# The dynamic fitness landscape of ageing haematopoiesis through clonal competition

Nathaniel Mon Père[1,*], Francesco Terenzi[1,*] & Benjamin Werner[1,†]

[1]Evolutionary Dynamics Group, Centre for Cancer Genomics and Computational Biology, Barts Cancer Centre, Queen Mary University of London, Charterhouse Square, London, United Kingdom EC1M 6BQ.

[*] These authors contributed equally,

[†]Correspondence should be addressed to b.werner@qmul.ac.uk

## Abstract

Clonal haematopoiesis (CH) – the ubiquitous expansion of variants in blood with age – is a prime example of ongoing somatic evolution in otherwise normal human tissues. The details of the underlying evolutionary processes and how they drive clonal dynamics and intra-tissue genetic heterogeneity remain to be understood. Here we show that clonal competition can explain the complex dynamics of CH with age. According to the model numerous clones of differing fitness continually appear and compete. This clonal competition implies an emerging fitness landscape of haematopoietic stem cells (HSCs) that changes over time in a highly stochastic manner within and between individuals. It naturally explains the presence of decreasing variant trajectories and declining average fitness in the elderly, the varying fitness of identical variants across individuals, and characteristic transitions of site frequency spectra with age. Comparisons of stochastic simulations of clonal competition-driven CH to time resolved clone trajectories in 385 individuals, and to changes of the HSC site frequency spectrum from newborns to octogenarians allow us to estimate the shape of the innate CH fitness distribution and the arrival rate of advantageous clones. We find a non-exponential fitness distribution with mean 0.07 to 0.09 and standard deviation 0.025 per year. We estimate an arrival rate of 2 to 20 advantageous clones per year, but the number of detectable clones above 1% frequency in the HSC population to remain between 3 to 20 by the age of 80.

## Introduction

Genetic heterogeneity pervades somatic tissues at all stages of human life [1–5]. It emerges from ubiquitous mutational processes and expansion of novel clones by random drift and clonal selection [6–10]. This phenomenon has been well-characterised in blood, where largely expanded clones are common late in life, a condition termed clonal haematopoiesis (CH) [11–16]. Still, obtaining a detailed

quantitative description of the evolutionary forces of CH and how these change with age, mutational load, and tissue regulation remains challenging.

The simplest evolutionary models of CH assume that positively selected somatic mutations in haematopoietic stem cells (HSCs) are rare and their fitnesses are context-independent [7, 8, 17]. However, recent observations suggest a more complex scenario [5, 15, 16, 18]. Clone frequencies measured at multiple time points exhibit saturating growth curves despite possessing clear positive fitness [15]. While by itself such behaviour is consistent with a size-constrained HSC population [19, 20], other observations remain unexplained: clones typically saturate below fixation, many clone trajectories decrease late in life, and the average HSC fitness appears to decrease with age [15, 16]. Stratifying inter-individually further complicates this picture, as the fitness of identical driver mutations is inconsistent across donors [7, 15, 21].

Here we show that these phenomena naturally emerge in an evolutionary model of clonal competition in which multiple clones of differing fitness consistently appear, compete, and change the HSC fitness landscape over time in a way that is highly stochastic between individuals (Figure 1A). The clonal competition model generates testable predictions on the expected change of clone trajectories and variant allele frequency distributions with age that are corroborated by *in vivo* observations. Furthermore, the underlying evolutionary parameters of CH – the yearly clone arrival rate and the mean and standard deviation of the clone fitness distribution – become identifiable and converge to the same quantitative estimates in two distinct data.

# Results

## Trajectories of clonal haematopoiesis suggest a dynamic fitness landscape in HSCs

To quantify the changing haematopoietic fitness landscape with age, we first revisited longitudinal measurements obtained by Fabre et al. [15] of variant allele frequencies of 56 known driver mutations across 385 individuals. This resulted in 697 variant trajectories with 13 years median follow-up time (Figure 1B). We evaluated the selection strength for each variant by fitting its change in frequency over time to a logistic differential growth function $dx(t)/dt = s_{\log} x(t)(1 - x(t))$, where $x(t)$ denotes the frequency of a variant at age $t$. The logistic fitness coefficient $s_{\log}$ is either positive for increasing, negative for decreasing, or zero for saturating trajectories. We obtained best-fit fitness coefficients by implementing a maximum likelihood estimator that assumes the observed variant frequencies to be binomially sampled from unobserved true frequencies (Figure 1B, SI 1). This resulted in a distribution of logistic fitnesses for individuals ranging from 62 to 98 years of age (Figure 1E). Overall, 67% of all successfully fitted trajectories were increasing, 24% decreasing, and 8% saturating (Figure 1C-D). A similar partition of frequency changes was recently reported by Van Zeventer et al. in a set of 1,642 mutations across 1,573 individuals [21]. After stratifying trajectories by age, the relative proportion of decreasing trajectories increases from 18% at age 63 to 41% at age 88 (Spearman correlation=0.68, p=0.003), and the mean logistic fitness decreases with age (Spearman correlation=-0.36, p=0.16), (Figure 1F). Since variants in a population of approximately $10^5$ HSCs are unlikely to have reached detectable frequencies of 0.01 or higher solely by random drift [7, 8], those with decreasing trajectories must have been under positive selection in the past. This suggests an emerging dynamic HSC fitness landscape in which observable clone fitness can change with age.

## Clonal competition predicts varying CH trajectories within and across individuals

The logistic growth model allows quantification of a clone's observable selection strength in a brief time window through its fitness parameter $s_{\log}$. However, mathematically this model describes the expected growth of only a single clone in a background population of static fitness [20]. We extended the single-clone model to account for the effects of a shared carrying capacity on trajectories of multiple coexisting clones in a context-dependent fitness landscape. In this model, new clones arise at random times according to a Poisson process with constant rate $\mu$. Each clone is assigned an innate fitness randomly drawn from a Gamma distribution, a flexible choice that contains many families of probability distributions as limiting cases. We show that the fully stochastic birth-death process of an arbitrary number of clones obeys a set of stochastic differential equations given by

$$\mathrm{d}x_i(t) = x_i(t)\left(s_i - \sum_{\text{all clones}} s_j x_j(t)\right)\mathrm{d}t + \mathcal{B}(t)\,\mathrm{d}W_t\,, \tag{1}$$

where $x_i(t)$ denotes the frequency of clone $i$ at time $t$, and $\mathcal{B}(t)$ is a fluctuation term with Gaussian noise $\mathrm{d}W_t$ (see SI 2.1). We denote $s_i$ the innate fitness of clone $i$, defined as the increase of the clone's proliferation rate relative to the wild type in the absence of competition: $s_i = \text{rate}_i - \text{rate}_{\text{wt}}$. The deterministic part of Equation (1) describes a competitive Lotka-Volterra equation [22], however here the number of equations increases as clones emerge. We confirmed its validity with agent-based simulations of the underlying birth-death process using the Gillespie algorithm [23] (SI 4.2, Figure SI 1E). Equation (1) introduces an important distinction between a clone's innate fitness $s_i$ and its observable growth rate, i.e. its change in clonal frequency measurable via sequencing. In fact, because the latter depends on the presence of all other clones ($s_i - \sum_j x_j s_j$), a clone with positive innate fitness can become negatively selected if this fitness is surpassed by the increasing average fitness in the population – that is if $s_i < \sum_j x_j s_j$ (Figure 1A, G). This becomes more likely with age as more fit clones inhabit the system. Consequently, as the clonal composition of the stem cell population grows the observable fitnesses of individual clones become increasingly dynamic.

Using Equation (1) we simulated clone trajectories in virtual individuals that match the age and frequency of clones in the data. We then applied the same maximum likelihood estimation to obtain the predicted age-stratified distribution of fitted logistic fitness coefficients $s_{\log}$ from simulated trajectories (Figure 1H). In accordance with the data, simulations show the emergence of decreasing variant trajectories with an increasing abundance with age (Spearman correlation = 0.75, $p = 0.0004$), while the average logistic fitness across all variants decreases with age (Spearman correlation = - 0.53, $p = 0.028$, Figure 1I). The model furthermore predicts that the measurable logistic fitness of a variant varies strongly with its time of occurrence, the individual's age at measurement, and the unique landscape of co-occurring mutations (Figure SI 1F). This provides a natural explanation for the observed variation of fitnesses of identical CH driver mutations across individuals [15, 17, 21]. Overall, these findings suggest that clonal competition could underlie the dynamic and stochastic nature of observed clone trajectories.

## Characteristic transitions of site frequency spectra with age reflect HSC population dynamics

To further probe the consequences of clonal competition, we quantified HSC genetic heterogeneity using the site frequency spectrum (SFS) – the distribution of prevalences of all somatic variants within a population. The SFS is typically dominated by neutral mutations which arise far more frequently

than positively selected variants. However, neutral mutations hitchhike on expanding fit clones [24], suggesting that a varying fitness landscape within the HSC population may change the SFS with age (Figure 2A).

To derive predictions for the expected SFS in the presence of clonal competition we extended our model to include the accumulation of neutral variants in a population of $10^5$ HSCs through agent-based stochastic simulations (SI 4). From these simulations we observe two frequency domains of the SFS (Figure 2B-D, Figure SI 5). Variants at high frequencies ($\gtrsim 10^{-1}$) are sparse and highly stochastic across realisations. At young ages they are predominantly developmental in origin, whereas in the elderly they are a combination of developmental and positively-selected variants. In contrast, at lower frequencies ($\lesssim 10^{-1}$) variant distributions follow power laws that emerge from the underlying HSC dynamics (Figure 2A). At birth the number of mutations at frequency $f$ follows a $1/f^2$ power law. This is characteristic for expanding populations [25], and occurs here as a consequence of developmental growth. By early adulthood, the SFS resembles a wave asymptotically approaching a $1/f$ power law, which is associated with a neutral homeostatic population of fixed size [10]. In the elderly our simulations suggest the SFS shifts again, with variants at low frequency reverting to a $1/f^2$ scaling due to the expansion of positively selected clones. Thus our stochastic simulations predict two transitions between three characteristic stages of the SFS in HSCs. The first reflects a deterministic change of the population level dynamics from a growing into an approximately constant homeostatic HSC population, that is only weakly affected by the presence or absence of expanding clones. The second transition, however, is solely driven by expanding clones and would be absent in neutrally evolving HSC populations.

We tested these predictions in SFS reconstructed from single-cell whole genome sequencing data of HSCs published by Mitchell and colleagues [16] (Figure 2B-D). In brief, whole genome sequencing of 200 to 500 single HSCs in 9 healthy donors aged 0 to 81 allowed us to compute SFS with high resolution across the observed frequency spectrum. After adjusting for sampling, measured spectra of all ages agree well with stochastic simulations (Kolmogorov-Smirnov test $p_{KS} > 5\%$ for all SFS, Figure 2B-D, SI 5). Also noticeable is the increasing roughness at higher frequencies in the SFS in the elderly, suggesting unique evolutionary histories. Single realisations of our stochastic simulations reproduce the magnitude of these fluctuations that are absent in a population average (Figure SI 5). In summary, a clonal competition model precisely recovers the non-trivial dynamics of the SFS from newborns to octogenarians, capturing the changes in the genetic heterogeneity observed in clonal haematopoiesis at single-cell resolution.

## A Bayesian computation framework to estimate parameters of CH evolution

We then investigated the identifiability of the underlying evolutionary parameters of CH from *in vivo* data. The clonal competition model defines three parameters that determine the HSC fitness landscape: the arrival rate of advantageous clones $\mu$, and the mean $\eta$ and standard deviation $\sigma$ of the innate fitness distribution. We additionally require the size of the adult HSC population $N$ and the wild-type HSC inter-division time $\tau$, which describe healthy haematopoiesis in the absence of selection.

To test what combinations of these parameters $\mathcal{P}_i = \{\mu_i, \eta_i, \sigma_i, N_i, \tau_i\}$ best describe the data, we implemented approximate Bayesian computation (ABC) inference schemes [26]. In brief, we randomly drew 65'000 parameter combinations $\mathcal{P}_i$ from uninformed uniform prior distributions. For each $\mathcal{P}_i$ we generated *in silico* predictions from our modelling and quantitatively compared them to donor data to obtain a metric distance. We then retained the 1% quantile of ranked distances to derive posterior

distributions containing those $\mathcal{P}_i$ whose predictions are closest to the data (Figure 3A).

We applied this ABC scheme independently to the two distinct datasets describing clone trajectories (Fabre et al. [15]) and the SFS (Mitchell et al. [16]) respectively. To compare simulations of Equation (1) with observed clone trajectories we used two summary statistics: the distribution of clone sizes across all donors and the distribution of logistic fitness coefficients $s_{\log}$ (Figure 3B-C). To capture the change of the SFS with age we generated longitudinal samples of our agent-based simulations for comparison with donor data. As summary statistics we use best fits in 7 out of 8 time points of the cumulative SFS (Figure 3E-F donors aged 0, 38, and 76; Figure SI 8 all donors) and the number of reported detectable clones above 1% frequency (Figure 3D).

The posterior distributions for the competition parameters ($\eta$, $\sigma$, and $\mu$) converge across both datasets and deviate significantly from imposed uninformed priors in both ABC inference schemes (Figure 3H-J, and Figures SI 2, SI 9). We report mode and 90% credibility intervals for best parameter values. From the inference on variant trajectories (Figure 3B-C) we find for the mean and the standard deviation of the fitness distribution $\eta = 0.07^{+0.04}_{-0.03}$ per year and $\sigma = 0.025^{+0.002}_{-0.012}$ (Figure 3H-I). For the arrival rate of advantageous variants in the stem cell population we find $\mu = 3^{+21}_{-1}$ per year (Figure 3J). The inference on the SFS (Figure 3D-G) estimates $\eta = 0.09^{+0.03}_{-0.04}$ per year, $\sigma = 0.025^{+0.011}_{-0.013}$ (Figure 3H-I), and an arrival rate of advantageous variants of $\mu = 3^{+23}_{-0}$ per year (Figure 3J). The posterior parameter distributions for the HSC population size $N$ and the HSC wild type inter-division time $\tau$ converge less sharply (Figure SI 2, SI 9). This is in accordance with our validation of the ABC framework on synthetic data (Figure SI 3B, SI 3.1) and implies that the competition model is compatible with a wider range for these parameters.

The parameter inference finds an arrival rate of fit clones on the order of 2 to 20 per year, suggesting that by age 50 an individual may have harbored up to a thousand clones under positive selection. Though this may appear high, the fraction of fitness-conferring mutations is low overall. In a population of $10^5$ HSCs with a total rate of 20 mutations per year per stem cell [5], this estimate results in 1 to 10 in $10^6$ mutations conferring a selective advantage. Furthermore, the estimated fitness distribution has its average at a relatively low 7-9% per year. Thus the combined effects of low fitness and competition prevent most clones from escaping extinction and reaching detectable sizes. Finally, not all clones necessarily originate from mutations, as epigenetic and environmental effects can contribute to the overall clonal arrival rate. Because our theoretical framework is agnostic to underlying genetic and epigenetic mechanisms, it therefore quantifies the compound cell-intrinsic fitness distribution.

The inferred innate fitness distribution peaks at a positive non-zero value (Figure 3K), implying that clones with very low innate fitness ($s \approx 0$) are rare. To ensure this result is not due to limited data resolution – since low fitness clones are less likely to expand to observable frequencies – we performed ABC inferences on simulated data with an imposed exponential innate fitness distribution measured at similarly limited resolution. The ABC successfully identified the correct innate fitness distribution despite the lowest fitness variants not being directly observed (Figure SI 3A). This demonstrates that even low fitness variants convey a measurable effect on the system dynamics through compound interactions. It is finally worth noting that due the nature of the data the inferred distribution (Figure 3K) describes the prevalence of advantageous variants that have not yet transformed into aggressive cancers, and thus cannot describe the fittest variants that likely occur at high $s$. This agrees with recent estimates of AML subclonal driver mutations with estimated fitnesses between 0.4 to 0.8 [27], and estimates of the most common mosaic chromosomal aberrations in blood that aggregate at the upper bound of 0.1 to 0.2 per year of the fitness distribution [28].

## Discussion

The converging inferences of clonal competition in CH in two different *in vivo* datasets suggests that clonal competition naturally explains complex and contrasting observations in ageing haematopoiesis. It also suggests that CH and age-related expansions are ubiquitous in the elderly. Our model explains the sudden transition in the number of detectable clones between the ages 50-60 [13, 16], and provides sharp bounds on the possible number of detectable clones in the elderly. Above the age of 70, between 5 and 20 clones should be visible at a resolution of 1% VAF (Figure 3D). This does not exclude the possibility that with higher resolution additional effects like age-related deregulation, immune interactions, inflammation, epistasis, or frequency-dependent effects may become identifiable.

Lastly, clonal competition may act as a barrier to some malignant expansions since widespread competition amongst a large population of HSCs limits growth of individual clones. Its exact manifestation however may be highly tissue-specific, and the extent to which it impedes the transition from healthy tissue to malignancies remains an open question [29].

## Acknowledgments

N.V.MP. and F.T. acknowledge support by the Barts Charity. B.W. is supported by a Barts Charity Lectureship (grant no. MGU045) and a UKRI Future Leaders Fellowship (grant no. MR/V02342X/1). We are grateful to David Dingli and Andrea Sottoriva for discussions and comments at different stages of this work.

## Code availability

The HSC dynamics are implemented in the Rust programming language and the code is available at `https://github.com/fraterenz/hsc`. The numerical implementation of the SDE system in Equation (S18) was written in the Julia programming language and is available at `https://github.com/natevmp/competitive-selection`. The Python code to recreate the figures is hosted at `https://github.com/fraterenz/hsc-draft`.

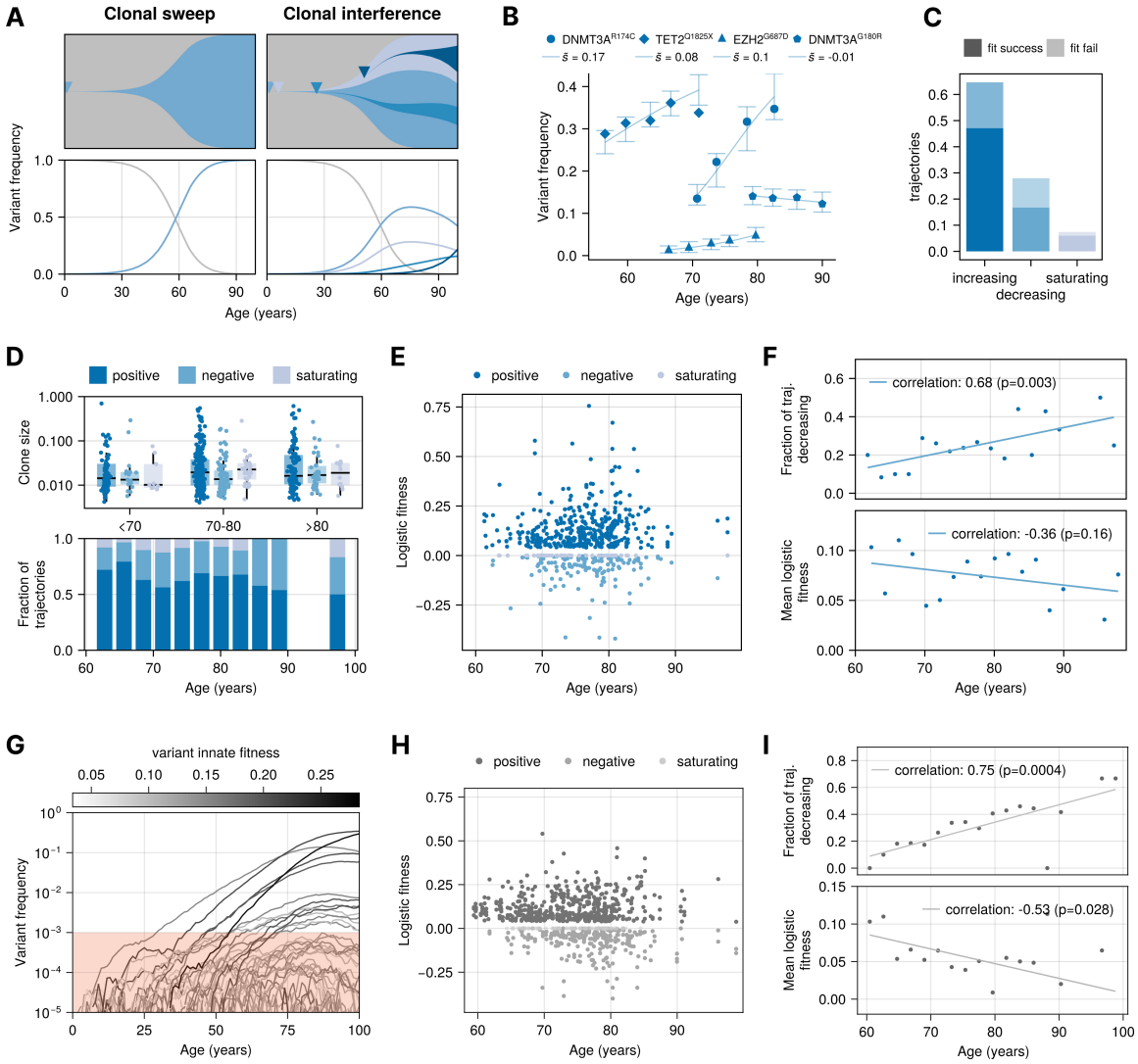


**Figure 1. Changing fitness landscape of human haematopoiesis with age. A)** Schematic of the expected trajectories of CH as clonal sweeps (fixation) or patterns of clonal interference. **B)** Variant allele frequency trajectories of CH variants with age from donor data (blue symbols), logistic fits (blue line), and standard errors. **C)** Fraction of increasing, decreasing, and saturating logistic fitnesses over 697 fitted variant trajectories. **D)** Clone sizes and trajectory composition with age. **E)** Logistic fitness coefficients $s_{\log}$ obtained from maximum likelihood estimation on all 697 variant trajectories. **F)** The fraction of decreasing trajectories increases with age (Spearman correlation = 0.68, $p = 0.003$). Mean logistic fitness decreases with age (Spearman correlation = -0.36, $p = 0.16$). **G)** Realisation of a stochastic simulation of CH from Equation (1). **H)** Logistic fitness coefficients $s_{\log}$ measured from simulations of CH with parameters $\eta = 0.11$, $\sigma = 0.032$, and $\mu = 5$. **I)** The fraction of decreasing trajectories also increases with age in stochastic simulations (Spearman correlation = 0.75, $p = 0.0004$) and mean fitness decreases with age (Spearman correlation = - 0.53, $p = 0.028$).

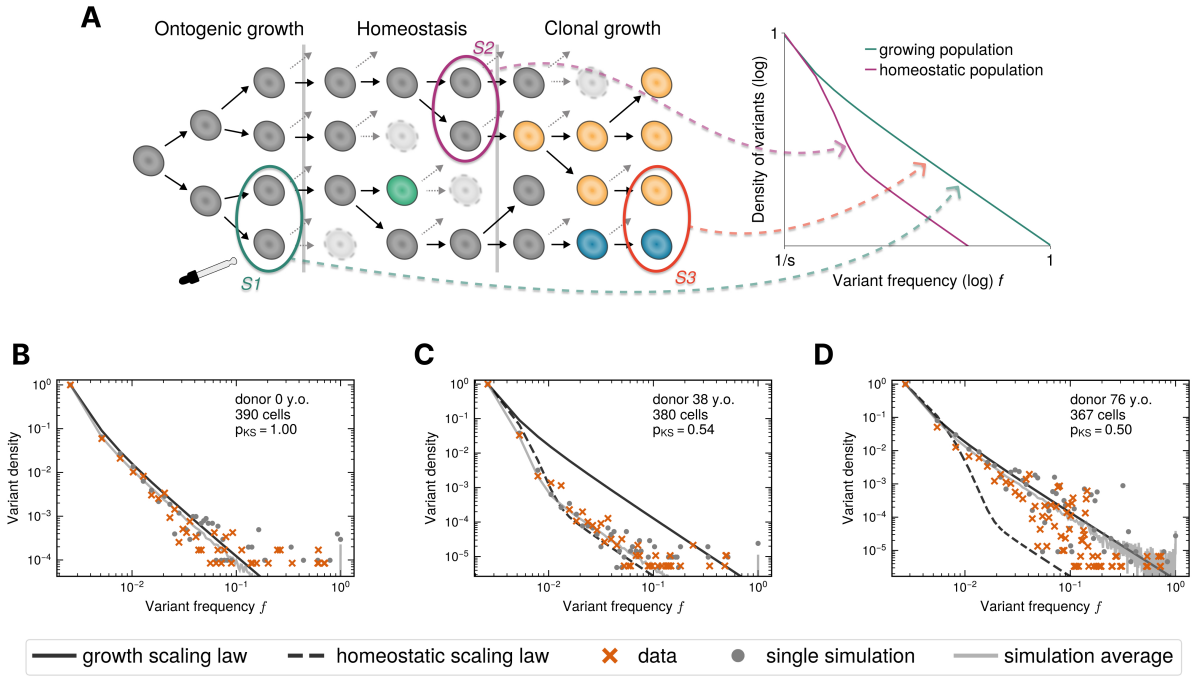


**Figure 2. Transitions of HSC site frequency spectra with age. A)** Schematic of the expected scaling shift of the SFS under ontogenic growth, homeostatisis, and clonal expansion of HSCs. Theory predicts three distinct phases of the SFS with age in the presence of clonal competition. **B)** The SFS of a newborn reconstructed from 390 whole genome sequenced single cells (orange crosses), a single realisation of the agent-based simulation (grey dots), averages over 500 simulations (grey line) and the expected scaling of an exponentially growing population (black line). **C)** SFS in a 38 year old donor constructed from 380 single HSCs. The SFS has shifted towards the expected scaling of a homeostatic population of constant size (black dashed line). **D)** SFS in a 76 year old donor constructed from 367 single HSCs. The SFS reverts to the scaling of a growing population at frequencies below $< 0.1$ (black line) and subclonal peaks appear.

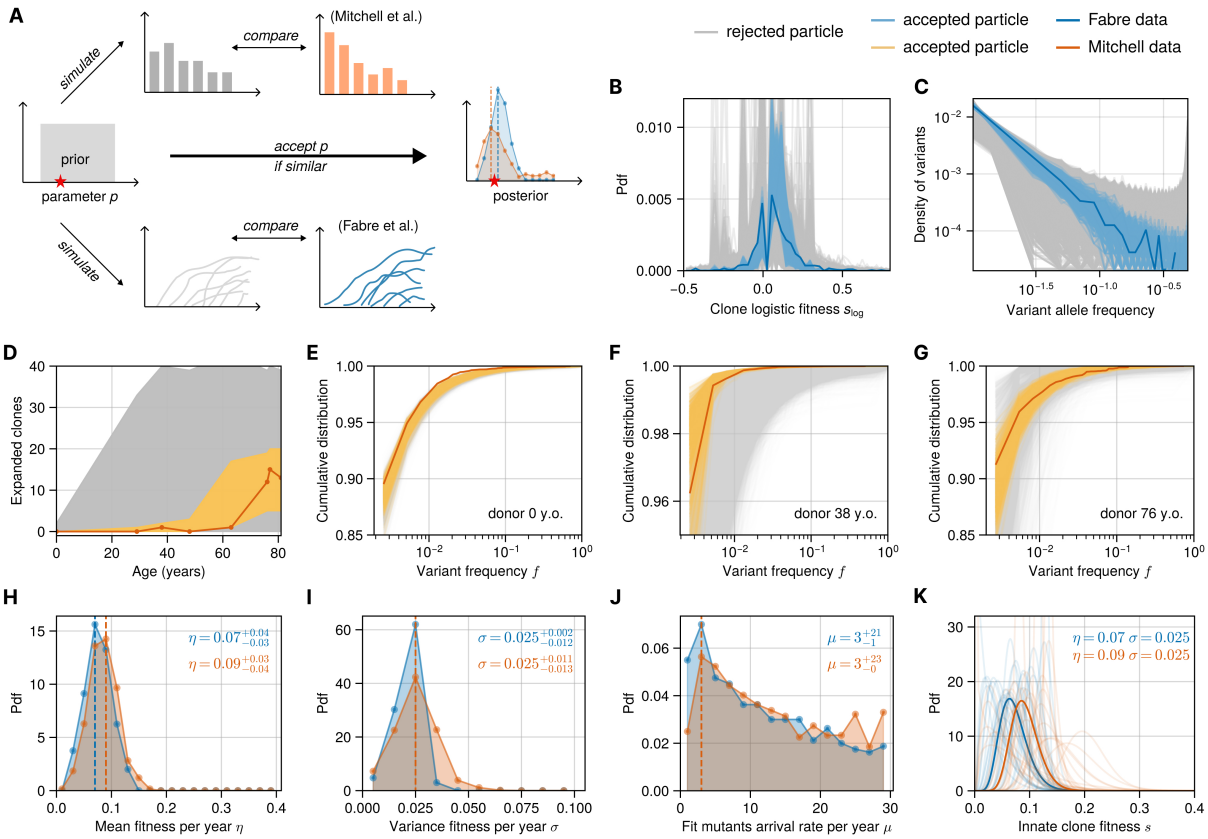


**Figure 3. Bayesian estimates for the fitness distribution and arrival rate of advantageous clones in CH. A)** Schematic of the approximate Bayesian computation (ABC) framework. **B)** Logistic fitness distribution $s_{\log}$ *in vivo* (dark blue lines), in ABC accepted runs (light blue lines), and in ABC rejected runs (grey lines). **C)** Clone size distribution *in vivo* (dark blue lines), in ABC accepted runs (light blue lines), and in ABC rejected runs (grey lines) runs. **D)** Number of visible clones at 1% resolution *in vivo* (orange dots), in ABC accepted runs (orange shaded area), and in ABC rejected runs (grey shaded area). **E-G)** Cumulative variant allele frequency distribution of HSCs in newborn, 38 year old, and 76 year old donors (dark orange lines), in ABC accepted runs (light orange lines), and in ABC rejected runs (grey lines). **H-J)** Posterior ABC parameter estimates based on trajectories (blue) and SFS (orange) for the mean $\eta$ and the standard deviation $\sigma$ of the innate fitness distribution of advantageous variants, and the HSC population arrival rate $\mu$ per year of advantageous variants. **K)** Best estimates of the innate fitness distribution of advantageous variants based on trajectories (blue) and SFS (orange). Thin lines represent individual accepted runs. Thick lines correspond to the estimated mean and standard deviation.

## References


[1] Francis Blokzijl et al. "Tissue-specific mutation accumulation in human adult stem cells during life". en. In: *Nature* 538.7624 (Oct. 2016), pp. 260–264.

[2] Henry Lee-Six et al. "Population dynamics of normal human blood inferred from somatic mutations". en. In: *Nature* 561.7724 (Sept. 2018), pp. 473–478.

[3] Iñigo Martincorena et al. "Somatic mutant clones colonize the human esophagus with age". en. In: *Science* 362.6417 (Nov. 2018), pp. 911–917.

[4] Benjamin Werner et al. "Measuring single cell divisions in human tissues from multi-region sequencing data". en. In: *Nat. Commun.* 11.1 (Feb. 2020), p. 1035.

[5] Federico Abascal et al. "Somatic mutation landscapes at single-molecule resolution". en. In: *Nature* 593.7859 (May 2021), pp. 405–410.

[6] Inigo Martincorena. "Somatic mutation and clonal expansions in human tissues". en. In: *Genome Med.* 11.1 (May 2019), p. 35.

[7] Caroline J Watson et al. "The evolutionary dynamics and fitness landscape of clonal hematopoiesis". en. In: *Science* 367.6485 (Mar. 2020), pp. 1449–1454.

[8] Gladys Y P Poon et al. "Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues". en. In: *Nat. Genet.* 53.11 (Nov. 2021), pp. 1597–1605.

[9] Amir Jassim et al. "Cancers make their own luck: theories of cancer origins". en. In: *Nat. Rev. Cancer* 23.10 (Oct. 2023), pp. 710–724.

[10] Marius E Moeller et al. "Measures of genetic diversification in somatic tissues at bulk and single-cell resolution". In: *eLife* 12 (2024), RP89780.

[11] Siddhartha Jaiswal et al. "Age-related clonal hematopoiesis associated with adverse outcomes". In: *New England Journal of Medicine* 371.26 (2014), pp. 2488–2498.

[12] Giulio Genovese et al. "Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence". In: *New England Journal of Medicine* 371.26 (2014), pp. 2477–2487.

[13] Florian Zink et al. "Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly". In: *Blood, The Journal of the American Society of Hematology* 130.6 (2017), pp. 742–752.

[14] Siddhartha Jaiswal and Benjamin L Ebert. "Clonal hematopoiesis in human aging and disease". en. In: *Science* 366.6465 (Nov. 2019), eaan4673.

[15] Margarete A Fabre et al. "The longitudinal dynamics and natural history of clonal haematopoiesis". en. In: *Nature* 606.7913 (June 2022), pp. 335–342.

[16] Emily Mitchell et al. "Clonal dynamics of haematopoiesis across the human lifespan". en. In: *Nature* 606.7913 (June 2022), pp. 343–350.

[17] Neil A Robertson et al. "Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects". In: *Nature medicine* 28.7 (2022), pp. 1439–1446.

[18] Hamish AJ MacGregor, Douglas Easton, and Jamie R Blundell. "Modelling the age-related deceleration of clonal haematopoiesis in UK Biobank". In: *bioRxiv* (2023), pp. 2023–12.

[19] Warren John Ewens. *Mathematical population genetics: theoretical introduction*. Vol. 1. Springer, 2004.

[20] Peter Czuppon and Arne Traulsen. "Understanding evolutionary and ecological dynamics using a continuum limit". In: *Ecology and Evolution* 11.11 (2021), pp. 5857–5873.

[21] Isabelle A van Zeventer et al. "Evolutionary landscape of clonal hematopoiesis in 3,359 individuals from the general population". en. In: *Cancer Cell* 41.6 (June 2023), 1017–1031.e4.

[22] Karen M Page and Martin A Nowak. "Unifying evolutionary dynamics". en. In: *J. Theor. Biol.* 219.1 (Nov. 2002), pp. 93–98.

[23] Daniel T. Gillespie. "Stochastic Simulation of Chemical Kinetics". In: *Annu. Rev. Phys. Chem.* 58 (2007), pp. 35–55.

[24] Gregory I Lang et al. "Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations". In: *Nature* 500.7464 (2013), pp. 571–574.

[25] Einar Bjarki Gunnarsson, Kevin Leder, and Jasmine Foo. "Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability". en. In: *Theor. Popul. Biol.* 142 (Dec. 2021), pp. 67–90.

[26] Katalin Csilléry et al. "Approximate Bayesian Computation (ABC) in practice". en. In: *Trends Ecol. Evol.* 25.7 (July 2010), pp. 410–418.

[27] Marc J Williams et al. "Quantification of subclonal selection in cancer from bulk sequencing data". In: *Nature genetics* 50.6 (2018), pp. 895–903.

[28] Caroline J Watson and Jamie R Blundell. "Mutation rates and fitness consequences of mosaic chromosomal alterations in blood". In: *Nature genetics* 55.10 (2023), pp. 1677–1685.

[29] Zahraa Rahal, Paul Scheet, and Humam Kadara. "Somatic Mutations in Normal Tissues: Calm before the Storm". In: *Cancer Discovery* 14.4 (Apr. 2024), 605–609. ISSN: 2159-8290. DOI: 10.1158/2159-8290.cd-23-1508. URL: http://dx.doi.org/10.1158/2159-8290.CD-23-1508.

# Supplemental Information

## Contents


**SI 1 Maximum likelihood estimation of logistic trajectories** **12**

**SI 2 Clonal competition as a multi-type capacity-limited birth-death process** **13**
SI 2.1 The SDE system for clonal competition . . . 14
SI 2.1.1 Deriving the SDE system . . . 14
SI 2.1.2 Numerical simulations of the SDE system . . . 15

**SI 3 Bayesian inference on clonal trajectories to estimate parameter values** **16**
SI 3.1 Testing the ABC convergence with simulated data . . . 16

**SI 4 Accumulation of neutral mutations in the multi-type birth-death process** **17**
SI 4.1 Ontogenic growth of the HSC population . . . 17
SI 4.2 Agent-based model with inclusion of neutral mutations . . . 17
SI 4.3 Model parameterisation . . . 18

**SI 5 Bayesian inference on site frequency spectra to estimate the fitness distribution** **20**
SI 5.1 Summary statistic . . . 20
SI 5.2 Longitudinal inference using the SFS and the number of clones . . . 20
SI 5.3 Testing the Bayesian framework on simulated data . . . 21


## SI 1 Maximum likelihood estimation of logistic trajectories

The logistic function – obtained as the solution to the differential equation $\mathrm{d}x(t)/\mathrm{d}t = s_{\log}\, x(t)[1 - x(t)]$ – describes the expected growth curve of a mutant under selection in a constant-size population with fixed background fitness [1]. Because the observed variant trajectories are sigmoidal, it can be used to approximate a trajectory in a short time window. Integrating the differential, the clone size is given by

$$x(t) = \frac{K}{1 + \frac{K - x_0}{x_0} e^{-s_{\log}(t - t_0)}} \tag{S2}$$

where $x_0$ is the clone's initial size at time $t_0$ and $K$ is the carrying capacity. Importantly this expression holds for both positively and negatively selection mutants, since in case of the latter flipping the sign of $s_{\log}$ is akin to flipping the sign of $t$, and at any point in time both subpopulations must have their sizes sum to $K$. If a trajectory is well-described by Equation (S2), the rate parameter $s_{\log}$ functions as a measure for the selection strength acting on the clone in the observed time window.

We fit all trajectories reported by Fabre et al. [2] to Equation (S2) using a maximum likelihood estimation as follows. We assumed each measurement of a variant allele's frequency $\tilde{x}_i$ to be a binomial sampling, where the success probability is given by the (unknown) true frequency $x_i$ and the number of trials is given by the coverage $V$ of that site. Thus for any proposed trajectory $y(t|\beta)$ (where $\beta$ is the set of parameters to fit), we could compute the probability of obtaining the observed sequence of

measurements $\tilde{X}_i = \tilde{x}_i V$ at times $t_i$ as

$$\mathcal{L}(\beta) = \prod_i \mathbb{P}\{\tilde{X}_i \mid y(t_i|\beta), V\} \tag{S3}$$

where

$$\mathbb{P}\{\tilde{X}_i \mid V, y(t_i|\beta)\} = \binom{V}{\tilde{X}_i} y(t_i|\beta)^{\tilde{X}_i}[1 - y(t_i|\beta)]^{V-\tilde{X}_i} \tag{S4}$$

which is the binomial probability of sampling successfully $\tilde{X}_i$ times out of $V$ trials with success probability $y(t_i|\beta)$. Equation (S3) is known as the likelihood function, and is maximized with respect to $\beta$ to obtain the best fit. For each trajectory we maximized three different likelihood functions: one assuming the underlying trajectory is increasing, another assuming it is decreasing, and finally one assuming it is constant ($y(t) = x_0$). In the case of an increasing trajectory on a non-sex chromosome we took $K = 0.5$ and $x_0 = 1/2N$ as fixed parameters and – assuming $s_{log} > 0$ – maximized with respect to $\beta = \{\tau, s_{log}\}$. For a decreasing trajectory we took $\beta = \{\tau, x_0, s_{log}\}$ ($x_0$ is then the size at which the clone became negatively selected) and instead assumed $s_{log} < 0$. Finally for the saturating trajectory $s_{log} = 0$ and the single free parameter $\beta = \{x_0\}$ (the fixed true clone frequency) was fit. The trajectory with the highest maximum likelihood was then taken as the best fit.

# SI 2 Clonal competition as a multi-type capacity-limited birth-death process

We model the haematopoietic stem cell compartment as a collection of cells, each carrying a set of genetic mutations. These cells individually undergo divisions and death as independent Poisson processes in time, where in case of the former the cell is replaced by two daughter cells inheriting the set of mutations from the parent, and in case of the latter the cell is removed from the population. We consider first only fitness-granting mutations which alter the birth rate of a cell. We take these to occur as a Poisson process in time, and assume each new mutation is distinct from those already existing, a supposition typically referred to as the *infinite sites model*. Denoting the set of all fitness carrying mutations as $\mathcal{I}$, a cell of type $i \in \mathcal{I}$ as $\mathcal{X}_i$, and the birth, death, and mutation rates as $\beta$, $\delta$, and $\mu$ respectively, the model is summarized through the following reactions:

$$\begin{aligned} \mathcal{X}_i &\xrightarrow{\beta_i} \mathcal{X}_i + \mathcal{X}_i \\ \mathcal{X}_i &\xrightarrow{\delta_i} \emptyset \\ \begin{cases} \mathcal{X}_i \\ \mathcal{I}, j \notin \mathcal{I} \end{cases} &\xrightarrow{\mu} \begin{cases} \mathcal{X}_j \\ \mathcal{I} \cup j \end{cases} \end{aligned} \tag{S5}$$

An important constraint is that the total stem cell population is expected to remain at a constant size throughout most of the human lifespan [3, 4]. We take this to occur by enforcing an equality of the total population-level birth and death rates:

$$\mathcal{B} = \sum_{c \in \text{cells}} \beta_c = \mathcal{D} = \sum_{c \in \text{cells}} \delta_c \tag{S6}$$

with $\beta_c$ and $\delta_c$ the birth and death rates of cell $c$. If all cells are of the same wild type – having identical birth and death rates – this is trivially realized by setting $\beta_c = \delta_d = \beta = \delta$, $\forall c, d$. This amounts to a

completely neutral population, where $\mathcal{B} = N\beta = \mathcal{D} = N\delta$ for a population of size $N$. If there are $K$ types the population-level rates can instead be written as

$$\mathcal{B} = \sum_{k=1}^{K} \beta_k X_k = \mathcal{D} = \sum_{k=1}^{K} \delta_k X_k \tag{S7}$$

where $X_k$ denotes the number of cells of type $k$, and all cells of that type share the birth and death rates $\beta_k$ and $\delta_k$. We define selection acting on a type $k$ as an altered birth rate with respect to the wild type through a fitness parameter $s_k$, and a fixed death rate across all types $\forall k : \delta_k = \mathcal{D}/N$:

$$\beta_k = \beta + s_k \tag{S8}$$

$$\delta_k = \beta + \frac{\sum_l s_l X_l}{N} \tag{S9}$$

## SI 2.1 The SDE system for clonal competition

### SI 2.1.1 Deriving the SDE system

From the rates (S8) and (S9) a master equation can be constructed for the probability distribution of the size of type $k$, where we use the shorthand notation $P_n^{(k)} = \mathbb{P}\{X_k = n\}$:

$$\begin{aligned} \frac{\mathrm{d}P_n^{(k)}}{\mathrm{d}t} &= \sum_m \left[ T_{m\to n}^{(k)} P_m^{(k)} - T_{n\to m}^{(k)} P_n^{(k)} \right] \\ &= T_{n-1,\uparrow}^{(k)} P_{n-1}^{(k)} + T_{n+1,\downarrow}^{(k)} P_{n+1}^{(k)} \\ &\quad - T_{n,\uparrow}^{(k)} P_n^{(k)} - T_{n,\downarrow}^{(k)} P_n^{(k)} \end{aligned} \tag{S10}$$

This expression contains the transition rates $T_{n,\uparrow\downarrow}^{(k)}$ for the size of clone $k$ to increase or decrease by 1:

$$\begin{cases} T_{n,\uparrow}^{(k)} = n\beta_k \\ T_{n,\downarrow}^{(k)} = n\delta_k \end{cases} \tag{S11}$$

Assuming that $N$ is large, meaning that there are many possible values of $X_k/N$ between 0 and 1, we construct an analogous stochastic process where the size is given by a continuous variable $x_k \in [0,1] \subset \mathbb{R}$, representing the frequency of clone $k$ in the population. One convenient way to do this (see for example [5]) is by writing the master equation as a Kramers-Moyal expansion. First, noting that the transition rates here describe displacements of distance $\Delta x = 1/N$, i.e. $T_{n,\uparrow\downarrow}^{(k)} \to T_{\pm\Delta x}^{(k)}(x) = T_{\uparrow\downarrow}^{(k)}(x)$, we rewrite (S10) in terms of a Taylor expansion of the continuous parameter $x$:

$$\begin{aligned} \frac{\mathrm{d}P^{(k)}(x)}{\mathrm{d}t} &= \left[ \sum_{i=0}^{\infty} \frac{(-\Delta x)^i}{i!} \left( \frac{\partial}{\partial x} \right)^i - 1 \right] T_{\uparrow}^{(k)}(x) P^{(k)}(x) \\ &\quad + \left[ \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} \left( \frac{\partial}{\partial x} \right)^i - 1 \right] T_{\downarrow}^{(k)}(x) P^{(k)}(x) \\ &= \sum_{i=1}^{\infty} \left( -\frac{\partial}{\partial x} \right)^i D_i(x) P^{(k)}(x) \end{aligned} \tag{S12}$$

where the final equality immediately describes the Kramers-Moyal expansion with coefficients

$$D_i(x) = (\Delta x^i / i!)\left[T_{\uparrow}^{(k)}(x) + (-1)^i T_{\downarrow}^{(k)}(x)\right] \tag{S13}$$

Using our assumption of small steps $\Delta x = 1/N$ we truncate after the second term, to obtain a Fokker-Planck equation with coefficients

$$\begin{aligned} D_1(x) &= x(\beta_k - \delta_k) \\ D_2(x) &= \frac{x}{2N}(\beta_k + \delta_k) \end{aligned} \tag{S14}$$

These can be used to construct an equivalent stochastic differential equation

$$\mathrm{d}x = \mu(x,t)\,\mathrm{d}t + \sigma(x,t)\,\mathrm{d}W_t \tag{S15}$$

where in the Itô formulation the drift and diffusion coefficients take the form

$$\mu(x,t) = D_1(x,t) \tag{S16}$$

$$\sigma(x,t) = \sqrt{2D_2(x,t)} \tag{S17}$$

Thus, plugging in (S8) and (S9), we have

$$\mathrm{d}x_k = s_k x_k \left(1 - \sum_l \frac{s_l}{s_k} x_l\right)\mathrm{d}t + \sqrt{\frac{1}{N}\left(2\beta + s_i + \sum_j s_j x_j\right) x_i}\,\mathrm{d}W_t \tag{S18}$$

which describes a set of coupled stochastic differential equations for $k = 1, 2, \ldots, K$.

#### SI 2.1.2 Numerical simulations of the SDE system

With Equation (S18) describing the time-evolution of existing variants clones, all that remains to obtain the model described by Equation (S5) is the introduction of newly arriving clones in time. It suffices to take the total number of equations $K$ to be a Poisson process in time, and have newly introduced clones initiate at size $1/N$:

$$\begin{cases} \mathrm{d}x_k = s_k x_k \left(1 - \sum_l \frac{s_l}{s_k} x_l\right)\mathrm{d}t + \sqrt{\frac{1}{N}\left(2\beta + s_i + \sum_j s_j x_j\right) x_i}\,\mathrm{d}W_t\,, \quad 0 < x_k < 1 \wedge t > t_k \\ \mathrm{d}x_k = 0, \qquad x_k \leq 0 \vee x_k \geq 1 \\ \quad x_k = 0, \qquad t < t_k \\ \quad x_k = 1/N, \qquad t = t_k \\ \quad t_k = t_{k-1} + \tau_k;\ \tau_k \sim \mathrm{Exp}(\mu) \end{cases} \tag{S19}$$

Upon introduction, a clone $k$ has its fitness $s_k$ drawn from a gamma distribution $s_k \sim \mathrm{Gamma}\left(\frac{\eta^2}{\sigma^2}, \frac{\sigma^2}{\eta}\right)$ with mean $\eta$ and variance $\sigma^2$. Numerical solution trajectories of this system were calculated using the Julia computing language and the package 'DifferentialEquations.jl' [6–8].

# SI 3 Bayesian inference on clonal trajectories to estimate parameter values

The Bayesian inference on the clonal trajectories (Fabre et al. [2]) was performed in the following manner. First 40'000 parameter combinations $\{P_i\}$ – hereafter referred to as particles – were drawn randomly from uniform priors for each of the $n$ parameters to be investigated. Next, for each particle a simulated dataset was generated from the SDE system, from which the variant size-distribution and logistic fitness distribution were measured and compared to those of the donor data using a chisquare metric. Finally a threshold was selected such that only the 1% best performing particles (i.e. having the highest similarity with the Fabre data according to the metric) were retained, resulting in a $n$-dimensional posterior distribution of the parameter space (Figure SI 2). Sensitivity of the posteriors was tested by varying the number of particles accepted, at 0.1%, 0.5%, and 5% (Figure SI 4).

To construct the simulated dataset for a single particle, 300 realizations of the SDE system were performed, representing a large cohort of virtual donors. Each run was measured at time $T$, and of the variants exceeding a detection threshold $x_{min} = 0.002$, only three were chosen for tracking, to obtain a similar number of total variants as well as variants per donor in the data. Because trajectories in the dataset were captured through targeted sequencing of known driver genes, it is possible that the observed variants would be biased towards higher innate fitness compared to a randomly selected set of drivers, since knowledge of a driver may correlate with its fitness. To account for this we introduced another parameter $q$ describing the probability of an observed driver possessing the highest innate fitness in the system. In other words, after ranking all variants in a simulated individual by their innate fitness, each of the observed variants has a probability $q$ of being chosen in ranked order, and a probability $1 - q$ of being selected randomly. Thus for $q$=1, observed variants are maximally biased towards high innate fitness, whereas for $q$=0 variant observation is entirely random. From the pooled set of selected variants across all simulated individuals, the size distribution was constructed by binning variant sizes at the time of measurement into 25 bins covering the full VAF range $[x_{min}/2, 0.5]$ (the size of a clone is twice its variant allele frequency). For the fitness distribution, each sampled variant was measured at the four equidistant time points spanning 10 years centered around T: $\{T - 5, T - 1.66, T + 1.66, T + 1.66\}$. The maximum likelihood fitting described in Section SI 1 was then performed to obtain all variants' logistic fitness coefficients $s_{\log}$. which were partitioned into 50 equidistant bins constituting the fitness distribution. Because both the size- and logistic fitness distributions are expected to change with the age of observed individuals, the donor dataset was partitioned into age groups covering 6 year spans, and the comparison with simulations was performed with $T$ at the center of the respective age group, for example variants measured between 70-76 years of age were compared with simulations measured at $T = 73$.

## SI 3.1 Testing the ABC convergence with simulated data

To test the ability of the ABC framework to accurately identify unknown parameter values, we performed the inference on 1000 simulated datasets with known parameters. As inferred values we took the modes of the posterior distributions. Taking the Euclidean distance between inferred and true parameter value as the error, we then measured the distribution of errors across all inferences for each parameter (Figure SI 3B). For all parameters the error distributions have modes near 0 and mean errors less than 1 fold-change.

# SI 4 Accumulation of neutral mutations in the multi-type birth-death process

To study site frequency spectra (SFS) we must include the accumulation of neutral mutations in individual haematopoietic stem cells, as these are expected to form the majority of the cell mutational burden. Current evidence suggests mutations can occur both during cell division due to replication errors, as well as in between divisions due to persistent background stresses on the genome (e.g. radiation, oxidation, etc.) and imperfect DNA repair [9]. We must distinguish between these two forms of mutation as they lead to different distributions of the mutational burden in single cells [10]. The number of division-associated mutations in a cell scales with the number of divisions its parental lineage has undergone, whereas the number of background-associated divisions scales directly with time. We take division-associated mutations to be acquired as a Poisson distributed number during each division, and background mutations as a Poisson process in time. To this end we expand upon the events described in Equation (S5). Denoting $\mathcal{X}_{k,\{u\}}$ a cell of type $k$ with set of mutations $\{u\}$, and $U$ the set of all existing mutations, division-associated mutations are introduced during birth events

$$\begin{cases} \mathcal{X}_{k,\{u\}} \\ U \cap \{v\}, \{w\} = \emptyset \end{cases} \xrightarrow{\beta_k} \begin{cases} \mathcal{X}_{k,\{u\}\cup\{v\}} + \mathcal{X}_{k,\{u\}\cup\{w\}} \\ U \cup \{v\}, \{w\} \end{cases} \tag{S20}$$

and background mutations establish the additional event

$$\begin{cases} \mathcal{X}_{k,\{u\}} \\ U, z \notin U \end{cases} \xrightarrow{\nu} \begin{cases} \mathcal{X}_{k,\{u\}\cup z} \\ U \cup z \end{cases} \tag{S21}$$

## SI 4.1 Ontogenic growth of the HSC population

Because population expansion drives neutral mutations to high frequency, the shape of the SFS reflects changes in a population's size [11]. These changes can occur gradually, such that SFS measured from blood in adult humans still contain signatures of the early development of the HSC pool. In order for a model to accurately capture the changing SFS in an individual, it must therefore also incorporate early ontogenic growth [10]. To this end we include an exponential growth phase preceding the previously-described constant population dynamics, starting from a single cell and lasting until the adult size of the HSC pool $N$ is reached, which we take as the time of birth (0 years old). We take this phase to occur as a pure-birth process without cell death.

## SI 4.2 Agent-based model with inclusion of neutral mutations

We implemented the clonal competition model with accumulation of neutral mutations as a Moran model [12] via the stochastic simulation algorithm (SSA, also referred to as the Gillespie algorithm [13]). The SSA generates stochastic trajectories of the model by iteratively simulating the events described in Equations (S5), (S20), and (S21) in a system of cells. Both the timing between events and the events themselves are drawn randomly, so that each simulation is a unique realisation of the underlying stochastic process. In our model we assume a *pure-birth process* for the ontogenic growth phase: each event describes a single cell division and results in the appearance of a new cell in the system. In the constant-population phase we assume *Moran* dynamics: this has each event consisting of a birth and a simultaneously occurring death, which enforces an equality of the total birth and death rates and thus

keeps the population size constant over time. All events are assumed to occur as Poisson processes in time, i.e. having exponentially distributed waiting times. This means that following an event at time $t$, from Equation (S20) the time until the next event is given by

$$\Delta t \sim \text{Exponential}\left(\sum_k \beta_k x_k(t)\right) \tag{S22}$$

with $x_k$ denoting the size of clone $k$. The clone to which the proliferating cell belongs is then drawn randomly from the weighted set of existing clones, with the likelihood of clone $k$ being selected given by $x_k s_k / \sum_l x_l s_l$. Repeating this sampling procedure generates a sequence of proliferating cells and inter-division times, representing a random realisation of the process described by Equation (S10) without requiring an analytical solution for $P_n^{(k)}$.

We implemented the SSA in the Rust programming language [14, 15] as an agent-based model. Each agent represents a single HSC, carries a set of mutations $\{u\}$, and belongs to a clone $k$. The proprieties of the agents define the state of the system at certain time, and we are interested in understanding the changes over time of the state of the system. Upon proliferation, a random cell belonging to the $k$-th clone is selected and assigned $m_b \sim \text{Pois}(\lambda_b \Delta t)$ background neutral mutations, where $\Delta \tilde{t}$ is the difference between the current time and the last time the cell divided. Then a Bernoulli trial with probability of success $\mu \Delta t / N$ is performed to determine whether to allocate the cell to a new empty clone. The dependence on the stochastic inter-division time decouples the occurrence of fit clones from proliferation events, ensuring that our framework is agnostic to the type of event that generates new clones: they do not acquire a proliferative advantage exclusively via division-associated genetic mutations, but also due to background mutations, epigenetic alterations, or other environmental factors. When a novel clone is created its fitness $s_k$ is drawn from a Gamma distribution. This quantifies how fast cells in the clone proliferate according to Equation (S8). Finally, the dividing cell is duplicated and each daughter cell is assigned a set of random mutations $m_d \sim \text{Pois}(\mu_d)$. If the system has passed the growth phase, a random cell drawn uniformly across all clones is then removed from the population.

Repeating these steps many times allowed us to simulate both clonal dynamics and the accumulation of neutral mutations that act as lineage markers in individual cells. Each simulations started with a single wild-type HSC growing exponentially until a population size of $N$ HSCs was reached, at which point the dynamics switched to the constant-population (Moran) phase. In order to match the mutational burden distributions obtained from HSCs from cord blood in 2 neonates (Figure SI 6D) We took the exponential phase to last for 5 months, so that the constant-size phase initiated 4 months before birth (see SI 4.3).

## SI 4.3 Model parameterisation

Having extended the model to incorporate distinct mutations in individual cells, we aimed to understand the dynamics of clonal haematopoiesis from single-cell resolution data published in Mitchell et al. [4]. All parameters required to perform a simulation of the ABM are shown in Table 1. Four of these were estimated from a Bayesian model (see SI 5), while the rest were either fixed from the literature or parameterised using additional information from the data. From current knowledge we set the rate of neutral mutations upon division $\mu_d$ to 1.14 [3, 16], and fixed the number of stem cells $N$ to $10^5$ cells, based on recent estimates [3, 4, 10, 17] and the fact that the ABC inference performed on clone trajectories displayed limited sensitivity to this parameter within a wide range. Finally, we parameterised the rate of accumulation of neutral background mutations using single-cell mutational

| Symbol | Description | Units | Inference |
|---|---|---|---|
| $N$ | Number of haematopoietic stem cells at maturity | cell number | Fixed to $10^5$ |
| $\tau$ | Inter-division time for wild-type cells | year | Inferred |
| $\eta$ | Mean of the Gamma distribution of innate fitnes coefficients | year$^{-1}$ | Inferred |
| $\sigma$ | Standard deviation of the Gamma distribution of innate fitness coefficients | year$^{-1}$ | Inferred |
| $\mu$ | Arrival rate of advantageous mutations | year$^{-1}$ | Inferred |
| $\mu_{\bar{b}}$ | Arrival rate of prenatal neutral background mutations | year$^{-1}$ | Parameterised to 35.64 |
| $\mu_b$ | Arrival rate of postnatal neutral background mutations | year$^{-1}$ | Parameterised to $\mu_b = 14.35 - \mu_d/\tau$ |
| $\mu_d$ | Arrival rate of neutral mutations upon division | division$^{-1}$ | Fixed to 1.14 |

**Table 1.** Parameters of the agent-based model. Parameters were inferred either through the ABC framework described in SI 5 or by regression of the mean of the number of mutations as described in the text.

burden information from the data (shown in Figure SI 6A). To accurately capture the dynamics of the burden observed in the data, we assumed different values of this parameter before and after birth, denoted $\mu_{\bar{b}}$ and $\mu_b$ respectively.

For the prenatal rate we took mutations to occur as following: an initial growth phase has mutations occurring through both divisions and background processes until the maximal population size $N$ is reached, then for the remaining time until birth only background mutations are added. Having the growth process complete before birth in this manner is required in order for the simulations to match the SFS of the newborns (Figure SI 8A). The value of $\mu_{\bar{b}}$ is then obtained from

$$\tilde{m} = a\mu_d + \mu_{\bar{b}} t_{\text{birth}} \tag{S23}$$

where $\tilde{m}$ is the expected burden distribution upon birth, $t_{\text{birth}}$=9/12 is the total time elapsed before birth, and $a = 2\log((N+1)/2)$ the expected number of divisions per lineage once the population has reached size $N$ [10]. Taking for $\tilde{m}$ the mean of the burden distributions measured from the two newborns in the data, this results in $\mu_{\bar{b}} = 35.64$. The time $t_{\text{growth}} < t_{\text{birth}}$ at which the growing population reaches size $N$ determines how much time the system spends in a constant phase before birth. We chose this value such that simulations match the SFS of the neonates in the data, which we found to be at $t_{\text{birth}}$=5 months (Figure SI 8A).

To compute the postnatal background mutation rate $\mu_b$, we regressed the means of the burden distributions for all individuals with no detected clones (i.e. individuals with 0, 29 and 48 years). We found an increase in the average burden of 14.35 mutations per year per cell (Figure SI 6B), and used as rate $\mu_b = 14.35 - \mu_d/\tau$. In this manner our model accurately captures the mean of the burden in data, however it greatly underestimates the variance, with a error that increases drastically with age. This suggests a high heterogeneity across cells in the number of mutations due to other processes not modelled in our framework (Figure SI 6C-D).

# SI 5 Bayesian inference on site frequency spectra to estimate the fitness distribution

We then investigated the identifiability of the underlying evolutionary parameters of CH from *in vivo* single-cell mutational data [4]. We first investigated the qualitative behaviour of the clonal competition model based on values of the arrival rate of advantageous clones $\mu$, and the mean $\eta$ and standard deviation $\sigma$ of the innate fitness distribution. Faster dynamics can be obtained by higher values of $\eta$ and $\mu$, and lower values of $\tau$ (Figure SI 7).

To test what combinations of these parameters $\{\tau, \mu, \eta, \sigma\}$ best describe the data, we implemented approximate Bayesian computation (ABC) inference schemes [18]. ABC is a Monte Carlo likelihood-free method that samples different parameter sets and stores those generating the simulations which are close to the data according to a metric of similarity.

## SI 5.1 Summary statistic

We implemented ABC by computing two summary statistics: the SFS $S$ and the number of clones $C$ with frequency greater than 1%. We quantified the similarity between the data $y$ and the simulations $z$ by computing the Wasserstein distance for the SFS $w(S_y, S_z)$ and the relative Euclidean distance for the number of detected clones $d(C_z, C_y) = \left|C_z - C_y\right| / C_z$. The Wasserstein distance, also called the Earth mover's distance or the optimal transport distance, is a similarity metric between two probability distributions which computes the unsigned average distance between the inverse cumulative distribution of $S_z$ and $S_y$ [19].

## SI 5.2 Longitudinal inference using the SFS and the number of clones

We performed a longitudinal inference to estimate the values of $\{\tau, \mu, \eta, \sigma\}$ from the single-colony data published in Mitchell et al. [4]. We generated 65'000 realisations of our agent-based model (of which 64'286 ran successfully) on the high-performance cluster [20] drawing different sets of parameters from uniform prior distributions. We saved the state of the simulations after subsampling the stem cell population to the number of cells found in the 8 donors at their respective ages (i.e. age 0, 29, 38, 48, 63, 76, 77, 81), and quantified the similarity between the simulations and data at the timepoint using the two summary statistics described above. We obtained the posterior distributions by storing the parameter sets that generated simulations with both metric distances within the 1% quantile for at least 7 out of 8 timepoints $i$, i.e. the parameter sets for which the simulations had $d(C_y^i, C_z^i) < q$ and $w(S_y^i, S_z^i) < q$, with $q$ being the 1% quantile. These filtered parameters are approximated samples of the posterior distributions that describe the data best. We summarised the posterior distributions into point estimates by taking their modes after binning them into histograms, reporting the 90% credible interval. The summary statistics for the 8 donors are shown in Figure SI 8 and Figure 3D.

The inferred posterior distributions show estimates of mean and standard deviation per year of $\eta = 0.09^{+0.03}_{-0.04}$ and $\sigma = 0.025^{+0.011}_{-0.013}$, and an arrival rate of advantageous variants of $\mu = 3^{+23}_{-0}$ (Figure SI 9). Taking the point estimates of the posterior distributions of $\eta$ and $\sigma$ allows us to plot the Gamma distribution of the fitness effects inferred from the data by mapping $\eta$ and $\sigma$ to the scale $\alpha$ and shape $\theta$: $\alpha = \eta^2/\sigma^2$ and $\theta = \sigma^2/\eta$. The recovered distribution of fitness effects is bell-shaped, suggesting that

most clones expand at similar rates (Figure 3K). Moreover, we find that fit clones expand slowly, with an average growth that is 9% faster per year compared to wild-type cells. The findings do not depend on the number of runs accepted by the Bayesian framework, the estimates being consistent across a range of accepted runs of 0.1%, 0.5% and 5% (Figure SI 11).

### SI 5.3 Testing the Bayesian framework on simulated data

To validate our inferential framework, we tested our implementation on synthetic data generated by the ABM simulations with $\eta = 0.10$, $\sigma = 0.723$, $\mu = 4.86$ and $\tau = 1.16$. We found $\eta = 0.09^{+0.11}_{-0.05}$, $\sigma = 0.055^{+0.02}_{-0.02}$, $\mu = 1^{+23}_{-0}$ and $\tau = 2.5^{+5.9}_{-0.09}$ (Figure SI 12), confirming the validity of the inferential framework for this parameter range.

## References


[1] Peter Czuppon and Arne Traulsen. "Understanding evolutionary and ecological dynamics using a continuum limit". In: *Ecology and Evolution* 11.11 (2021), pp. 5857–5873.

[2] Margarete A Fabre et al. "The longitudinal dynamics and natural history of clonal haematopoiesis". en. In: *Nature* 606.7913 (June 2022), pp. 335–342.

[3] Henry Lee-Six et al. "Population dynamics of normal human blood inferred from somatic mutations". en. In: *Nature* 561.7724 (Sept. 2018), pp. 473–478.

[4] Emily Mitchell et al. "Clonal dynamics of haematopoiesis across the human lifespan". en. In: *Nature* 606.7913 (June 2022), pp. 343–350.

[5] Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*. 2nd ed. Springer, 1996. Chap. 4.5.

[6] Christopher Rackauckas and Qing Nie. "Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in julia". In: *Journal of Open Research Software* 5.1 (2017), p. 15.

[7] Christopher Rackauckas and Qing Nie. "Adaptive methods for stochastic differential equations via natural embeddings and rejection sampling with memory". In: *Discrete and continuous dynamical systems. Series B* 22.7 (2017), p. 2731.

[8] Chris Rackauckas and Qing Nie. "Stability-optimized high order methods and stiffness detection for pathwise stiff stochastic differential equations". In: *2020 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE. 2020, pp. 1–8.

[9] Federico Abascal et al. "Somatic mutation landscapes at single-molecule resolution". en. In: *Nature* 593.7859 (May 2021), pp. 405–410.

[10] Marius E Moeller et al. "Measures of genetic diversification in somatic tissues at bulk and single-cell resolution". In: *eLife* 12 (2024), RP89780.

[11] Einar Bjarki Gunnarsson, Kevin Leder, and Jasmine Foo. "Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability". en. In: *Theor. Popul. Biol.* 142 (Dec. 2021), pp. 67–90.

[12] Warren John Ewens. *Mathematical population genetics: theoretical introduction*. Vol. 1. Springer, 2004.

[13] Daniel T. Gillespie. "Stochastic Simulation of Chemical Kinetics". In: *Annu. Rev. Phys. Chem.* 58 (2007), pp. 35–55.

[14] Nicholas D. Matsakis and Felix S. Klock. "The Rust Language". In: *ACM SIGAda Ada Letters* 34 (Oct. 2014), pp. 103–104. ISSN: 1094-3641. DOI: 10.1145/2692956.2663188.

[15] Jeffrey M. Perkel. "Why Scientists Are Turning to Rust". In: *Nature* 588 (Dec. 2020), pp. 185–186. ISSN: 1476-4687. DOI: 10.1038/d41586-020-03382-2.

[16] Benjamin Werner et al. "Measuring single cell divisions in human tissues from multi-region sequencing data". en. In: *Nat. Commun.* 11.1 (Feb. 2020), p. 1035.

[17] Caroline J Watson et al. "The evolutionary dynamics and fitness landscape of clonal hematopoiesis". en. In: *Science* 367.6485 (Mar. 2020), pp. 1449–1454.

[18] Mark A. Beaumont. "Approximate Bayesian Computation in Evolution and Ecology". In: *Annual Review of Ecology, Evolution, and Systematics* 41 (2010), pp. 379–406. DOI: 10.1146/annurev-ecolsys-102209-144621.

[19] Aaditya Ramdas, Nicolas Garcia, and Marco Cuturi. *On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests*. Oct. 2015. DOI: 10.48550/arXiv.1509.02237. arXiv: 1509.02237 [math, stat].

[20] Thomas King, Simon Butcher, and Lukasz Zalewski. "Apocrita - High Performance Computing Cluster for Queen Mary University of London". In: (Mar. 2017). DOI: 10.5281/zenodo.438045.

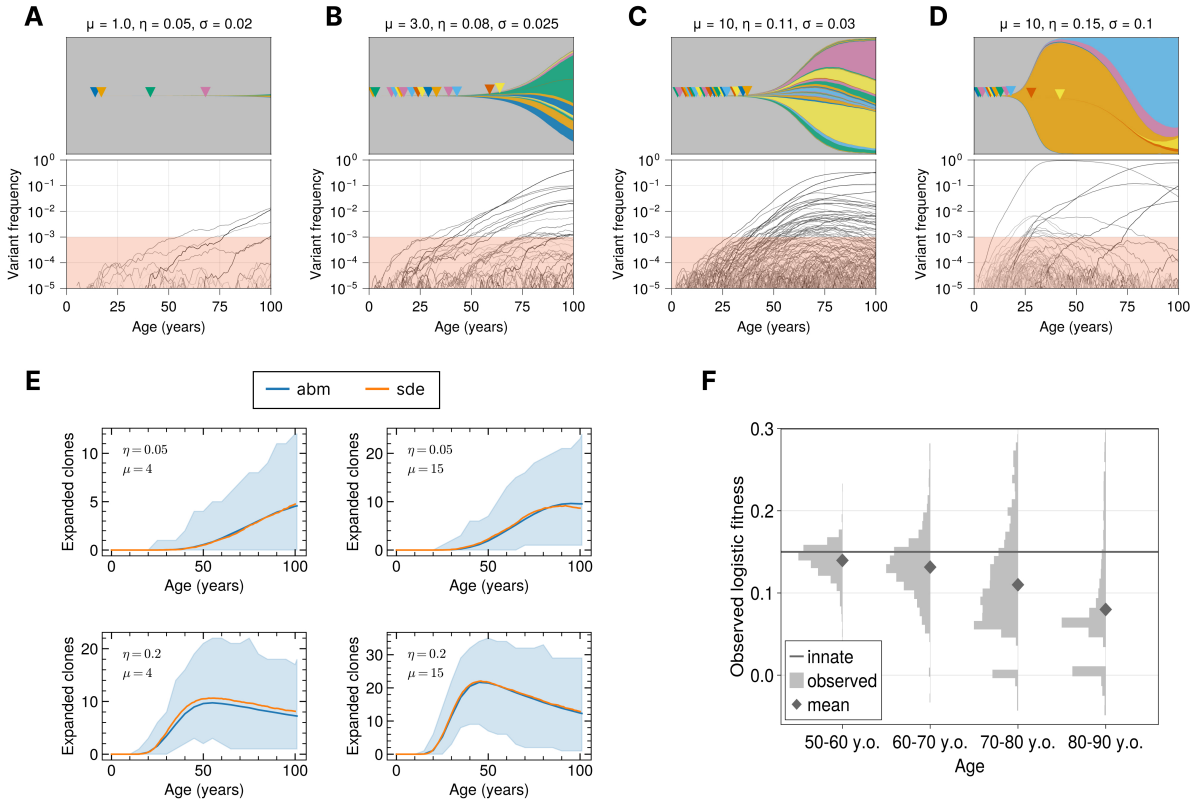


**Figure SI 1. Model dynamics testing and exploration. A-D)** Example realisations of the SDE model (Equation (S19)) for different parameter combinations. In the Muller plots (top row) the mutant arrival times are indicated by triangles. **E)** The number of detectably expanded (>0.5%) clones with age averaged over 500 simulations agrees between the SDE (SI 2.1) and ABM (SI 4.2) implementations of the model. The blue shaded area shows the minimum and maximal values of the expanded clones for the ABM model. **F)** In 10'000 simulations a variant with innate fitness $s = 0.15$ is introduced at age 10 and observed at different time points later in life. The distribution of measured logistic fitnesses becomes wider and has its mean decrease with age.

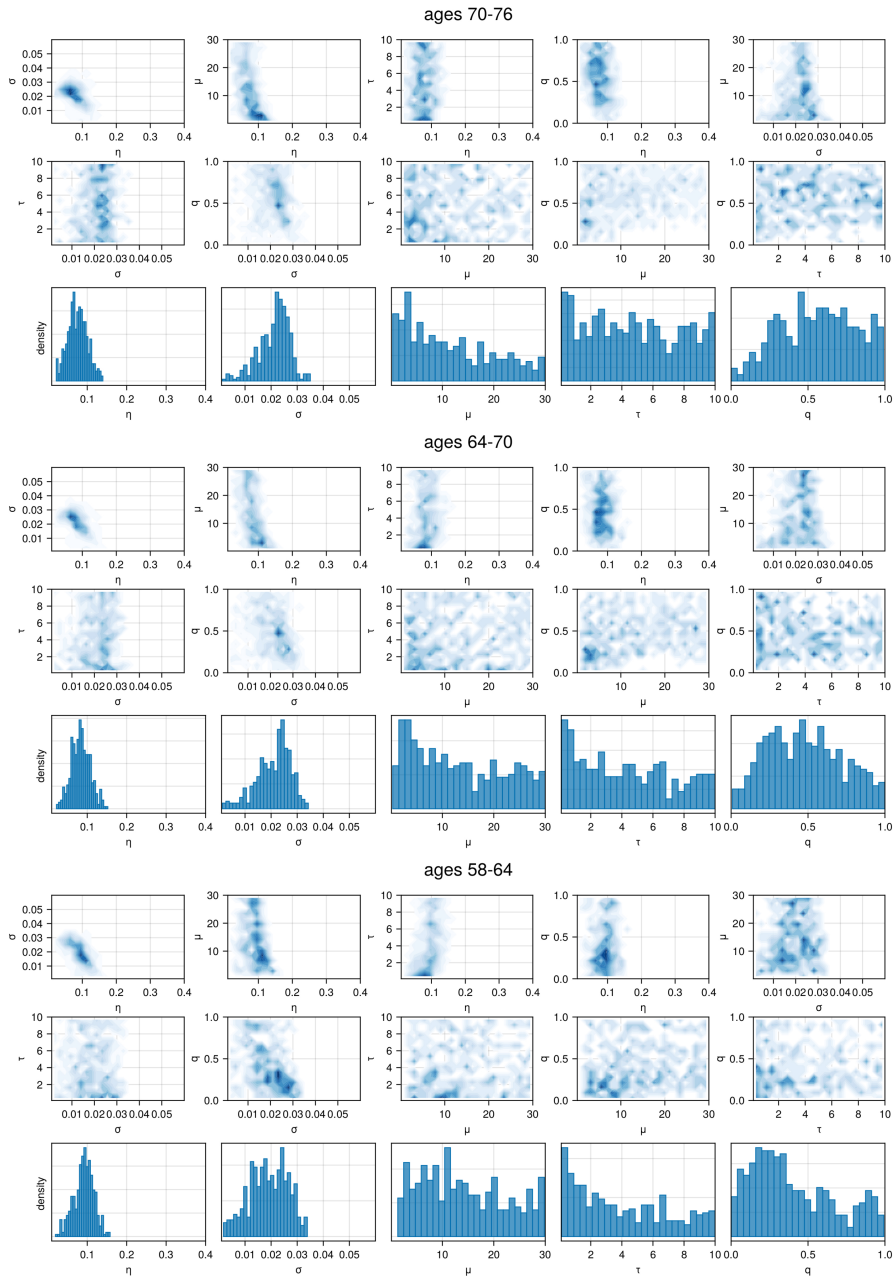


**Figure SI 2. Parameter posterior distributions obtained via approximate Bayesian computation on time-resolved clone trajectories from Fabre et al. donor data.** For each age group 40′000 virtual patients were simulated with parameters randomly drawn from uniform priors, with only the 1% most similar retained in the posterior. Top rows: 2D projections of the full 5-parameter posterior distribution for parameter pairs. Bottom row: marginal distributions for each parameter.

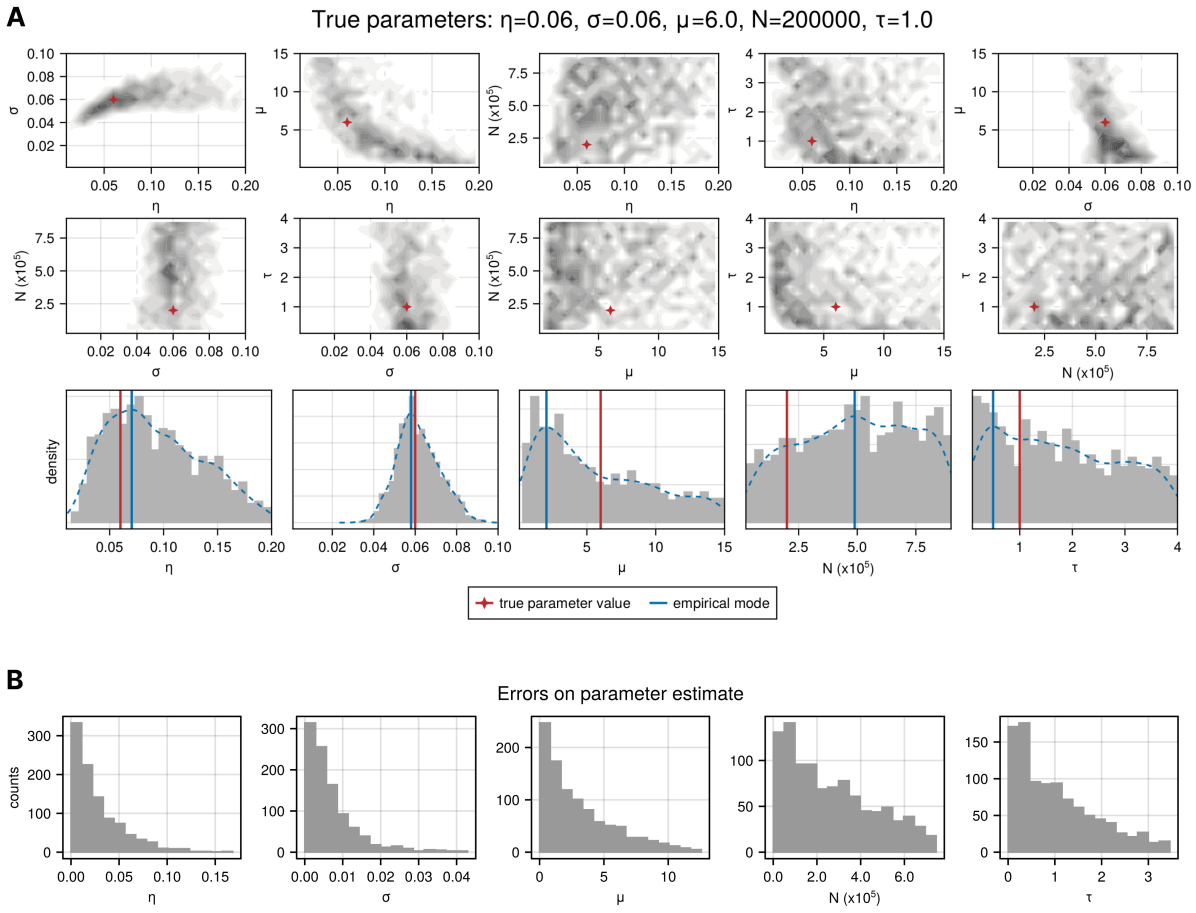


**Figure SI 3. Validation of approximate Bayesian computation on variant trajectories from simulated datasets. A)** Posterior distributions of evolutionary parameters for a single simulated dataset with known parameter values. Top rows: 2D projections of the full 5-parameter posterior distribution for parameter pairs. Bottom row: marginal distributions for each parameter, with estimated value (mode of the empirical distribution) and true value indicated as red and blue vertical lines respectively. **B)** Distribution of errors on estimated parameters from ABC for 1000 simulated datasets.

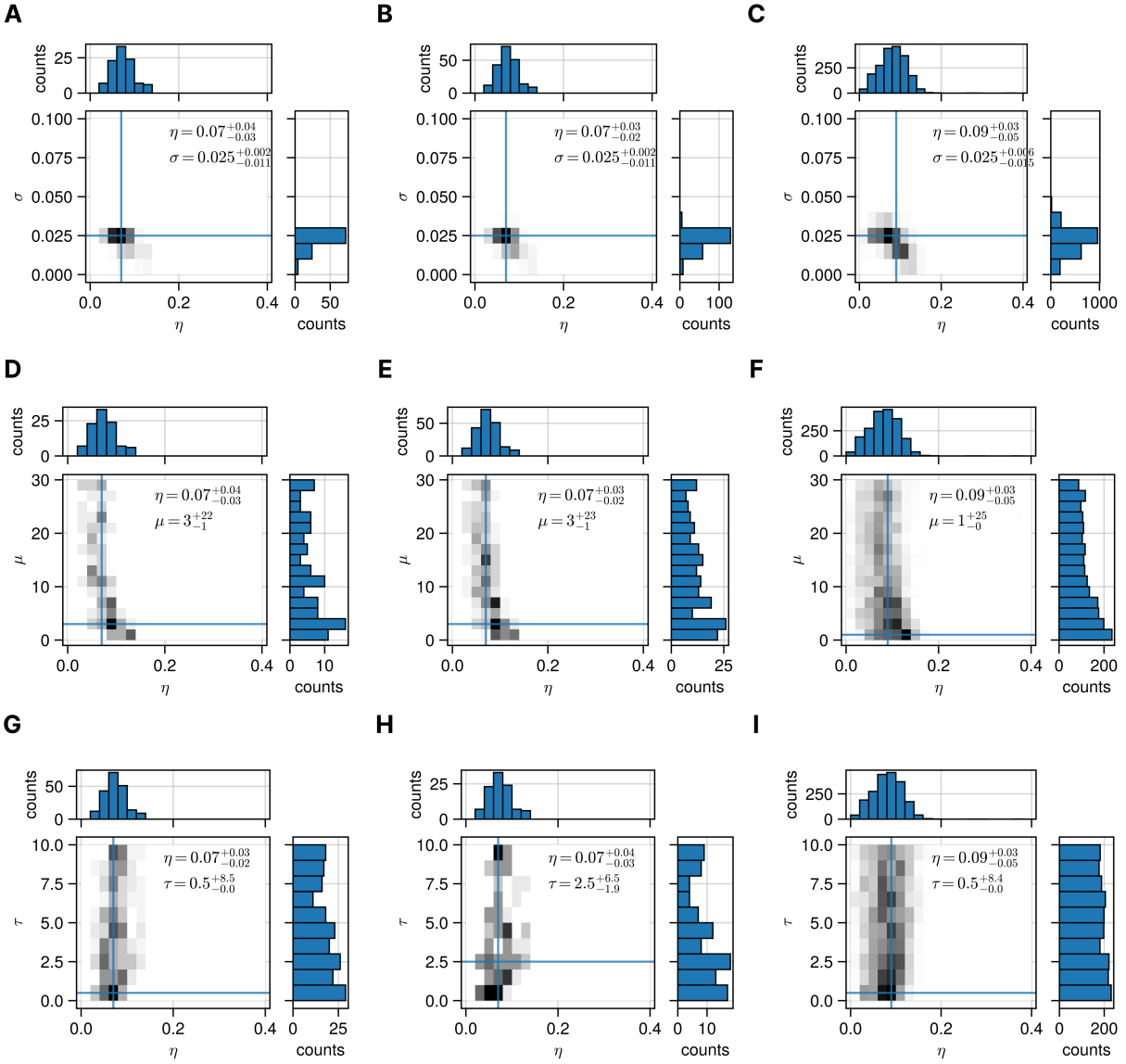


**Figure SI 4. Sensitivity analysis for the posterior distributions inferred from the time-resolved clone trajectories from Fabre et al. donor data.** The posterior distributions and their modes show consistent parameter values inferred for ($\eta$, $\sigma$, $\mu$) from different percentages of the total 40′000 realisations of the HSC dynamics. This suggests that the point values inferred are not influenced by the number of samples retained. **A-C)** The 2D posterior distributions of the pair $\eta$-$\sigma$ considering 0.25%, 0.5% and 5% of the runs along with the marginal posterior distributions. **D-F)** The 2D posterior distributions of the pair $\eta$-$\mu$ considering 0.25%, 0.5% and 5% of the runs along with the marginal posterior distributions. **G-I)** The 2D posterior distributions of the pair $\eta$-$\tau$ considering 0.25%, 0.5% and 5% of the runs along with the marginal posterior distributions.

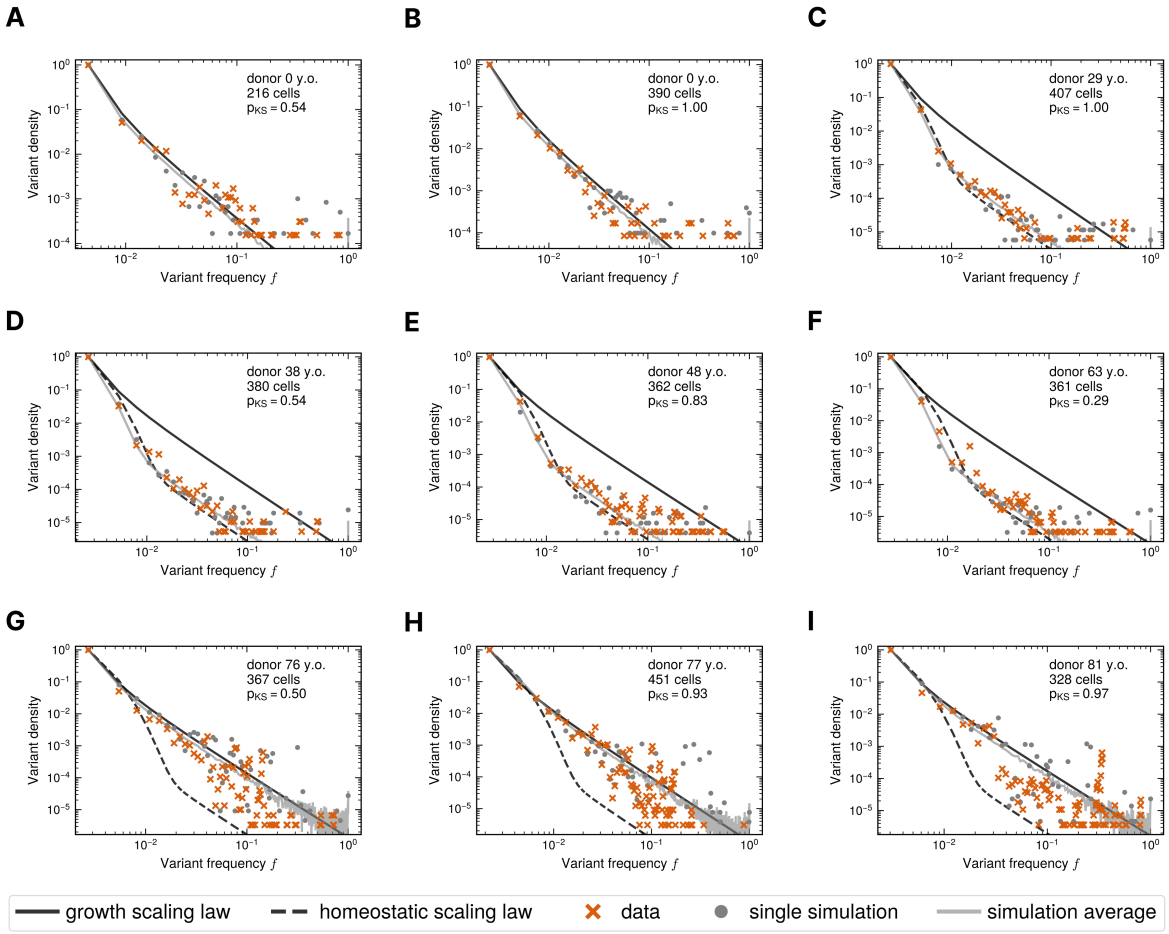


**Figure SI 5. Transitions of HSC site frequency spectra (SFS) with age in the donor cohort.** The SFS computed from a number of cells ranging from 216 to 451 are shown with orange crosses, whereas simulations are shown in grey (the line is the average over 500 and the dots is the SFS of one realisation). The expected value of the SFS for a growing population (solid black line) and fixed-size population (dashed black lines) are also plotted.

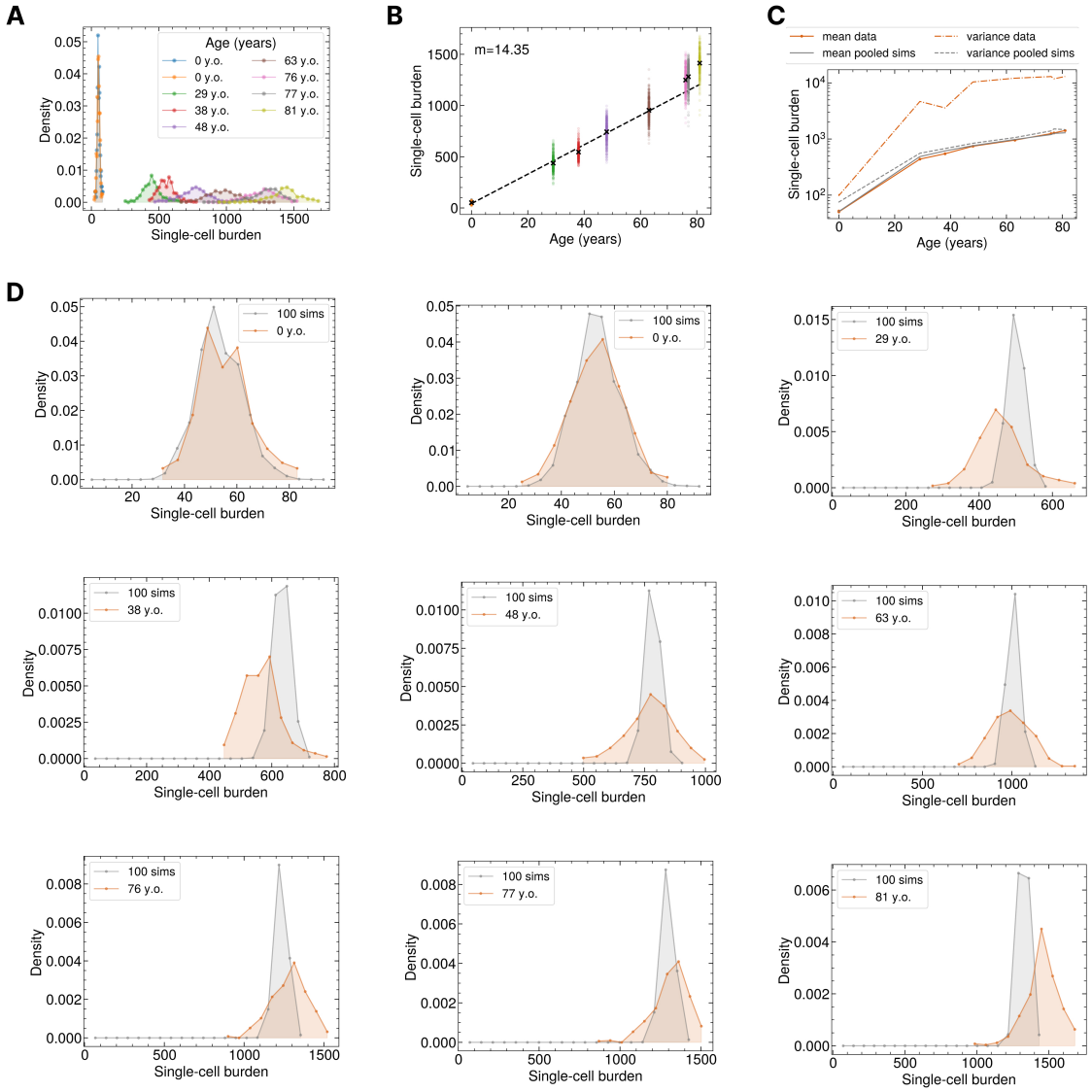


**Figure SI 6. Agent-based model parametrisation with Mitchell's data. A)** Distribution of the number of mutations (indels and single-nucleotides) in individual cells across the donor cohort. **B)** Estimation of the average burden increase per year considering only donors without any clone at frequency greater than 1%. **C)** The mean (solid lines) and the variance (dashed lines) of the single-cell mutational burden in the data (orange) and those computed from 100 pooled simulations (grey). **D)** The single-cell mutational burden distributions in the data (orange) and those computed from 100 pooled simulations (grey).

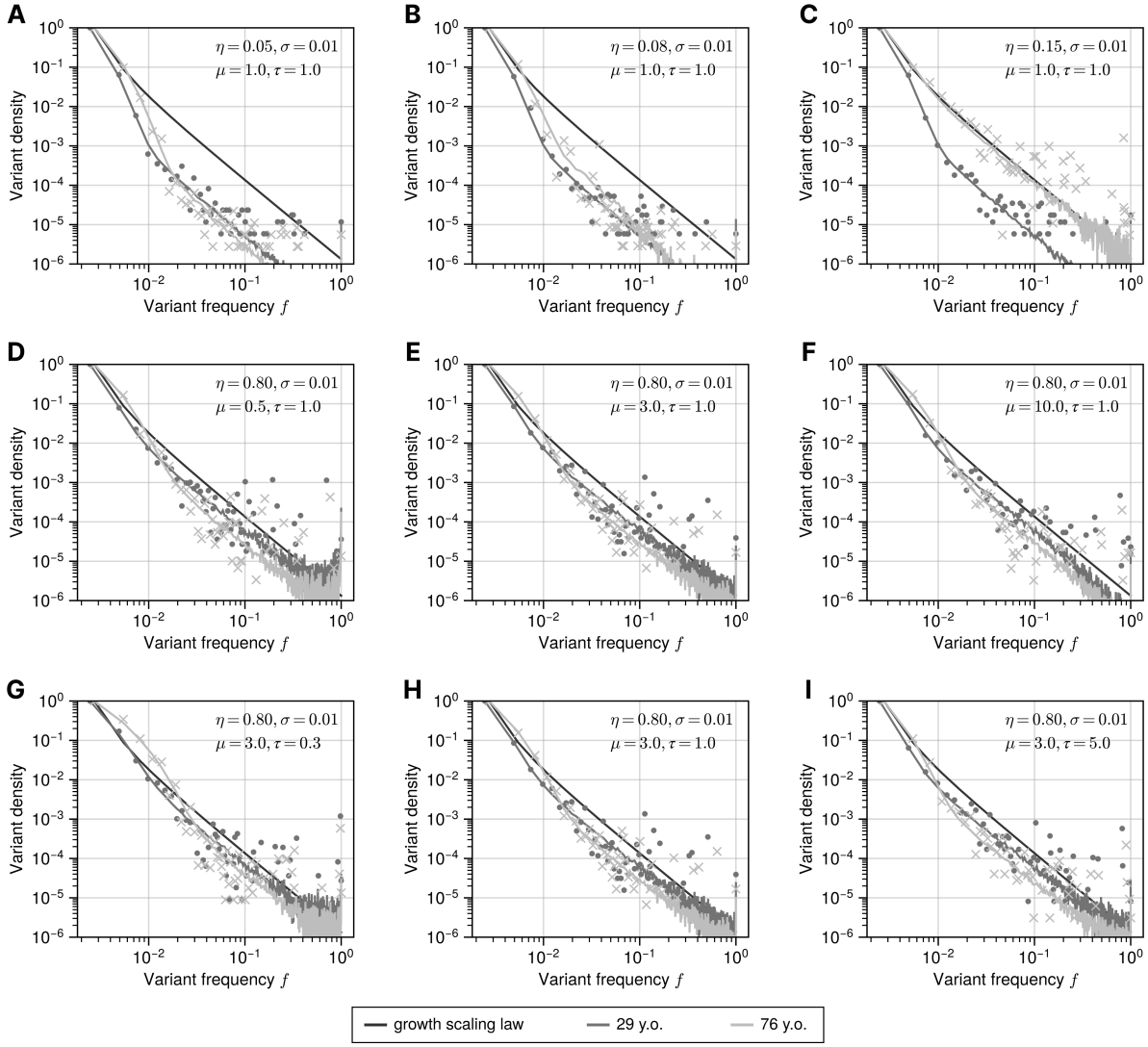


**Figure SI 7. The dynamics of the simulated site frequency spectrum (SFS)**. Simulated SFS constructed by taking a sample of size 407 cells and 367 cells from population of $10^5$ HSCs at age 29 and 76 respectively. Grey lines show the SFS averaged over 500 simulations, dots and crosses show individual simulations, and the solid black line shows the growth scaling law considering 367 cells. **A-C)** The dynamics of the SFS varying the value of the mean of the Gamma distribution of fitness effects $\eta$. **D-F)** The dynamics of the SFS varying the value of the arrival rate of fit clones per year $\mu$. **G-I)** The dynamics of the SFS varying the value of the inter-division time for of wild-type cells per year $\tau$.

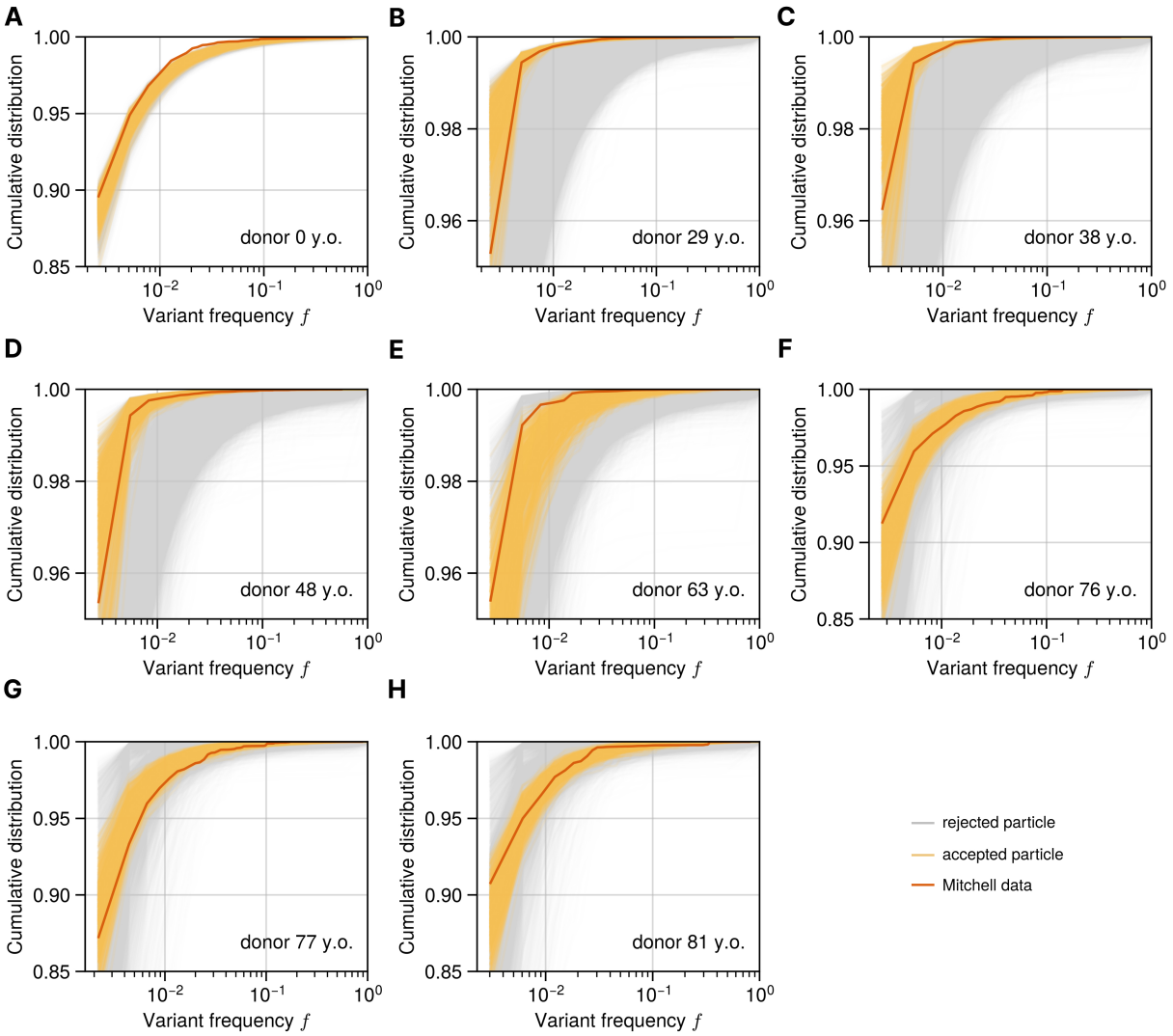


**Figure SI 8. Summary statistics used to infer to parameters from the data via approximate Bayesian computation. A-H)** The cumulative distribution (cdf) computed from the site frequency spectra of the data (dots) is compared against the simulations. The simulations that are close to the data according to the average distance in the cdf for 7 out of 8 donors are kept (yellow) as samples of the posterior distributions. The remaining simulations (grey) are discarded. **I)** The number of detectable expanded clones (frequency $\geq 1\%$) of the data (dark orange), the accepted simulations (yellow) and the rejected simulations (grey).

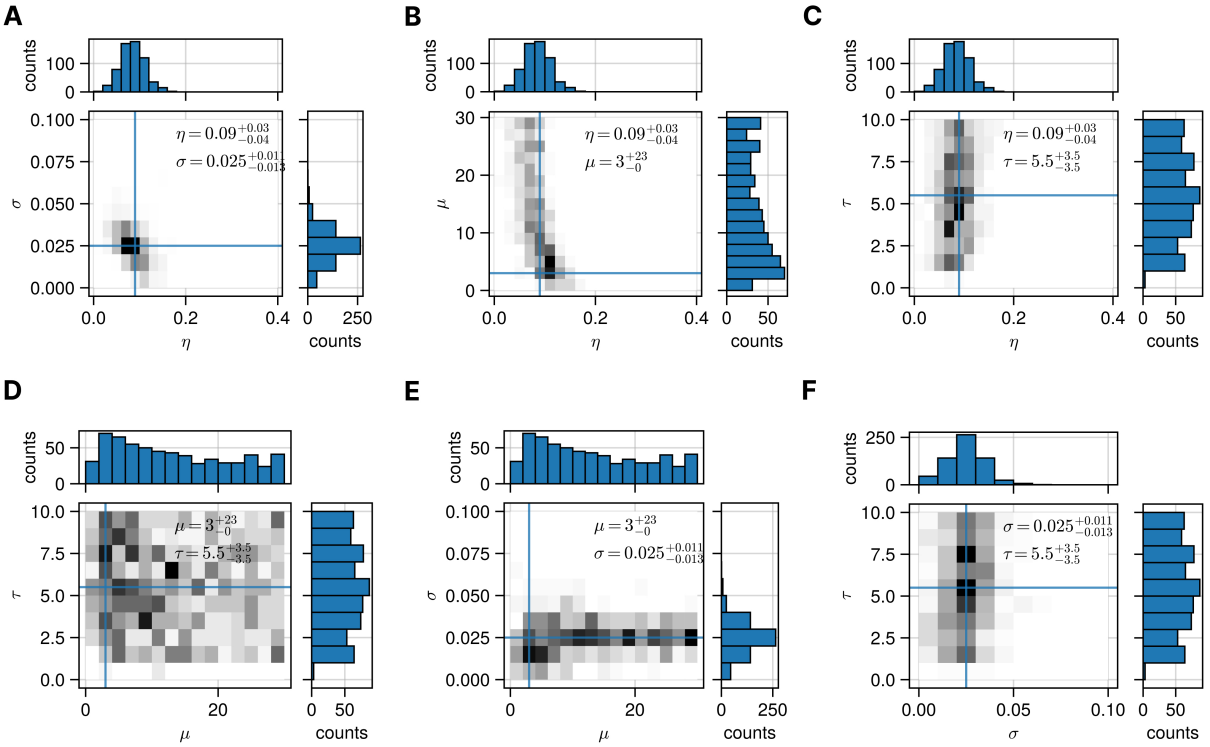


**Figure SI 9. Parameter posterior distributions obtained via approximate Bayesian computation on the SFS and the number of detectable expanded clones from Mitchell et al. donor data.** 65′000 virtual patients were simulated with parameters randomly drawn from uniform priors, with only the 1% most similar retained in the posterior. For pairs of parameters 2D projections of the posterior are shown alongside individual marginal distributions and their modes (point estimates).

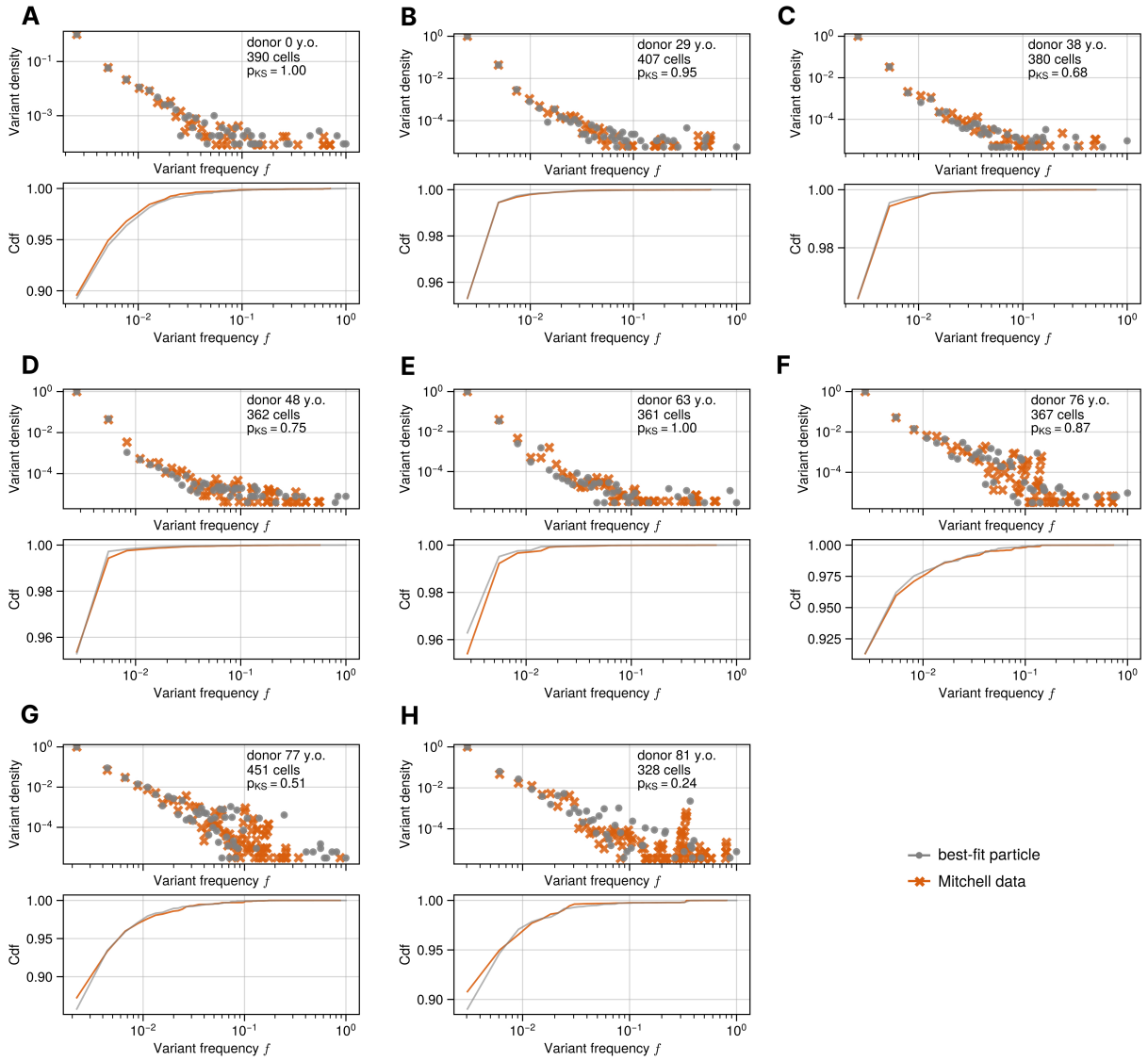


**Figure SI 10. The SFS summary statistic of the simulation that fits the data best according to the inferential framework.** Donors (orange) are shown in ascending age along with the simulation that fits the data best (grey) according to both the SFS and the number of clones. The SFS (top) and its cumulative distribution (bottom) are shown in every panel.

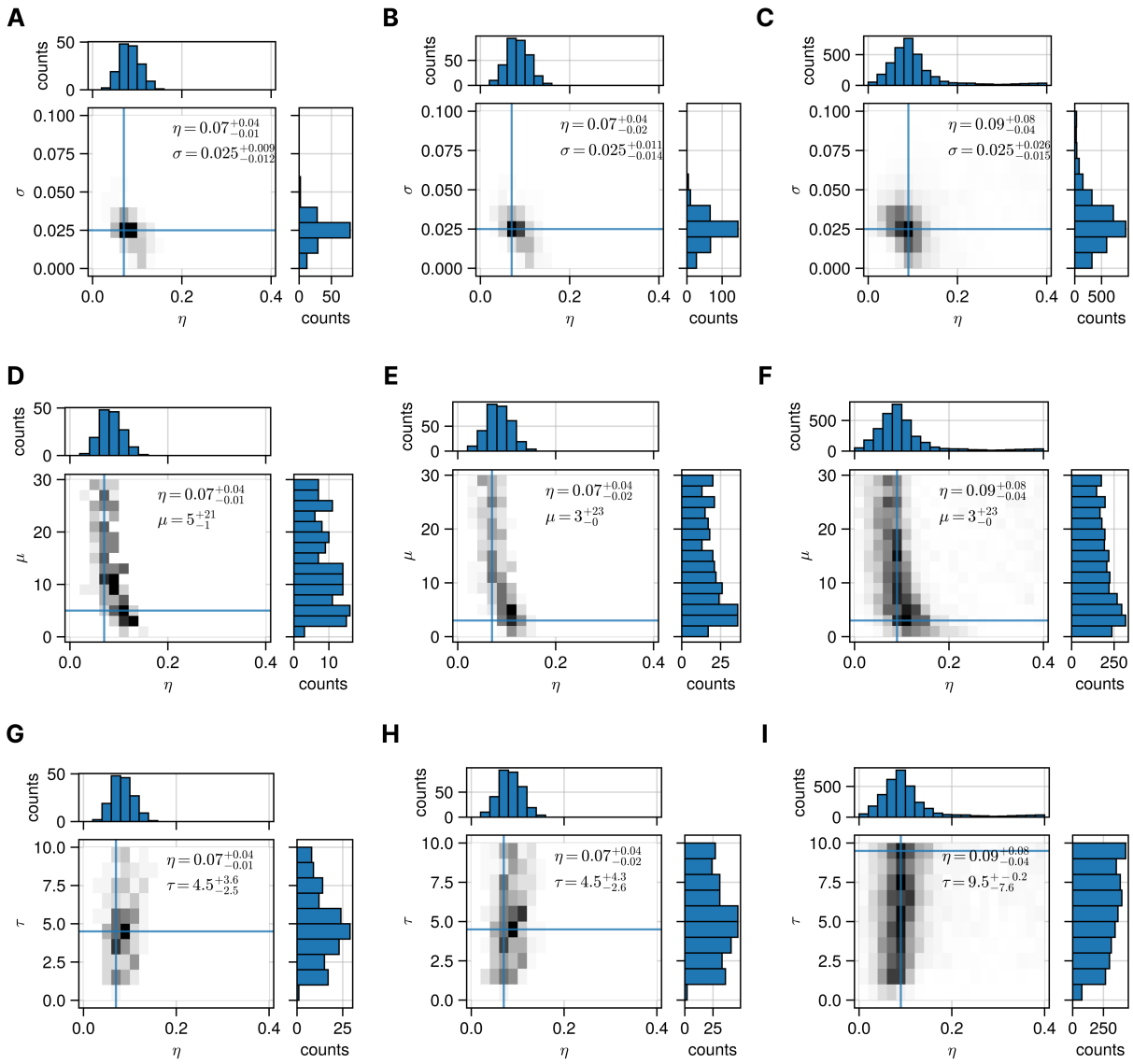


**Figure SI 11. Sensitivity analysis for the posterior distributions inferred from the SFS and the number of detectable expanded clones from Mitchell et al. donor data.** The posterior distributions and their modes show consistent parameter values inferred for ($\eta$, $\sigma$, $\mu$) from different percentages of the total 65′000 realisations of the HSC dynamics. This suggests that the point values inferred are not influenced by the number of samples retained. **A-C)** The 2D posterior distributions of the pair $\eta$-$\sigma$ considering 0.25%, 0.5% and 5% of the runs along with the marginal posterior distributions. **D-F)** The 2D posterior distributions of the pair $\eta$-$\mu$ considering 0.25%, 0.5% and 5% of the runs along with the marginal posterior distributions. **G-I)** The 2D posterior distributions of the pair $\eta$-$\tau$ considering 0.25%, 0.5% and 5% of the runs along with the marginal posterior distributions.

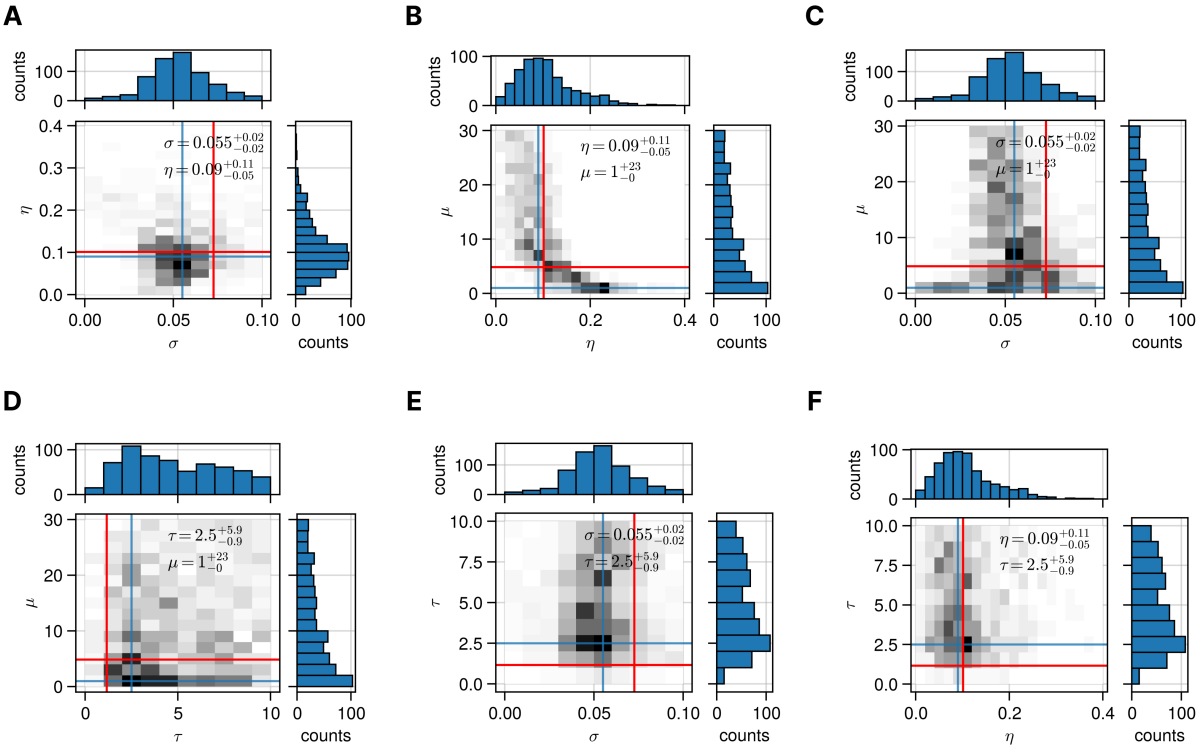


**Figure SI 12. Validation of approximate Bayesian computation on SFS and number of clones on simulated data.** A set of parameter values (red) have been used to generate a single realisation of the HSC dynamics, which is considered as ground truth and used to test the accuracy of the inferential framework. The posterior distributions computed from 1% of the 65'000 realisations. The inferred point estimates (modes) are shown (blue). The target values are: $\eta = 0.10, \sigma = 0.723, \mu = 4.86, \tau = 1.16$.