Detecting and quantifying clonal selection in somatic stem cells

Verena Körber[1], Naser Ansari-Pour[2], Niels Asger Jakobsen[2], Rachel Moore[2], Nina Claudino[1], Marlen Metzner[2], Franziska Hörsch[1], Batchimeg Usukhbayar[2], Mirian Angulo Salazar[2], Simon Newman[3,4], Benjamin JL Kendrick[3,4], Adrian H Taylor[3,4], Rasheed Afinowi-Luitz[4], Roger Gundle[3,4], Bridget Watkins[3], Kim Wheway[3], Debra Beazley[3], Stephanie G Dakin[3], Andrew J Carr[3], Paresh Vyas[2*†] Thomas Höfer[1*†]

Affiliations:[1]Division of Theoretical Systems Biology, German Cancer Research Center (DKFZ); Heidelberg, Germany. [2]MRC Molecular Haematology Unit, Oxford Biomedical Research Center, Hematology Theme, Oxford Centre for Haematology, Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford; Oxford UK. [3]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford; Oxford UK. [4]Nuffield Orthopaedic Centre, Oxford University Hospitals NHS Foundation Trust; Oxford UK.

[†]These authors contributed equally.

*Correspondence: paresh.vyas@imm.ox.ac.uk (PV), t.hoefer@dkfz.de (TH)

Abstract

As DNA variants accumulate in somatic stem cells, become selected or evolve neutrally, they may ultimately alter tissue function. When, and how, selection occurs in homeostatic tissues is incompletely understood. Here, we introduce SCIFER, a scalable method that identifies selection in an individual tissue, without requiring knowledge of the underlying driver event. Moreover, SCIFER infers the self-renewal and mutation dynamics of the tissue's stem cells, and, if selection is present, the size and growth rate of the largest selected clone. We benchmark SCIFER with published data and then probe bone marrow of 22 non-leukemic individuals for clonal hematopoiesis (CH), identifying CH with known and unknown driver events. Unexpectedly, we find accelerated division of all stem cells in CH, compared to age-matched non-CH individuals, suggesting that the bone marrow environment alters stem cell dynamics in individuals with CH. SCIFER is broadly applicable to renewing somatic tissues to detect and quantify selection.

Main

In phenotypically normal stem cells, somatic variation accumulates throughout development [1] and post-natal life [2-6] through diverse genetic mechanisms[7,8]. Somatic variants may alter stem cell function, with the potential to ultimately change tissue function. Over time, stem cell clones vary in size due to genetic drift[9,10], or selection[4-6]. Understanding how somatic mosaicism is shaped by selection and genetic drift, and quantifying the underlying proliferation, differentiation and mutation dynamics of tissue stem cells, are fundamental questions.

Hematopoiesis is a paradigm for somatic mosaicism. In elderly individuals, clonal expansions occur in the hematopoietic stem cell (HSC) population[11-16]. This clonal hematopoiesis (CH) is associated with a heightened risk of developing blood cancer[12,13,15,17], excess cardiovascular mortality[18] and chronic infection[19,20]. Interestingly, a recent study suggests that atherosclerosis can drive CH by accelerating HSC proliferation and hence mutation acquisition[21]. Typically, CH is detected by probing blood samples for recurrent mutations in genes associated with leukemia, such as *DNMT3A*, *TET2* and *ASXL1*. More recently, blood cell clones, with known driver mutations, have been characterized in longitudinal studies[22,23]. When in life CH clones are born, and under which conditions driver mutations are generated and selected, remain key open questions.

Mathematical analyses of driver mutation patterns in large cohorts have shown that clonal expansion in CH occurs through selection[24] and predict that clonal selection of HSCs is far more prevalent in humans than suggested by the observed number of CH cases with known drivers[25]. Hence, a large fraction of CH cases and their drivers may remain unidentified. Indeed, recent studies of human hematopoiesis found frequently expanded clones in elderly individuals for which the driver event was unknown[22,26]. Here, clonal selection was detected by sequencing 3050 whole genomes of single hematopoietic stem and progenitor cells (HSPCs) expanded ex vivo, from 11 donors; phylogenetic trees were reconstructed using somatic variants in these cells[2].

These seminal observations raise the following questions: (a) Do CH drivers emerge by chance during normal HSC turnover, or does increased turnover or/and somatic mutation rate facilitate acquisition of driver events that promote selection? (b) Do stem cell clones with known CH driver genes, and clones with unknown drivers, differ in their frequency in the human population and in their clinical phenotypes? (c) When do selection events, with known or unknown drivers, occur in life, and (d) how fast do the selected clones grow with known and unknown driver events? These questions are more easily addressed with a more cost effective, scalable method, than phylogenetic tree reconstruction, to identify clonal selection and quantitate stem cell dynamics. Ideally, such a method would also work in tissues where stem cells cannot be cultured ex vivo. A recent approach used WGS of single blood samples

to search for CH, exploiting the fact that selection of a mutated HSC will often increase the total count of somatic variants detectable by WGS[27]. This method uses prior knowledge of driver mutations and cannot reconstruct the underlying evolutionary dynamics.

Here, we show that neutral evolution and clonal selection leave characteristic imprints in the statistics of somatic variants detected by bulk WGS at a single snapshot, irrespective of whether the driver of selection is known or not. Using this observation, we develop population-genetics theory for the accumulation and spread of somatic variants in development and subsequent homeostatic tissue renewal, either in the absence, or presence, of selection. Based on this theory, we devise a Bayesian inference framework that discriminates clonal expansion due to selection from neutral evolution, and quantifies the underlying stem cell parameters in a homeostatic tissue. We validate our method with synthetic data and published experimental data, and then apply it to a new cohort of bone marrow samples of non-leukemic individuals.

Results

Quantifying clonal selection in somatic mosaics

The vast majority of genetic variants acquired in somatic stem cells are neutral[28]. We reasoned that the quantitative distribution of these variants measured in stem cells of an individual provides information on the underlying dynamics of genetic drift and selection. To derive the expected distribution and subsequently compare it to experimental data, we considered developmental expansion of a renewing tissue, such as the blood, the skin and the gut epithelium, followed by a phase of homeostatic turnover. During development, stem cell divisions feed tissue expansion, whilst in adulthood stem cell self-renewal usually balances loss, by differentiation and cell death (Fig. 1a and Methods for a mathematical description of the model). We focussed specifically on the abundant somatic single-nucleotide variants (SSNVs) acquired during both development and homeostasis and inherited to progeny. As the number of developing tissue stem cells is initially small, few SSNVs acquired early in life (Fig 1b; variants A and B); but they will have high variant allele frequencies (VAFs). Later, when the stem cell numbers are larger, more variants are acquired in total but now each variant will be at a lower VAF (Fig 1b; variants C, D and E). Owing to the inherent stochasticity of self-renewing divisions and loss of stem cells, neutral SSNVs will drift with time so that some variants will randomly increase their VAF (Fig. 1b; variants B, C, D and F) while others either decrease, or are extinguished (Fig. 1b; variant E). To quantify the acquisition of SSNVs and the changes in their frequencies, we model the neutral clonal dynamics during development and homeostasis by combining expanding

102 (supercritical) and homeostatic (critical) birth-death processes (see Methods, Population-
103 genetics model). We find an analytical solution for the VAF distribution as a function of time
104 that lends itself to parameter inference from experimental data.

105 How do the two principal phases – development and homeostasis – shape the VAF
106 distribution measured in adulthood? During developmental expansion, genetic drift is
107 negligible irrespective of the total number of stem cells, resulting in an invariant linear relation
108 between cumulative SSNV count and 1/VAF (Fig. 1c; to zoom in on small VAFs, we plot the
109 cumulative SSNV distribution against 1/VAF)[29,30]. During homeostasis, genetic drift might
110 alter the cumulative VAF histogram by increasing the number of SSNVs with low VAF (Fig.
111 1d). The extent of homeostatic drift depends on the number of stem cells and their self-
112 renewal rate. If there are $5x10^3$ stem cells dividing 5 times per year, drift will increase the
113 variant count with high VAF over the human lifespan (Fig. 1e; given that the development of
114 HSCs and major increase in number occurs in the fetus, in the model we assumed transition
115 from development to approximate homeostasis around birth). For $5x10^4$ stem cells and the
116 same division rate, the variant count will increase with age at lower VAFs (right part of the
117 VAF distribution) and remain stable at high VAFs (left part of the VAF distribution) (Fig. 1f).
118 Finally, with $5x10^5$ stem cells, the distribution is nearly stable with age for all VAFs >1% (Fig.
119 1g). Prior work has inferred, for a 59-year old-human, that $4x10^4$ - $2x10^5$ HSCs divide 0.6 - 6
120 times per year[2]. With these values, our model suggests the VAF histogram at high VAF
121 (about 10% or higher) is insensitive to genetic drift over the human lifespan but might be
122 modified by drift at low VAF.

123 Next, we modelled how clonal selection alters the VAF distribution. We consider the leading
124 selected clone, born at time t_s consequent to a driver event with selective advantage s,
125 promotes mutant clone expansion and concomitant normal stem cell displacement (Fig. 1h
126 and i). With clonal selection, neutral variants may have originated: (i) in normal stem cells
127 before acquisition of the driver event; (ii) within the selected clone or (iii) in non-selected
128 stem cells after the driver event occurred. We find analytical expressions for the evolution of
129 all three types of variants, over time, which together yield the cumulative VAF distribution of
130 SSNVs within a tissue, in the presence of a selected clone (Methods). Expansion of a
131 selected clone will increase the VAF of the founding driver mutation (mutation D in Fig. 1i)
132 and of all other neutral SSNVs in the founding cell of the selected clone. As a consequence,
133 the selected clone generates a shoulder in the cumulative VAF distribution (Fig. 1j), as
134 opposed to the gradual increase of the cumulative SSNV number in the case of neutral
135 evolution (Fig. 1d). Importantly, the shape of the distribution discriminates between selection
136 and neutral evolution, whereas the mere variant count detectable in WGS may fail to do so
137 (compare the VAF histograms in Fig. 1j, shaped by selection, with Fig. 1e-f, shaped by drift).

138 Hence, the shoulder in the cumulative VAF distribution provides a robust signature of clonal
139 selection agnostic of the identity of the driver.

140 The height of the shoulder increases with the age at which the selected clone originated, t_s,
141 (Fig. 1k), since the number of neutral variants in the clone's stem cell of origin is the higher,
142 the older the individual as more prior cell divisions have occurred. As the selected clone
143 expands with time, the shoulder reaches higher VAFs (i.e., the shoulder moves to the left on
144 the x-axis the greater the time interval between the birth of the clone and measurement of
145 SNNVs and their VAFS by WGS; Fig. 1l). The VAF reached in the time interval between
146 clone origin and sample acquisition yields the selective advantage *s* (Fig. 1h, lower panel).
147 Thus, height and position of the shoulder allow inference of the clone's age and selective
148 advantage.

149 **Selection detected in noisy WGS data**

150 WGS measures the VAF of a variant with an error that is approximately binomially
151 distributed, with the variance depending on sequencing depth[31]. To understand how this
152 measurement error influences the ability to identify clonal selection in bulk WGS data, we
153 simulated WGS data with our models of neutral evolution (Fig. 1a,b) and selection (Fig. 1h,i).
154 Briefly, we performed repeated stochastic simulations of the two models, generating 10 data
155 sets with neutral evolution and 70 data sets with selection (with a driver mutation acquired at
156 20 years of age causing stochastic clonal expansion with an average growth rate of 22% per
157 year). To account for the WGS measurement error, we drew variant reads from binomial
158 distributions for average sequencing depths of 30x, 90x or 270x (Methods; Supplementary
159 Table 1). We then used approximate Bayesian computation (ABC) to fit both the model of
160 neutral evolution and the selection model to each data set (Extended Data Fig. 1), yielding
161 posterior probabilities for the neutral model and the selection model to reproduce the noisy
162 data.

163 To determine sensitivity and specificity in detecting selection, we used the posteriors to
164 compute receiver-operating curves (ROC) for identifying clones of different sizes. At 90x
165 sequencing depth, we found that selected clones with VAF ≥ 5% were detected reliably (Fig.
166 2a; area under the curve, AUC = 0.88 for VAF = 5% and AUC ~ 1 for VAF = 7.5%), whereas
167 smaller clones were not (Fig. 2a; VAF = 2.5%; AUC = 0.5). Discrimination between selection
168 and neutral evolution was optimal when using a threshold of 15% for the posterior probability
169 of the selection model, such that a posterior >15% identified selection (Fig. 2a, the operating
170 points correspond to the 15% threshold, achieving ~100% sensitivity and 80% specificity for
171 clone size 5% VAF and ~100% sensitivity and specificity for clone size 7.5% VAF).
172 Exemplary selection posteriors for simulated 90x WGS data sets show discrimination
173 between neutral evolution (0% VAF) and selection of clones ≥5% VAF (Fig. 2b). Further, our

174 ROC analysis predicts that 30x WGS will allow detection of large selected clones > 20%
175 VAF, while deeper WGS will enable detection of smaller clones, with VAF > 1% at 270x (Fig.
176 2c). Taken together, our method for **s**elected-**c**lone **i**n**fer**ence, termed SCIFER, is suitable for
177 detecting clonal selection in bulk WGS data.

178 **Benchmarking SCIFER with phylogenetic trees of human hematopoiesis**

179 Next, we benchmarked SCIFER with published WGS data from human HSPC single-cell
180 clones, where the absence or presence of expanded clones is visible in reconstructed
181 phylogenetic trees[2,22,26]. For a 59-year-old human, where hematopoiesis evolved neutrally,
182 key HSC parameters (HSC number, self-renewal rate and the rate of acquiring somatic
183 variants) were inferred based on WGS of 140 HSPC clones and targeted sequencing of
184 peripheral blood cell populations[2]. This provided a quantitative reference for SCIFER. In
185 addition to the experimental error in VAF measurement (cf. Extended Data Fig. 1), a further
186 source of error in WGS is the calling of genome-wide somatic variants. Therefore, we used a
187 stringent calling approach for identifying low-VAF variants in bulk WGS, intersecting the
188 results of Mutect2 and Strelka[32] (Extended data Fig. 2a). To examine the robustness of
189 SCIFER with respect to variant calling, we also performed our analyses on the original
190 variant calls obtained with Caveman[2], which includes a substantial number of additional
191 SSNVs (Extended data Fig. 2b-c). First, the reconstruction of the phylogenetic tree using our
192 calling approach recapitulated the original work, yielding long branches with coalescent
193 events located mainly at the top of the tree (Fig. 3a). Second, we generated pseudo-bulk
194 WGS data from the single-cell variants and used SCIFER to fit the cumulative VAF
195 distribution (Fig. 3b; Extended Data Fig. 2d). Irrespective of the calling algorithm, and in
196 agreement with the original work, SCIFER inferred neutral evolution (Fig. 3c). Third, SCIFER
197 inferred HSC number (Fig. 3d) and rate of self-renewing divisions (Fig. 3e), with close
198 estimates for the two variant callers ($6x10^4$-$2x10^5$ HSCs dividing 3 to 8 times per year using
199 the variants from Caveman; $2x10^4$-$5x10^4$ HSCs dividing 0.4 to 6 times per year using the
200 variants from Mutect2/Strelka; ranges are 80% credible intervals), which both overlapped
201 with the estimate of the original work ($4x10^4$-$2x10^5$ HSCs dividing 0.6 to 6 times per year) [2].
202 Finally, the rate of SSNV acquisition inferred by SCIFER (3-5 SSNVs/cell division) (Fig. 3f)
203 overlapped with the estimate in the original work (3-28 SSNVs per division) when using the
204 variants from Caveman. As expected, this rate was smaller (0.6-1 SSNV per division) with
205 our conservative variant calls using the intersect of Mutect2 and Strelka. Interestingly,
206 SCIFER could infer HSC number, self-renewal rate and mutation rate separately (Extended
207 Data Fig. 2e, f). This feature relies on our use of absolute variant counts for each individual,
208 as many rapidly dividing stem cells accumulate more variants in total than few rarely dividing
209 stem cells (although the ratio of stem cell number and self-renewal rate may remain
210 unchanged[2,24]), and the presence of both expansion and homeostatic phases

211 (Supplementary Note 1 and Supplementary Fig. 1). In conclusion, SCIFER detected neutral
212 evolution and quantified HSC number and division rate from a single pseudo-bulk sample.
213 The inference was robust with respect to the variant calling algorithm, indicating that key
214 aspects of stem cell dynamics can be robustly inferred from the VAF histogram.

215 To further validate SCIFER's ability to identify neutral evolution, we generated pseudo-bulk
216 data from three further neutrally evolving, published human HSPC single-cell WGS datasets
217 (AX001, KX001 and KX002) for which separate estimates of HSC number and division rate
218 were not available[26] (Extended Data Fig. 2g,h). In all cases, SCIFER correctly detected
219 neutral evolution (Fig. 3c) and inferred HSC parameters broadly consistent with the estimate
220 for the data set from Lee-Six et al [2] (Fig. 3d,f). Interestingly, for the oldest individual (AX001,
221 63 years of age, compared to KX001 and KX002, 29 and 38 years of age, respectively), the
222 inferred HSC division rate was higher than for the other individuals (Fig. 3e). These data
223 further support SCIFER's ability to detect and quantify neutral evolution.

224 Next, we asked whether SCIFER detects clonal selection. WGS data from 448 single HSPCs
225 from two published samples KX0004[26] and id2259[22] was used to construct phylogenetic
226 trees. In both samples, CH driver mutations had been identified. KX004 harbored a major
227 clone with a *DNMT3A* mutation (~8% VAF) and several smaller clones (Fig. 3g); id2259 had
228 an expanded clone with a mutation in the splicing factor *SF3B1* (with a size of nearly 50%
229 VAF) containing a small *CBL* subclone (Extended Data Fig. 2g). In pseudo-bulk WGS data
230 generated from single-cell SSNVs for both samples, SCIFER detected the respective leading
231 *DNMT3A* (VAF 6%) and *SF3B1* (VAF 49%) clones (Fig. 3h; Extended Data Fig. 2i). SCIFER
232 also inferred that these clones had expanded through selection (Fig. 3i). Using coalescent
233 theory, the original studies estimated that the *SF3B1* mutation arose 25-30 years, and the
234 *DNMT3A* mutation 45-50 years, prior to tissue sampling. SCIFER determined very similar
235 clone ages (Fig. 3j). The estimated clonal growth rates were 68% per year for the *SF3B1*
236 mutation and 18% for the *DNMT3A* mutation (Fig. 3k), in line with previous results
237 suggesting stronger selection of *SF3B1* mutant clones[23]. The inferred stem cell number,
238 division rate and rate of acquired SSNV/cell division (Fig. 3l) were in the same range as
239 estimates for individuals without CH (Fig. 3d-f). Finally, we applied SCIFER to a previously
240 published sample, KX003[26], where phylogenetic reconstruction showed a large expanded
241 clone (~17% VAF) that harbored no known CH driver and that contained smaller clones with
242 known drivers (Fig. 3m). Applying SCIFER to WGS pseudo-bulk data indeed detected a
243 leading clone of 18% VAF (Fig. 3n) and inferred selection (Fig. 3o). The quantitative
244 parameters – HSC number, division rate, rate of acquired SSNV/cell division, clonal growth
245 rate and age of the clone – were in the same ranges as in the two samples above with
246 known CH driver mutations (Fig. 3p, compare with Fig. 3j-l).

247 In summary, SCIFER robustly identified and quantified neutral evolution and selection in
248 pseudo-bulk WGS data of human HSPCs, without using prior knowledge of the presence of a
249 driver.

250 **SCIFER uncovers complexity of somatic mosaicism in bulk samples**

251 Having established the performance of SCIFER, we applied it to genuine bulk WGS data
252 from 22 humans, 30-89 years of age (Supplementary Table 2). We used human bone
253 marrow (BM) samples collected from individuals at total hip replacement surgery. The
254 samples were screened for CH by targeted sequencing of DNA from bone marrow
255 mononuclear cells using a gene panel covering 97 known driver genes sequenced at an
256 average of 800x to detected mutations with a VAF of ≥1% VAF(Supplementary Table 3)[33]. All
257 individuals had normal blood counts and blood films, without an antecedent history of blood
258 or inflammatory or autoimmune disorder[33].

259 We selected 10 individuals without known drivers and 12 individuals with at least one
260 mutation in either *DNMT3A* (6 individuals), *TET2* (4 individuals) or *ASXL1* (2 individuals). 90-
261 120x WGS was performed on FACS-sorted $Lin^{-}CD34^{+}$ HSPCs (Extended Data Fig. 3a-d). A
262 comparison with WGS results for mature BM mononuclear cells is given in Supplementary
263 Note 2 and Supplementary Fig. 2. 30x WGS from hair follicle DNA was used as a germline
264 control. We called SSNVs, small indels, copy number variants (CNVs), and structural
265 variants (SVs) (see Methods). The WGS data confirmed either the presence, or absence, of
266 known CH driver mutations in each sample that was originally noted in panel sequencing.

267 Applying SCIFER to the WGS data of the 10 cases without known CH drivers, seven cases
268 were classified (labelled N1-N7) as neutrally evolving. However, in the remaining 3 cases
269 (U1-U3) SCIFER detected clonal selection, with unknown driver events. For the 12
270 individuals with known CH drivers, SCIFER detected a leading clone in 10 of them, that could
271 be associated with the CH driver mutation found by panel sequencing (A1-A2, with *ASXL1*
272 mutation; D1-D5, with *DNMT3A* mutation; T1-T3, with *TET2* mutation). In the remaining two
273 samples (U4 and U5), SCIFER detected a larger additional clone, without a known CH driver,
274 whilst the driver identified by panel sequencing was associated with a smaller and different
275 clone. Hence, SCIFER uncovered complexity of somatic mosaicism beyond the insights
276 offered by panel sequencing of CH drivers.

277 **Quantifying neutral evolution**

278 We now report the SCIFER results in detail. We focus first on the seven cases (N1-N7)
279 without known CH driver mutations where neutral evolution was inferred by SCIFER (Fig. 4a-
280 b; Extended Data Fig. 5; 2 males and 5 females, ranging from 30 to 76 years of age). The
281 somatic single-base-pair substitution profiles agreed with neutral mutation accumulation

282 during normal ageing (Extended Data Fig. 4a)[2,26,34]. Moreover, we found neither copy
283 number alterations (CNAs), nor structural variants (SVs) in driver genes in these samples
284 (Extended Data Fig. 4b, Supplementary Tables 4 and 5), consistent with neutral evolution.

285 In samples N1-N3 and N6-N7 there were around 20-30 SSNVs with VAF ≥5% whilst N4 and
286 N5 stood out with >60 SSNVs (Fig. 4c). To understand this, we computed the ratio of HSC
287 number to HSC division rate, N/λ (Fig. 4d). This parameter characterizes the timescale over
288 which neutral evolution will cause a variant to reach fixation. For all our neutral samples,
289 except N4 and N5, SCIFER inferred N/λ to be ~10^5 to 10^6 years, supporting the notion that
290 clones do not drift to large size in normal human hematopoiesis[24]. However, individuals N4
291 and N5 had smaller N/λ, of ~10^4 years, consistent with genetic drift increasing SSNV counts
292 at low VAF (cf. Fig. 1e-f). To probe the differences in HSC behavior between N4/N5 and the
293 other individuals in more detail, we inferred stem cell number ($5x10^4$-$1.5x10^6$ HSCs), HSC
294 division rate (1-2 times a year for N1-N3 and N6-N7) and rate of acquired SSNV/cell division
295 (1-3) (Fig. 4e). Interestingly, these estimates were broadly consistent with our inferences
296 from the pseudo-bulk data from published samples (Fig. 3d-f), although the credible intervals
297 were substantially larger. However, N4 and N5, were above 70 years of age (Fig. 4b),
298 showed larger HSC division rates than the other individuals (Fig 4e), about 5 self-renewing
299 divisions per year, suggesting that the drift effects seen at low VAF could be caused by more
300 rapid HSC proliferation.

301 In summary, the analysis of the individuals with neutral HSC dynamics supports the notion
302 that large hematopoietic clones do not arise by drift. However, in two of the seven cases we
303 inferred acceleration of HSC division, by unknown causes. This likely accounts for increased
304 cumulative number of SSNVs at lower VAFs, which we interpret as a signature of clonal drift.

305 **Clonal selection frequently linked with accelerated stem cell self-renewal**

306 Next, we quantified selection for the 10 individuals that had at least one mutation in either
307 *DNMT3A*, or *TET2,* or *ASXL1*, genes (2 males, 8 females, ranging between 62 and 89 years
308 of age; Supplementary Table 2) and where SCIFER identified clonal selection (Fig. 5a-b;
309 Extended Data Fig. 7a) consistent with the known CH driver mutation. The clone size
310 inferred by SCIFER, without knowledge of the driver, agreed with the VAF of the driver (Fig.
311 5c). We identified no CNVs or SVs associated with CH (Extended Data Fig. 6a;
312 Supplementary Tables 4 and 5). The single base-pair variant profiles were consistent with
313 those observed in patients without CH driver mutations, indicating that the same mutational
314 processes were active (Extended Data Fig. 6b).

315 We asked whether HSC parameters differed between cases with distinct driver genes. There
316 was no association between the CH driver gene and the total number of HSCs (Fig. 5d, left

317 panel), the HSC division rate (Fig. 5d, middle panel) and the rate of acquisition of new SSNV
318 per cell division (Fig. 5d, right panel). Two samples with a *DNMT3A* mutation (D2 and D4)
319 stood out with a smaller HSC number (~$5x10^3$) and one of them, D4, also had a lower N/λ
320 ratio (Fig. 5e). Otherwise, the HSC number estimate was rather uniform, whereas the HSC
321 division rate showed considerable variation.

322 Next, we asked whether there were systematic differences compared to HSC parameters in
323 individuals with neutral evolution (N1-N7; Fig.4d,e). The rate of acquisition of new SSNV per
324 cell division and HSC numbers were overall not distinguishable between individuals who
325 showed neutral evolution versus those with CH (Fig. 5f,g). By contrast, the HSC division
326 rates were more variable in individuals with CH. The majority of samples with selected clones
327 (7 out of 10) had a higher HSC division rate than the neutral-evolution group, while three CH
328 cases had HSC division rates comparable to cases of neutral evolution (Fig. 5h, left panel).
329 When plotting HSC division rate as a function of age, we noticed that larger division rates
330 were concentrated in the older individuals (≥65 years of age). The two individuals exhibiting
331 neutral evolution with a higher HSC division rate also fell into this older age group (N4 and
332 N5, cf. Fig. 3e). However, the largest division rates (≥10 divisions per year) were exclusively
333 found in cases where selection operates. Hence, our data suggest that variability in stem cell
334 division rate increases with age, and that clonal selection events are frequently linked with
335 accelerated HSC division.

336 **Clonal selection without known drivers**

337 Finally, we analysed the five cases (U1-U5) where SCIFER identified major selected clones
338 that did not have a corresponding driver mutation at the same VAF in the panel of 97 known
339 CH-associated genes. In all five samples, copy number profiles (Extended Data Fig. 8a) and
340 single base pair substitution profiles (Extended Data Fig. 8b) were as for all the other
341 samples. For U1, U3, U4 and U5, SCIFER found clear evidence of selection (Fig. 6a-b). By
342 contrast, the model support for selection for sample U2 was just above the threshold for
343 selection; it should be kept in mind that the false-positive rate is ~20% for small clones at the
344 detection threshold of 5% VAF (Fig. 2a).

345 Samples for U1-U3 did not have CH mutations on panel sequencing. Samples for U4 and U5
346 had recurrent mutations in known CH driver genes (*TET2* for U4 and *DNMT3A* for U5).
347 However, the size of the leading clone inferred by SCIFER differed strongly from the VAF of
348 the respective CH driver. In U4 the SCIFER-inferred clone (14% VAF) was much larger than
349 the VAF of the *TET2* mutation (1%). Hence, a large HSC clone had already grown before the
350 *TET2* mutation occurred. In U5, the *DNMT3A* mutation was acquired very early in life, most
351 likely during embryogenesis as there were only 6 clonal SSNVs (Fig. 6a, amplified right
352 panel) and almost reached fixation (43% VAF). Within this early selected clone, SCIFER

353 inferred a subclone with a VAF of 7.5% without a known CH driver. This interesting result
354 indicates that SCIFER may miss selection events occurring in embryogenesis (such as the
355 *DNMT3A* mutation in U5) because of the small clonal SSNV count (note that a clone of
356 similar size, but much later origin and hence larger SSNV count, was readily detected by
357 SCIFER in the pseudo-bulk sample id2259[22]; Extended Data Fig. 2g).

358 Assessing the HSC parameters (Fig. 6c), we found little variation in HSC number (Fig. 6c,
359 left panel) and overall similar numbers as in the CH cases with known drivers (Fig. 5d).
360 However, three individuals (U3-U5) had accelerated HSC divisions, with a rate close to 10
361 divisions per year, while the other two cases had a rate similar to the majority of non-CH
362 cases (about 2 divisions per year). The accelerated rate for U3-U5 agreed with the HSC
363 division rate of the majority of CH cases with known drivers (Fig. 5h). These data further
364 support the link between clonal selection events, with known or unknown drivers, and
365 accelerated HSC turnover.

366 **Homogeneous occurrence of selection events across life**

367 Finally, we asked when clones with or without known CH drivers originated in life and how
368 fast they expand. The clone ages inferred by SCIFER showed that clones of both kinds
369 typically emerged several decades before BM sampling (Fig. 7a); moreover, there were no
370 significant differences in clone age between associated driver mutations in *DNMT3A*, *TET2*
371 and *ASXL1* and unknown drivers (Fig. 7b). The time-averaged clonal growth rates showed
372 substantial inter-individual variation, even within the five individuals with a *DNMT3A* mutation
373 and the three individuals with a *TET2* mutation (Fig. 7c). We did not detect systematic
374 differences in growth rates for clones associated with different drivers or unknown drivers
375 (Fig. 7d).

376 A key question is when driver mutations of CH arise in life; previous work based on
377 phylogenetic reconstruction has inferred that expanded clones originate before 40 years of
378 age [26]. Combining all our cases with clonal selection, which were selected in our cohort
379 without reference to the age of the individual, we deduced the age-incidence distribution for
380 the onset of clonal selection (Fig. 7e). Intriguingly, we found that selected clones were born
381 at a fairly constant rate from early childhood to about 50 years of age. This limit does not
382 imply that selection events no longer occur beyond this age; it may rather be a limit of clone
383 detection (e.g., a clone that grows by 25% per year requires 40 years to grow to detectable
384 size, ~0.05×10^5 cells, for 90x bulk WGS; the oldest individual in our cohort is 89 years of age,
385 and hence such a clone born at age 49 would just be detectable). Hence our data suggest
386 that selection events in HSCs occur quite uniformly across life.

387

Discussion

Selection of mutant somatic cells is central to cancer evolution. In contrast, neutral evolution was initially emphasized in non-malignant tissues such as the normal gut epithelium by quantitative studies of stem cell dynamics[10]. However, more recent data show that, over the long human lifespan, selection also shapes somatic mosaicism in non-malignant tissues, and this can be accompanied by altered tissue function[22,24-26,35]. Here, we present a method to detect clonal selection and quantify the dynamic parameters of tissue stem cells from snapshot WGS data using a single bulk tissue sample. Our population-genetics approach, applied to populations of cells, provides a quantitative framework to model homeostatic competition of clones arising during development or adult homeostasis. Akin to recent quantitative approaches to cancer evolution[29-31,36-38], our theory delineates how the allele frequency spectra of somatic variants are shaped by neutral evolution versus selection. In contrast to previous work on distinguishing genetic drift from selection in expanding tumors [30,36], our study of somatic evolution in homeostatic tissues requires the explicit treatment of genetic drift and its interplay with clonal selection. This allows us to identify selection in an individual homeostatic tissue without knowledge of the underlying driver events.

An alternative for detecting selection in a normal tissue agnostic of the driver is provided by phylogenetic reconstruction from the genomes of individual stem and progenitor cells[2]. With ~300 HSPCs per individual analyzed, this method has yielded detailed insight into clonal structure of hematopoiesis in 12 individuals, thus far[22,26]. These studies have highlighted multiple selected clones in an individual (each selected clone corresponding to an expanded clade in the phylogeny and the smallest expanded clade having a frequency of ~1.3%). The majority of expanded clones had no known driver events. While less sensitive at sequencing depths of 90x we used here, our method is cost-effective, and hence scalable, enabling a search for unknown drivers in larger population cohorts. Moreover, if we sequence at 270x our simulations suggest SCIFER would achieve a similar level of sensitivity as phylogenetic reconstruction. SCIFER can also be applied to tissues where stem-cell clones cannot be grown easily from stem cells.

Moreover, the agreement of basic quantitative parameters of hematopoiesis – HSC number, turnover and mutation rate – inferred with the help of the phylogenetic tree[2] and from SCIFER is remarkable, as our method strongly relies on variants with higher VAF emerging relatively early, whereas the phylodynamic inference strongly relies on the large number of low-VAF SSNVs generated later. This suggests that these HSC parameters are rather uniform across time, at least before selection events occur.

Our study is the most comprehensive to date, of quantitative parameters of human HSC dynamics in humans. We find neutrally evolving hematopoiesis in individuals of all ages up to

our clonal detection limit. In individuals without CH, we find uniform dynamics of HSCs across life, except for two older individuals (72 and 74 years, respectively) with smaller numbers of stem cells. For these two individuals, we infer noticeable drift of variants, generated in fetal development and early in life, to allele frequencies detectable in bulk WGS. This observation cautions against solely using the number of SSNVs per se, as evidence for selection[27]. In several individuals with selected clones, HSC division and mutation rates are as in the cases of neutral evolution, indicating that CH can arise from normal HSC dynamics.

Remarkably, however, we found an enrichment for cases with increased time-averaged HSC division rate of HSC in individuals with CH. We also found a uniform rate of acquiring CH across life. This includes CH driver events occurring early in life or even in embryogenesis (such as the *DNMT3A* mutation in case U5), akin to blood cancer (myeloproliferative disorder[39]).There are at least three avenues of future research that arise from this observation. First, the accelerated HSC division may aid the emergence of CH by increasing the rate of driver mutation acquisition[21]. Second, as SCIFER estimates an average division rate for all HSC in an individual, it leaves open the question of whether the HSC division rate is different between HSCs that are selected, compared to those that are not selected, within a CH marrow. Recent single cell data suggest that all HSCs within in a CH bone marrow environment have transcriptional signatures of inflammation but that this is less pronounced in selected HSCs[33]. This raises the hypothesis that selected HSCs with a dampened response to inflammation, may function better in an inflammatory environment[40]. Third, as our data are time averaged from single timepoint measurements it would be fascinating to apply SCIFER suggest to measure HSC parameters (division rates and birth of clones) serially in longitudinal samples across an individual's lifetime. Specifically, it would be important to test the impact of specific stressful events such as intercurrent inflammation, infection and disease and therapeutic interventions. Given the scalability of SCIFER, this is now feasible.

Consistent with previous studies[22,25,26] we identified five cases with clonal selection in the absence of known drivers. The selective advantage conferred by unknown driver events was indistinguishable from known CH drivers, and unknown drivers were also frequently associated with increased HSC self-renewal rate. Although in-depth analyses of the genome-wide mutation profiles revealed potential candidates for driver mutations in four of these cases (Supplementary Table 6), the statistical power of the current study is not large enough to identify new recurrent driver mutations with confidence. However, by applying SCIFER to large cohorts it may be possible to more accurately determine the frequency of individuals with unknown driver events, where selection is operative in blood or other tissues. With this information in hand, the phenotype and clinical significance of these events could be assessed. Combined with in depth genome-wide multi-ome data and functional screens, this

may also provide insight into the mechanisms by which these currently unknown driver events mediate selection and shape tissue function.

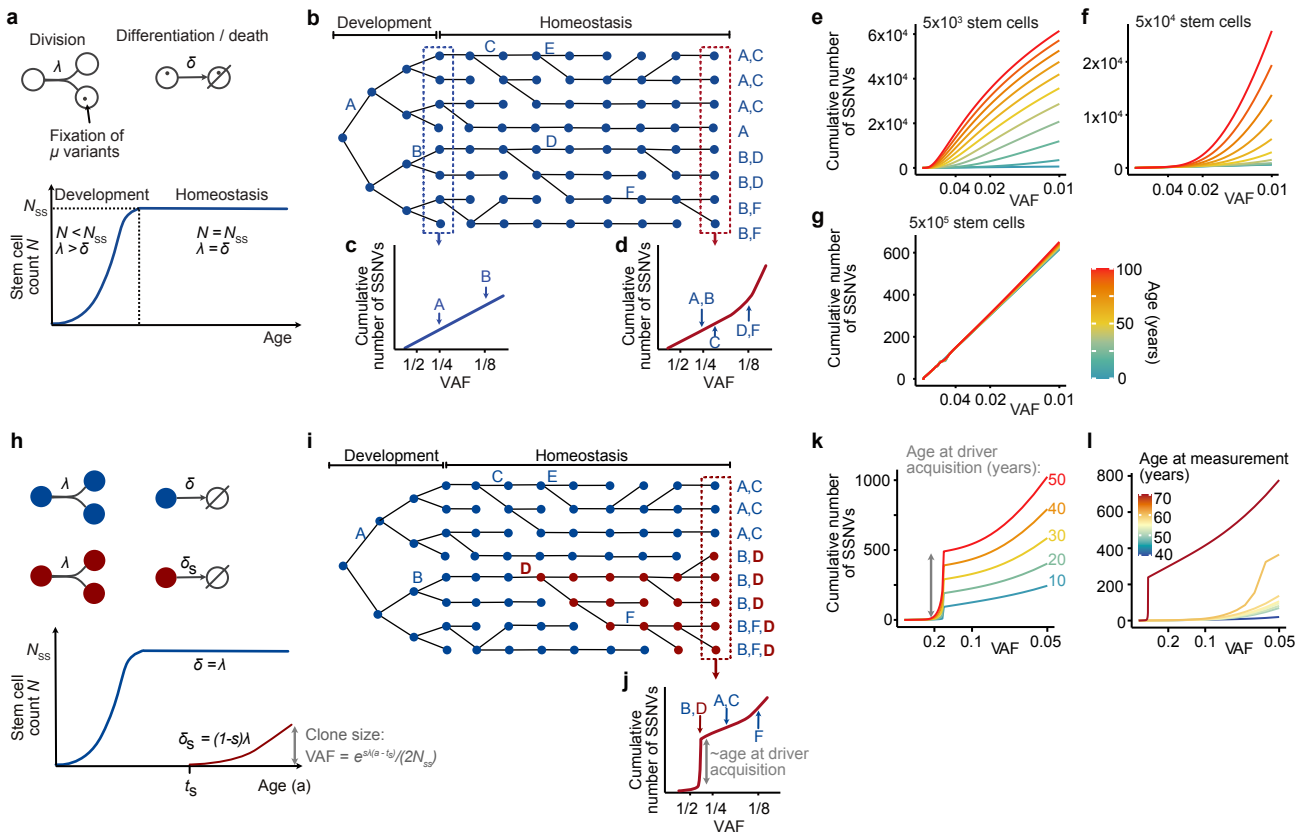


Fig. 1. Population-genetics model of drift and selection in homeostatic tissues. a, Modelled processes and associated parameters in the model of drift. Stem cells either divide symmetrically with rate λ, or exit the stem cell compartment by differentiating (or dying), with rate δ. The blood stem cell count (N) increases in development ($\lambda > \delta$) until reaching steady state numbers (N_{ss}) and remains constant during adulthood ($N = N_{ss}$ and $\lambda = \delta$). On average, cells acquire μ neutral variants during each cell division. **b**, Schematic illustrating variant accumulation during development and subsequent homeostasis. **c**, Cumulative number of somatic single nucleotide variants (SSNV) versus variant allele frequency (VAF) in development (where the scaling of the x-axis is transformed to 1/VAF to spread out low-frequency variants). **d**, Cumulative number of SSNV versus VAF in adult life (where the scaling of the x-axis is transformed to 1/VAF to spread out low-frequency variants). **e**, Simulated cumulative VAF distribution of SSNVs at selected ages between 0 and 100 years for $5x10^3$ stem cells; λ=5/year, μ=10/division. **f**, Simulated cumulative VAF distribution of SSNVs at selected ages between 0 and 100 years for $5x10^4$ stem cells; λ=5/year, μ=10/division. **g**, Simulated cumulative VAF distribution of somatic SSNVs at selected ages between 0 and 100 years for $5x10^5$ stem cells; λ=5/year, μ=10/division. **h**, Model of clonal selection. A selective driver event reduces the loss rate (differentiation or death) by a factor s, causing selective outgrowth of the mutant clone (red); the remaining parameters are defined in **a**. The VAF of the selected clone increases exponentially with the age at measurement, a. **i**, As in **b**, but here an acquired driver mutation (D) causes the selective

485 outgrowth of the mutant clone (red). **j**, All variants in the selected clone's cell of origin are
486 inherited to its progeny and hence reach a high VAF during clonal expansion, reflected in a
487 shoulder in the cumulative VAF distribution. **k**, Simulated cumulative VAF distributions when
488 a driver mutation is acquired at different ages, and the SSNVs are measured 45 years later,
489 when the clone has reached a size of 32%. Note that, the later in life the selected clone is
490 born, the more somatic variants have accumulated in its founder cell, and hence the larger
491 the height of the shoulder. In the simulation, the selected clone grows by 22% per year
492 (s=0.02, λ=10/year, μ=1/division, N_{ss}=25,000). **l**, Simulated cumulative VAF distributions
493 measured at varying ages after a driver mutation was acquired at 20 years of age. As the
494 time interval between the acquisition of the selective driver mutation and the sampling of the
495 tissue increases, the selected clone will be larger and eventually reach fixation. Therefore,
496 SSNVs defining the founder cell of the selected clone will have higher VAFs, and the position
497 of the shoulder on the x-axis will progressively move to the left. As in **k**, the selected clone
498 grows by 22% per year (s=0.02, λ=10/year, μ=1/division, N_{ss}=25,000).
499

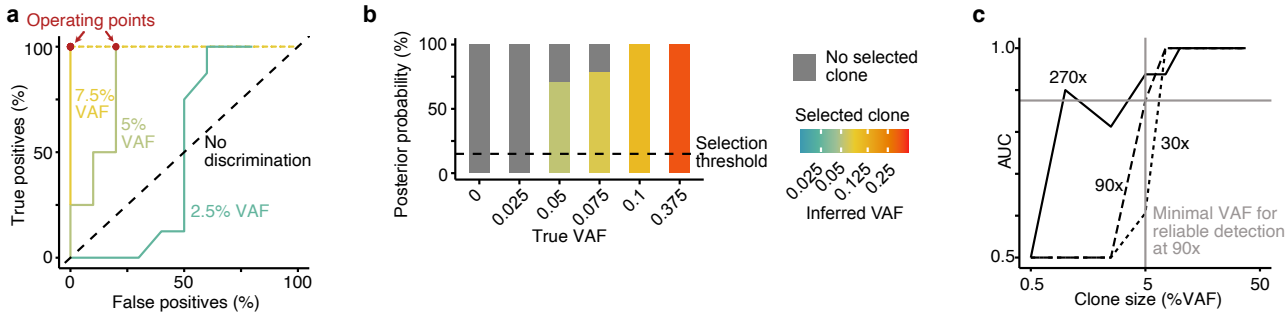
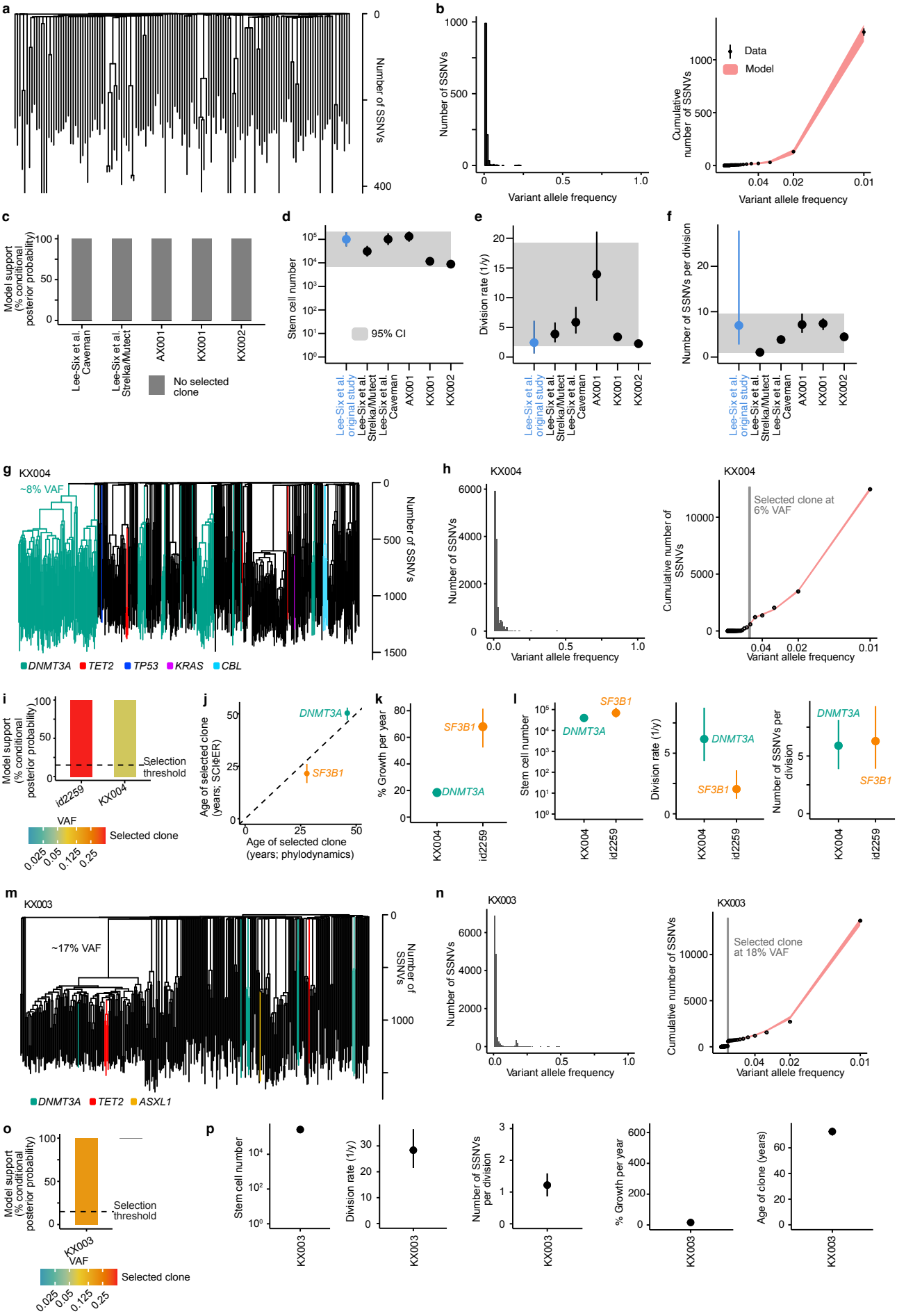


Fig. 2. Benchmarking SCIFER with simulated data. a, Receiver-operating curves (ROCs) quantifying the ability of SCIFER to accurately detect clonal selection for different sizes of a selected clone (color encoded). ROCs were generated by applying SCIFER to simulated data, generated with a stochastic birth-death process with (s=0.02, corresponding to a selective advantage of 2% increase in birth versus death) or without (s=0) selection of a clone initiated at 20 years-of life (λ=10/year, μ=1/division, N_{ss}=25,000, and assuming sequencing with an average coverage of 90x). In total 63 cases with selected clone sizes of VAF 0%, 2.5%, 5%, 7.5%, 10%, 25% and 37.5%, were generated. Models with and without clonal selection were fit to the data using approximate Bayesian computation. True positives and false positives were evaluated for varying posterior probability thresholds of clonal selection. For selected clones with VAFs ≥5%, the difference between true positives and false positives was maximal for a selection threshold of 15% (operating points, shown in red). **b**, Posterior probability for clonal selection (colored bars) and neutral evolution (grey bars) conditioned on selected clones with VAF ≥5% for six simulated cases (STN8, STS25, STS39, STS46, STS53, STS84) with varying clone size. Shown are true and inferred VAFs of the selected clone. The dashed line marks the selection threshold at 15% conditional posterior probability. **c**, Accuracy of SCIFER to distinguish clonal selection from genetic drift in simulated whole genome sequencing data. Shown are areas under the curve (AUC) computed from the receiver-operating curves (ROCs) shown in (**a**) and from ROCs obtained in analogy for simulated sequencing depths of 30x and 270x. The simulated data were generated as in (**a**). For sequencing depths of 30x and 90x, 63 cases with selected clone sizes of VAF 0%, 2.5%, 5%, 7.5%, 10%, 25% and 37.5%, were generated. For 270x sequencing depth, an additional 17 cases with selected clone sizes of 0.5% VAF and 1% VAF were used for model evaluation.


a
Number of SSNVs
b
Number of SSNVs
Variant allele frequency
Data
Model
Cumulative number of SSNVs
Variant allele frequency
c
Model support (% conditional posterior probability)
Lee-Six et al. Caveman
Lee-Six et al. Strelka/Mutect
AX001
KX001
KX002
No selected clone
d
Stem cell number
95% CI
Lee-Six et al. original study
Lee-Six et al. Strelka/Mutect
Lee-Six et al. Caveman
AX001
KX001
KX002
e
Division rate (1/y)
f
Number of SSNVs per division
g
KX004
~8% VAF
Number of SSNVs
DNMT3A
TET2
TP53
KRAS
CBL
h
KX004
Number of SSNVs
Variant allele frequency
KX004
Selected clone at 6% VAF
Cumulative number of SSNVs
Variant allele frequency
i
Model support (% conditional posterior probability)
Selection threshold
id2259
KX004
VAF
Selected clone
0.025
0.05
0.125
0.25
j
Age of selected clone (years; SCiΦER)
DNMT3A
SF3B1
Age of selected clone (years; phylodynamics)
k
% Growth per year
SF3B1
DNMT3A
KX004
id2259
l
Stem cell number
SF3B1
DNMT3A
Division rate (1/y)
DNMT3A
SF3B1
Number of SSNVs per division
DNMT3A
SF3B1
KX004
id2259
m
KX003
~17% VAF
Number of SSNVs
DNMT3A
TET2
ASXL1
n
KX003
Number of SSNVs
Variant allele frequency
KX003
Selected clone at 18% VAF
Cumulative number of SSNVs
Variant allele frequency
o
Model support (% conditional posterior probability)
Selection threshold
KX003
VAF
Selected clone
0.025
0.05
0.125
0.25
p
Stem cell number
Division rate (1/y)
Number of SSNVs per division
% Growth per year
Age of clone (years)
KX003

526

527 **Fig. 3. Benchmarking SCIFER with published pseudobulk data. a**, Reconstructed single-
528 cell phylogenies after re-calling SSNVs and indels from single-cell WGS data[2]. **b**, Left, VAF
529 distribution of SSNVs shown in **a**, truncated at 1%. Right, model fit to the cumulative 1/VAF
530 distribution (points and error bars, mean and standard deviation of the measured data; red
531 area, 95% posterior probabilities of the model fit). **c**, Posterior probability for neutral evolution
532 for pseudo-bulk WGS data from[2] (labelled Lee-Six) and three samples (AX001, KX001 and
533 KX002) from[26]; SCIFER did not support selection for a clone size >5% VAF. SCIFER was
534 applied twice to the data from Lee-Six et al.[2], using the SSNV counts obtained with Caveman
535 or with the intersection of Mutect and Strelka. **d**, Inferred hematopoietic stem cell number
536 (left), division rate (middle) and mutation rate (right) for the cases shown in **c** (median and
537 80% credible intervals for each sample; grey areas, 95% confidence band for all samples).
538 Estimates of these parameters from[2] are given for comparison. **e**, Single-cell phylogeny of
539 published sample KX004[26]. Clones with variants in *DNMT3A* (green), *TET2* (red), *KRAS*
540 (purple) and *CBL* (cyan) are shown. **f**, Left, VAF distribution of SSNVs of sample shown in **e**,
541 truncated at 1% VAF. Right, model fit to the cumulative 1/VAF distribution (points and error
542 bars, mean and standard deviation of the measured data; red area, 95% posterior
543 probabilities of the model fit; grey area, 80% credible interval of the estimated clone size). **g**,
544 Posterior probabilities for selection (conditioned on clones with VAF ≥5%) and neutral
545 evolution for pseudo-bulk WGS data of two samples: id2259 described in Extended Data Fig.
546 2g, with known CH driver mutations in *SF3B1* (id2259, ref.[22]) *CBL* and (id KX004, ref.[26]). The
547 color scale shows the inferred VAF of selected clone. The dotted line shows the 15%
548 selection threshold. **h**, Comparison of age of the selected clones in the cases shown in **g**,
549 estimated by SCIFER and by phylodynamic modelling in the original publications (points,
550 median; error bars, 80% credible intervals). **i**, Estimated growth rates of the selected clones
551 in the samples shown in **g** (points, median; error bars, 80% credible intervals). **j**, Estimated
552 hematopoietic stem cell number for the two cases shown in **g**. Shown are median and 80%
553 credible interval. **k**, Estimated HSC division rate for the two cases shown in **g**. Shown are
554 median and 80% credible interval. **l**, Estimated number of new SSNVs acquired per cell
555 division for the two cases shown in **g**. Shown are median and 80% credible interval. **m**,
556 Single-cell phylogenies in published sample KX003 (ref.[26]), where the large expanded clone
557 had an unknown driver event. Small clones with variants in *DNMT3A* (green), *TET2* (red) and
558 *ASXL1* (orange) are also shown. **n**, Left, VAF distribution of SSNVs shown in **m**, truncated at
559 1%. Right, model fit to the cumulative 1/VAF distribution (points and error bars, mean and
560 standard deviation of the measured data; red area, 95% posterior probabilities of the model
561 fit; grey area, 80% credible interval of the estimated clone size). **o**, Posterior probabilities for
562 clonal selection (conditioned on clones with VAF ≥5%) for pseudo-bulk WGS data of sample
563 KX003 (ref.[26]). The color scale shows the inferred VAF of selected clone. The dotted lines

564 show the 15% selection threshold. **p**, Estimated stem cell parameters (hematopoietic stem
565 cell number, stem cell division rate and number of new SSNVs acquired per cell division) and
566 selection parameters (growth rate and age of selected clone) for the sample shown in **m**.
567 Shown are median and 80% credible intervals.
568

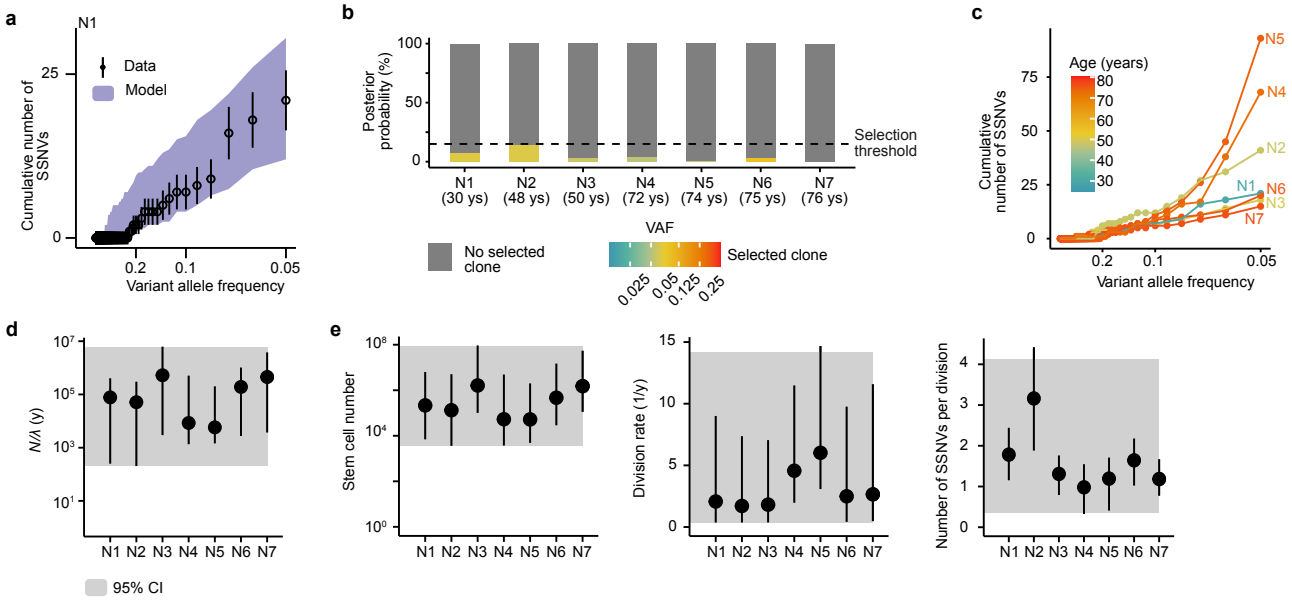


Fig. 4. Neutral evolution in physiological hematopoiesis measured by bulk WGS. a, Model fit to the cumulative 1/VAF distribution measured in DNA from CD34$^+$ HSPCs of sample N1 (points and error bars, mean and standard deviation of the measured data; purple area, 95% posterior probabilities of the model fit). **b**, Model support for clonal selection (posterior probabilities conditioned on clones with VAF ≥5%) and neutral evolution across seven samples without known CH driver mutation (N1-N7). The color scale shows the inferred VAF of selected clone. The dotted lines shows the 15% selection threshold. All cases are classified as neutrally evolving. **c**, Cumulative VAF distributions for samples N1-N7 shown in **b**. Age of individuals from whom samples N1-N7 were obtained are color coded. **d**, Inferred ratio between stem cell number and division rate (N/λ). **e,** Inferred parameter estimates for the number of hematopoietic stem cells (left panel), division rate of a stem cell per year (middle panel) and the number of new SSNVs acquired per cell division for samples N1-N7 shown in **a**. Shown are median and 80% credible intervals for each sample; the grey areas give the 95% confidence interval across the lower and upper bounds of all samples.

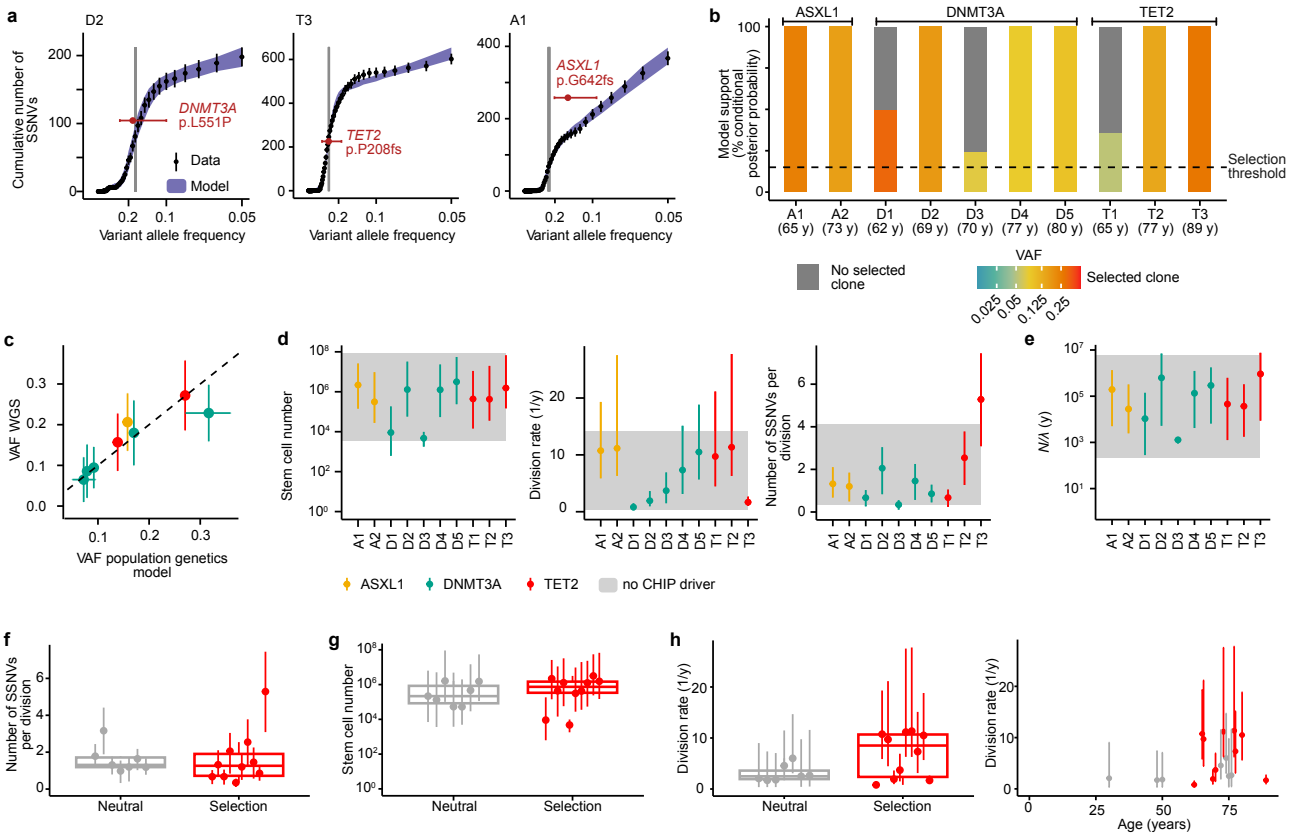


Fig. 5. Clonal selection in hematopoiesis measured by bulk WGS. a, Model fit to the cumulative VAF distribution measured in CD34$^+$ HSPCs of samples D2, T3 and A1 (points and error bars, mean and counting errors of the measured data; purple area, 95% posterior probabilities of the model fit; grey area, 80% credible interval of the estimated clone size; red points and error bars, mean and 95% confidence interval of the VAF of known CH drivers). **b**, Model support for clonal selection (posterior probability conditioned on clones >5% VAF) and neutral evolution in ten cases classified as selected and with a CH driver mutation in *AXSL1* (A), *DNMT3A* (D) and *TET2* (T). The age of the individuals is shown below each bar; the color scale shows the inferred VAF of the selected clone and the dotted lines shows the 15% selection threshold. **c**, Estimated sizes of the selected clones (median and 80% credible intervals) versus measured VAF of the known CH driver (mean and 95% CI according to a Binomial distribution). **d**, Estimated number of HSCs contributing to hematopoiesis (left), their division rate (middle) and number of SSNVs acquired per cell division (right) for the ten cases shown in **b**. Shown are median and 80% credible intervals for each sample; the grey areas give the 95% confidence interval across the lower and upper bounds of the seven neutrally evolving samples without known CH driver for comparison (c.f. Fig. 4). **e**, As in **d**, but showing the ratio between stem cell number and division rate (*N*). **f,** Estimated number of newly acquired SSNVs per HSC division in the 7 neutrally evolving cases (introduced in Fig. 4) as compared to 10 cases with selection for a known CH driver (introduced in this figure; points and error bars, median and 80% credible intervals for each sample; boxplots, median

607 and interquartile range, whiskers extend to the largest and smallest value no further than 1.5
608 times the interquartile range). **g,** Estimated number of HSCs contributing to hematopoiesis in
609 the 7 neutrally evolving cases (introduced in Fig. 4) as compared to 10 cases with selection
610 for a known CH driver (introduced in this figure; points and error bars, median and 80%
611 credible intervals for each sample; boxplots, median and interquartile range, whiskers extend
612 to the largest and smallest value no further than 1.5 times the interquartile range). **h**,
613 Estimated HSC division rate in the 7 neutrally evolving cases (introduced in Fig. 4) as
614 compared to 10 cases with selection for a known CH driver (introduced in this figure; Points
615 and error bars, median and 80% credible intervals for each sample; boxplots, median and
616 interquartile range, whiskers extend to the largest and smallest value no further than 1.5
617 times the interquartile range). The right panel plots the estimated division rate against the
618 age of the individuals from whom samples were obtained. Samples that are neutrally
619 evolving are shown in grey, samples with evidence of clonal selection are shown in red.

620

621

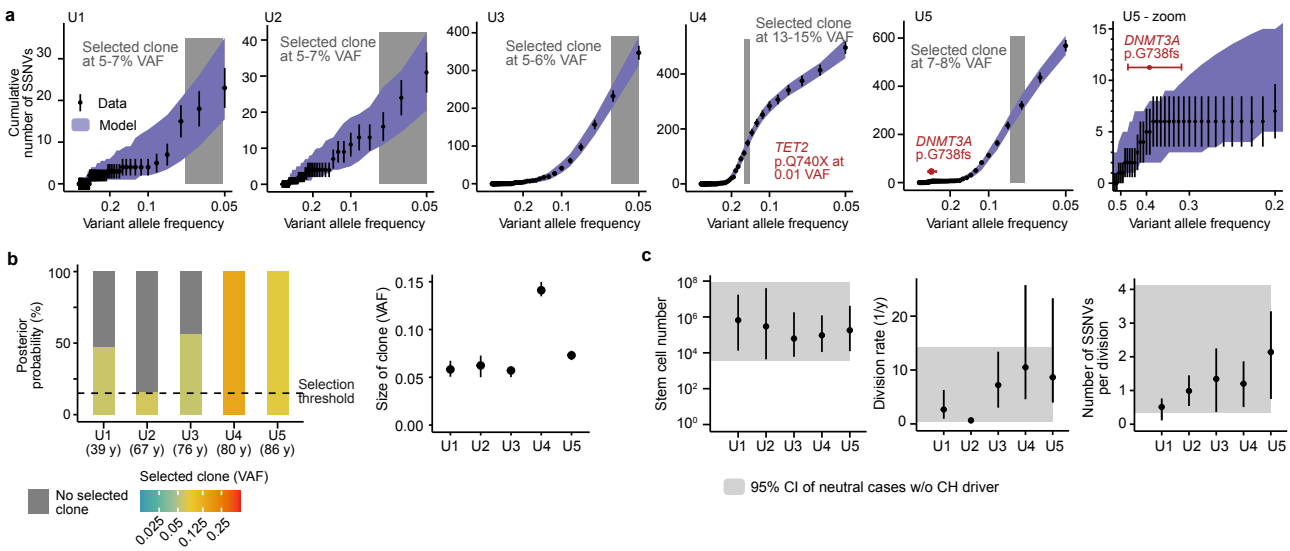


Fig. 6. Clonal selection for unknown drivers. a, Model fits to the cumulative 1/VAF distribution measured in CD34$^+$ HSPCs of samples U1-U5 (points and error bars, mean and standard deviation of the measured data; purple area, 95% posterior probabilities of the model fit; grey area, 80% credible interval of the estimated clone size). In U4 and U5, red points and error bars show mean and 95% confidence interval of the VAF of known CH driver mutations (*TET2* for U4 and *DNMT3A* for U5). In both U4 and U5 selection was not associated with the known CH driver mutations (U4, inferred selected clone at 14% VAF and *TET2* mutation at 1% VAF; U5, inferred selected clone at 7% VAF and *DNMT3A* mutation at 43% VAF). The panel on the right shows an expanded view of the cumulative VAF distribution and the model fitting to the data at high VAF. At most 6 SSNVs were acquired prior to the DNMT3A mutation, suggesting that the mutation was acquired early during development. **b**, Model support for clonal selection (posterior probability conditioned on selected clones with ≥5% VAF) and neutral evolution across samples U1-U5. The age of the individuals is shown below each bar; the color scale shows the inferred VAF of selected clone and the dotted lines shows the 15% selection threshold. The right panel gives the estimated clone size (median and 80% credible intervals). **c**, Estimated number of HSC contributing to hematopoiesis stem cell number (left), HSC division rate (central) and number of newly SSNVs acquired per cell division (right). Shown are median and 80% credible intervals for each sample; for comparison, the grey areas give the 95% confidence interval across the lower and upper bounds of neutrally evolving samples without known CH driver (as shown in Fig. 4).

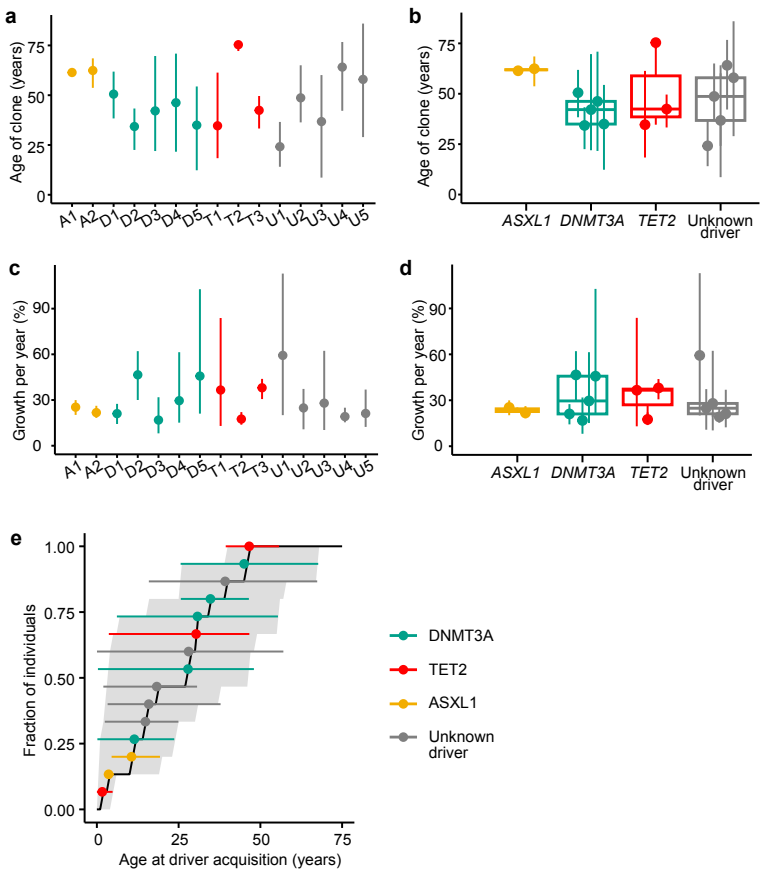
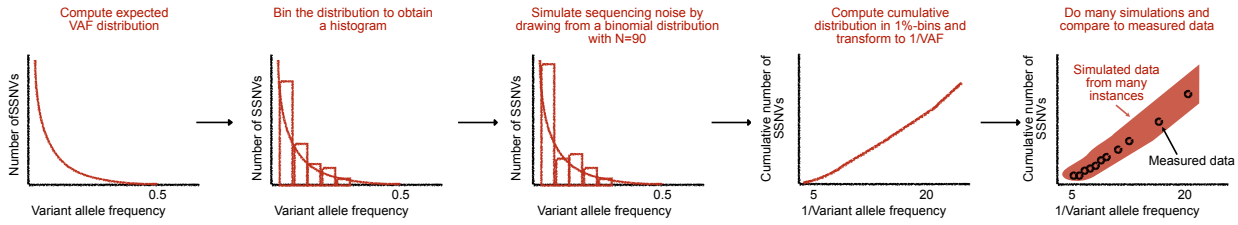
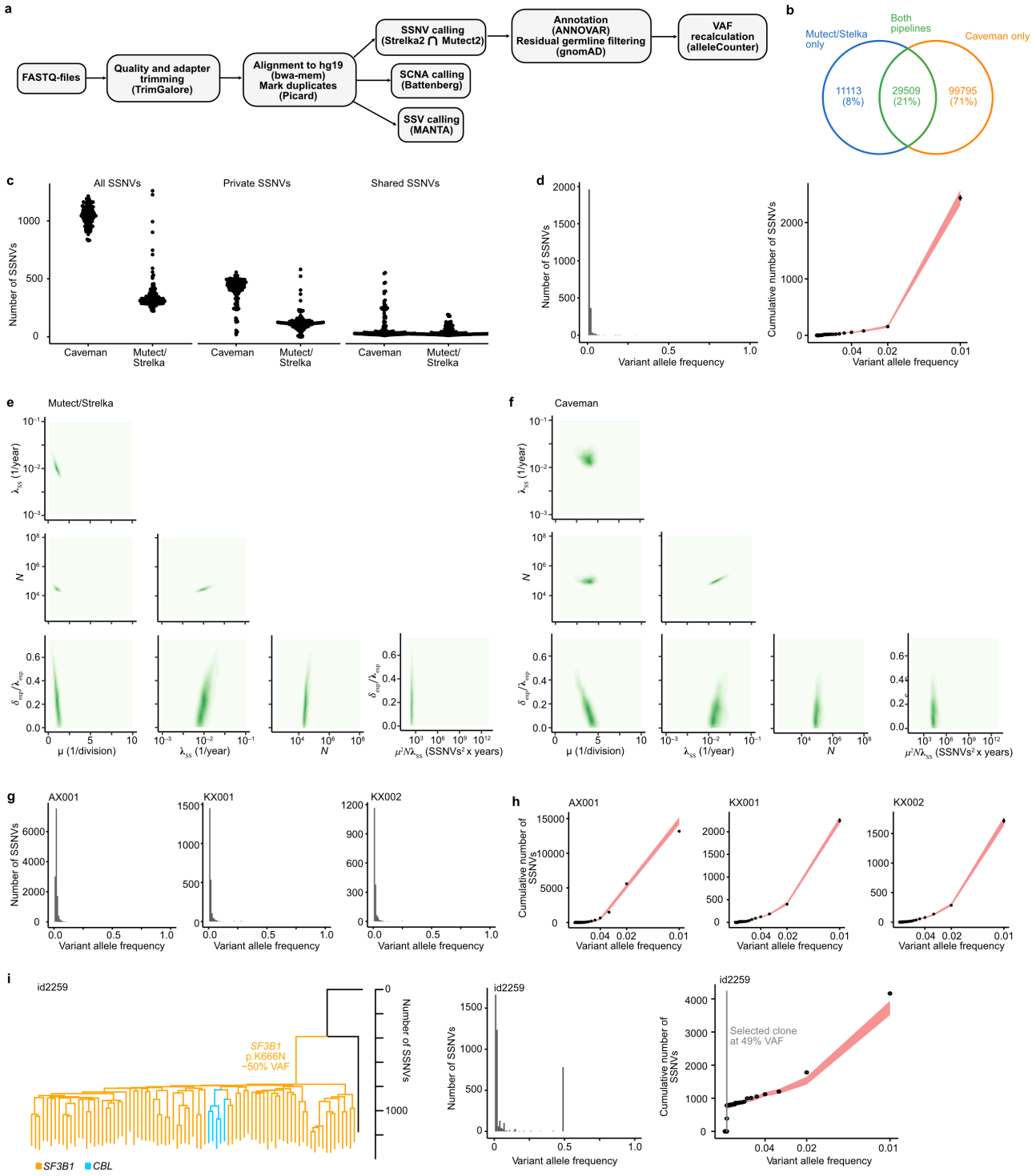


Figure 7. Selection dynamics with and without known drivers. a, Estimated age of the selected clone (for the ten cases with known drivers introduced in Fig. 5 and the five cases with unknown drivers introduced in Fig. 6; points, median; error bars, 80% credible intervals). **b**, Per-driver-gene comparison of the estimated age of the selected clone; data are the same as in **a** (points, median; error bars, 80% credible intervals; boxplots, median and interquartile range, whiskers extend to the largest and smallest value no further than 1.5 times the interquartile range). **c**, Estimated growth rates for the fifteen cases introduced in **a** (points, median; error bars, 80% credible intervals). **d**, Per-driver-gene comparison of the estimated growth rates for the fifteen cases introduced **a** (points, median; error bars, 80% credible intervals; boxplots, median and interquartile range, whiskers extend to the largest and smallest value no further than 1.5 times the interquartile range). **e**, Cumulative distribution of estimated age at CH driver acquisition (median); shaded area, lower and upper bounds of the cumulative distribution of estimated age at CH driver acquisition based on 80% credible intervals of the estimated parameters; points and error bars, median and 80% credible intervals for the per-sample estimates. Data are from ten cases with selection of known CH driver mutations and five cases with selection of unknown drivers introduced in **a**.

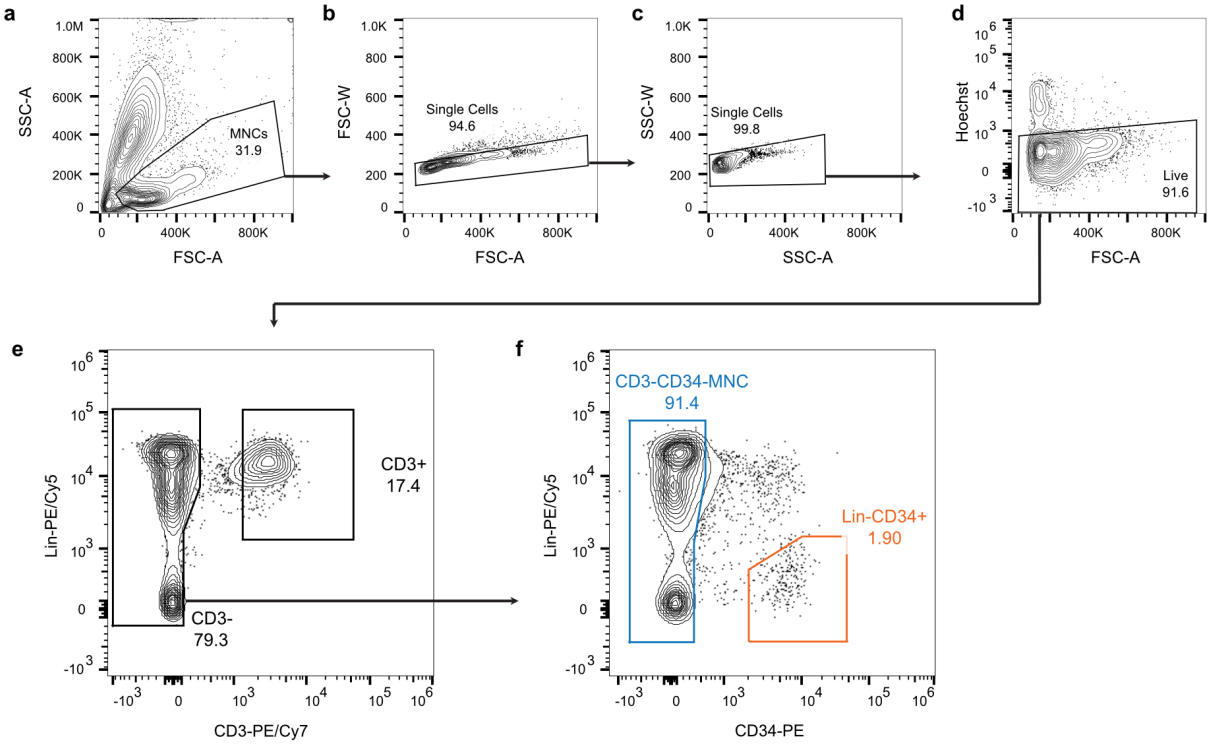


Extended Data Fig. 1. Quantifying selection and drift from bulk whole genome sequencing data. Parameter estimation with approximate Bayesian computation. First, the expected variant allele frequency histogram is analytically computed. Then, sequencing noise is simulated by drawing from a binomial distribution with average 90x coverage. The modelled cumulative distribution is compared to the measured data for varying VAFs (1% step size).

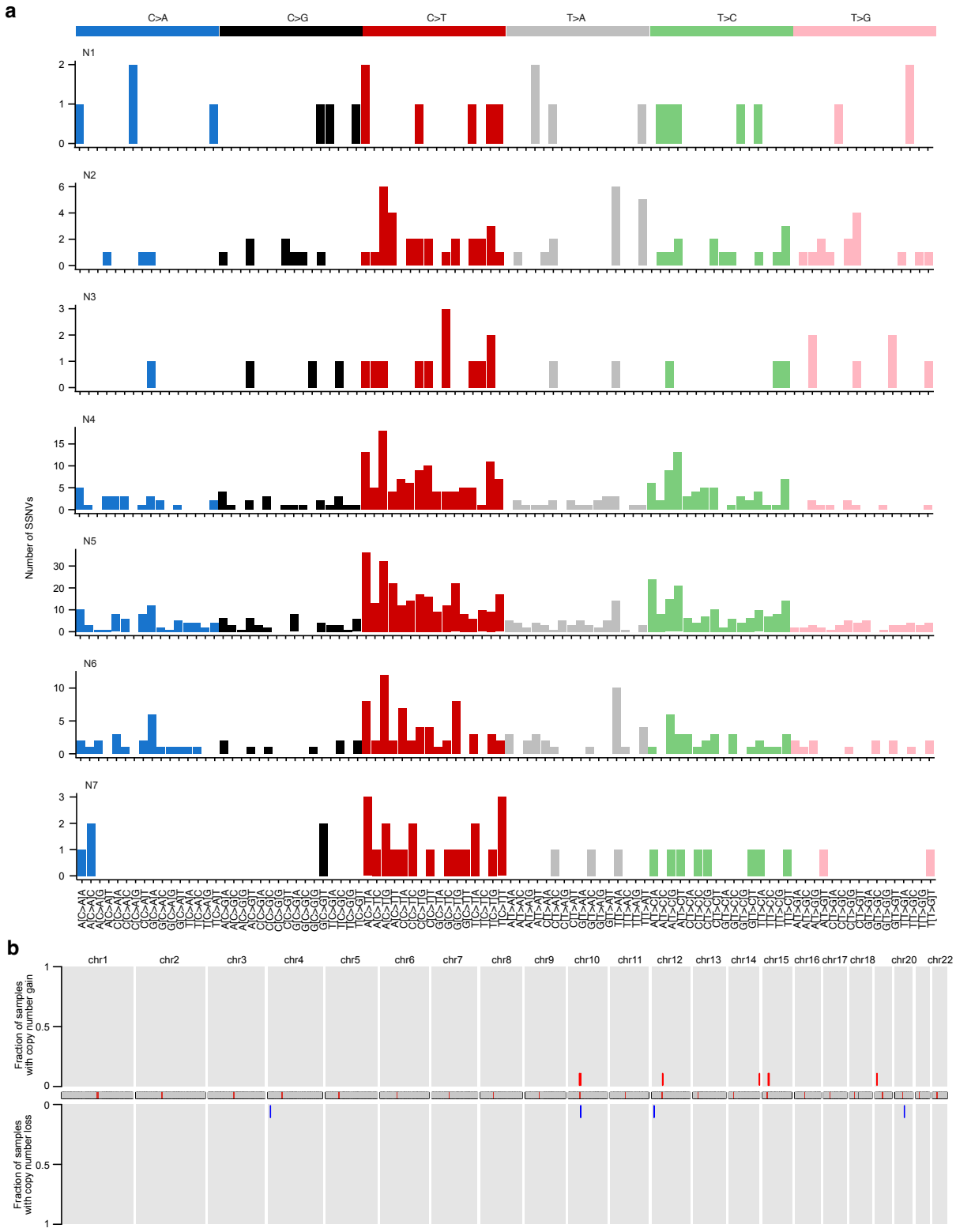


Extended Data Fig. 2. Quantifying drift and selection in published WGS data from single HSPC clones. a, Somatic variant calling pipeline (using hg19). SSNV, somatic single nucleotide variant. SCNA, somatic copy number abnormality. SSV, somatic structural variant. **b**, Number of shared and unique SSNVs identified in 140 whole genomes from HSPC clones from published data[2] identified by Caveman (original publication) and our pipeline (intersection of Mutect and Strelka shown in **a**). **c,** Number of SSNVs of different classes

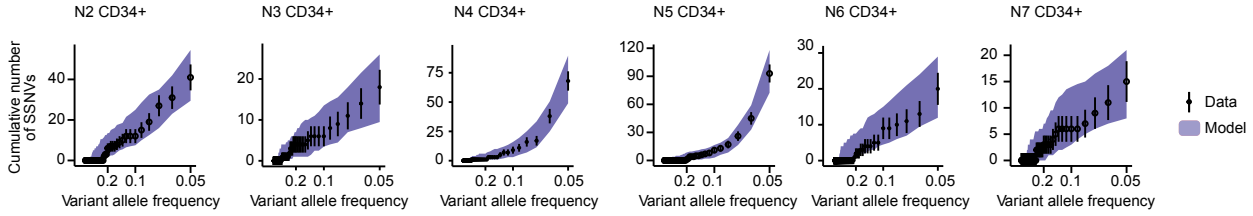
identified by each variant caller. **d**, Left panel, variant allele frequency distribution (truncated at 0.01 VAF) of SSNVs identified by Caveman from published data[2]. Right panel, model fit to the cumulative 1/VAF distribution shown in left panel (points and error bars, mean and standard deviation of the measured data; red area, 95% posterior probabilities of the model fit). **e,** Two-dimensional posterior probability distributions for the parameters estimated by SCIFER from the cumulative 1/VAF distribution of SSNVs identified with our pipeline (intersection of Mutect and Strelka) in published data[2]. Axis limits show the range of the prior distributions. **f,** Two-dimensional posterior probability distributions for the parameters estimated by SCIFER from the cumulative 1/VAF distribution of SSNVs identified with Caveman in published data[2]. Axis limits show the range of the prior distributions. Note that while the posterior for the fraction of cell loss, $\delta_{\mathrm{exp}}/\lambda_{\mathrm{exp}}$, is relatively broad, this has only a small effect on the values of the mutation rate and also the other parameters (see Supplementary Note 1). **g**, VAF distribution (truncated at 0.01 VAF) of SSNVs from 1150 genomes from three samples (AX001, KX001 and KX002) that did not show signs of clonal selection (SSNVs taken without re-call from published data[26]). **h**, Model fit to the cumulative VAF distribution shown in **g** (points and error bars, mean and standard deviation of the measured data; red area, 95% posterior probabilities of the model fit). **i**, Left, phylogenetic tree of published sample id2259, containing a large selected clone with a *SF3B1* pK605N mutation (and a smaller *CBL* subclone)[22] (reconstruction from the original publication). Middle panel, VAF distribution (truncated at 0.01 VAF) of SSNVs (taken without re-call from[22]). Right panel, model fit to the cumulative 1/VAF distribution (points and error bars, mean and standard deviation of the measured data; red area, 95% posterior probabilities of the model fit; grey area, 80% credible interval of the estimated clone size).



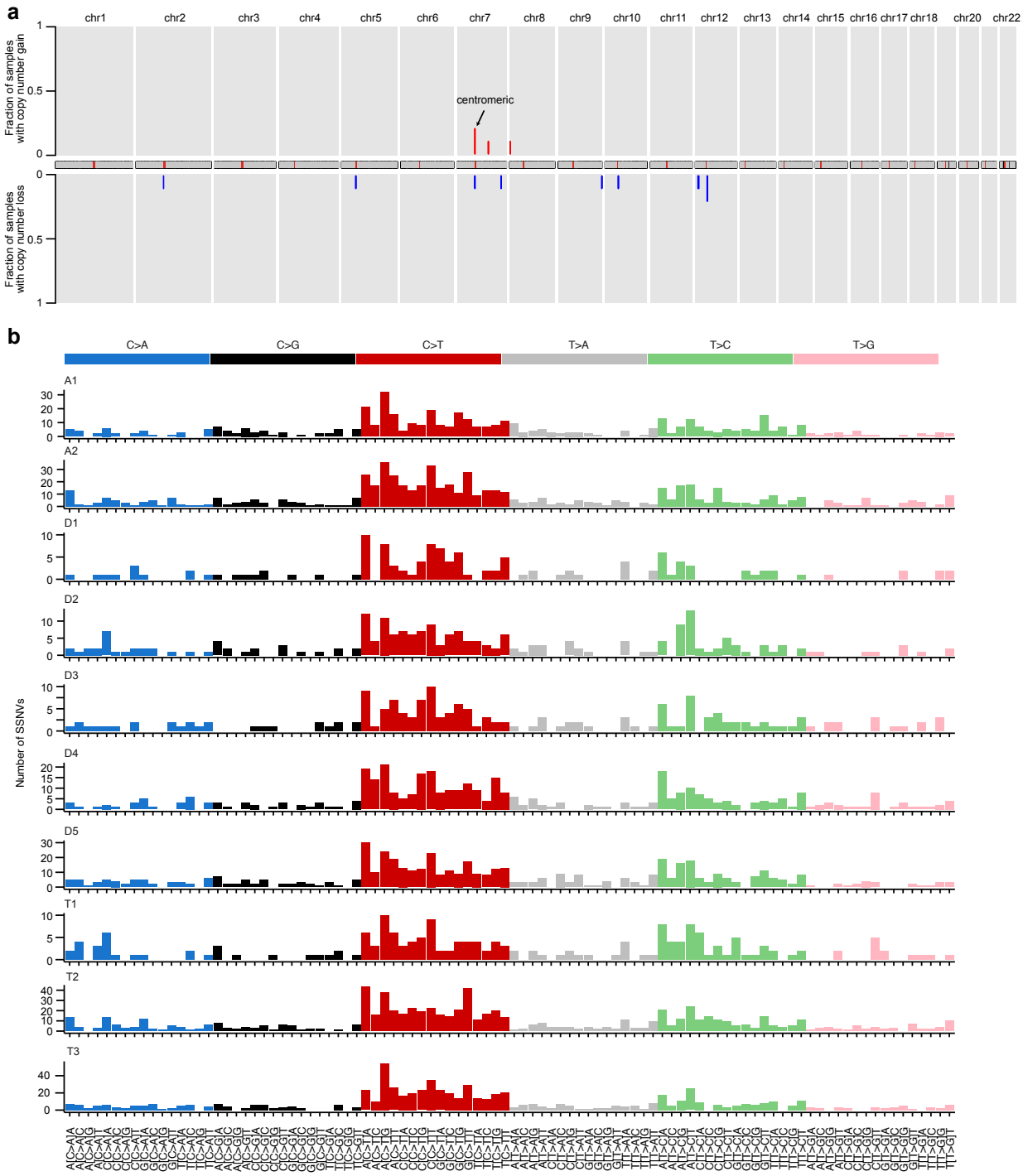
Extended Data Fig. 3. FACS gating for hematopoietic stem and progenitor cells. a-f, Cells were stained with a panel of antibodies (Methods), then single and live mononuclear cells were sorted (MNCs, **a-d**). T cell depleted MNCs (MNC(–T)) were sorted from the CD3-, CD34+ cell fraction (**e** and **f**), and, hematopoietic stem and progenitor cells were sorted from the CD3-Lin-CD34+ cell fraction (**f**). Image created with FlowJo v10.8.1; values in **a-f** give percentages.



Extended Data Fig. 4. Somatic variants in individuals with neutral evolution. a, Trinucleotide substitution patterns across the genome in CD34$^+$ HSPC in samples N1-N7. **b**, Genome-wide (chromosomes indicated and arrayed across x-axis) profiles of copy number gains (red) and losses (blue) in CD34$^+$ HSPC across the 7 same samples as in **a**.



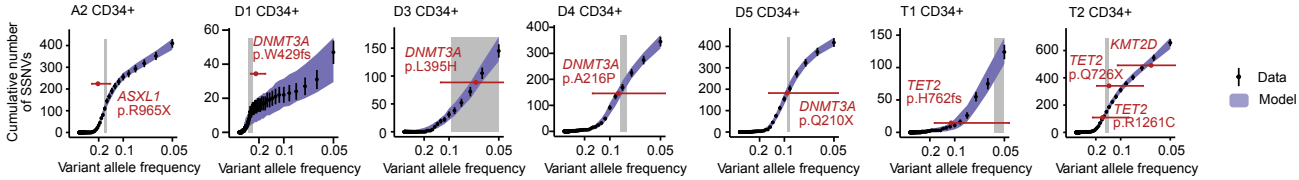
Extended Data Fig. 5. Quantifying neutral drift and mutation accumulation from bulk whole genome sequencing. Model fits to the cumulative VAF distributions measured by whole genome sequencing (WGS) of $CD34^+$ HSPC samples from six individuals (N2-N7) without known CH driver mutation (points and error bars, mean and counting errors of the measured data; purple areas show the 95% posterior probabilities of the model fit). The model fits show no evidence of selection.



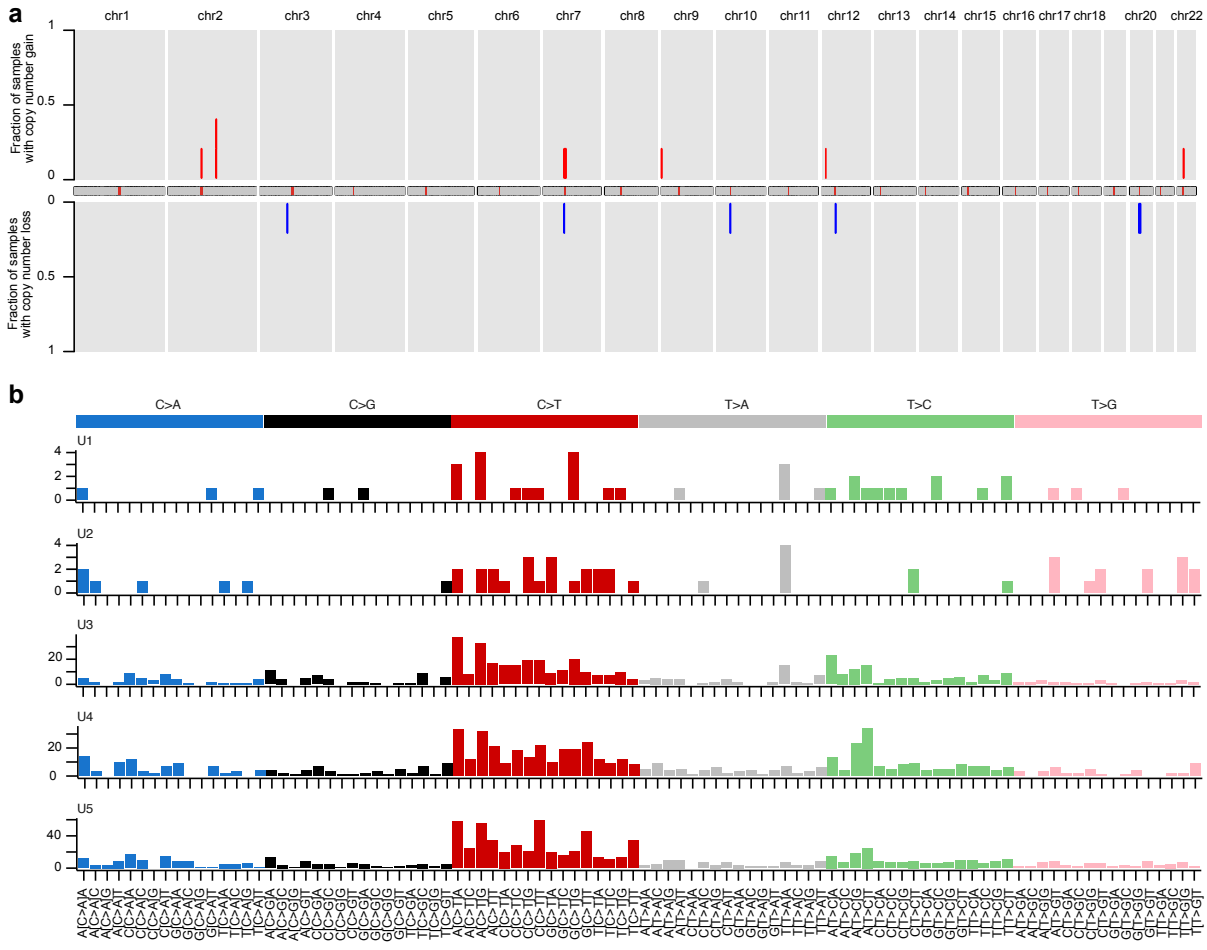
725

726 **Extended Data Fig. 6. Somatic mutations in clonal hematopoiesis. a**, Genome-wide
727 (chromosomes indicated and arrayed across x-axis) profiles of copy number gains (red) and
728 losses (blue) in CD34$^+$ HSPC from 10 individuals (A1-A2, D1-D5 and T1-3) with a known CH
729 driver mutation. **b**, Genome wide trinucleotide substitution patterns in CD34$^+$ HSPC across
730 the 10 same samples as in **a**.

731



Extended Data Fig. 7. Quantifying clonal selection from bulk whole genome sequencing. The cumulative VAF distributions measured by whole genome sequencing (WGS) of CD34^{+} HSPC from seven individuals with known CH driver mutations and evidence for clonal selection (A2, D1, D3-5 and T2-3; A=individual with *ASXL1* mutation; D=individual with *DNMT3A* mutation; T=individual with *TET2* mutation) (points and error bars, mean and counting errors of the measured data). The violet areas show the 95% posterior probabilities of the model fit. Grey areas, 80% credible interval for the estimated clone size; red points and error bars, mean and 95% confidence interval of the VAF of known CH drivers



743

744 **Extended Data Fig. 8. Somatic mutations in clonal hematopoiesis with selection of**
745 **unknown drivers. a**, Genome-wide (chromosomes indicated and arrayed across x-axis)
746 profiles of copy number gains (red) and losses (blue) in CD34$^+$ HSPC from 5 individuals (U1-
747 5) where expanded selected clones were detected but no known genetic driver was
748 identified. **b**, Genome wide trinucleotide substitution patterns in CD34$^+$ HSPC across the 5
749 same samples as in **a**.
750

751 **Methods**

752 **Ethical approval**

753 Patient samples were collected with informed consent under the Mechanisms of Age-Related
754 Clonal Haematopoiesis (MARCH) Study. Written informed consent was obtained from all
755 participants in accordance with the Declaration of Helsinki. This study was approved by the
756 Yorkshire & The Humber - Bradford Leeds Research Ethics Committee (REC Ref:
757 17/YH/0382).

758 **Study samples**

759 Participants were recruited from individuals undergoing elective total hip replacement surgery
760 at the Nuffield Orthopaedic Centre, Oxford. Exclusion criteria were: history of rheumatoid
761 arthritis or other inflammatory arthritis, history of septic arthritis in the limb undergoing
762 surgery, history of hematological cancer, bisphosphonate use, and oral steroid use. Patient
763 characteristics are summarized in Supplementary Table 2. At the time of surgery, trabecular
764 bone fragments and BM aspirates were obtained from the femoral canal and collected in
765 anticoagulated buffer containing acid-citrate-dextrose, heparin sodium and DNase. Samples
766 of peripheral blood were collected in EDTA vacutainers. Hair follicle samples were collected
767 from participants as a germline control. Peripheral blood and BM mononuclear cells (MNCs)
768 were isolated by Ficoll density gradient centrifugation and viably frozen. Peripheral blood
769 granulocytic cell pellets were frozen for later DNA extraction. Genomic DNA was extracted
770 from bone marrow MNCs, peripheral blood granulocytes and hair follicles using a DNeasy
771 Blood & Tissue Kit (Qiagen).

772 **Cell sorting**

773 Thawing media was prepared with IMDM medium (Gibco) supplemented with 20% fetal
774 bovine serum (FBS) and 110 µg/mL DNase. BM samples were thawed at 37°C in a water
775 bath, 1 mL warm FBS was added, and the suspension then diluted by dropwise addition of 8
776 mL thawing media. The suspension was centrifuged at 400 g for 10 mins, cells were
777 resuspended in flow cytometry staining medium (IMDM with 10% FBS and 10 µg/mL
778 DNase), filtered through a 35 µm cell strainer, and placed on ice.

779 Cells were stained with the following antibodies: anti-CD34-PE (1:160, Biolegend, clone
780 581), anti-CD3-PE/Cy7 (1:100, Biolegend, clone HIT3a), anti-CD2-PE/Cy5 (1:160,
781 Biolegend, clone RPA-2.10), anti-CD4-PE/Cy5 (1:160, Biolegend, clone RPA-T4), anti-CD7-
782 PE/Cy5 (1:160, Biolegend, clone CD7-6B7), anti-CD8a-PE/Cy5 (1:320, Biolegend, clone
783 RPA-T8), anti-CD11b-PE/Cy5 (1:160, Biolegend, clone ICRF44), anti-CD14-PE/Cy5 (1:160,
784 eBioscience, clone 61D3), anti-CD19-PE/Cy5 (1:160, Biolegend, clone HIB19), anti-CD20-
785 PE/Cy5 (1:160, Biolegend, clone 2H7), anti-CD56-PE/Cy5 (1:80, Biolegend, clone MEM188),

786 and anti-CD235ab-PE/Cy5 (1:320, Biolegend, clone HIR2). Following antibody incubations,
787 cells were washed with 1 mL flow cytometry staining buffer, centrifuged at 350 g for 5 min
788 and resuspended in flow cytometry staining buffer containing 1:10,000 Hoechst 33342
789 live/dead stain.

790 Cell sorting was performed on a BD FACSAria Fusion or Sony MA900 equipped with a 100
791 µm nozzle or sorting chip. Unstained, single-stained, and FMO controls were used to
792 determine background staining and compensation in each channel. Doublets and dead cells
793 were excluded. The following populations were sorted with a mean purity > 95%: $Lin^{-}CD34^{+}$
794 HSPCs, $Lin^{+}CD3^{+}$ T cells, and $Lin^{+/-}CD34^{-}CD3^{-}$ MNCs. Genomic DNA from sorted cell
795 populations was extracted using a QIAamp DNA Micro Kit (Qiagen).

796 **Whole genome sequencing**

797 *Library preparation and sequencing*

798 Sequencing libraries were prepared using the Illumina DNA Prep Kit (#20018704) and
799 Nextera DNA CD Indexes (#20018707) according to manufacturer's guidelines. The Qubit
800 HS DNA Assay Kit (Invitogen, #Q32854) and a tape station run with the Agilent HS D5000
801 Assay Kit (#5067-5593) were used for quality control. Thereafter, libraries were diluted to
802 10nM, pooled and sequenced on either HiSeq X PE150 or on NovaSeq 6000 PE150.

803 Low quality bases on raw sequencing reads were trimmed using Trim Galore v0.5.0
804 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with cutadapt v2.8[41] with
805 the following settings: –quality 30, –illumina –length 32 –trim-n –clir_R1 2 –clip_R2 2 –
806 three_prime_clip_R1 2 –three_prime_clip_R2 2. Trimmed read pairs were mapped to the
807 human reference genome (build 37 with the GRCh37.75 genome annotation) using bwa
808 mem v0.7.12[42]. Mapped reads were coordinate-sorted with samtools sort v1.5[42], followed by
809 marking duplicate read pairs with gatk MarkDuplicates v4.0.9.0[43] and indexing with samtools
810 index.

811 *Detection of SSNVs and indels*

812 Somatic SNVs and small insertions/deletions in BM and blood samples were called with
813 Strelka v2.9.2[44] and Mutect2, GATK v4.2.0.0[43], using the matched hair follicle samples as
814 germline control. Variants in repeat regions and simple repeat regions (downloaded from
815 UCSC table browser, setting the assembly to hg19, the track to "RepeatMasker" or "Simple
816 Repeats"; accession date: December 5th, 2018) were filtered using bedtools intersect
817 v2.24.0[45]. To generate a set of high-confidence variants, only variants that passed the default
818 filters of both Strelka and Mutect2 were retained. To identify remaining germline variants,
819 variants were looked up in dbSNP (v150) and population frequency of reported variants was
820 annotated based on the Genome Aggregation Database (gnomAD v2.1.1;

821 https://gnomad.broadinstitute.org/). Variants were then filtered to exclude potential
822 contamination of germline variants based on the global allele frequency (AF) of gnomAD.
823 Specifically, only variants with AF <0.01 were retained as somatic variants. In a few cases, a
824 CH driver was identified by panel sequencing at low VAF, but was not recovered by Strelka
825 and Mutect; here, we re-examined the respective position using bcftools mpileup v1.10.2[46].
826 All variants were annotated with ANNOVAR (version May2018;
827 http://annovar.openbioinformatics.org/)[47] according to the human genome build version 19.
828 For SSNVs, VAFs were re-calculated directly from the BAM files using alleleCounter v.4.0.2
829 (https://github.com/cancerit/alleleCount; with default base and mapping quality thresholds)
830 and MaC (https://github.com/nansari-pour/MaC).

831 *Detection of copy number variants*

832 Genome-wide subclonal copy number aberrations (CNA) were identified with Battenberg
833 (v2.2.10; https://github.com/Wedge-lab/battenberg), which has been described in detail
834 previously[48].

835 *Detection of structural variants*

836 Structural variants (SVs) from whole genome data were identified using MANTA v1.6.0[49] with
837 default tumor-normal pair settings where samples were compared to the germline control to
838 identify somatic structural rearrangements. We kept variants that passed Manta's default
839 filter, had a minimal SOMATICSCORE of 30, were not classified as IMPRECISE, had at
840 most 2 variant reads in the control sample, and had VAF <5% in the control sample.

841 *Driver analysis*

842 To comprehensively search for candidate driver mutations, we concatenated published lists
843 of putative driver genes in clonal hematopoiesis and leukemia from

844 - The cancer driver database Intogen[50] (subsetting genes described in "ALL", "AML",
845 "CLL", "CML" or "MM" from release 2020.02.01, and the clonal hematopoiesis gene
846 list from Pich et al., 2022[51]),
847 - The cancer driver database Cosmic[52] v94 (subsetting on genes with annotation in
848 leukemia),
849 - Bick et al., 2019[53] and Jaiswal et al., 2014[12],
850 - A curated list of clonal hematopoiesis-associated mutations compiled from five large
851 studies[12,13,17,54,55].

852 Mutations identified by Strelka and targeting any of these genes were looked up in ClinVar[56]
853 (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf, downloaded on March 31st, 2023) and annotated
854 accordingly. Using the R package drawProteins[57] v1.18.0 we then collected information on

855 the protein domain targeted by each variant. Moreover, we manually annotated variant
856 information (involvement in disease, functional evidence for mutations at this site, SNPID,
857 association with genetic discorders and whether the variant targets a functional domain)
858 using the manually curated variant information "homo_sapiens_variation.txt" from Uniprot[58]
859 (downloaded on April 4th, 2023), variant information provided on www.uniprot.org, and variant
860 information on www.omim.org[59]. Thereafter, we computed SIFT[60] and Revel[61] prediction
861 scores for each substitution. We kept variants

862 - causing a "stopgain", "stoploss", "frameshift_insertion", or "frameshift_deletion", or
863 - with "Conflicting interpretations of pathogenicity", "Likely pathogenic",
864 "Pathogenic/Likely pathogenic" or of "uncertain significance" according to ClinVar, or
865 - annotated to a related disease (based on uniprot.org or omim.org) or
866 - annotated as pathogenic (uniprot.org) or
867 - targeting a "binding site" or a "DNA binding" site (uniprot.org) or
868 - with a REVEL score ≥0.75 or
869 - targeting either *ASXL1*, or *DNMT3A* or *TET2*.

870 This yielded a set of variant positions, based on which we recalled variants in all blood and
871 control samples with bcftools mpileup[46] v 1.10.2, using the option -p, and bcftools call, using
872 the option -mA. In the final set, we kept variants targeting any of *ASXL1*, *DNMT3A* or *TET2*
873 and variants that were found on at least 2 reads in the sample and on <5 reads in the
874 controls.

875 **Analysis of single-cell WGS data**

876 We tested our population genetics model on published single-cell WGS data from refs.[2,22,26].
877 To this end, we computed VAFs from the single-cell phylogenies as

878 $$VAF = \frac{n_{\text{variant cells}}}{2n_{\text{cells}}}$$

879 where $n_{\text{variant cells}}$ is the number of sequenced cells harboring a variant n_{cells} is the total number
880 of sequenced cells; the factor 2 accounts for diploidy.

881 *Parameter estimation*
882 We fit our population genetics model to the cumulative VAF distribution truncated at 0.01, as
883 detailed below and using the prior probabilities outlined in Supplementary Table 7.

884 *Re-analysis of SSNV and indels*

885 To exclude potential differences between our data and the published data that may arise due
886 to differences in the variant calling pipelines, we re-called SSNVs and indels in the data from

887 ref.[2]. In particular, we used the intersection between Mutect2[43] and Strelka2[44], filtered
888 remaining germline variants using gnomAD (retaining variants with AF <0.01) and re-
889 calculated VAFs directly from the BAM files using alleleCounter v.4.0.2
890 (https://github.com/cancerit/alleleCount; with default base and mapping quality thresholds)
891 and MaC (https://github.com/nansari-pour/MaC). We filtered the remaining variants as stated
892 in ref.[2]. In brief, we removed variants occurring in >120 of the 140 colonies, variants that fell
893 within 10bp of each other and variants with a coverage <6 on autosomes or <3 on sex
894 chromosomes in >5 samples. Moreover, we retained only variants with a mean VAF > 0.3
895 across all samples with at least one mutant read. Finally, we excluded sites at which >10% of
896 the samples with at least one mutant read had a VAF<0.1.

897 *Phylogenetic reconstruction*

898 We reconstructed single-cell phylogenies based on the re-called SSNVs and Indels from
899 Lee-Six et al.[2] by converting the mutation table into a fasta file, learning the phylogenetic tree
900 using MPBoot v1.1.0[62] and plotting the tree using custom scripts from[26]
901 (https://github.com/emily-mitchell/normal_haematopoiesis/).

902 **Population genetics model**

903 **i. Theory**

904 We modeled the evolution of variant allele frequencies mechanistically, accounting for
905 accumulation, drift and selection of somatic variants in a homeostatic tissue. The model is
906 parametrized by the rate at which HSCs divide during adulthood, λ, the number of SSNVs
907 acquired between two cell divisions, μ, the number of HSCs during adulthood, N_{ss}, as well as
908 the time of origin of the selected clone, t_{s}, and its selective advantage expressed by r. We
909 bundled the model functions in an R package, SCIFER, available on
910 https://github.com/VerenaK90/SCIFER/tree/paper.

911 *Modeling the site frequency spectrum of somatic variants generated by neutral evolution.*

912 The time-dependent site frequency spectrum $S_i(t)$ gives the number of variants with clone
913 size i, where i ranges from some minimally observable clone size (owing to WGS
914 sequencing depth) up to the total number of stem cells $N(t)$. We derive analytical
915 expressions for the site frequency spectra resulting from neutral evolution and clonal
916 selection and compare these with measured VAF histograms.

917 To begin with neutral evolution, we develop a stochastic model for accumulation and drift of
918 neutral somatic variants during developmental expansion and subsequent homeostasis of
919 the stem cell pool. The model assumes that stem cells proliferate via symmetric self-
920 renewing divisions, with rate λ, and are lost by differentiation and cell death, with rate δ.
921 Between two subsequent stem cell divisions, on average μ new variants are introduced into

each daughter cell. These variants are inherited to daughter cells, and, depending on the dynamics of the corresponding stem cell clone, may either go extinct, or drift to variable frequencies. The site frequency spectrum generated by these dynamics is

$$S_i(t) = \int_0^t \lambda\mu N(t')p_{1,i}(t,t')\mathrm{d}t',$$

describing the generation of new variants at time t', in a population of size $N(t')$, and their drift to clone size i up to the time of measurement t with probability $p_{1,i}(t,t')$. To compare $S_i(t)$ with measured VAF histograms, we transform from clone size to VAF:

$$VAF = i/2N\ .$$

We will see below that the transformed site frequency spectrum depends on the parameters of the birth-death process and mutation acquisition as follows:

$$S_i(t;\lambda,\delta,N,\lambda\mu N)$$

Proliferation and loss rate, as well as the total cell number, appear as individual parameters algebraically and can in principle be inferred from the data as individual parameters. Hence also the mutation rate, appearing in the product $\lambda\mu N$, can be inferred. Practical identifiability will depend on the sequencing depth (Fig. 2c). Thus our approach will aim to qualitatively distinguish between neutral evolution and selection in experimental data and determine the underlying parameters quantitatively.

To describe developmental expansion of the stem cell pool followed by homeostasis, we concatenated linear birth-death processes. During development, the division rate will exceed the loss rate, $\lambda_{\text{exp}} > \delta_{\text{exp}}$, hence defining a supercritical birth-death process. At time t_1, the system reaches its steady-state with a constant number of active stem cells, N_{ss}. The cellular dynamics are now appropriately described by a critical birth-death process with steady-state rate $\lambda_{\text{ss}} = \delta_{\text{ss}}$.

To compute the site frequency spectrum during developmental expansion, we consider the probability that a cell acquiring a new variant will expand to a clone of size a within time t:[63]

$$P_{\text{exp},1,a}(t) = \begin{cases} x(t), \text{ if } a = 0 \\ \big(1 - x(t)\big)\big(1 - y(t)\big)y(t)^{a-1}, \text{ if } a \geq 1' \end{cases} \quad (1)$$

with

$$\boldsymbol{x(t)} = \frac{\boldsymbol{\delta}_{\text{exp}}\boldsymbol{e}^{(\boldsymbol{\lambda}_{\text{exp}}-\boldsymbol{\delta}_{\text{exp}})\boldsymbol{t}} - \boldsymbol{\delta}_{\text{exp}}}{\boldsymbol{\lambda}_{\text{exp}}\boldsymbol{e}^{(\boldsymbol{\lambda}_{\text{exp}}-\boldsymbol{\delta}_{\text{exp}})\boldsymbol{t}} - \boldsymbol{\delta}_{\text{exp}}}, \ \boldsymbol{y(t)} = \frac{\boldsymbol{\lambda}_{\text{exp}}\boldsymbol{e}^{(\boldsymbol{\lambda}_{\text{exp}}-\boldsymbol{\delta}_{\text{exp}})\boldsymbol{t}} - \boldsymbol{\lambda}_{\text{exp}}}{\boldsymbol{\lambda}_{\text{exp}}\boldsymbol{e}^{(\boldsymbol{\lambda}_{\text{exp}}-\boldsymbol{\delta}_{\text{exp}})\boldsymbol{t}} - \boldsymbol{\delta}_{\text{exp}}}. \quad (2)$$

The measured VAF histograms report on the number of variants with a given frequency. To calculate variant number in the model, we note that between time t' and $t' + \mathrm{d}t'$ on average $\mu\, \lambda_{\text{exp}} N(t')\mathrm{d}t'$ variants are generated. Hence,

$$\mu\, \lambda_{\text{exp}} N(t')\mathrm{d}t' \times P_{\text{exp},1,a}(t - t') \quad (3)$$

variants introduced at time t' each occur in a cells at time t.

Next, we consider the fate of these variants during tissue homeostasis. Then stem cell division and loss will be balanced, occurring both at steady-state rate λ_{ss}. Accordingly, the drift of already existing variants is described by the critical birth-death process. Recall that t_1 is the time at which the homeostatic stem cell number is reached. A variant in a cells at t_1 will drift to occur in b cells within time t with probability[63]

$$P_{\text{ss},a,b}(t) = \begin{cases} p(t)^a, \text{if } b = 0 \\ \sum_{j=0}^{\text{b}} \frac{j}{b}\binom{a}{j} p(t)^{a-j}(1 - p(t))^j \binom{b}{j} p(t)^{b-j}(1 - p(t))^j, \text{ else,} \end{cases} \quad (4)$$

With

$$p(t) = \frac{\lambda_{\text{ss}} t}{1 + \lambda_{\text{ss}} t}. \quad (5)$$

Thus, at $t \geq t_1$, the number of variants occurring exactly in *b* cells is

$$\sum_{a=1}^{N_{\text{ss}}} \mu \lambda_{\text{exp}} N(t') P_{\text{exp},1,a}(t_1 - t')\, P_{\text{ss},a,b}(t - t_1)\ \mathrm{d}t' \quad (6)$$

for variants generated during developmental expansion ($t' < t_1$).

Finally, we consider variants acquired during homeostasis, which evolve according to the critical birth-death process entirely. Hence, the number of such variants occurring exactly in *b* cells is

$$\mu\lambda_{\text{ss}}N_{\text{ss}} \times P_{\text{ss},1,b}(t-t')\,\mathrm{d}t' \tag{7}$$

where $t' \geq t_1$. Combining the contribution of variants originating in expansion or homeostasis, we arrive at the site frequency spectrum of neutral variants in a homeostatic tissue without selection:

$$S_i(t) = \underbrace{\int_0^{t_1} \sum_{a=1}^{N_{\text{ss}}} \mu\lambda_{\text{exp}} N(t') P_{\text{exp},1,a}(t_1 - t')\, P_{\text{ss},a,i}(t-t_1)\,\mathrm{d}t'}_{\text{variants generated in development}} + \underbrace{\int_{t_1}^{t} \mu\lambda_{\text{ss}} N_{\text{ss}} P_{\text{ss},1,i}(t-t')\,\mathrm{d}t'}_{\text{variants generated in homeostasis}}, \tag{8}$$

Eq. (8) generalizes a result by Ohtsuki and Innan[64] for expanding tissues.

Modeling the site frequency spectrum of somatic variants under selection

Clonal selection will modify the site frequency spectrum. Consider that a positively selected mutation is acquired at time t_{s} and reduces the rate of stem cell loss by the factor $r < 1$ (the alternative for imparting selective advantage, an increase in the rate of stem cell division, yields very similar results). The cell number of mutated stem cells, $C(t)$, expands at the expense of the number of normal stem cells, $U(t) = N_{\text{ss}} - C(t)$, according to the competition model

$$\begin{aligned} \frac{dU}{dt} &= \lambda_{\text{ss}} U(t)\left(1 - \rho_{U,U}\frac{U(t)}{N_{\text{ss}}} - \rho_{U,C}\frac{C(t)}{N_{\text{ss}}}\right) = g(U,C), \\ \frac{dC}{dt} &= \lambda_{\text{ss}} C(t)\left(1 - \rho_{C,C}\frac{C(t)}{N_{\text{ss}}} - \rho_{C,U}\frac{U(t)}{N_{\text{ss}}}\right) = h(U,C). \end{aligned} \tag{9}$$

The ρ parameters denote phenomenological competition coefficients between and within the mutant clone and the normal stem cells, maintaining homeostasis. We have no further growth if either normal or mutated stem cells fill the entire compartment, $g(N_{\text{ss}},0) = 0 = h(0,N_{\text{ss}})$ and $g(0,N_{\text{ss}}) = 0 = h(N_{\text{ss}},0)$, implying that $\rho_{U,U} = 1 = \rho_{C,C}$. Moreover, $\rho_{C,U} = r$ and hence

$$\frac{dU}{dt} = \lambda_{\text{ss}} U(t) \left(1 - \frac{U(t)}{N_{\text{ss}}} - (2-r)\frac{C(t)}{N_{\text{ss}}}\right), \tag{10}$$

$$\frac{dC}{dt} = \lambda_{\text{ss}} C(t) \left(1 - \frac{C(t)}{N_{\text{ss}}} - r\frac{U(t)}{N_{\text{ss}}}\right). \tag{11}$$

The number of mutant stem cells $C(t)$ will remain much smaller than the number of normal stem cells for extended periods of time, and hence we approximate Eq. (11) by

$$\frac{dC}{dt} = \lambda_{\text{ss}} C(t)(1-r). \tag{12}$$

Eqs (10) and (12) will be used when computing the site frequency spectrum.

With clonal selection, the site frequency spectrum has three principal contributions for somatic variants originating (i) before the positively selected mutation occurred, $S_{i,1}$, (ii) after this mutation occurred and happening in the mutant clone, $S_{i,2}$, and (iii) after the mutation occurred but happening in normal stem cells, $S_{i,3}$:

$$S_i(t) = S_{i,1}(t) + S_{i,2}(t) + S_{i,3}(t) \tag{13}$$

We specify each contribution in turn. The site frequency spectrum of variants in the mutant clone, $S_{i,2}$, is

$$S_{i,2}(t) = \int_{t_{\text{S}}}^{t} \mu\lambda_{\text{ss}} C(t') P_{\text{exp},1,i}(t - t' | \lambda_{\text{ss}}, r\lambda_{\text{ss}})\, \mathrm{d}t', \tag{14}$$

where $P_{\text{exp},1,i}$ is given by Eq. (1).

The site frequency spectra of variants acquired in normal stem cells, $S_{i,1}$ and $S_{i,3}$, are shaped by the decline of the number of normal stem cells after the emergence of the selected clone at t_{S} (Eq. 10). We approximate this process by a subcritical birth-death process with division rate λ_{ss} and effective loss rate $\delta_{\text{eff}}(t)$. The loss rate is chosen such that the expectation of the decline of normal stem cells in the birth-death process equals the number of normal stem cells lost by the competition dynamics, D. (Eq. 10) implies that

$$D = \int_{t_{\text{S}}}^{t} \lambda_{\text{ss}} \left(\frac{U(t)}{N_{\text{ss}}} + (2-r)\frac{C(t)}{N_{\text{ss}}}\right) U(t)\mathrm{d}t. \tag{15}$$

Therefore, the effective loss rate $\delta_{\text{eff}}(t)$ is defined by

$$\int_{t_S}^{t} \delta_{\text{eff}} N_{\text{ss}} \exp\big((\lambda_{\text{ss}} - \delta_{\text{eff}})t\big) \mathrm{d}t = D. \tag{16}$$

The site frequency spectrum of variants acquired in normal stem cells after t_S, $S_{i,3}$, is the superposition of $(N_{\text{ss}} - 1)$ independent linear subcritical birth-death processes and is given by

$$S_{i,3}(t) = (N_{\text{ss}} - 1) \int_{t_S}^{t} \mu \lambda_{\text{ss}} e^{(\lambda_{\text{ss}} - \delta_{\text{eff}})(t' - t_s)} P_{\text{exp},1,i}(t - t' | \lambda_{\text{ss}}, \delta_{\text{eff}})\, \mathrm{d}t'. \tag{17}$$

Finally, to compute the site frequency spectrum of variants acquired prior to the generation time of the selected mutation t_S, $S_{i,1}$, we denote by k the size of the clone harbouring a particular variant at t_s. With probability k/N_{ss}, the founder cell of the mutant clone will have this variant (as well as $k - 1$ normal stem cells at t_S); with probability $1 - k/N_{\text{ss}}$, the variant will not be in the founder cells of the mutant clone (as well as k normal stem cells will harbor it at t_S). Noticing that clones of size $i < C(t)$ can only contain normal stem cells, we find that

$$S_{i,1}(t) = \begin{cases} \displaystyle\sum_{k=1}^{N_{\text{ss}}} S_k(t_s) \left[\frac{k}{N_{\text{ss}}} P_{\text{exp},k-1,i-C(t)}(t - t_s | \lambda_{\text{ss}}, \delta_{\text{eff}}) + \left(1 - \frac{k}{N_{\text{ss}}}\right) P_{\text{exp},k,i}(t - t_s | \lambda_{\text{ss}}, \delta_{\text{eff}}) \right], i \geq C(t), \\ \displaystyle\sum_{k=1}^{N_{\text{ss}}} S_k(t_s) P_{\text{exp},k,i}(t - t_s | \lambda_{\text{ss}}, \delta_{\text{eff}}), i < C(t), \end{cases} \tag{18}$$

where $S_k(t_s)$ is the site frequency spectrum at t_s and $P_{\text{exp},a,b}$ is the clone size distribution generated by a subcritical birth-death process initiated by *a* cells:[63]

$$P_{\text{exp},a,b}(t) = \begin{cases} x(t)^a, & \text{if } b = 0, \\ \displaystyle\sum_{j=0}^{\min(a,b)} \binom{a}{j} \binom{a+b-j-1}{a-1} x(t)^{a-j} y(t)^{b-j} \big(1 - x(t) - y(t)\big)^j, & \text{if } b \geq 1. \end{cases} \tag{19}$$

Thus, we have specified the contributions to the site frequency spectrum of a stem cell population containing a selected clone, $S_i(t) = S_{i,1}(t) + S_{i,2}(t) + S_{i,3}(t)$.

ii. Parameter estimation

Parameter estimates were obtained for each donor separately, using approximate Bayesian computation as implemented in pyABC[65] (posterior sample size of 1000 and termination criterion $\varepsilon_{\text{min}} = 0.05$). We performed the following steps to compare the cumulative number of variants, $M_i(t) = \sum_{j=i}^{N_{\text{ss}}} S_{i,1}(t)$, between measured and simulated data:

1. [Experimental data] Determine $M_i(t)$ for the experimental data. For bulk WGS data, include variants if they are supported by at least 3 reads, absent in the germline control sample, and if the locus is covered by at least 10 and at most 300 reads. For pseudo-bulk WGS data, include all variants. Denoting the measured variant allele frequencies with VAF, compute the cumulative number of variants, $M_{\text{experimental},f} = \sum n_{\text{VAF}\geq f}$, where n_j denotes the number of variants with VAF given by j. We use bins for f of width 0.01, with $0.05 \leq f \leq 1$ for bulk WGS data (30x and 90x), and $0.01 \leq f \leq 1$ for pseudo-bulk WGS data and bulk WGS data (270x, simulated).
2. [Model] Sample parameters from their prior distributions (Supplementary Table 7).
3. [Model] Simulate the expected cumulative number of variants, $M_{\text{sim},f}(t_a)$, where t_a is the age of the patient (for numerical implementation of the model see **Section iv** below). The bins are as with the experimental data. Depending on the following criterion, based on the prior parameter sample, the cumulative VAF histogram is simulated with either the neutral model or the selection model: If a selected clone could grow above the detection limit with the prior parameter sample, the selection model is used; otherwise the neutral model is used. Formally, if $e^{\lambda_{\text{SS}}^{\text{prior}}(1-r^{\text{prior}})(t_a - t_{\text{S}}^{\text{prior}})} \geq \gamma N_{\text{SS}}^{\text{prior}}$ then the selection model is used. The detection limit depends on the sequencing depth, with $\gamma = 0.05$ for 90x and 30x WGS data, and $\gamma = 0.01$ for pseudo-bulk WGS data and simulated 270x bulk WGS data. Smaller clones are not accessible by the respective sequencing approach and can hence be neglected. Importantly, note that the selection model can return a posterior corresponding to neutral evolution.
4. [Model] Add the number of variants present in the founder cell of the hematopoietic system, $\Delta_{\text{clonal}}^{\text{prior}}$, to $M_{\text{sim},1}$ (corresponding to mutations acquired during early development).
5. [Model] Simulate experimental error of sequencing. To this end, generate the expected site frequency spectrum, $S_{\text{sim}}(f_i) = M_{\text{sim}}(f_i) - M_{\text{sim}}(f_{i+1})$. For bulk WGS data, sample for each simulated variant a sequencing coverage c according to $Pois(\hat{c})$, $\hat{c} \in \{30, 90, 270\}$, depending on the sequencing depth. Thereafter, sample VAFs for each variant according to $B\left(\frac{f}{2}, c\right)/c$, where B denotes the Binomial

distribution. Discard variants supported by less than 3 reads. For pseudo-bulk WGS data, sample VAFs for each variant according to $B(f\ n_{\text{cells}})/(2n_{\text{cells}})$, where n_{cells} is the number of sequenced single-cell clones. Compute the sampled cumulative number of variants $M_{\text{sim},f,\text{sampled}}$.

6. [Model vs Experimental data] Determine the distance function for ABC

$$d = \sum_f \left(M_{\text{sim},\ f,\ \text{sampled}} - M_{\text{experimental},\ f}\right)^2.$$

iii. Classification of cases as neutrally evolving or as selected

We classified cases as selected if $P\left(\text{selection}\middle|\frac{C(t_a)}{N_{\text{ss}}} \geq 0.1\right) \geq 0.15$, that is if the selected clone has at least 5% VAF and the corresponding posterior probability was at least 15%. These thresholds were defined based on in silico generated test data, and correspond to the WGS depth of 90x (Figure 2). Upon sample classification, we computed the 80% highest density intervals for each parameter on the parameter subsets supporting neutral evolution or clonal selection, respectively.

iv. Numerical computation of the cumulative number of variant allele frequencies

To facilitate numerical computation of the cumulative number of variants, $M_i(t)$, in particular if N becomes large, we implemented numerical approximations that are detailed in the following.

Drift of somatic variants acquired during expansion

First, we assessed the site frequency spectrum after exponential expansion, $M_{i,\text{exp}}(t_1)$. Recall that t_1 denotes the time point of transition between expansion and homeostasis. Accordingly,

$$M_{i,\text{exp}}(t_1) = \sum_{j=i}^{N_{\text{ss}}} S_{i,1}(t_1) = \int_0^{t_1} \mu\lambda_{\text{exp}} N(t') \sum_{j=i}^{N_{\text{ss}}} P_{\text{exp},1,j}(t_1 - t')\, dt'. \quad (20)$$

As N_{ss} is large, we approximate the sum in Eq. (20) by integration, yielding

$$M_{i,\text{exp}}(t_1) \approx \int_0^{t_1} \mu\lambda_{\text{exp}} N(t') \frac{P_{\text{exp},1,N_{\text{SS}}}(t_1-t') - P_{\text{exp},1,i}(t_1-t')}{\log y(t_1-t')} dt'. \quad (21)$$

The full site frequency spectrum is obtained by evaluating Eq. (21) for every clone size i with $1 \leq i \leq N_{\text{ss}}$. However, for large N_{ss}, this becomes numerically costly (the upper bound of the prior for N_{ss} is 10^8). Hence, we evaluate $M_{h_j,\text{exp}}(t_1)$ according to Eq. (21) for a series of bins

with logarithmic bin sizes for small clones (<10%), and constant bin sizes of 5% for clones > 10%, as follows:

$$\{h_l\} = \{u_l\} \cup \{v_l\}$$
$$\{u_l\}_{l=0,\ldots,\frac{(\log_{10} N_{\mathrm{ss}})-1}{0.05}} = 10^{0.05l}$$
$$\{v_l\}_{l=0,\ldots,17} = N_{\mathrm{ss}}(0.105 + 0.05l). \tag{22}$$

We determined a binned site frequency spectrum, $S_{h_j}(t_1) = M_{h_j}(t_1) - M_{h_{j+1}}(t_1)$, at the end of the expansion phase. Then the cumulative site frequency spectrum of variants acquired during expansion at a later time point $t_2 > t_1$, is obtained as:

$$M_{i,\exp}(t_2) \approx \frac{\sum_{h_j} w_{h_j} S_{h_j}(t_1) \sum_{k=i}^{N_{\mathrm{ss}}} P_{\mathrm{ss},h_j,k}(t_2-t_1) + S_{h_{j+1}}(t_1) \sum_{k=i}^{N_{\mathrm{ss}}} P_{\mathrm{ss},h_{j+1},k}(t_2-t_1)}{1+S_{h_j}(t_1)/S_{h_{j+1}}(t_1)}. \tag{23}$$

To further facilitate computation, we approximated $P_{\mathrm{ss},h_j,k}$ with a Γ-distribution for clone sizes $k > 20$. The Γ-distribution was parametrized by mean and variance, given by $m(t) = h_j$ and $\sigma^2(t) = 2h_j\lambda_{\mathrm{ss}}t$, respectively.[63]

Drift of somatic variants acquired during homeostasis

We next computed the site frequency spectrum of variants acquired during homeostasis, $M_{i,\mathrm{ss}}(t_2)$ (where 'ss' labels variants acquired during steady state), by evaluating

$$M_{i,\mathrm{ss}}(t_2) = \int_0^{t_2-t_1} \mu\lambda_{\mathrm{ss}} N_{\mathrm{ss}} \sum_{j=i}^{N_{\mathrm{ss}}} P_{\mathrm{ss},1,j}(t-t')\, dt'. \tag{24}$$

We approximated $P_{\mathrm{ss},1,j}$ with a Γ-distribution if $p(t)(1-p(t)) \geq 9 \wedge jp(t)(1-p(t)) \geq 9$. The Γ-distribution was parametrized by mean and variance, given by $m(t) = 1$ and $\sigma^2(t) = 2\lambda_{\mathrm{ss}}t$, respectively.[63] Moreover, if $i \geq 100$, we approximated the sum in Eq. (24) with integration, yielding

$$M_{i,\mathrm{ss}}(t_2) \approx \int_0^{t_2-t_1} \mu\lambda_{\mathrm{ss}} N_{\mathrm{ss}} \frac{\left(p(t')^{2N_{\mathrm{ss}}-1} - p(t')^i\right)\left(1-p(t')\right)^2}{\log p(t')}\, dt'. \tag{25}$$

Finally, we computed the full site frequency spectrum generated by drift at t_2 by summing up the variants contributed from both phases, exponential expansion (Eq. 21) and homeostasis (Eq. 25):

$$M_i(t_2) = M_{i,\text{exp}}(t_2) + M_{i,\text{ss}}(t_2; t_1). \tag{26}$$

Frequency of somatic variants under selection

To model the site frequency spectrum of variants under selection, we first assessed the cumulative site frequency spectrum at t_s, $M_{h_j}(t_s) = M_{h_j,\text{exp}}(t_s) + M_{h_j,\text{ss}}(t_s - t_1)$ using the numerical implementation outlined above. To this end, we defined the sequence h_j, running from clones of size 1 to N_{ss} as follows:

$$\begin{aligned}
&\{h_l\} = \{u\} \cup \{v_l\} \cup \{w_l\} \\
&\{u\} = (1, 10, 25, 50, 75, 100) \\
&\{v_l\}_{l=0,\ldots,\frac{(\log_{10} N_{\text{ss}})-3}{0.05}} = 10^{2+0.05l} \\
&\{w_l\}_{l=0,\ldots,17} = N_{\text{ss}}(0.105 + 0.05l).
\end{aligned} \tag{27}$$

Note that we here resolved small clones at high resolution, because they may reach large frequencies due to selection. From the cumulative site frequency spectrum, we determined the histogram by computing $S_{h_j}(t_s) \approx M_{h_j}(t_s) - M_{h_{j+1}}(t_s)$. We then computed the drift of these mutations during expansion of the selected clone at both borders of each bin, and took a weighted average:

$$M_{i,\text{unmutated cells},1}(t_2)$$

$$\approx \sum_{k=i}^{N_{\text{ss}}} \begin{cases} \sum_{h_j} \frac{1}{1+w_{h_j}} \begin{cases} w_{h_j} S_{h_j}(t_s) \begin{bmatrix} \frac{h_j}{N_{\text{ss}}} P_{\text{exp},h_j-1,k-C(t_2)}(t_2 - t_s|\lambda_{\text{ss}}, \delta_{\text{eff}}) + \\ \left(1 - \frac{h_j}{N_{\text{ss}}}\right) P_{\text{exp},h_j,k}(t_2 - t_s|\lambda_{\text{ss}}, \delta_{\text{eff}}) \end{bmatrix} + \\ S_{h_{j+1}}(t_s) \begin{bmatrix} \frac{h_{j+1}}{N_{\text{ss}}} P_{\text{exp},h_{j+1}-1,k-C(t_2)}(t_2 - t_s|\lambda_{\text{ss}}, \delta_{\text{eff}}) + \\ \left(1 - \frac{h_{j+1}}{N_{\text{ss}}}\right) P_{\text{exp},h_{j+1},k}(t_2 - t_s|\lambda_{\text{ss}}, \delta_{\text{eff}}) \end{bmatrix} \end{cases}, k \geq C(t_2) \\ \sum_{h_j} \frac{1}{1+w_{h_j}} \begin{Bmatrix} w_{h_j} S_{h_j}(t_s) P_{\text{exp},h_j,k}(t_2 - t_s|\lambda_{\text{ss}}, \delta_{\text{eff}}) + \\ S_{h_{j+1}}(t_s) P_{\text{exp},h_{j+1},k}(t_2 - t_s|\lambda_{\text{ss}}, \delta_{\text{eff}}) \end{Bmatrix}, k < C(t_2). \end{cases} . \tag{28}$$

where $w_{h_j} = \frac{S_{h_j}}{S_{h_{j+1}}}$. We further facilitated computation by approximating $P_{\text{exp},h_j,k}$ with a Γ-distribution if $h_j + k > 10$. The Γ-distribution was parametrized by mean and variance of P_{exp},

given by $m(t) = h_j e^{(\lambda_{\text{exp}} - \delta_{\text{eff}})t}$ and $\sigma^2(t) = h_j \frac{\lambda_{\text{exp}} + \delta_{\text{eff}}}{\lambda_{\text{exp}} - \delta_{\text{eff}}} e^{(\lambda_{\text{exp}} - \delta_{\text{eff}})t} \left(e^{(\lambda_{\text{exp}} - \delta_{\text{eff}})t} - 1 \right)$, respectively.[63]

We next computed the site frequency spectrum of variants that were newly acquired during the expansion of the mutant clone by evaluating $M_{i,\text{mutant cells}}(t_2 - t_s | \lambda_{\text{ss}}, r\lambda_{\text{ss}})$ and $M_{i,\text{unmutated cells},2}\,(t_2 - t_s | \lambda_{\text{ss}}, \delta_{\text{eff}})$ according to Eq. (9). As before, we approximated $P_{\text{exp,a,b}}$ with a Γ-distribution, parametrized by mean and variance, if $a + b > 10$.

Finally, we computed the full site frequency spectrum at t_2 by summing up the contributions of variants acquired prior to and during expansion of the CH clone:

$$M_i(t_2) = M_{i,\text{unmutated cells},1}(t_2) + M_{i,\text{mutant cells}}(t_2 - t_s) + (N_{\text{ss}} - 1) M_{i,\text{unmutated cells},2}(t_2 - t_s | \lambda_{\text{ss}}, \delta_{\text{eff}}). \quad (29)$$

Simulation of phylogenetic trees and in-silico evaluation of model performance

Stochastic simulation

To validate the population genetics model, we simulated the accumulation of variants, and the drift and selection of stem cell clones using stochastic birth-death processes. We made the functions used to simulate these processes available as part of the R package SCIFER (https://github.com/VerenaK90/SCIFER/tree/paper). During the initial expansion phase, we neglected cell loss ($\delta_{\text{exp}} = 0$) and, in each simulated time step, randomly selected a cell to divide. Moreover, in order to mimic neutral variant accumulation, we introduced $\mu = 1$ new variant in each daughter cell and memorized the mother-daughter relationship. We abrogated the expansion phase and initiated the homeostatic phase once the population reached a size of $N_{\text{ss}} = 25{,}000$ cells. Assuming that the expansion phase ends early in life, we began to track time when the system transitioned to homeostasis and defined $t = 0$ at this time point.

During the homeostatic phase, we simulated cell division and loss stochastically, using τ-leaping at constant $\tau = 250$ (i.e., summarizing 1% of the reactions occurring in 25,000 stem cells). In each time step, we simulated τ reactions, randomly distributed to births and deaths, associated with equal probability of 0.5 due to tissue homeostasis. We first simulated the birth events by iteratively sampling random cells to divide, while introducing 1 new neutral variant per daughter cell and memorizing the phylogenetic relationships, as before. Thereafter, we simulated the death events by iteratively sampling random cells to be lost from the system. We sampled the duration of each simulated time step, $\mathrm{d}t$, from an

exponential distribution with rate $N(t)2\lambda_{\mathrm{ss}}/\tau$, where we set $\lambda_{\mathrm{ss}} = 10/\mathrm{year}$. We reported the system state at 0, 25, 50 and 75 years and terminated the simulation at 75 years.

To simulate clonal selection, we modified the simulation procedure by introducing a driver mutation with selective advantage $r = 0.98$ to a randomly selected cell at 20 years of simulation time. We stably inherited the selective advantage to the progeny of this founder cell and favored the mutant cells over normal cells for division, while assigning equal probability for them to be selected for death. To be specific, the mutant cells were selected for division with probability $\frac{2-r}{C(t)(1-r)+N(t)}$, whereas normal cells were selected with probability $\frac{1}{C(t)(1-r)+N(t)}$, where $C(t)$ denotes the number of mutant cells at time t. In case the mutant cells got randomly extinct, we re-introduced the driver in the next time step. When simulating selection we used $\tau = 2500$ and reported the system state at 0, 25, 50 and 75 years and, in addition, once the selected clone reached clone sizes of multiples of 5%. As before, we terminated the simulation after 75 years.

Parameter inference

To infer the dynamic parameters, $(\mu,\ \delta_{\mathrm{exp}}, \lambda_{\mathrm{ss}}, N_{\mathrm{ss}}, t_s, r)$, we subsampled 10,000 cells from the simulated trees at time points specified in Supplementary Table 1 and computed the simulated VAF distribution from subsampled trees. To account for technical noise, we simulated sequencing by sampling for each variant a sequencing coverage $c \propto Pois(\hat{c})$, where $\hat{c}$ is the average sequencing coverage and thereafter sampling mutant reads according to $B(VAF, c)$, where B is the binomial distribution. We simulated VAFs for coverages of 30x, 90x and 270x, and fitted the population genetics model to the simulated bulk WGS data as described above for the real data.

Evaluation of sensitivity and specificity

To evaluate sensitivity and specificity of out model, we analyzed the posterior probability of clonal selection. In accordance with the minimal clone sizes used in the inference setup (see above), we computed the probability of clonal selection as $P(\mathrm{selection}) = \frac{\sum_i \frac{C(t_a|\theta_i)}{N_{\mathrm{ss},i}} \geq 0.05}{\sum_i 1}$ for 30x and 90x sequencing depths, and as $P(\mathrm{selection}) = \frac{\sum_i \frac{C(t_a|\theta_i)}{N_{\mathrm{ss},i}} \geq 0.01}{\sum_i 1}$ for 270x sequencing depth, where $C(t_a|\theta_i)$ is the size of the selected clone at the patient age, a, given the *i*-th parameter sample, θ_i, and $N_{\mathrm{ss},i}$ is the i-th estimate of the stem cell number. Among all parameter sets with $\frac{C(t_a|\theta_i)}{N_{\mathrm{ss},i}} \geq 0.05$ (30x and 90x coverage) or $\frac{C(t_a|\theta_i)}{N_{\mathrm{ss},i}} \geq 0.01$ (270x coverage), we determined

median clone size and rounded it with 5% accuracy if it was at least 5% or with 1% accuracy otherwise. Thereafter, we computed sensitivity and specificity of our approach, by varying β between 1% and 100% and classifying cases as selected if $P(\text{selection}) \geq \beta$. For each of the true clone sizes (1%, 2%, 5%, 10%, 15%, 20%, 25%, 50%, 75%; see Supplementary Table 1), we computed the number of true positives as the number of cases where the actual clone size was correctly inferred (where we classified inferred clone sizes between 0.5 and 1.5 of the actual clone size as correct). Conversely, we computed the number of false positives as the number of cases where the respective clone size was falsely inferred whereas the actual clone size was zero.

The difference between true positives and false positives was maximal for $\beta = 15\%$. At this threshold, clones of size ≥5% VAF were reliably inferred.

Statistics and reproducibility

Bayesian parameter inference and statistical analyses were conducted with pyABC[65] v0.12.6, using python v3.10.1 and R (v4.2.0 and v4.2.1). We used the following R packages: ape[66] v5.6-2, phytools[67] v1.2-0, phangorn[68] v2.10.0, castor[69] v1.7.5, TreeTools v1.8.0, deSolve[70] v1.33, openxlsx v4.2.5, cdata v1.2.0, ggpubr v0.4.0, RRphylo[71] v2.7.0, ggplot2[72] v3.4.2, cgwtools v3.3, ggVennDiagram v1.2.2, ggbeeswarm v0.6.0, ggsci v2.9, Hmisc v4.7.1, lemon v0.4.5, data.table v1.14.2, RColorBrewer v1.1.3, ggridges v0.5.4, doParallel v1.0.17, foreach v1.5.2, parallel v2.1, wesanderson v0.3.6, bedr v1.0.7, ggformula v0.10.2, HDInterval v0.2.2, reshape2 v1.4.4[73], dplyr v1.0.9, scales v1.2.1.

Data availability

.bam files have been deposited on EGA (EGAS00001007558). Variant calls and model fits have been made available on Mendeley data (doi: 10.17632/yvxdb7t3yk.1). Patient information and driver mutations are available as Supplementary Data to this manuscript. If excess human DNA used in this study is available it will be made available.

Code availability

Code to reproduce the analysis is available on https://github.com/VerenaK90/Clonal_hematopoiesis/tree/main and https://github.com/VerenaK90/SCIFER/tree/paper.

Acknowledgments

We thank the subjects and allied health care professionals who supported collection of bone marrow and blood samples. We thank Nils Becker and all members of the Höfer and Vyas groups for discussions.

Author contributions

V.K., N.A.J., P.V. and T.H. conceived the project. S.N., B.J.L.K., A.H.T., R.A.L., R.G., B.W., K.W., D.B., S.D. and A.J.C. collected bone marrow and blood samples. N.A.J. performed flow cytometry; N.A.J., M.M., R.M., B.U. and M.S. performed DNA extraction; N.C. prepared DNA sequencing libraries. V.K. and N.A.-P. analysed whole genome sequencing data. F.H. helped with phylogenetic inference. V.K. developed the mathematical model, performed parameter estimation and all computations. V.K., T.H. and P.V. wrote the manuscript with input from all coauthors. T.H. and P.V. acquired funding and administered the project.

Competing interests

The authors declare no competing interests.

Supplementary Table 7. Prior probabilities to model site frequency spectra.

Parameter	Description	Distribution	Lower limit	Upper limit	
				Bulk WGS data / simulated bulk WGS data at 30x or 90x	**Pseudo-bulk WGS / simulated bulk WGS data at 270x**
δ_{exp}	Relative loss during development	uniform	0	0.75	
N_{ss}	Active stem cell count during homeostasis	Log-uniform	$10^{2.5}$	10^{8}	
μ	Number of neutral variants	uniform	0.1	10	

	acquired per division				
λ_{ss}	Division rate during homeostasis	Log-uniform	10^{-3}	10^{-1}	
t_s	Time point of acquiring the driver	uniform	0	$t_a - \frac{\log 0.01 N_{\text{ss}}}{\lambda_{\text{ss}}}$	$t_a - \frac{\log 0.001 N_{\text{ss}}}{\lambda_{\text{ss}}}$
r	Selective advantage associated with driver	uniform	$1 - \frac{\log N_{\text{ss}}}{\lambda_{\text{ss}}(t_a - t_s)}$	$1 - \frac{\log 0.01 N_{\text{ss}}}{\lambda_{\text{ss}}(t_a - t_s)}$	$1 - \frac{\log 0.001 N_{\text{ss}}}{\lambda_{\text{ss}}(t_a - t_s)}$
Δ_{clonal}	Number of neutral variants present in the founder cell of the hematopoietic system	uniform	0	10	

References

1 Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85-90 (2021).

2 Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473-478 (2018).

3 Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).

4 Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538-542 (2019).

5 Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640-646 (2020).

6 Grossmann, S. *et al.* Development, maturation, and maintenance of human prostate inferred from somatic mutations. *Cell Stem Cell* **28**, 1262-1274. e1265 (2021).

7 De, S. Somatic mosaicism in healthy human tissues. *Trends in Genetics* **27**, 217-223 (2011).

8 Vijg, J. & Dong, X. Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* **182**, 12-23 (2020).

9 Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822-825 (2010).

10 Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144 (2010).

11 Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature genetics* **44**, 1179-1181 (2012).

12 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine* **371**, 2488-2498 (2014).

13 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine* **371**, 2477-2487 (2014).

14 Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350-355 (2018).

15 Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine* **20**, 1472-1478 (2014).

16 Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130-135 (2020).

17 Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400-404 (2018).

18 Jaiswal, S. *et al.* Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *New England Journal of Medicine* **377**, 111-121 (2017).

19 Dharan, N. J. *et al.* HIV is associated with an increased risk of age-related clonal hematopoiesis among older adults. *Nature Medicine* **27**, 1006-1011 (2021).

20 Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nature medicine* **27**, 1012-1024 (2021).
21 Heyde, A. *et al.* Increased stem cell proliferation in atherosclerosis accelerates clonal hematopoiesis. *Cell* **184**, 1348-1361. e1322 (2021).
22 Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335-342 (2022).
23 Robertson, N. A. *et al.* Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *Nature medicine* **28**, 1439-1446 (2022).
24 Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449-1454 (2020).
25 Poon, G. Y., Watson, C. J., Fisher, D. S. & Blundell, J. R. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nature Genetics* **53**, 1597-1605 (2021).
26 Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343-350 (2022).
27 Weinstock, J. S. *et al.* Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. *Nature*, 1-11 (2023).
28 Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029-1041. e1021 (2017).
29 Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS computational biology* **12**, e1004731 (2016).
30 Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature genetics* **48**, 238-244 (2016).
31 Körber, V. *et al.* Evolutionary trajectories of IDHWT glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer cell* **35**, 692-704. e612 (2019).
32 Tanner, G., Westhead, D. R., Droop, A. & Stead, L. F. Benchmarking pipelines for subclonal deconvolution of bulk tumour sequencing data. *Nature Communications* **12**, 6396 (2021).
33 Jakobsen, N. A. *et al.* DNMT3A and TET2 mutations lead to distinct and shared perturbations in human clonal hematopoiesis. *in revision with Cell Stem Cell* (2023).
34 Osorio, F. G. *et al.* Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell reports* **25**, 2308-2316. e2304 (2018).
35 Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
36 Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nature genetics* **50**, 895-903 (2018).
37 Körber, V. *et al.* Neuroblastoma arises in early fetal development and its evolutionary duration predicts outcome. *Nature Genetics* **55**, 619-630 (2023).
38 Gabbutt, C. *et al.* Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nature Biotechnology* **40**, 720-730 (2022).
39 Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162-168 (2022).
40 Zeng, A. G. X. *et al.* A hematopoietic stem cell subset that retains memory of prior inflammatory stress accumulates in aging and clonal hematopoiesis. *bioRxiv*, 2023.2009.2011.557271, doi:10.1101/2023.09.11.557271 (2023).

41 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
42 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754-1760 (2009).
43 Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).
44 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods* **15**, 591-594 (2018).
45 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
46 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
47 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164 (2010).
48 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
49 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
50 Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods* **10**, 1081-1082 (2013).
51 Pich, O., Reyes-Salazar, I., Gonzalez-Perez, A. & Lopez-Bigas, N. Discovering the drivers of clonal hematopoiesis. *Nature Communications* **13**, 4267 (2022).
52 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).
53 Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763-768 (2020).
54 Acuna-Hidalgo, R. *et al.* Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *The American Journal of Human Genetics* **101**, 50-64 (2017).
55 Desai, P. *et al.* Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature medicine* **24**, 1015-1023 (2018).
56 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **46**, D1062-D1067 (2018).
57 Brennan, P. drawProteins: a Bioconductor/R package for reproducible and programmatic generation of protein schematics. *F1000Research* **7** (2018).
58 Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506-D515 (2019).
59 Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM. org: leveraging knowledge across phenotype–gene relationships. *Nucleic acids research* **47**, D1038-D1043 (2019).
60 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-3814 (2003).
61 Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).

62 Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC evolutionary biology* **18**, 1-11 (2018).
63 Bailey, N. The elements of stochastic processes John Wiley & Sons. *Inc., New York* (1964).
64 Ohtsuki, H. & Innan, H. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theoretical Population Biology* **117**, 43-50 (2017).
65 Klinger, E., Rickert, D. & Hasenauer, J. pyABC: distributed, likelihood-free inference. *Bioinformatics* **34**, 3591-3593 (2018).
66 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019).
67 Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, 217-223 (2012).
68 Schliep, K., Potts, A. A., Morrison, D. A. & Grimm, G. W. Intertwining phylogenetic trees and networks. Report No. 2167-9843, (PeerJ Preprints, 2016).
69 Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053-1055 (2018).
70 Soetaert, K., Petzoldt, T. & Setzer, R. W. Solving differential equations in R: package deSolve. *Journal of statistical software* **33**, 1-25 (2010).
71 Castiglione, S. *et al.* A new method for testing evolutionary rate variation and shifts in phenotypic evolution. *Methods in Ecology and Evolution* **9**, 974-983 (2018).
72 Villanueva, R. A. M. & Chen, Z. J. (Taylor & Francis, 2019).
73 Wickham, H. Reshaping data with the reshape package. *Journal of statistical software* **21**, 1-20 (2007).