A simple and direct method to define clonal selection in somatic mosaicism

Verena Körber[1,†], Naser Ansari-Pour[2,†], Niels Asger Jakobsen[2], Rachel Moore[2], Nina Claudino[1], Marlen Metzner[2], 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], 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 Center, 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)

One sentence summary: We present an approach to distinguish genetic drift from selection, as subclones arise in somatic tissues, in individuals.

Abstract: Dividing somatic stem cells acquire DNA changes marking different clones. With time, clones can become large, either stochastically through neutral drift, or increased fitness and consequent selection. We present a simple, direct, and general approach that distinguishes between these two processes in normal somatic tissue in individuals. The method relies on single time point whole genome sequencing to study somatic mosaicism as tissues age. Using this method, we show that in human clonal hemopoiesis (CH), clones with CH driver mutations, that comprise a median of 24% of hematopoiesis originate decades before they are detected. They expand, through selection by a median of 26% per year. Overall, there is a 3-fold increased rate of stem cell division and an 8.6-fold increase in active long-term stem cells.

Whole genome sequencing (WGS) has quantified the gradual accumulation of somatic DNA variation in phenotypically normal tissue stem cells throughout development (*1, 2*) and post-natal life (*3-7*). This variation occurs through DNA replication errors, endogenous DNA damage with unfaithful DNA repair, telomere attrition, and movement of mobile repeat elements (*8, 9*). These genetically distinct stem cell clones can also vary epigenetically (*10*) and respond differentially to their environment (*11*). Over time, the heterogeneous stem cell clones vary in size, either stochastically through neutral drift (*12, 13*) when variants do not provide a selective advantage, or through selection when somatic variant(s) or epigenetic change(s) allow a stem cell clone to outcompete other stem cell clones (*4-7*). As a consequence of somatic mosaicism stem cell numbers, and the size of stem cell clones, may alter with the potential to ultimately change tissue function. Therefore, it is important to define when, and how, somatic variants arise in long-lived stem cells, how they modulate stem cell number and stem cell clone size over time, and whether neutral drift and/or selection drives stem cell dynamics.

Hematopoiesis is a highly accessible, informative and well-studied model to answer these questions. Somatic mosaicism in blood stem cells can lead to clonal expansion, called clonal hematopoiesis (CH) (*14-19*). CH increases in prevalence with age (*15-17, 20*) and is associated with a heightened risk of developing blood cancer (*15-17, 21*), excess cardiovascular mortality (*22*) and chronic infection (*23, 24*). The spectrum of CH somatic variants includes single nucleotide variants (SNV) and small insertions/deletions (indels) (*14-17*), as well as large-scale mosaic chromosomal alterations (mCAs) (*18, 19*). Many of the SNVs and indels are in driver genes frequently mutated in preleukemia and leukemia, such as *DNMT3A*, *TET2* and *ASXL1*. Longitudinal studies of blood sampled from subjects with CH have shown that growth rates of expanded hemopoietic clones vary markedly between different driver genes (*20*), and can vary with time (*25*). This is concordant with population genetic modelling methods on large cohorts of subjects, showing varied fitness of different leukemic mutations in mediating clonal expansion (*26*). Reports have also suggested that CH may occur in the absence of known driver mutations (*16*) either through selection of as yet unidentified drivers (*27*), or by neutral drift (*28*). However, distinguishing between selection and drift as a mechanism for clonal expansion remains a challenge. The normalized ratio of non-synonymous to synonymous mutations (dN/dS) has been used to distinguish between selection and drift. A clear example of the value of this approach is in cancer where known driver genes show a dN/dS ratio of up to 4.5 with an average of 2 (*29*). In contrast, in somatic tissues, including hematopoietic tissues, the dN/dS ratios are close to 1 (*29*) supporting the need for alternative methods to distinguish drift from selection in somatic tissues.

To tackle the key questions of distinguishing clonal expansion due to either selection or drift, and to provide a simple method to measure average clonal growth for individual subjects, as opposed to whole cohorts, we developed a new population genetics method that leverages deep whole-genome sequencing from a single time point bone marrow, or blood, sample.

Results

Description of the MARCH normal and clonal hematopoiesis cohort

To develop and experimentally test a mechanistic model for neutral drift and selection in hematopoietic somatic tissue, we studied individuals from a cohort of 195 subjects who had a total hip arthroplasty from whom we collected paired peripheral blood (PB) and bone marrow (BM) samples (fig. S1). The clinical characteristics of the cohort are listed in Data S1. Notably, we excluded subjects receiving medications that could alter blood cell output or with auto-immune or inflammatory arthritis, or hematological conditions. We identified 96 subjects with CH driver mutations using a 347kb panel covering 97 genes (Data S2 and S3). The clinical characteristics of subjects with and without CH driver mutations are summarized in Data S4.

Next, we compared our cohort with a recently described, large CH cohort (Abelson cohort) (*21*). Both cohorts were similar with respect to the incidence of CH (Fig. 1A), the mean variant allele frequency (VAF) of the major CH clone (Fig. 1B) and incidence of a second CH mutation (Fig. 1 C), all as a function of age. This is consistent with the notion that the MARCH cohort is representative of the CH population at large.

Next, using the age demographic and VAFs of CH drivers in the MARCH cohort, and a population genetics-based approach, we modelled the rate of acquiring a driver mutation in one of the three most common mutated genes in CH (*DNMT3A*, *TET2*, *ASXL1*) and determined the kinetics of clonal growth (summarized in Fig. 1D, details in Supplementary Methods). Briefly, the model estimates the rate of growth of a CH clone in the context of a steady state number of blood stem cells (N_{ss}) based on three rate parameters: λ for symmetrical self-renewing normal blood stem cell division, $\lambda\mu_d$ for acquiring a CH driver and λs for symmetrical self-renewing division of stem cell with a CH driver mutation.

By fitting the model to the MARCH cohort (fig. S2A-C) using approximate Bayesian computation (ABC), we estimate that the median rate of acquiring a *DNMT3A*, *TET2* and *ASXL1* driver mutation was once every $3x10^5$ stem cell divisions (80% CI, $3x10^7$ to $2x10^4$) (Fig 1E). The median clonal growth rate/year was a little higher for *DNMT3A*-mutant clones, 9% (80% CI, 3–15%) compared with *TET2*-mutant clones, 5% (80% CI, 2–8%) and *ASXL1*-mutant clones, 4% (80% CI, 2–5%) (Fig. 1F). Conversely, the age at which a CH clone would have reached a VAF of 1% was younger for *DNMT3A*-mutant clones (median, 29 years; 80% CI, 14–45 years) compared with *TET2*-mutant clones (median, 51 years; 80% CI, 29–70 years), with the *ASXL1*-mutant clones reaching this age at an even later time point (median, 67 years; 80% CI, 52–80 years).

Generation and validation of a model of neutral drift in somatic tissue

We studied how neutral somatic variants accumulate in a tissue in the absence of genetic selection. The mathematical theory of population genetics provides a framework to identify genetic drift (*30*)Williams, 2016 #6186} and clonal selection (*31*) in a growing tumor. However, this theory cannot be applied to a normal homeostatic tissue as the growth history differs from that of a tumor. In a non-cancerous tissue, there are two principal phases in which somatic variants are generated; the initial developmental expansion of the stem cell pool and subsequent homeostasis (Fig. 2A). The allele frequency of a given variant (VAF) will be shaped by when the variant arose and for how long it drifted for, before the VAF was measured. To describe these processes, we developed a fully stochastic model of variant accumulation and drift in a developing, and subsequently homeostatic, tissue (Fig. 2B, see Supplementary Methods), thus progressing beyond previous approximate deterministic (*32*), or stochastic (*30*), formulations for expanding tumors.

Our model predicts how the variant burden increases with age (Fig. 2C). Somatic single nucleotide variants (SSNVs) acquired early in life, will be relatively few and have a high VAF. With time, the stem cell pool expands and reaches its homeostatic size. As each stem cell undergoes more cell divisions, it accumulates more variants, but now at lower VAF. When the total number of SSNVs is plotted against 1/VAF (Fig. 2D), in young individuals there is a shallow slope but as more variants accumulate with age, the curve bends upwards. This homeostatic shape is distinct

from the straight line obtained for growing tumors (*32*). Our model shows that the precise shape of the VAF distribution depends on three parameters: the homeostatic number of active HSCs, N; the HSC division rate, λ; and the average number of SSNVs acquired per cell cycle μ (Fig. 2B). We observed that each parameter affects the VAF distribution in a specific manner (fig. S3A-B).

To test if this model is valid, we performed deep WGS (90X on bone marrow mononuclear cells (BMMNCs) and 30X on germline control, hair follicle) on ten normal individuals from the MARCH cohort, aged between 30 and 76 years, who did not have known CH driver mutations. Before feeding SSNVs into the model, we stringently filtered to retain only true SSNVs, even at low VAF (to 5%). First, we used the intersection of two SSNV callers (Strelka and Mutect) (fig. S4A) to remove potential sequencing errors. Second, we discarded germline variants not detected in the control by filtering against known SNPs (gnomAD) and, third, we removed potentially spurious calls of SSNVs in repeat regions and those that were detected in more than one individual. SSNV signatures were consistent with clock-like accumulation of variants (fig. S4B). Finally, to account for false negatives in our model, we determined the frequency of false negative calls by analyzing two individuals where we had WGS data from BMMNCs and peripheral blood mononuclear cells (PBMNCs). At a VAF of greater than 25% both populations were virtually identical implying virtually no false negative calls. Between 5%-25% VAF, there was a proportion of false negative calls (fig. S4C) and we took this rate into account as experimental error in our model (Supplementary Methods). The full list of CH driver variants, and somatic variants, identified by WGS are in Data S5 and S6, respectively.

We then fitted the model to the experimentally derived cumulative number of SSNVs against 1/VAF (Fig. 2E) for each of the ten individuals using ABC (fig. S4D). The increase in SSNV burden with age was consistent with our model of drift (Fig. 2F). The best-fit parameters then allowed us to estimate the number of long-term self-renewing HSC contributing to BMMNC (median 3,000 interquartile range 1,000–7,000, Fig. 2G), their division rate per year (median 1.1 and interquartile range 0.9–1.9, Fig. 2H) and the number of SSNVs/HSC division (median 1.1 and interquartile range of 0.7–1.7, Fig. 2I).

Robust detection of clonal selection from genome-wide somatic variants in individuals

To understand how the selection of a hematopoietic stem cell by a CH driver alters the VAF distribution, we extended our model to allow for selective expansion of a stem cell clone (Fig. 3A) (see Supplementary Methods for the extension of the model). The selected clone is characterized by the time its founding driver CH mutation arises (t_s) and the associated selective advantage (s), which determines its rate of expansion (Fig. 3B). Selection will expand not only the VAF of the founding driver mutation but also of all the other SSNVs in that subclone, causing a shoulder in the plot of cumulative number of SSNVs versus VAF (Fig. 3A) and 1/VAF over time (Fig. 3A and fig. S5A). This shoulder will also vary in size depending on the age of the clone, t_s (fig. S5B) and degree of clonal selective advantage (s) (fig. S5C). Importantly, the existence of the shoulder in the cumulative distribution of the SSNVs versus 1/VAF suffices to distinguish selection, rather than drift as a mechanism of clonal expansion. Finally, the model does not require the identity of the driver mutation to be known.

To test the selection model, we performed 90X WGS on BMMNCs from twenty-one individuals from the MARCH cohort (fig. S1D-G) with driver mutations in genes commonly mutated in CH (*DNMT3A*, *TET2*, *ASXL1*, *PPM1D*, *GNB1*, *KMT2E*). Within the selected samples, 12 had more than one CH driver mutation. In four individuals we also sequenced PBMNCs. Hair follicle DNA was used as germline control. From the MARCH cohort we selected samples that had slightly larger clones: the median VAF of the dominant CH mutation was 12% in the 21 selected cases (range 4.6 to 36.6%) versus 3.2% for the whole MARCH cohort (Data S3). As in the normal samples, the majority of SSNVs were assigned to the clock-like signature SBS5 (fig. S6A). However, CH samples overall had a higher burden of variants at high VAF (≥5%) than normal

samples, at all ages tested (Fig. 3C) and across the spectrum of different CH driver genes with mutations (fig. S6B).

In eighteen out of the twenty-one cases, the data fitted the model of selection as a mechanism for clonal expansion (Fig. 3D-I, fig. S7A). In one case in which the selection was not inferred (CH9, fig. S7A), a *DNMT3A* variant was measured at 1%VAF and hence below the detection limit of our method. The two remaining cases that did not fit the selection model will be discussed later. In fourteen of the eighteen cases that fitted the selection model, there was a visible shoulder in the plot of cumulative SSNVs against 1/VAF; in descending order of selected clone size: CH15, 13, 14 (Fig. 3 D-I) and CH4, CH10, CH 21, CH8, CH5, CH2, CH16, CH18, CH11, CH19, CH17 (fig. 7SA). In four cases the shoulder was not visible but inferred by the posterior distribution of the clone size; CH7, CH6, CH1, CH12. By contrast, the selected CH clone was visible in the conventional (non-cumulative) SSNV VAF distribution only in ten out of 21 cases (CH15, 13, 14; Fig. 3D-I and CH4, CH10, CH21, CH8, CH5, CH2, CH16; fig. S7A) suggesting that the cumulative SSNVs 1/VAF distribution and the model are more sensitive at identifying selection as a mechanism of clonal expansion.

To further validate the model, we used multiple orthogonal approaches. First, for three of the eighteen cases above, CH15, CH14 and CH18, we had WGS data from BMMNCs and peripheral blood (PB) cells. We compared the fit of the model to the data between BM and PB samples in each case (fig. S7B). In two cases, CH15 and CH14, the cumulative 1/VAF distribution of SSNVs in both BM (Fig. 3D and H) and PB (fig. S7B) fitted the model of selection. In one case, CH18, the distribution did not fit the model of selection in PB (fig. S7B) but did in BM (fig. S7A). Notably, the VAF of the driver mutation in *DNMT3A* was 2.1% in PB and 7.2% in BM (Data S5), consistent with the sensitivity of the model to detect selection when the VAFs of selected variants (including driver mutations) was greater than 5%. Second, we compared the inferred clone sizes from the model with two orthogonal experimental methods of determining clone size. Our clone sizes inferred from the population genetics model, without knowledge of CH drivers, agreed with the clone size measured by driver gene VAFs from panel sequencing (Fig. 3J) and with the binomial somatic variant clustering by clonal fraction (*33, 34*) (Fig. 3K), demonstrating the quantitative accuracy of our selection model applied to individual cases.

Inferred hematopoietic stem cell dynamics in the presence of clonal selection

We then compared the rate of acquisition of SSNVs per HSC division in the 19 cases of CH shown in fig. S7A (Fig. 3L) with 10 individuals without CH (Fig. 2I). There were no significant differences, consistent with the notion that CH driver mutations do not increase the underlying mutation rate and the fact that the mutational signatures were similar between CH and non-CH cases. Next, we inferred the number of active long term self-renewing HSCs in CH (median 26,000 inter quartile range 9,000–40,000, Fig. 3M). Though the number was not different across the different driver genes we studied, the number was significantly higher than in individuals without CH (Fig. 3N) and also increased to a greater extent with age (Fig. 3O). We also inferred that the rate of cell division in long term self-renewing blood stem cells in subjects with CH (median 3.1, and interquartile range 2.9–4.2, Fig. 3P) was on average 3-fold higher than in normal subjects (Fig. 3P). Finally, we timed the age of the CHIP clones and found that in all cases, CH clones originated several decades before CH was detected in the bone marrow sample (median 34 years and interquartile range 27–44 years Fig. 3Q). This was further validated by an alternative method of dating clone age via the CpG>TpG molecular clock (Fig. 3R, fig. S8, Data S7). Taken together, in all eighteen cases, the CH driver mutations emerged in mid-life and were associated with large clonal growth (median 26%/year and interquartile growth rate 22%–34%/year, fig. 3S).

Selected variants acquired early in life

In two CH cases, with bona fide driver mutations, where the WGS data did not fit the model of selection, there were small visible shoulders in the cumulative 1/VAF distributions of SSNVs

(CH20, Fig. 4A; CH3, Fig. 4C). In CH20, aged 86 years, a driver *DNMT3A* frameshift mutation (pG890fs) predicted to delete the distal C-terminal helix of the protein and thus very likely causing loss of function (Data S3 and S5), was mapped to a small cluster of 6 SSNVs with high clonal fraction in both BMMNCs and PB cells (Fig. 4B). CH3, aged 62 years, also had a *DNMT3A* frameshift mutation (pW581fs) predicted to be loss-of-function that mapped to a cluster of 12 SSNVs again at high clonal fraction (Fig. 4D). Both mutations and variant clusters defined very large clones but fell short of reaching fixation. In the case of CH20 the VAF of the *DNMT3A* mutation was 45% and 43% by WGS in BMMNCs and PB cells respectively, and 37% by panel sequencing of BMNNCs (Data S3 and S5). In CH3, the VAF of the *DNMT3A* mutation was 28% by WGS in BMMNCs, and 23% by panel sequencing of BMNNCs. Interestingly, by using the CpG>TpG molecular clock timing method the inferred origin of the *DNMT3A* driver mutations was at the age of 1-2 years (CH20) and between the ages of 11-25 years (CH3) (Supplementary Methods, Data S7). Both inferred HSC number and division rate were consistent with the values found for the other CH cases (Fig. 4E and F). Given the estimated median clonal growth for *DNMT3A* CH mutations is 33% per year (interquartile range 24-36% per year; Fig. 3S), the estimated median active HSC number for *DNMT3A* mutant cases is 23,000 (interquartile range 7,000–38,000; Fig 3M) and the age of acquisition of both *DNMT3A* clones in CH20 and CH3, if there had been a constant growth rate the *DNMT3A* clone in CH20 and CH3 should have reached fixation at age 30–51 and at age 40–74 years respectively. Given the ages of the patients (CH20, 72 years; CH3, 63 years), it is notable that neither CH clone reached fixation.

Discussion

We present a new approach that combines stringent detection of SSNVs from WGS with new mathematical models to distinguish clonal expansion that occur by neutral genetic drift versus increased fitness and consequent clonal selection in an individual. Using hematopoietic tissue as an exemplar, we show that individuals with no known driver mutations accumulate somatic variants in an age-dependent fashion consistent with a model of neutral drift, whereas in individuals with known CH driver mutations, the distribution of genome-wide somatic variants demonstrates evidence of clonal selection. By fitting experimental data from 31 human subjects to a population genetics model, this approach allowed us to estimate the number of hematopoietic stem cells that contribute to bone marrow mononuclear cells, the stem cell division rate, the rate of acquisition of SSNVs per stem cell division and in cases of known CH, the growth rate of the major CH subclone.

One key variable in identifying SSNVs that input into the model is the sequencing depth which sets the limit for the clone size detected. By sequencing to 90X, rather than the more common 60X, we reliably detect SSNVs at a threshold of VAF 3-4%. This meant we focused on individuals with larger clones within the MARCH cohort. Our data would suggest that these individuals with larger clones have clones that grow at a faster rate (range 22–34% per year, Fig. 3S) compared with individuals across the whole MARCH cohort (range 2–15%, Fig. 1F). This observation suggests that large CH clones grow particularly fast. Concordant with this, the division rates of long-term self-renewing stem cell are 3-fold higher in individuals with CH compared to those who do not have CH. This marked degree of increase results in a mean 8.6-fold increase in stem cell number in subjects with CH (median 26,000) compared with individuals without CH (median 3,000). This is consistent with evidence from murine models showing that the most common CH driver mutations, loss-of-function mutations in *DNMT3A* and *TET2,* lead to increased self-renewing stem cell divisions (*35*-*37*). Given that previous work has shown that individuals progressing from CH to acute myeloid leukemia (AML) have large CH clones (*21*), our findings on the dynamics of stem cell growth and cell division are particularly relevant to these at-risk individuals. It will also be important to test if deeper WGS, reducing the VAF detection threshold, may enable detection of even smaller clones and give an even richer understanding of clonal dynamics.

We also find that the number of SSNVs acquired per stem cell division is similar in cases with and without CH, and that the mutational signatures are similar in the two groups, providing evidence that most CH is not associated with acquisition of new mutagenic processes. Third, we estimate that most CH clones are acquired two-three decades before the variants in a CH clone reach a VAF of 2% and some cases arise within the first decade of life. In this regard, we note that neither of the two *DNMT3A*-mutant clones detected early in life reached fixation, suggesting that clones may grow variably, and possibly decelerate later in life, as has been suggested recently (*25*).

Our approach has several notable features. First, a method distinguishing neutral drift from selection as a mean of clonal expansion in somatic, non-cancerous tissues. Second, our approach can be applied to a variety of non-hematopoietic tissues where somatic clonal expansion is very common. Third, it provides a cost-effective approach to calculate stem cell numbers and growth rates. Whilst other methods have been recently developed for calculating these parameters, they rely on WGS of hundreds DNA samples from colonies, grown from single cells (*3, 25*), as opposed to WGS from bulk tissue. Fourth, our method infers clonal growth rates without relying on longitudinal sampling, which are often not possible to obtain in non-hematopoietic tissues. Fifth, when selection is identified as a mechanism of subclone expansion, it does not require *a priori* knowledge of the mechanism of selection, for example knowledge of a genetic driver mutation or epigenetic or environmental mechanisms that impart selective advantage. Finally, our approach can be used for individual cases.

Taken together, we provide a general approach to study the quantitative dynamics of stem cell somatic mosaicism. This approach also provides a platform for further iteration to take into account clonal diversity, tissue-specific stem cell dynamics and longitudinal changes that occur from extrinsic perturbation. By performing bulk WGS on a single tissue sample, our work lends itself to uncovering clonal selection and determining clonal growth rate in a personalized medicine setting for early detection of individuals at high risk for developing aggressive malignancy and other phenotypes from rapid clonal growth.

References

1. M. Spencer Chapman *et al.*, Lineage tracing of human development through somatic mutations. *Nature* **595**, 85-90 (2021).
2. T. H. H. Coorens *et al.*, Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387-392 (2021).
3. H. Lee-Six *et al.*, Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473-478 (2018).
4. H. Lee-Six *et al.*, The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).
5. S. F. Brunner *et al.*, Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538-542 (2019).
6. L. Moore *et al.*, The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640-646 (2020).
7. S. Grossmann *et al.*, Development, maturation, and maintenance of human prostate inferred from somatic mutations. *Cell Stem Cell* **28**, 1262-1274 e1265 (2021).
8. S. De, Somatic mosaicism in healthy human tissues. *Trends Genet* **27**, 217-223 (2011).
9. J. Vijg, X. Dong, Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**, 12-23 (2020).
10. R. N. Delgado *et al.*, Maintenance of neural stem cell positional identity by mixed-lineage leukemia 1. *Science* **368**, 48-53 (2020).
11. A. Heyde *et al.*, Increased stem cell proliferation in atherosclerosis accelerates clonal hematopoiesis. *Cell* **184**, 1348-1361 e1322 (2021).
12. C. Lopez-Garcia, A. M. Klein, B. D. Simons, D. J. Winton, Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822-825 (2010).

13. H. J. Snippert *et al.*, Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144 (2010).
14. L. Busque *et al.*, Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet* **44**, 1179-1181 (2012).
15. S. Jaiswal *et al.*, Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498 (2014).
16. G. Genovese *et al.*, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487 (2014).
17. M. Xie *et al.*, Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472-1478 (2014).
18. P. R. Loh *et al.*, Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350-355 (2018).
19. C. Terao *et al.*, Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130-135 (2020).
20. T. McKerrell *et al.*, Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell reports* **10**, 1239-1245 (2015).
21. S. Abelson *et al.*, Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400-404 (2018).
22. S. Jaiswal *et al.*, Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med* **377**, 111-121 (2017).
23. N. J. Dharan *et al.*, HIV is associated with an increased risk of age-related clonal hematopoiesis among older adults. *Nat Med* **27**, 1006-1011 (2021).
24. S. M. Zekavat *et al.*, Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* **27**, 1012-1024 (2021).
25. M. A. Fabre *et al.*, The longitudinal dynamics and natural history of clonal haematopoiesis. *bioRxiv* **doi: https://doi.org/10.1101/2021.08.12.455048**, (2021).
26. C. J. Watson *et al.*, The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449-1454 (2020).
27. G. Y. P. Poon, C. J. Watson, D. S. Fisher, J. R. Blundell, Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nat Genet* **53**, 1597-1605 (2021).
28. F. Zink *et al.*, Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742-752 (2017).
29. I. Martincorena *et al.*, Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021 (2017).
30. H. Ohtsuki, H. Innan, Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theor Popul Biol* **117**, 43-50 (2017).
31. M. J. Williams *et al.*, Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet* **50**, 895-903 (2018).
32. M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, A. Sottoriva, Identification of neutral tumor evolution across cancer types. *Nat Genet* **48**, 238-244 (2016).
33. N. Bolli *et al.*, Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).
34. R. Rabbie *et al.*, Multi-site clonality analysis uncovers pervasive heterogeneity across melanoma metastases. *Nat Commun* **11**, 4306 (2020).
35. G. A. Challen *et al.*, Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* **44**, 23-31 (2012).
36. K. Moran-Crusio *et al.*, Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11-24 (2011).
37. C. Quivoron *et al.*, TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25-38 (2011).

Acknowledgments: We thank the subjects and allied health care professionals who supported the collecting of bone marrow and blood samples.

Funding:
NA-P is supported by the Molecular Diagnostic Theme of the Oxford Biomedical Research Centre

NAJ is supported by the Medical Research Council and Leukaemia UK (Clinical Research Training Fellowship; grant number MR/R002258/1).
AJC is supported by the Wellcome Trust, UKRI/MRC and NIHR.
PV is supported by MRC Molecular Haematology Unit (MC_UU_12009/11), NIHR the Hematology Theme of the Oxford Biomedical Research Centre, Blood Cancer Specialist Program Grant 13001.
TH is supported by the German Federal Ministry for Education and Research (031L0238, INFER-NB), Deutsche Forschungsgemeinschaft (SFB 873, P11) and DKFZ core funding.

Author contributions:
Conceptualization: VK, NA-P, NAJ, PV, TH.
Methodology: VK, NA-P, NAJ, TH.
Sample processing and experimentation: NAJ, NC, MM, RM, BU, MS.
Sample collection: SN, BJLK, AHT, RAL, RG, BW, KW, DB, AJC.
Funding acquisition: TH, PV, NAJ.
Project administration: PV, TH.
Supervision: TH, PV, AJC, SN, NAJ.
Writing – original draft: PV, VK, NA-P, NAJ, TH.
Writing – review & editing: PV, VK, NA-P, NAJ, TH, SN, BJLK, AHT, RAL, RG, AJC.

Competing interests: There are no competing interests.

Data and materials availability: All data, code, and sequence of the variants used in the analysis are available in the Supplementary Methods for purposes of reproducing or extending the analysis.

Note accession numbers to any data relating to the paper and deposited in a public database; include a brief description of the data set or model with the number.

Supplementary Document

This contains Materials and Methods, Supplementary Methods and supplementary tables and Figures S1-S7. Tables containing primary data (Data S1 to S7) are in separate Excel files.

Figure Legends

Figure. 1. Characterization of the MARCH clonal hematopoiesis cohort and population modelling of clonal advantage.

A-C, Comparison of the MARCH cohort and a cohort of 414 individuals reported by Abelson et al., 2018. (**A** and **C**) show the incidence of CH with at least one (**A**) or two (**C**) CH drivers (VAF≥0.01) with age; error bars represent bootstrapped 95% confidence intervals. (**B**), Mean and standard error of the variant allele frequency (VAF) measured for the largest CH clone. The data are binned by decades.

D, Schematic of the population genetics model of CH. The number of hematopoietic stem cells (HSC) at steady state (N_{ss}) divides at rate λ. CH drivers are acquired at rate $\lambda\mu_d$ and confer a selective advantage, s. The model reports the incidence and the mean size of the CH clone.

E-G, Median and 80% credible intervals of the model parameters with (**E**) showing the estimated probability to acquire a CH driver per HSC division; (**F**) illustrates the estimated clonal growth per year as a percentage increase; (**G**) depicts the estimated age of a CH clone when reaching 1% VAF.

Figure 2. Using somatic variants to build and test a model of drift in normal hematopoiesis.

A: Schematic illustrating variant accumulation during development, when tissue growth occurs, and adulthood, when there is tissue homeostasis.

B: Modeled processes and associated parameters in the model of drift. Blood stem cells either undergo symmetrical self-renewing divisions (left) with rate λ or exit the stem cell compartment through differentiation, or death, with rate δ (right). Below, schematic graph of blood stem cell count (N) increasing in development ($N<N_{ss}$ and $\lambda>\delta$) and during adulthood, when $N=N_{ss}$ and $\lambda=\delta$.

C: Schematic illustrating subclonal diversification shaped by variant accumulation (exemplary variants are depicted by a letter) and genetic drift in growing (HSC expansion) and homeostatic tissues (steady state HSC numbers). At the right the genotype of each stem cell is depicted. The bottom panels illustrate the cumulative variant allele frequencies (VAF) distribution of SSNVs through development, and at two stages of adult life. The lower panels show the changes in VAF of variants A-F through drift during development and adulthood.

D: Modelled, expected cumulative 1/VAF distribution of SSNVs at selected ages between birth (0 years) and 75 years. Right, an expanded view of the drift in 1/VAF in early life (0–25 years). The modelling assumes that the hematopoietic stem cell pool consists of 10^3 cells that divide once per year and acquire on average 1 mutation per division (values were based on the fit of the model to the data, see Fig. 2G-I, were the inferred median number of HSCs was ~3000, median division rate ~ 1.1/year and median number of SSNVs/division 1.1).

E: Measured cumulative 1/VAF distribution of SSNVs in 10 individuals, without known CH driver mutations, aged between 30 years and 76 years. Color code indicates age of the individual.

F: Comparison of the number of SSNVs with VAF≥0.05 (y-axis) as a function of age (x-axis) in the model (shaded grey area) with actual data from the 10 individuals shown in **E**. The grey area represents upper and lower limits of the model prediction across the cohort, obtained by evaluating the model with the best parameter set of each patient. The solid line corresponds to the model prediction at the median parameter values across the cohort. Colored dots show the actual data points from the 10 individuals.

G-I, Estimated model parameters (x-axes) for the individuals shown in **E** (y-axis). **G**, the number of active HSCs. **H**, the stem cell division rate per year. **I**, the estimated number of SSNVs acquired per HSC division. Shown are the median and the 80% highest density intervals as blocks. Grey areas represent the 95% confidence interval of the median parameter values across the cohort.

Figure 3. Using somatic variants to build and test a model of selection with drift in clonal hematopoiesis.

A: Schematic illustrating variant accumulation (exemplary variants are depicted by a letter) during development and adulthood (tissue homeostasis). Here subclonal diversification is shaped by both drift and when a driver mutation (D) promotes selection during adult life. The bottom panels illustrate the cumulative variant allele frequencies (VAF) distribution of somatic single nucleotide variants (SSNVs) through development, and at two stages of adult life. The bottom panels show changes in VAF of variants A-F through drift and selection during development and adulthood.

B: Modeled processes and associated parameters in the model of selection and drift. Top, in adult hemostatic tissue, blood stem cells, without a clonal hematopoiesis (CH) driver mutation (blue dots, above) or with a CH driver mutation (below, red dots), undergo symmetrical self-renewing divisions (left) with rate λ. These stem cells exit the stem cell compartment through differentiation, or death, with rate δ for normal stem cells and at a decreased rate δ_s in stem cells with a CH driver mutation (right). Below, schematic graph of blood stem cell count (N), which increases in development ($N<N_{ss}$ and $\lambda>\delta$) and remains constant during adulthood ($N=N_{ss}$ and $\lambda=\delta$). In addition, the model incorporates the acquisition of a CH clone at time t_s.

C: Measured and modeled number of SSNVs with VAF≥0.05 of individuals with (colored points) and without (grey points) clonal hematopoiesis. The color code for the different CH driver mutations is shown on top. The grey area represents upper and lower limits of the model

prediction across the cohort without clonal hematopoiesis, obtained by evaluating the model with the best parameter set of each patient. The solid line corresponds to the model prediction at the median parameter values across the cohort without clonal hematopoiesis.

D-I: Three examples of the cumulative VAF distribution of SSNVs, and respective model fits, in individuals with clonal hematopoiesis (CH15, CH13, and CH14). Red areas show 95% predictive posterior bounds, error bars represent sampling errors; the lines show model simulations at the best fit. Maximum likelihood and 95% confidence interval of the measured VAF of CH drivers are indicated by horizontal lines (**D**, **F**, **H**). For each of these three examples, posterior distributions of the estimated subclone size are shown in **E, G, I** respectively.

J: Plot of the VAF of CH driver mutations inferred by the population genetics model versus that measured by the targeted resequencing panel. Shown are the 80% credible intervals of the model estimates (horizontal lines) and the 95% confidence interval of the measurement (for panel resequencing). The colors correspond to the driver mutations in different genes as set out in panel **C**. Pearson's correlation coefficient is given in the plot.

K: Plot of the VAF of the CH driver mutations inferred by the population genetics model versus that estimated by DPClust. The colors correspond to the driver mutations in different genes as set out in panel **C**. Pearson's correlation coefficient is given in the plot.

L: Estimated number of SSNVs per stem cell division (x-axis) for the different individuals with CH (y-axis).

M: Estimated number of active stem cells (x-axis) for the different individuals with CH (y-axis).

N: Comparison of the number of active stem cells in individuals without (normal) and with CH driver mutation (CH). The difference is statistically significant at a p value of 0.0005, Wilcoxon rank sum exact text.

O: Plot of the number of active stem cells in individuals without (grey dot) and with CH driver mutation (colored dot) shown as a function of age of the subject. The colors of the dots correspond to the driver mutations in different genes as set out in panel **C**.

P: Estimated number of stem cell division rate (x-axis) for the different individuals with CH (y-axis). The colors of the bars correspond to the driver mutations in different genes as set out in panel **C**. The stem division rate is significantly higher in individuals with CH driver mutation as compared to individuals without (p=0.001, Wilcoxon rank sum exact test).

Q: Inferred age of the CH clone estimated clone age by population genetics model normalized for the time point at which the CH clone reached 2% VAF. Shown are median and 80% credible intervals.

R: Comparison of the estimated age of the CH clone by the population genetics model (blue line) and the CpG>TpG molecular clock approach (green line). Shown are median and the 80% credible interval (population genetics model) and 95% confidence intervals (CpG>TpG molecular clock).

S: Left, Estimated percentage clonal growth per year (x-axis) of the major CH clone in different subjects with CH (y-axis). The colors of the bars correspond to the driver mutations in different genes as set out in panel **C**. Right, the percentage clonal growth per year of CH clones with different mutations is shown.

In **L**, **M**, **P**, **Q**, **S**, the median and 80% credible intervals are shown per sample. The 95% confidence interval of the median parameter values across the cohort are visualized by the grey boxes.

Figure 4. Dynamics of clones with CH drivers acquired early in life and demonstrating selection.

A: Measured and modeled number of SSNVs in bone marrow (BM) and peripheral blood (PB) of subject CH20. Top panels show measured VAF distributions of SSNVs. Middle panels show measured (dots) and modeled (black line and red area) 1/VAF distribution of SSNVs with VAF≥0.05. Red areas show 95% predictive posterior bounds, the lines show a model simulation at the best fit, error bars represent sampling errors. Maximum likelihood and 95% confidence interval of the measured VAF of CH drivers are indicated by horizontal lines. The bottom panels show a zoom of the measured 1/VAF distribution at high VAF.

B: Two-dimensional distribution of clonal cell fraction of SSNVs in BM and PB of patient CH20. Red areas show SSNV densities. The *DNMT3A* variant falls into a small cluster of 6 SSNVs at high clonal cell fraction (top right).

C: As in (**A**) but for the BM sample of subject CH3.

D: Distribution of clonal cell fraction of SSNVs in BM of patient CH3. Turquoise area shows SSNV densities. The *DNMT3A* variant falls into a small cluster of 12 SSNVs at high clonal cell fraction.

E and **F:** Estimated number of active stem cells (**E**) and of the stem cell division rate (**F**) (x-axes) for subjects CH3 and CH20 (y-axis). Boxes represent median and 80% credible intervals of the parameters. Grey areas represent the range of the median parameter values across the two subjects.