event (where the number of lines of descent falls from n to $n-1$), measured in Moran model generations. More generally, for $r = n$, $n-1$, ..., 2, $W_M\left(r\right)$ is the waiting time (measured in Moran model generations) for the coalescent event where the number of lines of descent falls from r to $r-1$.

The pair of lines of descent which coalesce at each coalescent event is chosen at random from the surviving lines. These geometric waiting time distributions ensure that this genealogical process is also a Markov chain.

If the duration of one Moran model generation is δ_M (in years), then the waiting time in years, is

$$T(r) = W_M(r)\,\delta_M.$$

The waiting time in years $T(r)$, has expectation

$$\mathbb{E}[T(r)] = \frac{N(N-1)}{r(r-1)}\delta_M = \frac{2(N-1)}{r(r-1)}\delta_{cell} \approx \frac{2}{r(r-1)}N\delta_{cell} = \frac{1}{\lambda(r)}.$$

If the population size N is large, then to a close approximation, the waiting time $T(r)$ (here measured in years), has an exponential distribution, with density

$$f(t(r);\lambda(r)) = \lambda(r)\exp(-\lambda(r)\,t(r)),$$

and rate parameter $\lambda(r)$. This is the coalescent process discovered by Kingman (Kingman, 1982). Notice that the rate parameter $\lambda(r)$ depends on the population size N only through the product $N\delta_{cell}$.

We have given a description of the Moran model which applies naturally to a population of cells. According to this description, *birth* events in the Moran model represent cell divisions in which a mother cell is replaced by a pair of daughter cells. We have defined (in equation 1) a parameter δ_{cell}, which is the mean life-span of a cell (in years). We now clarify how we apply the Moran model to a population of stem cells. Every cell in the population of size N, is now a stem cell. Each birth event now represents a *symmetric* cell division, in which a stem cell divides to leave two daughter stem cells. A *death* event represents the removal of a stem cell from the population. This may occur by the occasional death of a stem cell, or more usually by a stem cell differentiating into a progenitor cell, or (what amounts to the same thing) by the stem cell dividing to leave two daughter stem cells which then both differentiate into progenitor cells. These two processes are assumed to be in balance (stem cell *birth* by symmetric cell division, and stem cell *death* or removal from the population). The asymmetric cell divisions (in which a stem cell produces one stem cell and one progenitor cell) have no effect on the composition of the population of stem cells. When

applying the Moran model to a stem cell population in this way, it is potentially misleading to refer to the parameter δ_{cell} as the mean life-span of a cell (because a number of asymmetric cell divisions may occur before a stem cell undergoes a symmetric cell division). Here the appropriate interpretation of the parameter δ_{cell} is the mean time (in years) between two symmetrical cell divisions along a line of descent. A glossary of terms is provided in table 1.

Population genetics term	Stem cell biology term	Term used in article	Meaning
Clade (asexual)	Clone	Clone	Set of descendant cells (asexually individuals) of a common ancestor
Line of descent		Line of descent	Sequence of ancestors of an (asexual) individual or cell
	Lineage	Lineage	Cells belonging to one functional cell type or to a set of hierarchically related cell types
Fisher-Wright generation time δ_{WF}	Mean time between two symmetrical cell divisions along a line of descent δ_{cell}	Both terms used here together with the term *cell life-span parameter*	These terms are equivalent as explained in Section 4
	Colony	Colony	Cells derived *in vitro* from a single stem cell progenitor

Table 1: Glossary

4 Equivalence between the Moran model and the Wright-Fisher model

Another widely used population genetic model for the process of random drift is the Wright-Fisher model. In the Wright-Fisher model, individuals do not persist from one generation to the next. Instead, each individual in the current generation is the daughter of an individual in the preceding generation. Each individual in the current generation is assigned a mother by sampling an individual at random from the preceding generation, and furthermore, this random assignment occurs independently for all individuals in the current generation.

This independence of the daughter-to-mother assignment across daughter individuals leads to a relatively simple reverse-time genealogical process. If we label the N individuals in the current population 1, 2, ..., N, and do the same for each proceeding generation, then we can represent the daughter-to-mother assignment as a mapping ϕ from the set $\mathbb{N}(N) = \{1, 2, \ldots, N\}$ into itself. More generally, for any sub-set $S \subseteq \mathbb{N}(N)$ of size n, the probability of a particular mapping from S into $\mathbb{N}(N)$ is

$$\mathbb{P}(\phi \,|S, N) = \frac{1}{N^n}.$$

The probability that a sample of n individuals drawn at random (without replacement) from the population has exactly r distinct ancestors in the previous generation, is obtained by counting the mappings ϕ which satisfy the condition $|\phi(S)| = r$ (the set S is mapped to a set of r distinct elements), and is given by

$$g(r; n, N) = \left\{ {n \atop r} \right\} \frac{[N]_r}{N^n}, \tag{2}$$

for $r = 1, 2, \ldots, n'$, where $n' = \min\{n, N\}$ (Watterson (1975), page 265, Equation 2.12, Gladstein (1978), page 634, Kingman (1980), page 63). Here we use the notation $[x]_r$ for the rth descending (falling) factorial power of x (See for example Aigner (1979)),

which is defined as

$$[x]_r = x\,(x-1)\cdots(x-(r-1))\,.$$

We use the notation $\left\{{n \atop r}\right\}$ for the number of ways of partitioning a set of n distinct elements into exactly r non-empty disjoint (and unlabelled) subsets. This is a *Stirling number of the second kind*. (See Aigner (1979).) Johnson et al. (2005) discuss the probability distribution 2 in their section (10.4.1, pages 439–444) on *Classical occupancy and coupon collecting*, where the well-known *birthday problem* is given as an example (Johnson et al. (2005), page 440). (Watterson (1975) cites Johnson and Kotz (1969), pages 251–252.)

The distribution in Equation 2 has expectation

$$\begin{aligned} m\,(n;N) =& N\left[1-\left(1-\frac{1}{N}\right)^n\right] \\ \approx& N\left[1-\exp\left(-\frac{n}{N}\right)\right], \end{aligned} \tag{3}$$

and the approximation holds when the population size N is large (Fu, 2006). This is the mean number of distinct ancestors of a sample of size n. Notice in particular that the mean number of distinct ancestors (in the preceding generation) of the entire population is

$$\begin{aligned} m\,(N;N) &\approx N\ \left(1-e^{-1}\right) \\ &\approx 0{\cdot}632\ N. \end{aligned}$$

When the population size N is large relative to the sample size n, we have the following approximation

$$
\begin{aligned}
g(n;n,N) =& \left\{ {n \atop n} \right\} \frac{[N]_n}{N^n} \\
=& \prod_{r=1}^{n-1} \left(1 - \frac{r}{N}\right) \\
=& 1 - \frac{1}{N}\left(\sum_{r=1}^{n-1} r\right) + O\left(N^{-2}\right) \\
=& 1 - \frac{n(n-1)}{2N} + O\left(N^{-2}\right),
\end{aligned}
$$

for the probability that a sample of n individuals drawn at random (without replacement) from the population has exactly n distinct ancestors in the previous generation, and

$$
\begin{aligned}
g(n-1;n,N) =& \left\{ {n \atop n-1} \right\} \frac{[N]_n}{N^{n-1}} \\
=& \frac{1}{N}\binom{n}{2} \prod_{r=1}^{n-2} \left(1 - \frac{r}{N}\right) \\
=& \frac{n(n-1)}{2N} + O\left(N^{-2}\right),
\end{aligned}
$$

for the probability that a sample of n individuals drawn at random (without replacement) from the population have exactly $n-1$ distinct ancestors in the previous generation. Thus $g(n;n,N)$ is close to 1, while $g(n-1;n,N)$ is of order N^{-1}, $g(n-2;n,N)$ is of order N^{-2}, and so on.

So, when the population size N is large relative to the sample size n, at each generation of the Wright-Fisher model, the number of lines of descent either remains at its current value of r or falls to $r-1$.

Recall that the daughter-to-mother mapping is generated independently in each generation. Therefore the number of (Wright-Fisher model) generations $W_{WF}(n)$ for which the number of lines of descent tracing back from a sample (of size n), remains equal to n, has the geometric distribution

$$\mathbb{P}\left(W_{WF}\left(n\right)=t\right)=p_{WF}\left(n\right)\left(1-p_{WF}\left(n\right)\right)^{t-1},$$

for $t=1,2,\ldots$, with parameter

$$p_{WF}\left(n\right)=\frac{n\left(n-1\right)}{2N},$$

and expectation

$$\mathbb{E}\left[W_{WF}\left(n\right)\right]=\frac{1}{p_{WF}\left(n\right)}=\frac{2N}{n\left(n-1\right)}.$$

The variable $W_{WF}\left(n\right)$ is the waiting time to the first coalescent event (where the number of lines of descent falls from n to $n-1$), measured in Wright-Fisher model generations. More generally, $W_{WF}\left(r\right)$ is the waiting time for the coalescent event where the number of lines of descent to fall from r to $r-1$ (measured in Wright-Fisher model generations).

The pair of lines of descent which coalesce at each coalescent event is chosen at random from the surviving lines. These geometric waiting time distributions ensure that this genealogical process is also a Markov chain.

If the duration of one Wright-Fisher model generation is δ_{WF} (in years), then the waiting time in years, is

$$T\left(r\right)=W_{WF}\left(r\right)\delta_{WF}.$$

The waiting time in years $T\left(r\right)$, has expectation

$$\mathbb{E}\left[T\left(r\right)\right]\approx\frac{2N\delta_{WF}}{r\left(r-1\right)}=\frac{1}{\lambda\left(r\right)}.$$

If the population size N is large, then to a close approximation, the waiting time $T\left(r\right)$ (here measured in years), has an exponential distribution, with density

$$f\left(t\left(r\right);\lambda\left(r\right)\right)=\lambda\left(r\right)\exp\left(-\lambda\left(r\right)t\left(r\right)\right),$$

and rate parameter $\lambda(r)$. Notice that the rate parameter $\lambda(r)$ depends on the population size N only through the product $N\delta_{WF}$.

Now let us consider a Moran model and Wright-Fisher model which have the same mean waiting time in years $\lambda(r)$. So, we set

$$\frac{2N\delta_{WF}}{r(r-1)} = \frac{N^2\delta_M}{r(r-1)},$$

and find that this equality is met, for all values of r (= 2, 3, ...), when

$$\delta_{WF} = \frac{N}{2}\delta_M. \tag{4}$$

In other words, the Moran model for N individuals is equivalent to the Wright-Fisher model for N individuals, after scaling time by a factor of $N/2$.

Comparing Equations 1 and 4, we see that $\delta_{WF} = \delta_{cell}$, which is mean time (in years) between two symmetrical cell divisions along a line of descent. So, the cell life-span parameter δ_{cell} of our model has a further interpretation, as the generation time of a Wright-Fisher model which has the same population size N as our Moran model (having population size N, and generation time δ_M), and essentially the same behaviour (same rate of coalescent, and rate of random drift) as this Moran model.

As a consequence of this equivalence, we can perform the computations for our Moran model of the stem cell population, using the more convenient Wright-Fisher model. This will be used in sections 5, and 6.

5 Estimation of the product $\delta_{cell}N$

We have a sample of n cells, for each of which we have a single cell genome. From these genomes we have constructed a phylogenetic tree, which we denote by $\mathcal{T}$. The internal nodes of the phylogeny can be ranked in order of their estimated age. As we ascend the tree, from the root node to the terminal nodes, we pass through the internal nodes in age order, beginning at the earliest node (the root node) and ending at the most recent internal node. At the most recent internal node (the first coalescent event), the number

of lines of descent falls from n to $n-1$. We label this internal node $n-1$, because it marks the last (most recent) point on the genealogy where the number of lines of descent was $n-1$. In general, the internal node where the number of lines of descent falls from $r+1$ to r is labelled r. The *age* of internal node r (the age of the subject when this coalescent event occurred) is denoted $h(r)$.

We can estimate the product $N\delta_{cell}$, from the sequence of inter-coalescent interval durations. We can obtain the sequence of inter-coalescent interval durations directly from the phylogeny of cells $\mathcal{T}$. The inter-coalescent interval duration

$$t(r) = h(r) - h(r-1),$$

is the duration (in years) for which the number of lies of descent remains at r. From the phylogeny of n single cell genomes, we obtain the sequence of inter-coalescent interval durations $(t(n), t(n-1), \ldots, t(2))$.

Suppose that the sequence of inter-coalescent interval durations $(t(n), t(n-1), \ldots, t(2))$ was generated under a Moran model, with parameters N and δ_{cell}. The joint density of the durations $(t(n), t(n-1), \ldots, t(2))$ is

$$\prod_{r=2}^{n} \left(f\left(t(r); \lambda(r) \right) \right) = \prod_{r=2}^{n} \left(\lambda(r) \exp\left(-\lambda(r)\, t(r) \right) \right), \tag{5}$$

where

$$\lambda(r) = \frac{2}{r(r-1)}\nu,$$

and where $\nu = N\delta_{cell}$. From this, we can see that the sequence of inter-coalescent interval durations $(t(n), t(n-1), \ldots, t(2))$ is a *sufficient* statistic for the parameter $\nu = N\delta_{cell}$.

Furthermore, the joint density of the inter-coalescent interval durations is the same for all Moran models which have the same value for the product $\nu = N\delta_{cell}$. (The parameters N and δ_{cell} are *non-identifiable*.) Therefore, the joint density of the

inter-coalescent intervals is the same for any Wight-Fisher model which has $N\delta_{WF} = \nu$.

If we specify a prior density on the parameter ν, then we can compute a posterior density for ν using the likelihood function in Equation 5. In fact Bayesian computations can be done for more general models, in which the parameter ν varies over time. Lan et al. (2015) have developed Bayesian methods for a very general model.

Using the method of Lan et al. (2015) it is possible to sample form the joint posterior distribution of the values $(\nu(t_1), \nu(t_2), \ldots)$, at any specified sequence of time points $(t_1, t_2, \ldots)$. Hence we can obtain a point estimate of $\nu(t)$ at any time point t. (Note that if the inter-coalescent intervals $(t(n), t(n-1), \ldots, t(2))$ are measured in years, then the parameter δ_{cell} is also measured in years.) We can also estimate ratios such as $\nu(t_1)/\nu(t_2)$. However, we can not estimate the components $N(t)$, and $\delta_{cell}(t)$, of $\nu(t) = N(t)\,\delta_{cell}(t)$, using this approach.

6 Joint estimation of the stem cell population size N and the cell life-span parameter δ_{cell}

We can make inferences about the stem cell population size N and the cell life-span parameter δ_{cell}, by combining information from the phylogeny of cells $\mathcal{T}$, and the NGS read data from the sample of peripheral blood. In this section, we describe a probabilistic model for the phylogeny of cells $\mathcal{T}$, and the NGS read data from the sample of peripheral blood. We show that the likelihood function for this data depends on the parameter N, separately from the product $\delta_{cell}N$. We also point out the difficulty of computing this likelihood function (it involves summation over a very large set of genealogies), and hence the difficulty of sampling from the joint posterior distribution of the parameters N and δ_{cell}. We then describe the methods used for computing an approximate joint posterior distribution for these parameters.

6.1 Probabilistic model of genomic data

Recall that we have a sample of n cells obtained from a bone marrow aspirate, for each of which we have a single cell genome (obtained by whole genome sequencing). Using these single cell genomes, we have constructed a phylogenetic tree $\mathcal{T}$ for this sample of n cells. We can label the cells in this sample using the integers in the set $\mathbb{N}(n) = \{1, 2, \ldots, n\}$. Recall that a bait-set has been designed to target the sites of a set of m mutations (labelled 1, ..., m). For each cell $i \in \mathbb{N}(n)$ we can determine the subset of these bait-set mutations which are present in the genome of that cell. This set, which we denote by $X_S(i)$, is what we will refer to here as the *genotype* of cell i. Let $\boldsymbol{X_S} = \{X_S(1), \ldots, X_S(n)\}$ denote the sequence of single cell genotypes of the sample of cells obtained from the bone marrow aspirate. (Each genotype in this sequence is represented by a set of mutations. So we can suppose that the genotypes in this sequence have been arranged in lexicographical order.)

NGS read data was also generated from a sample of M granulocytes isolated from a sample of peripheral blood. Let $\boldsymbol{X_B}$ denote the sequence of M genotypes for the granulocytes in this sample. In fact we do not know the genotypes of the M granulocytes in this sample. However, in order to express the likelihood function for our model, we need to sum over the possible states of these genotypes. Recall that the set of m mutations have been placed on the phylogenetic tree of single cell genomes. The positions of these mutations on the phylogenetic tree determines the set of single cell genotypes which are allowed (assuming that each mutation has occurred only once).

Each granulocyte is the descendant of a progenitor cell, which in turn is the daughter of a stem cell. We assume that every time a new progenitor cell is produced by the division of a stem cell, each stem cell has an equal chance of being chosen as the parent cell. Thus each granulocyte is the descendant of a stem cell sampled *with* replacement from the stem cell population. Let $\boldsymbol{X_{SB}}$ denote the sequence of genotypes for the stem cells in this ancestral sample of stem cells from the stem cell population. Let R denote the number of stem cells in this ancestral sample. If the number of stem cells N is large in comparison to the sample size M, then we can expect the sample of

M granulocytes to contain a greater number of mutation represented by a single read, or by very small numbers of reads. Furthermore, we can expect a greater proportion of mutations to be private to a single biological replicate, rather than being shared among multiple (or all) biological replicates.

The sample of M granulocytes is split into L sub-samples (biological replicates) of approximately equal sizes M_1, ..., M_L (and which satisfy the constraint M_1 + ...+ $M_L = M$). These biological replicate are kept entirely separate throughout these DNA extraction and sequencing steps. We let $\boldsymbol{M} = (M_1, \ldots, M_L)$, denote the vector of replicate sample sizes. As before, $\boldsymbol{Y}$ denotes the array of read count data obtained from the sample of peripheral blood, with elements $Y_{j,k,0}$ ($Y_{j,k,1}$) specifying the count of reads, in biological replicate k, which cover the site of mutation j, and which do not (do, respectively) report mutation j. Recall that $\boldsymbol{n}$ denotes the array of read depth data, and that the element $n_{j,k}$ in row j and column k of this array is the read depth at the site of mutation j, in biological replicate k. We now introduce the notation $\boldsymbol{n}(k) = (n_{1,k}, \ldots, n_{m,k})$, for the kth column vector of this array, which contains the read depths (one element for each mutation) in biological replicate k. We will also use $\boldsymbol{Y}(k)$ to denote the sub-array of read counts which are specific to biological replicate k.

The likelihood function for the process which generated the read count data $\boldsymbol{Y}$ from the sample of peripheral blood (together with the phylogeny of cells $\mathcal{T}$, and the single cell genotypes $\boldsymbol{X_S}$ of the sample of cells obtained from bone marrow aspirate) has three components. The first component comes from the genealogical process occurring in the stem cell population. The parameter N enters this component of the likelihood only through the product $\nu = N\delta_{cell}$. The second component comes from the sampling *with* replacement of granulocytes from the population of stem cells. The parameter N enters this component of the likelihood separately from the product $N\delta_{cell}$. The third component comes from the process by which a collection of NGS reads is generated from a sample of granulocytes. We now write the likelihood function in the following form

$$
\begin{aligned}
&P\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) \\
&= \sum_{\boldsymbol{X_B}} \sum_{R} \sum_{\boldsymbol{X_{SB}}} P\left(\boldsymbol{Y}, \boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, \mathcal{T}, R \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) \\
&= \sum_{\boldsymbol{X_B}} \sum_{R} \sum_{\boldsymbol{X_{SB}}} P\left(\boldsymbol{Y} \,|\boldsymbol{X_B}, \boldsymbol{n}, \boldsymbol{M}\right) P\left(R \,|M, N\right) P\left(\boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, \mathcal{T} \,|n, M, R, \nu\right), \quad (6)
\end{aligned}
$$

where the sum extends over every possible sequence of genotypes $\boldsymbol{X_B}$ in the sample of M granulocytes, every possible value for the number of distinct stem cell ancestors R of this sample of granulocytes, and over every possible sequence of genotypes $\boldsymbol{X_{SB}}$ in the ancestral sample of R stem cells.

The factor $P\left(\boldsymbol{Y} \,|\boldsymbol{X_B}, \boldsymbol{n}, \boldsymbol{M}\right)$ is the probability of obtaining the array of read count data $\boldsymbol{Y}$, conditional on the sequence of genotypes $\boldsymbol{X_B}$, the array of read depth data $\boldsymbol{n}$, and on the vector $\boldsymbol{M}$ of replicate sample sizes. Before we can obtain an expression for this probability, we need to introduce some additional notation. We let $\boldsymbol{U}\left(k\right)$ denote the sequence of genotypes of the cells in biological replicate k, and we let $\boldsymbol{U} = \left(\boldsymbol{U}\left(1\right), \ldots, \boldsymbol{U}\left(L\right)\right)$, denote the sequence which contains these replicate-specific sequences of genotypes as its elements. We can now express the probability $P\left(\boldsymbol{Y} \,|\boldsymbol{X_B}, \boldsymbol{n}, \boldsymbol{M}\right)$ as

$$
P\left(\boldsymbol{Y} \,|\boldsymbol{X_B}, \boldsymbol{n}, \boldsymbol{M}\right) = \sum_{\boldsymbol{U}} P\left(\boldsymbol{Y} \,|\boldsymbol{n}, \boldsymbol{U}\right) P\left(\boldsymbol{U} \,|\boldsymbol{M}, \boldsymbol{X_B}\right), \quad (7)
$$

where the sum is over all possible sequences $\boldsymbol{U} = \left(\boldsymbol{U}\left(1\right), \ldots, \boldsymbol{U}\left(L\right)\right)$, of replicate-specific sequences of genotypes, which are compatible with the array of read counts $\boldsymbol{Y}$. The probability $P\left(\boldsymbol{U} \,|\boldsymbol{M}, \boldsymbol{X_B}\right)$ is given by the multi-variate hypergeometric distribution. Conditional on the genotypes $\boldsymbol{U}\left(k\right)$ of the cells in each biological replicate $k = 1, ..., L$, the replicate-specific arrays of read count data $\boldsymbol{Y}\left(k\right)$, are statistically independent across biological replicates, and the pairs of read counts $\left(Y_{j,k,0}, Y_{j,k,1}\right)$ are statistically independent across mutation sites. Therefore the probability $P\left(\boldsymbol{Y} \,|\boldsymbol{n}, \boldsymbol{U}\right)$

can be expressed as a product

$$P\left(\boldsymbol{Y}\,|\boldsymbol{n},\boldsymbol{U}\,\right)=\prod_{k=1}^{L}\prod_{j=1}^{m}P\left(Y_{j,k,0},Y_{j,k,1}\,|n_{j,k},\boldsymbol{U}\left(k\right)\right), \tag{8}$$

where $P\left(Y_{j,k,0},Y_{j,k,1}\,|n_{j,k},\boldsymbol{U}\left(k\right)\right)$ is the probability of obtaining the pair of read counts $(Y_{j,k,0},Y_{j,k,1})$, at mutation site j in biological replicate k, conditional on the read depth $n_{j,k}$, and the genotypes $\boldsymbol{U}\left(k\right)$ of the M_k cells in biological replicate k. These probabilities are given by the (univariate) hypergeometric distribution.

Returning to equation 6, the factor $P\left(R\,|M,N\,\right)$ is the probability that the sample of M granulocytes has exactly R distinct stem cell ancestors. This is given by

$$\begin{aligned}P\left(R\,|M,N\,\right)&=g\left(R;M,N\right)\\&=\left\{\begin{matrix}M\\R\end{matrix}\right\}\frac{\left[N\right]_R}{N^M},\end{aligned} \tag{9}$$

for $R=1,2,\ldots,n'$, where $n'=\min\left\{M,N\right\}$. Notice that the probability distribution in Equation 9 is the same as the probability distribution which occurred in Equation 2, for the number of distinct ancestors of a sample in the Wright-Fisher model. This is because the sample of granulocytes is generated from the population of stem cells by the process of sampling *with* replacement, in the same way as the ancestors of a sample are determined in the Wright-Fisher model. Notice also the occurrence of the population size parameter N in this contribution to the likelihood function.

The factor $P\left(\boldsymbol{X_B},\boldsymbol{X_{SB}},\boldsymbol{X_S},\mathcal{T}\,|n,M,R,\nu\right)$ is the joint probability of the sequence of genotypes $\boldsymbol{X_B}$ in the sample of M granulocytes, the sequence of genotypes $\boldsymbol{X_{SB}}$ in the ancestral sample of R stem cells, and the phylogeny of cells $\mathcal{T}$ together with the observed genotypes $\boldsymbol{X_S}$ in the sample obtained from the bone marrow aspirate. This probability can be expressed as a product, as follows

$$
\begin{aligned}
&P\left(\boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, \mathcal{T} \,|n, M, R, \nu\right) \\
&= P\left(\boldsymbol{X_B} \,|M, R, \boldsymbol{X_{SB}}\right) P\left(\boldsymbol{X_{SB}}, \boldsymbol{X_S}, \mathcal{T} \,|n, R, \nu\right). \qquad (10)
\end{aligned}
$$

Computation of the probability $P\left(\boldsymbol{X_{SB}}, \boldsymbol{X_S}, \mathcal{T} \,|n, R, \nu\right)$ requires the following summation

$$
P\left(\boldsymbol{X_{SB}}, \boldsymbol{X_S}, \mathcal{T} \,|n, R, \nu\right) = \sum_G P\left(G \,|n, R, \nu\right), \qquad (11)
$$

where the sum extends over every genealogy G of the ancestral sample (ancestral to the peripheral blood sample) of R stem cells, drawn from the stem cell population, together with the sample of n stem cells (obtained from the bone marrow aspirate), which is compatible with the phylogeny of cells $\mathcal{T}$, and with the observed sequence of genotypes $\boldsymbol{X_{SB}}$ in the ancestral sample, and the observed genotypes $\boldsymbol{X_S}$ in the sample obtained from the bone marrow aspirate. Notice the occurrence of the the product $\nu = N\delta_{cell}$ as a parameter in this contribution to the likelihood function. This probability also depends on the duration of the phase of population growth which occurs from the conception to the point where the stem cell population attains a stable size. We denote the duration of the population growth phase by t_0. However, in the Bayesian analysis which follows, we treat the value of this parameter as known. (We used a point estimate representing the duration in years for which the expected 100 mutations to accumulate along a line of descent.) For this reason, we have suppressed the parameter t_0 in our notation.

Using this notation, we can express the likelihood function (Equation 6) in the form

$$
\begin{aligned}
&P\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) \\
&= \sum_{\boldsymbol{X_B}} \sum_{R} \sum_{\boldsymbol{X_{SB}}} \sum_{G} P\left(\boldsymbol{Y}, \boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, R, G \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right), \qquad (12)
\end{aligned}
$$

where

$$
\begin{aligned}
&P\left(\boldsymbol{Y}, \boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, R, G \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) \\
&= P\left(\boldsymbol{Y} \,|\boldsymbol{X_B}, \boldsymbol{n}, \boldsymbol{M}\right) P\left(R \,|M, N\right) \\
&\quad P\left(\boldsymbol{X_B} \,|M, R, \boldsymbol{X_{SB}}\right) P\left(G \,|n, R, \nu\right), \qquad (13)
\end{aligned}
$$

and where the summations are as previously defined.

If we specify a joint prior density $p\left(N, \delta_{cell}\right)$ on the parameters N and δ_{cell}, then we can express the joint posterior density of the parameters N and δ_{cell}, as

$$
\begin{aligned}
&p\left(N, \delta_{cell} \,|\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T}\right) \\
&= \frac{P\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) p\left(N, \delta_{cell}\right)}{p\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}\right)} \\
&\propto P\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) p\left(N, \delta_{cell}\right), \qquad (14)
\end{aligned}
$$

where $P\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right)$ is the likelihood function, and $p\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T} \,|\boldsymbol{n}, n, \boldsymbol{M}\right)$ is the marginal likelihood.

Substituting Equation 12 into the right hand side of Equation 14, yields

$$
\begin{aligned}
&p\left(N, \delta_{cell} \,|\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T}\right) \\
&\propto \sum_{\boldsymbol{X_B}} \sum_{R} \sum_{\boldsymbol{X_{SB}}} \sum_{G} P\left(\boldsymbol{Y}, \boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, R, G \,|\boldsymbol{n}, n, \boldsymbol{M}, N, \delta_{cell}\right) p\left(N, \delta_{cell}\right), \qquad (15)
\end{aligned}
$$

where

$$
\begin{aligned}
&p\left(\boldsymbol{Y}, \boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, R, G, M, N, \delta_{cell} \,|\boldsymbol{n}, n\right) \\
=&P\left(\boldsymbol{Y}, \boldsymbol{X_B}, \boldsymbol{X_{SB}}, \boldsymbol{X_S}, R, G \,|\boldsymbol{n}, n, M, N, \delta_{cell}\right) p\left(N, \delta_{cell}\right), \qquad (16)
\end{aligned}
$$

is the the joint distribution of the parameters N, δ_{cell}, the genealogy G, the genotypes $\boldsymbol{X_S}$, $\boldsymbol{X_{SB}}$, $\boldsymbol{X_B}$, and the read count data $\boldsymbol{Y}$.

In principle, we could sample from the posterior density $p\left(N, \delta_{cell} \,|\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T}\right)$, by sampling from the joint posterior distribution of the parameters N and δ_{cell}, and the genealogies, using an MCMC (Markov chain Monte Carlo) sampler (Berthier et al., 2002; Beaumont, 2003).

6.2 Approximate Bayesian computation (ABC)

An alternative approach is to sample from the posterior density $p\left(N, \delta_{cell} \,|\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T}\right)$, using likelihood-free Bayesian computation, also referred to as ABC (approximate Bayesian computation) Beaumont et al. (2002). These are methods which use a rejection sampling algorithm to sample from an approximate posterior distribution. Methods of this type are appropriate for Bayesian inference problems where it is relatively easy to generate simulated data under the statistical model, while it computationally costly to compute the likelihood function.

In order to describe the ABC method, we need to introduce some additional notation. Let $\boldsymbol{\theta} = (N, \delta_{cell})$ denote a vector of parameters for our model. These are the parameters which we are treating as unobserved (their values are uncertain). Let $\mathbf{t} = T\left(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T}\right)$ denote a vector of summary statistics $(t_1, \ldots, t_D)$, which can be computed from the data (including the phylogeny of cells $\mathcal{T}$ which we have constructed from the available single cell genomes). From now on, we need to distinguish between the observed data set, and various simulated data sets. Let $\mathbf{T}_0$ denote the vector of summary statistics computed from the observed data set, and let $\mathbf{T}_i$ denote the vector

of summary statistics computed from the ith simulated data set.

In ABC we perform a large number of iterations N_{sim} of a rejection sampling procedure. These iterations can be performed in parallel on a computing cluster. Each iteration i begins with a proposal step in which a vector of parameter values Θ_i is drawn from the joint prior distribution. This vector of parameter values is then passed to a data simulation algorithm. The data simulation algorithm generates a simulated data set, from which we compute the vector of summary statistics $\mathbf{T}_i$. We now have an observation $(\Theta_i, \mathbf{T}_i)$ which is drawn from the joint prior distribution of the parameter vector $\boldsymbol{\theta}$ and the vector $\mathbf{t}$ of summary statistics. This is followed by an acceptance/rejection step, in which we compute the Euclidean distance $d(\mathbf{T}_i, \mathbf{T}_0)$ to determine if the vector $\mathbf{T}_i$ lies within a ball $B(\mathbf{T}_0; \epsilon)$ with centre $\mathbf{T}_0$ and radius ϵ. This ball is the acceptance region. Only observations $(\Theta_i, \mathbf{T}_i)$ for which the vector $\mathbf{T}_i$ lies within a ball $B(\mathbf{T}_0; \epsilon)$ are accepted.

Let N_{acc} denote the number of iterations where the vector $\mathbf{T}_i$ falls within the acceptance $B(\mathbf{T}_0; \epsilon)$. We number the accepted observations $i = 1, 2, ..., N_{acc}$, and let $(\Theta_i^*, \mathbf{T}_i^*)$ denote the ith accepted observation. The sequence of parameter vectors Θ_i^* ($i = 1, 2, ..., N_{acc}$) from the accepted observations, constitute a sample from a first approximation to the posterior distribution of the parameter $\boldsymbol{\theta}$. As also pointed out by Beaumont et al. (2002), it is possible to improve on this approximation, by using a regression method to adjust the parameter vectors Θ_i^* according to the displacement of the vector $\mathbf{T}_i^*$ from the observed vector of summary statistics $\mathbf{T}_0$.

The ABC rejection method outlined above involves two approximations, in addition to the inevitable Monte Carlo error of any simulation-based approach. First, since the observed data $(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T})$, only enters the algorithm via the vector of summary statistics $\mathbf{T}_0 = T(\boldsymbol{Y}, \boldsymbol{X_S}, \mathcal{T})$, we can only hope to compute the conditional density $p(\boldsymbol{\theta} | \mathbf{t} = \mathbf{T}_0)$, rather than the true posterior density of the parameter $\boldsymbol{\theta}$ (unless the vector of summary statistics $\mathbf{t}$ is a *sufficient* statistic for the parameter $\boldsymbol{\theta}$). The second approximation is a consequence of the acceptance region extending beyond the point $\mathbf{t} = \mathbf{T}_0$. Using the ABC rejection method, we compute the conditional density

$p(\boldsymbol{\theta}\,|\mathbf{t} \in B(\mathbf{T}_0; \epsilon))$, rather than the conditional density $p(\boldsymbol{\theta}\,|\mathbf{t} = \mathbf{T}_0)$. It is the error resulting from this second approximation which we hope to reduce by using an ABC regression method.

The insight behind these ABC regression methods (Beaumont et al., 2002; Blum and François, 2010) is that the sequence of parameter vectors Θ_i^* (i = 1, 2, ..., N_{acc}) from the accepted observations, can be represented in the form

$$\Theta_i^* = \boldsymbol{B}\mathbf{T}_i^* + \mathbf{Z}_i, \tag{17}$$

where $\boldsymbol{B}$ is a matrix of coefficients, and each $\mathbf{Z}_i$ is a vector of residuals. Furthermore, as a first approximation, we can assume that the residuals are drawn from a distribution which does not depend on the vector of summary statistics $\mathbf{T}_i^*$. This is a linear model. An observation from the conditional density $p(\boldsymbol{\theta}\,|\mathbf{t} = \mathbf{T}_0)$ could therefore be represented in the same way, as

$$\Theta_i^* = \boldsymbol{B}\mathbf{T}_0 + \mathbf{Z}_i. \tag{18}$$

Now, from the sequence of accepted observations $(\Theta_i^*, \mathbf{T}_i^*)$ (i = 1, 2, ..., N_{acc}), we can obtain a point estimate $\hat{\boldsymbol{B}}$ of the matrix of coefficients (using least squares regression, or an alternative method). We can now compute the empirical residuals

$$\hat{\mathbf{Z}}_i = \Theta_i^* - \hat{\boldsymbol{B}}\mathbf{T}_i^*, \tag{19}$$

for i = 1, 2, ..., N_{acc}. From this sample of residuals, we could estimate the distribution of residuals. Alternatively, we could use the sample of empirical residuals $\hat{\mathbf{Z}}_i$ directly. The resulting predicted value of the parameter vectors

$$\begin{aligned}\hat{\Theta}_i &= \hat{\boldsymbol{B}}\mathbf{T}_0 + \hat{\mathbf{Z}}_i \\ &= \Theta_i^* + \hat{\boldsymbol{B}}\left(\mathbf{T}_0 - \mathbf{T}_i^*\right),\end{aligned} \tag{20}$$

should more closely approximate a sample from the conditional density $p(\boldsymbol{\theta}\,|\mathbf{t} = \mathbf{T}_0)$.

More advanced regression methods (Blum, 2010; Blum and François, 2010; Csilléry et al., 2012) can also be used to approximate the conditional density $p(\boldsymbol{\theta}\,|\mathbf{t} = \mathbf{T}_0)$. In the implementation described below, we used a neural network method to perform the regression step.

The simulations of the genealogies could be performed using coalescent simulations for a model with parameters $\nu = N\delta_{cell}$ (N and δ_{cell} sampled from the prior density), and t_0 equal to our point estimate (corresponding to 100 mutations on the molecular clock). Here we chose to use forward-time simulations of the corresponding Wright-Fisher model.

6.3 Simulation of haematopoiesis model

We simulated haematopoiesis according to the following model:

- The adult stem cell population size N is drawn from a uniform distribution on the log scale between 1,096 (exp(7)) and 3,269,017 (exp(15)) stem cells. Once the adult stem cell population is reached, we assume that it stays constant over adulthood, which is consistent with our results from the phylodyn algorithm and with previous results in the literature (Catlin et al., 2011; Werner et al., 2015). We commence our simulations at the point where the adult stem cell pool has reached a constant size.

- We model drift in the adult stem cell population as a Wright-Fisher process, resampling the adult stem cell population with replacement with every generation. The Wright-Fisher generation time δ_{cell} is equal to the mean time between two symmetrical stem cell divisions along a line of descent, and is drawn from a uniform distribution on a log scale between 20 days (exp(3)) and 8103 days (exp(9) i.e. 22 years).

- We know that an average of 1000 mutations have been accumulated by each stem cell over the course of life in our volunteer. The number of mutations acquired

over the stable population size phase is therefore 1000 minus the number of mutations that were acquired in childhood. We know from the phylodynamic model that the adult stem cell population size is reached after approximately 100 mutations have been acquired (figure 3). Therefore, an average of 900 mutations have accumulated during the stable population phase. The number of mutations acquired by every stem cell in each Wright-Fisher generation was drawn from a Poisson distribution with mean equal to 900 divided by the total number of Wright-Fisher generations.

- We then recapitulate our experiment. First, we choose 155 colonies for whole genome sequencing and construct a phylogeny from them.
  - NB: of 200 colonies sequenced, only 140 were clonal, and these were the ones used to build the tree, and for all analyses except for the ABC. However, of the polyclonal colonies, we could salvage 15 because they had a dominant clone and shared more than ten mutations with a clonal colony that was on our tree. We could therefore graft these 15 extra colonies onto the tree of 140 clones. This is helpful because it provides an extra time point on each branch onto which an extra colony is grafted. Mutations can then be classified as being shared with the polyclonal colony that has been grafted on, or absent from the polyclonal colony, thus providing additional information about the timing of the mutation. No mutations that were present only in polyclonal colonies (and not in clonal colonies) were used, as we could not be sure where to place them on the tree.
- Second, we design a bait-set for the simulated tree, using the same criteria as were used to design the real bait-set (described in the methods section).
- We then simulate the sampling of peripheral blood granulocytes. We generate a sample of $M = 540{,}000$ granulocyte by sampling with replacement from the stem cell population. This simulated sample of 540,000 cells is then split into $L = 6$ sub-samples (biological replicates) each of size 90,000 (to simulate the 6 biological

replicates obtained from our volunteer at the nine month time point).

- We then simulate targeted sequencing of the mutations in the bait-set. In the observed dataset, there were 3952 mutations in the bait-set that – after our duplicate removal and consensus calling step (methods) – were covered by at least 4000 reads in the control cord blood DNA, but where no mutant reads were found in the cord blood DNA. We therefore used these 3952 mutations for analysis, and also only used 3952 mutations in the simulated bait-set.

- For every bait-set locus in every biological replicate (from the granulocyte sample), we randomly draw a sequencing depth from the empirical distribution of sequencing depths for the real targeted data from the nine month time point. We sample the chosen number of reads over this locus from the granulocytes without replacement, and count the number of reads that have the mutation.

- Sequencing errors were included in the simulation as follows. The sequencing error rate was learnt from the control cord blood. For all 7116 positions in the bait-set, the VAF in the cord blood was calculated. Where the VAF was zero (if there were no mutant reads), the VAF was set to 1/10,000. For each of the 3952 loci used in simulations, then, an error rate was drawn from these VAF distributions. The number of false positive reads was obtained by drawing from the binomial distribution, with parameter p equal to the randomly chosen error rate and parameter n equal to the sequencing depth. We also tested two other error models: one with no sequencing errors, and one with double the sequencing error rate observed in the cord blood controls. These made little difference to the median of the posterior distribution of the model, but affected its width. A separate false positive rate was included based on the estimated rate of homoplasmy, assuming that every granulocyte has 2000 mutations (double the number of mutations present in a stem cell, which seemed a reasonable upper bound for the number of additional mutations that a granulocyte could acquire) and that these are spread evenly across the genome.

- We then extract summary statistics from the resulting simulated data set. There are two categories of summary statistics used. Only the first category of summary statistics was extracted for the first set of simulations, as explained below.

- – Summary statistics that reflect the shape of the phylogeny, referred to as lineages through time or LTT summary statistics. We divided the molecular time scale, from mutation 100 to mutation 800, into bins of molecular time, each 100 substitutions wide. For each mutation we recorded the number of branches in the phylogeny which are contemporaneous with the specified mutation, and then for each bin, we computed the mean value accros all 100 mutations within the bin.

  – Summary statistics that use the targeted sequencing information from peripheral blood. We wanted statistics that would capture whether different granulocyte samples descended from the same stem cell population or not. As explained above, the less overlap in the contributing stem cell populations, the larger the stem cell population is likely to be. For a given mutant read count c, we define the shared mutation count statistic $SMC(c, k)$ to be the number of mutations (out of the 3952 mutations in the bait-set) which have c or more mutant reads in exactly k sub-samples. We compute these statistics for mutant read cutoffs of $c = 1, 2, ..., 6$ reads, and for $k = 0, 1, ..., 6$ sub-samples. So we have a 6×7 array SMC containing 42 individual statistics.

These summary statistics are recorded for every simulation. The same summary statistics were calculated for the observed data. Summary statistics were analysed as explained below.

6.4 Justification of bounds placed on prior

Total stem cell number. The adult stem cell population size was drawn from a uniform distribution on the log scale between 1,096 (exp(7)) and 3,269,017 (exp(15)) stem cells.

A minimum number of approximately 1000 stem cells was chosen because we knew from preliminary simulations that a smaller number of stem cells than 1000 could not produce the low VAF mutations that we observe. The maximum number of stem cells was chosen because, firstly, it was at the limits of what was computationally feasible with the resources available (each simulation at this upper limit requires approximately 150 GB of memory), and secondly, because it was an order of magnitude higher than any number of stem cells that has been proposed, to the best of our knowledge.

Mean time between symmetrical cell divisions for one stem cell (equal to the Wright-Fisher generation time). The mean time between two symmetrical stem cell divisions was drawn from a uniform distribution on a log scale between 20 days (exp(3)) and 8103 days (exp(9) i.e. 22 years). The minimum time of 20 days was chosen because stem cells are reportedly relatively quiescent (Arai and Suda, 2007; Orford and Scadden, 2008). Not all of a stem cells divisions need be symmetrical: a proportion are likely to be asymmetrical, producing one daughter stem cell and one progenitor cell. We are blind to asymmetrical divisions. Therefore, if a cell is dividing asymmetrically in addition to symmetrical divisions on average every 20 days, it will be dividing significantly faster than every 20 days, which seemed unlikely given prior knowledge of stem cell quiescence. Furthermore, shorter times between cell divisions mean that more generations need to be simulated, which is computationally costly. The maximum time between symmetrical cell divisions of 22 years was chosen because it required a very small number of HSCs to create a phylogeny of the right shape, and such a small number was not compatible with the observed range of VAFs.

6.5 Details of ABC implementation

Two sets of simulations were run. First, we generated 120,000 simulations drawing both the number of stem cells and the time between symmetrical stem cell divisions from a uniform prior on a log scale (extended figure 5a), and extracted summary statistics that reflect the shape of the phylogeny (the LTT summary statistics explained below). As explained in section 5 this allowed us to identify a relationship between stem cell

number and generation time, effectively targeting a diagonal line on the sample space that is more plausible based on the shape of the phylogeny (extended figure 5b). We therefore ran another 80,000 simulations targeting this area of the sample space (extended figure 5c). Both sets of simulations were run in the same way, but additional summary statistics were extracted from the latter set of 80,000.

ABC 1: simulations on log uniform prior. 121,329 simulations were generated as described above, resulting in a flat joint prior distribution of stem cell numbers and generation time (both on a log scale) (extended data figure 5a). For this first set of simulations, only the LTT summary statistics, which reflect the shape of the phylogeny, were used. Going backwards in time, the faster the rate of random drift, the more rapidly the number of lineages decreases. Thus simulations that have too rapid a random drift rate (simulations with a small population size and short generation time, as in figure 5g) have LTT statistics that are too low for early bins of molecular time, and simulations that have too slow a random drift rate (with a large population size and long generation time, as in figure 5h) have LTT statistics that are too high for early bins of molecular time. These were the only summary statistics that were used to define the plausible region of sample space in which to run the second set of simulations (extended figures 5b and 5c).

ABC 2: simulations restricted to the region of sample space identified in ABC 1. A further 80,762 simulations were run in the region of the plausible diagonal indicated by the first set of simulations. For this set of simulations, both the LTT and peripheral blood-derived summary statistics were extracted. Summary statistics were normalised, and for each simulation the Euclidean distance between the simulated and observed vector of summary statistics was calculated. To maximise the accuracy of our model, we cross-validated both the size of the acceptance region and the number of simulations included in the acceptance region. For each of 1000 cross validation samples, we draw one simulation to act as fake observed data, and remove it from the pack of simulations. We then analysed the data as though our fake observed data were the true data. We took the n (where n is the number of accepted simulations)

simulations that produced summary statistics that were most similar to our fake observed data, as determined by their Euclidean distance. We then calculated a number of statistics, as shown in extended figure 5e-i. First, we plotted the accepted simulations on a graph of stem cell numbers vs generation time (both on a log scale) and drew an ellipse that contained 90% of the n points inside it. We then saw whether the true value of the fake observed data fell inside this ellipse. The proportion of cross validation samples for which the fake observed data fell inside the ellipse is shown in extended figure 5e. We also measured the mean area of the ellipse (extended figure 5f), the distance between the median number of stem cells of the accepted simulations and the fake observed data number of stem cells (extended figure 5g), the distance between the median generation time of the accepted simulations and the fake observed data generation time (extended figure 5h), and finally the distance between median of the posterior of a neural network regression run on the accepted simulations and the fake observed stem cell number (extended figure 5i). We chose an LTT weighting of 1 and an acceptance region of 1000, since this resulted in an accurate prediction of the stem cell number from the neural network regression and a high proportion of the fake observed values falling in the ellipse, while keeping the size of the ellipse relatively small. We then analysed the true observed data, using the error model that took VAFs from the observed control data, weighting the LTT summary statistics by 1 and choosing the 1000 most similar simulations to fall in the accepted region. Neural network regression was performed on these best 1000 simulations using the R package abc (Csilléry et al., 2012), to find the number of stem cells that minimised the distance between the observed and simulated summary statistics (figure 5c, extended figure 5l). The neural network regression was run using the default of one hidden layer with five units. As predictions from different neural networks can vary, thirty neural networks were run and the median provided. To test the robustness of our analysis, we repeated it with the other two error models described, and also ignoring summary statistics that used a mutant read cut-off of 1, since these would be most sensitive to incorrect modelling of the sequencing error rate. Both of these additional analyses resulted in a

widening of the posterior distribution for the number of contributing stem cells, but did not significantly change its location (data not shown).

References

Aigner, M. (1979). *Combinatorial theory*. Springer-Verlag, New York.

Arai, F. and Suda, T. (2007). Maintenance of quiescent hematopoietic stem cells in the osteoblastic niche. *Annals of the New York Academy of Sciences*, 1106(1):41–53.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035.

Berthier, P., Beaumont, M. A., Cornuet, J.-M., and Luikart, G. (2002). Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics*, 160(2):741–751.

Blum, M. G. (2010). Approximate bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

Blum, M. G. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73.

Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P., and Abkowitz, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo. *Blood*, 117(17):4460–4466.

Csilléry, K., François, O., and Blum, M. G. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*, 3(3):475–479.

Fu, Y.-X. (2006). Exact coalescent for the Wright–Fisher model. *Theoretical population biology*, 69(4):385–394.

Gladstein, K. (1978). The characteristic values and vectors for a class of stochastic matrices arising in genetics. *SIAM Journal of Applied Mathematics*, 34:630–642.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. John Wiley & Sons, Hoboken, NJ, 3rd edition.

Johnson, N. L. and Kotz, S. (1969). *Discrete distributions: Distributions in statistics*. Houghton Mifflin, Boston, MA, first edition.

Kingman, J. F. C. (1980). *Mathematics of Genetic Diversity*, volume 34 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.

Kingman, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab.*, 19A:27–43.

Lan, S., Palacios, J. A., Karcher, M., Minin, V. N., and Shahbaba, B. (2015). An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics*, 31(20):3282–3289.

Moran, P. A. P. (1958). Random processes in genetics. *Proceedings of the Cambridge Philosophical Society*, 54(1):60–71.

Orford, K. W. and Scadden, D. T. (2008). Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation. *Nature Reviews Genetics*, 9(2):115.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7:256–276.

Werner, B., Beier, F., Hummel, S., Balabanov, S., Lassay, L., Orlikowsky, T., Dingli, D., Brümmendorf, T. H., and Traulsen, A. (2015). Reconstructing the in vivo dynamics of hematopoietic stem cells from telomere length distributions. *Elife*, 4.

Yu, F., Etheridge, A., and Cuthbertson, C. (2010). Asymptotic behavior of the rate of adaptation. *The Annals of Applied Probability*, 20(3):978–1004.