Modelling the age-related deceleration of clonal haematopoiesis in UK Biobank

Hamish AJ MacGregor[1,2] ✉, **Douglas F Easton**[2], **and Jamie R Blundell**[1] ✉

[1]Early Cancer Institute, University of Cambridge, United Kingdom
[2]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, United Kingdom

Somatic mutations acquired in haematopoietic stem cells can drive clonal expansions, a phenomenon called clonal haematopoiesis (CH). CH has been associated with higher risk of haematological cancer, cardiovascular disease and lower life expectancy and is strongly age-dependent, becoming almost ubiquitous in older adults. However, the dynamics of CH are not yet fully understood, and an accurate quantitative model of clonal expansion in healthy people could inform early detection strategies. Here, we analyse the age-dependence of the variant allele frequency (VAF) distributions of common CH driver variants from the exomes of 400,000 cancer-free UK Biobank participants. We find evidence of a substantial deceleration of clonal expansion with age taking place in younger people. This deceleration appears to be general: different CH hotspot variants all displayed a $\sim$3-fold reduction in growth rate between young and old. We develop several alternative models of clonal dynamics that might explain this age-related deceleration. First we show that variation in fitness of variants among individuals (e.g., due to genetic susceptibility) is unable to produce the required deceleration. Second, we consider a clonal competition model in which clonal expansion in later life is inhibited by competition with an unobserved background of other expanding clones. Finally, we consider an ageing model where HSC division slows down with increasing age. The observed age-specific VAF distributions are consistent with the ageing model with plausible parameters. To further interrogate these models, we consider a large data set of longitudinal growth rates obtained from serial samples of blood. These data suggest that a general slowdown of HSC division rates plays a key role in shaping the genetic diversity of ageing blood.

Correspondence: *jrb75@cam.ac.uk* **and** *hajm2@cam.ac.uk*

Introduction

The acquisition and clonal expansion of somatic mutations is a ubiquitous feature of human ageing, observed across a range of tissues (1–6). In blood, this phenomenon, termed clonal haematopoiesis (CH), causes dramatic changes to the clonal diversity of haematopoietic stem cells (HSCs) in later life and is associated with higher risk of haematological cancers (4, 5), cardiovascular disease (7, 8), and lower life expectancy (4). CH is rare in young adults but the prevalence increases with age and it is observed in almost all adults by age 65 (6)A rapid collapse in clonal diversity is often observed around the age of 70 (9). However, we do not yet have a full quantitative understanding of the factors that shape this age dependence.

A simple model of CH in which driver mutations occur at a constant average rate and expand exponentially has been shown to capture key features of the variant allele frequency (VAF) distributions of clones, enabling one to estimate the fitness effects conferred by key CH-associated genes and mutation 'hotspots' (10, 11). However, longitudinal studies of CH have yielded conflicting estimates of fitness, and initial evidence suggests that some variants may exhibit deceleration in later life (12). The extent to which this deceleration could be caused by differences in the fitness effects between individuals, competition between clones in the HSC pool or explicit age-related slow-down in HSC biology remains unknown.

UK Biobank (UKB) is a cohort of more than 400,000 individuals aged 37-73 at recruitment. Whole Exome Sequencing (WES) of blood samples from the cohort provides an opportunity to characterise CH at individual genomic hotspots across a large, healthy population spanning more than three decades. Here we use UKB to characterise VAF distributions of common CH driver mutations in people of different ages to investigate quantitatively how the dynamics of CH change with age. We observe that there is a fundamental inconsistency between the growth rates required to explain the clones size distribution observed in the youngest age-group in UKB and the growth rates required to explain how the clone size distribution evolves between the youngest and oldest age-groups. By exploring mathematical models of three hypotheses to explain this conundrum we find evidence for a general age-related deceleration of CH that suggests selection levels in the first four decades of life are substantially greater than they are in middle age.

Results

Pervasive age-related deceleration of CH. Somatic clone sizes are shaped by when variants arise and how fast they subsequently grow. The variation in clone sizes observed in CH is consistent with a model in which most CH variants arise at a uniform rate (μ) throughout life and confer an intrinsic fitness effect to cells (s) which remains constant throughout life (9, 10, 12, 13). To test whether this 'constant growth model' can quantitatively explain the age-dependence of CH, we analysed the VAF distributions across 416,171 individuals (aged 37-73) from the UK Biobank exome sequencing dataset. We focused our initial analysis on four common hotspot variants for which sufficient data existed to analyse VAF spectra (DNMT3A R882H: observed in 373 subjects, DNMT3A R882C: n=236, JAK2 V617F: n=145, SF3B1 K700E: n=50, see Supplementary Note 1B).

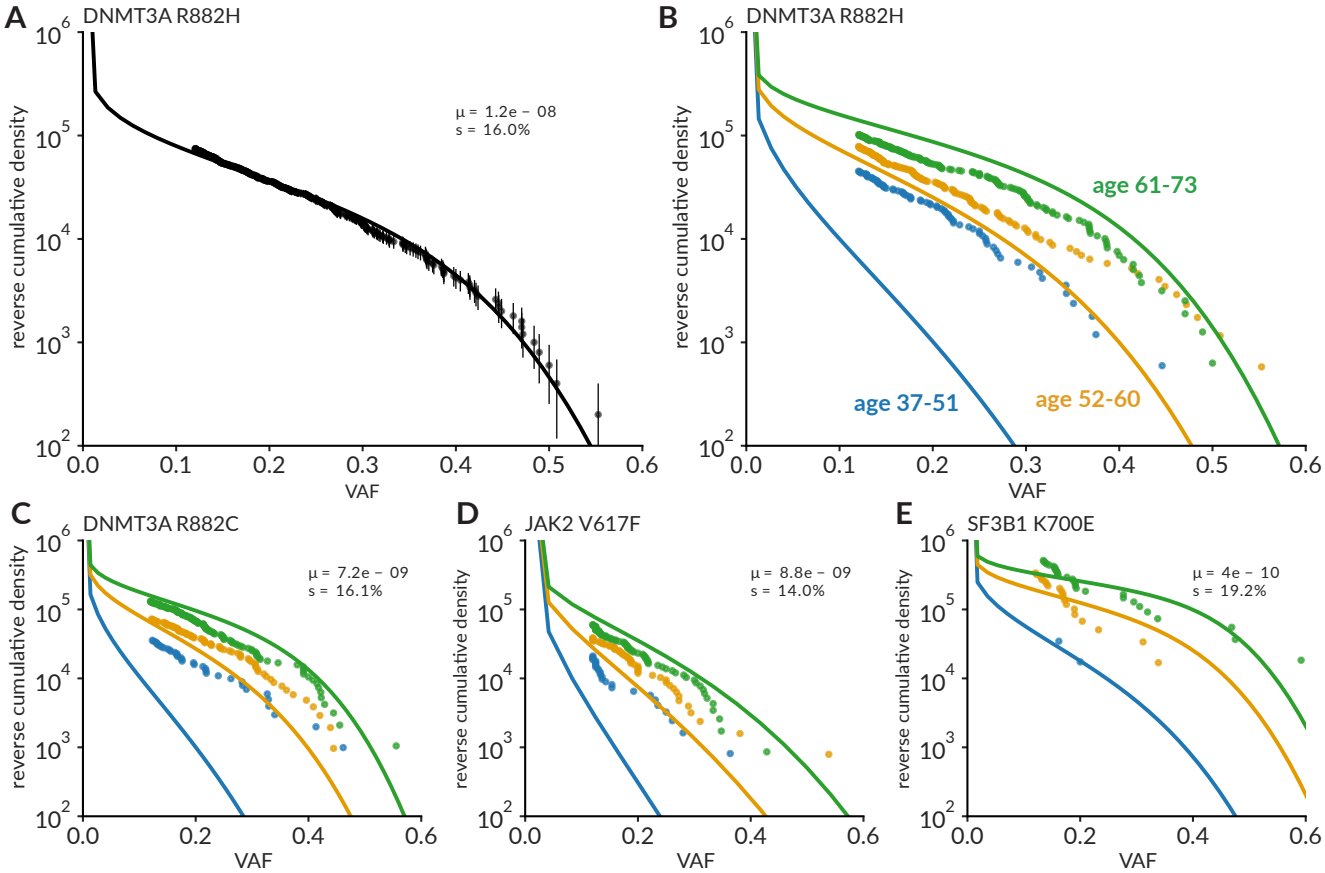

Fig. 1. The age-dependent behaviour of CH variants in UK Biobank is inconsistent with a constant growth-rate model. **(A)** Cumulative density of *DNMT3A* R882H variants in UK Biobank (points) compared to expected density based on the constant growth rate model (line), with mutation rate μ (per cell per year) and fitness effect s (% per year) estimated using least-squares minimisation. Error bars were derived assuming counts were drawn from a Poisson distribution (± 1 s.e.). **(B)** Cumulative densities from (A) stratified into three age tertiles each compared to its respective density predicted by the constant growth model. **(C-E)** The same age-stratified analysis as shown in (B) for other CH hotspot variants: DNMT3A R882C (C), JAK2 V617F (D) and SF3B1 K700E (E).

We applied an evolutionary framework we had previously developed (10) to estimate the fitness effect associated with each specific variant by fitting the model to the VAF distribution observed across all ages combined (Figure 1A). The VAF distribution predicted by the constant growth model is in very close agreement with the data and the fitness effect and mutation rates inferred from this fit are consistent with previous estimates (10, 12). However, stratifying the data into three age tertiles (37-51, 52-60, 61-73) shows that the VAF distributions clearly deviate from those predicted by the constant growth model. In the youngest age group, the fraction of somatic clones with VAF>12% is substantially greater than predicted by the constant growth rate model. In the oldest age group, in contrast, fraction of somatic clones with VAF>12% is underpredicted by the constant growth model. This behaviour is reproduced across all four hot-spot variants studied (Fig. 1B-E). This suggests a general slowdown in the growth rate of clones over a person's lifetime, a finding that has previously been reported for specific variants (11, 12) and observed in the age-dependence of synonymous variants (13).

We next modelled three hypotheses that might explain the weaker-than-expected age dependence of the observed VAF distributions.

1. **Variation in growth rates across individuals.** A number of genome-wide association studies (GWAS) have identified loci associated with both increased and decreased prevalence of CH (14–17). If the fitness effect of a variant varies by the germline genetics of the host individual, this will impact the expected age dependence.

2. **Clonal interference between CH clones.** Previous work has shown that the majority of clonal expansions in CH do not involve alterations in currently known cancer driver genes (9, 13, 18). The presence of other clonal expansions rising to high VAF within the same individual would lead to clonal competition and may slow, or even reverse, some clonal expansions later in life.

3. **Age-related slow down in HSC division.** Rates of clonal expansion are influenced by the symmetric division rates of HSCs (denoted here as $1/\tau$). Some studies have suggested that in adults the product $N\tau$ (total number of HSCs $\times$ symmetric division time) may in-

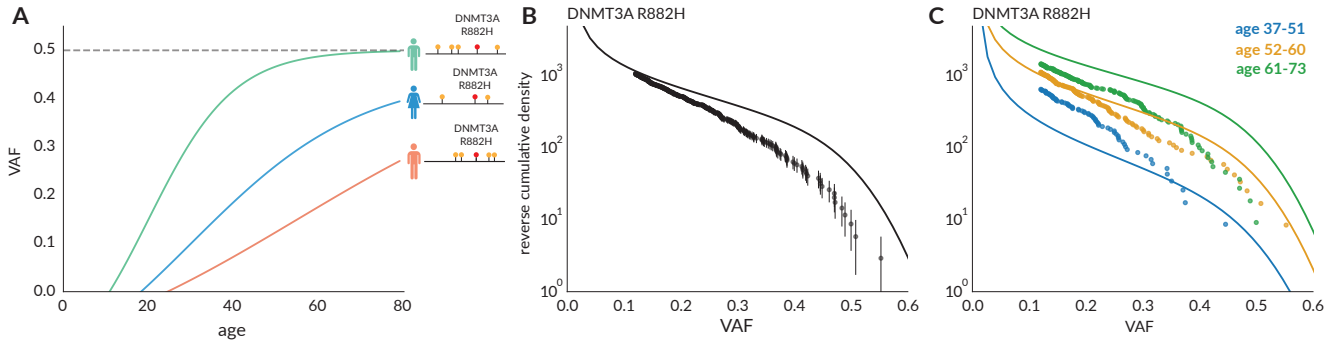

Fig. 2. Variation in fitness effects between individuals is insufficient to explain age-dependent behaviour. **(A)** Schematic view: clones in individuals carrying the same CH hotspot mutation (e.g. DNMT3A R882H) but with different germline backgrounds may experience different fitness effects. **(B-C)** Cumulative density of DNMT3A R882H variants in UK Biobank (points) compared to predicted density from a constant growth model with fitness varying between individuals, on aggregate (B) and stratified into three age tertiles (C). Error bars in B were derived assuming counts were drawn from a Poisson distribution.

crease with age (10, 19). If this increase were due to τ increasing with age, it would lead to the growth rate of all clones, regardless of mutation, slowing with age.

In order to evaluate these hypotheses quantitatively we developed a new maximum-likelihood approach which explicitly accounts for both age and clone size together in each individual (see Supplementary Note 2C). Our method also adjusts for the effect of sequencing at finite depth, an important source of noise in relatively low-depth sequencing data such as the UKB exomes (Supplementary Fig. S2).

Fitness variation cannot explain age deceleration. To test whether variation in fitness effects across individuals could explain the slower than expected age-dependence of CH, we modified the constant growth model to allow clonal fitness to vary between individuals. In this model the fitness effect of a particular variant in a given individual is drawn from a distribution of fitness effects across individuals, but remains constant throughout a given individual's lifetime (Figure 2A). Fitness effects were drawn from a family of stretched exponential distributions with characteristic fitness $\bar{s}$ and 'stretching parameter' p, motivated by the intuition that low-fitness drivers should be more common than high-fitness drivers (more detail in Supplementary Note 3). We showed that while such a distribution of fitness effects could produce an apparent slowing of clonal expansion, they also altered the form of the VAF distribution, producing 'flatter' distributions that were not consistent with UKB observations (Supplementary Fig. S5). We estimated the maximum-likelihood μ, $\bar{s}$ and p for each variant but found that no distribution of fitness effects was able to produce the observed variant distributions by age, regardless of the variation in fitness implied by the stretching parameter (Fig. 2B-C).

Clonal competition can cause deceleration of CH. Previous work suggests that the majority of events that drive CH occur outside currently known canonical CH driver genes (9, 13, 18). If present in the same individual, such 'hidden' clones could compete with visible clones and result in an apparent slowing of clonal expansion later in life. To investigate whether pervasive clonal competition could reproduce the deceleration observed in UK Biobank, we modelled the behaviour of an observable clone (e.g. harbouring a CH hotspot variant) in the presence of an unobserved generic 'background' clone (representing a collection of 'hidden' CH drivers) (Fig. 3A). Both the observed and background clones were assumed to arise at a constant mutation rate and have constant fitness effects, with the age-dependent behaviour emergent from the competition between the two. The model was constrained so that the parameters of the background clone remained constant across all variants studied.

The clonal competition model was able to capture more accurately the age-dependence of the VAF distributions observed in UKB (Fig. 3B-F). We inferred the mutation rates and fitness effects for the four CH hotspots and the unobserved background that were required to obtain a good fit to the UKB data (Fig. 3G). An important prediction of this model (and also of the 'stem cell ageing' model described below) was that the estimated fitness effects of the CH variants under this model are significantly greater than under the constant-growth model (e.g. DNMT3A R882H: $s = 22\%$ per year) – variants expand rapidly in early life before growth is suppressed by clonal competition. This arises from the fact that there are more variants in the younger UKB participants than predicted by the constant-growth model, therefore the average fitness at ages below that of the youngest UKB participants (approximately 40) must be greater.

The model, however, required a mutation rate to unseen background variants of 1.1×10^{-5} per cell per year and a fitness effect of 22% per year. These parameters are considerably larger than previously published estimates of 'missing' drivers (13), and are unlikely to be plausible (see discussion).

Ageing stem cells can cause deceleration of CH. We next explored a model in which changes in clonal growth are driven by a systemic change in the behaviour of HSCs over a lifetime. We modelled this 'stem cell ageing' as a two-stage process: clones grow a constant rate s_1 until age 40, then at a slower rate s_2 thereafter (Fig. 4A). The choice of age 40 for the transition from fast to slow growth was motivated by the fact that there are very few participants in the UK Biobank

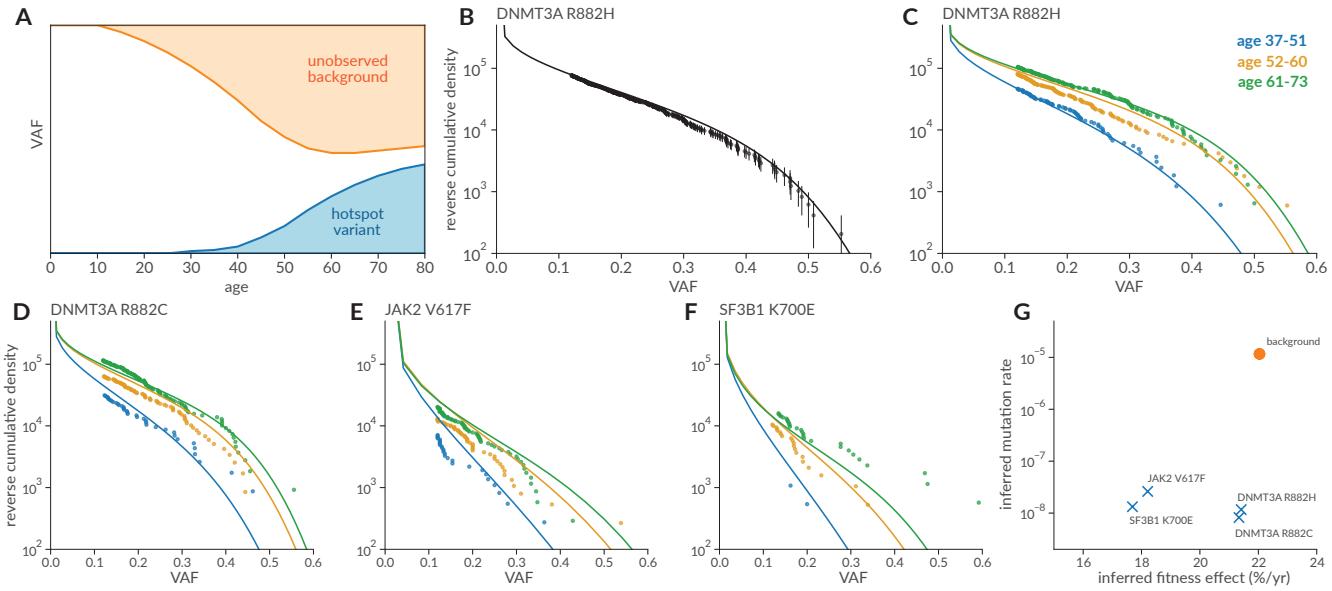

Fig. 3. Competition between clones can cause slowing expansion in later life. (A) Schematic representation of clonal competition. A CH hotspot clone (such as *DNMT3A* R882H) occurs alongside a collection of unseen 'background' mutations, modelled as a single variant with constant mutation rate and fitness effect. As the clones grow, they compete and inhibit their mutual expansion. Comparison of predictions of the clonal competition model (lines) with DNMT3A R882H variants (points) with best fit parameters, both overall (**B**, black) and stratified into three age tertiles (**C**, ages 37-52 (blue), 53-61 (yellow), 61-73 (green)). Error bars in B were derived assuming counts were drawn from a Poisson distribution (± 1 s.e.). **(D-F)** The same age-stratified analysis as shown in (C) for other CH hotspot variants: DNMT3A R882C (C), JAK2 V617F (D) and SF3B1 K700E (E) **(G)** Estimated mutation rates and fitness effects for the four CH hotspot variants (blue) and the common unobserved background (orange)

below this age, so for the period age 0-40 we are only able to infer an average growth rate. The model is deliberately simplified – our hypothesis is not that there is a drastic change at stem cell biology at age 40 – but it is sufficiently flexible to encompass a range of plausible scenarios (see discussion). To reflect the hypothesis that growth is slowed by an underlying change in HSC dynamics, the model was constrained so that the fold-change in fitness at age 40 (s_1/s_2) was constant across all variants. The values of s_1, s_2 and mutation rate μ were otherwise fitted independently for the four CH hotspot variants.

Of the models we investigated, the stem cell ageing model was able to reproduce the observed age-stratified VAF distributions most faithfully (Fig. 4B-F). As above, clones were estimated to expand at considerably faster rates in young people compared with current estimates (for example, for DNMT3A R882H $s_1 = 22.4\%$ per year). The fold-change in fitness at age 40 was estimated to be s_1/s_2 = 3.3 (Fig. 4G), which represents a substantial deceleration when considering that the variants are growing exponentially: twenty years after such a change, a DNMT3A R882H clone would be more than 15 times smaller than if it continued growing at s_1. For DNMT3A R882H and R882C (Fig. 4C-D), the more modest expansion rates predicted to occur in later life (e.g. DNMT3A R882H $s_2 = 6.7\%$ per year) are consistent with the lower growth rates observed in longitudinal studies in older people (12). Longitudinal estimates of JAK2 V617F (Fig. 4E) fitness are likely to be confounded by a common and likely fitness-enhancing loss of heterozygosity (see Supplementary Note 1C), making direct comparison difficult. For SF3B1 K700E (where existing studies generally predict higher growth rates) the stem cell ageing model fitted well at low VAF, but we observed a number of high-VAF outliers in older participants (see discussion) (Fig. 4F).

CH prevalence more consistent with HSC ageing model. An important difference between the 'clonal competition' and 'stem-cell ageing' models is in their respective predictions of total clonal prevalence. The stem cell ageing model, in which variants continue to expand at a reduced rate, predicts that prevalence should continue to increase with age. Under the competition model, however, lower fitness variants can be out-competed by the fitter 'background' variants and the prevalence can decline with age.

We examined the age-stratified prevalence of CH hotspot variants in UK Biobank. The prevalence of the four hotspots all increase monotonically with age, with no indication of declining prevalence over the observed period (ages 40-70), whereas the competition model predicts that the prevalence in all variants would begin to decline by age 70 (Fig. 5). This, together with the very high prevalence and fitness of hidden mutations required to produce the observed deceleration, suggests that the observed deceleration is unlikely to arise solely as a result of pervasive clonal competition.

Longitudinal analysis. Models of clonal expansion can be further investigated using longitudinal analysis of individuals with sequencing data at multiple time-points. We studied the growth rates of CH hotspot clones in 3,359 participants (aged 60-91) in the Lifelines study (20), each sequenced twice with median 46 months between samples.

For each model of clonal expansion ('constant growth',

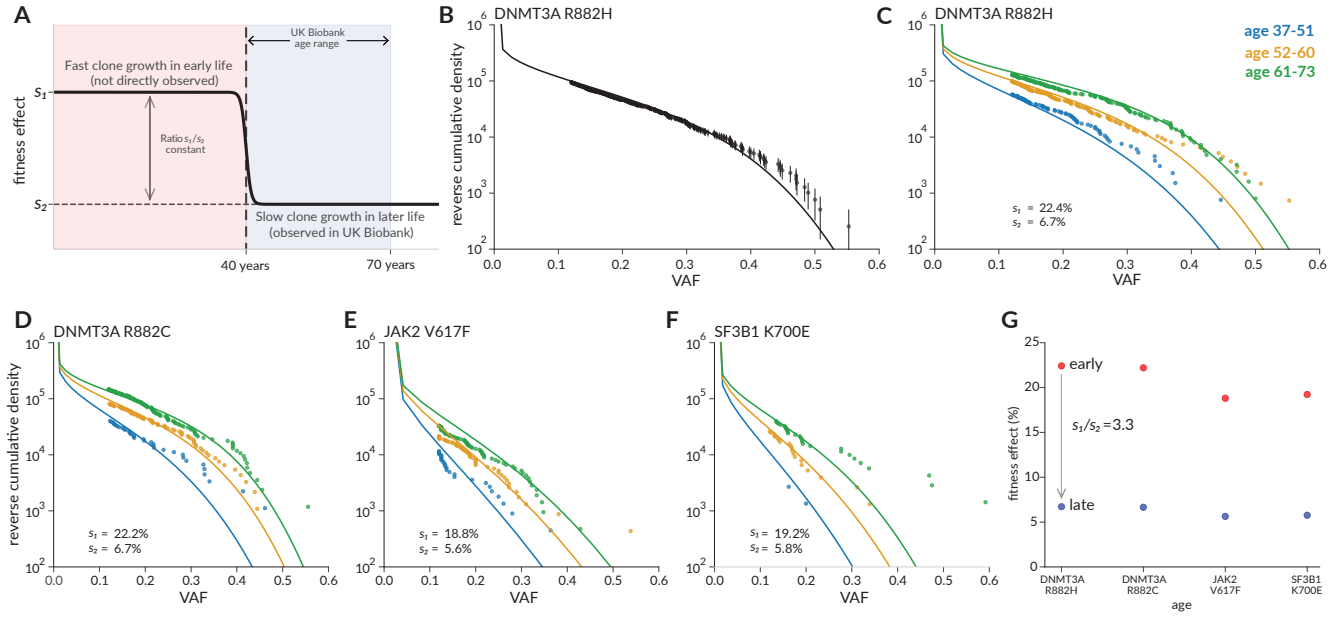

Fig. 4. Slowing of clonal expansion modelled by systemic stem cell ageing. **(A)** Schematic view of the stem cell ageing hypothesis. Clones expand with a fitness effect s_1 before age 40 and s_2 thereafter. The ratio s_1/s_2 is constrained to be common across all four hotspot variants **(B-C)** Comparison of predictions of the stem cell ageing model (lines) with DNMT3A R882H variants (points) with best fit parameters, both overall (B, black) and stratified into three age tertiles (C, ages 37-52 (blue), 53-61 (yellow), 61-73 (green)). Error bars in B were derived assuming counts were drawn from a Poisson distribution (± 1 s.e.). **(D-F)** The same age-stratified analysis as shown in (C) for other CH hotspot variants: DNMT3A R882C (C), JAK2 V617F (D) and SF3B1 K700E (E) **(G)** Summary of estimated early-life and late-life fitness effects for the four variants, linked by the common slowing factor $s_1/s_2 = 3.3$

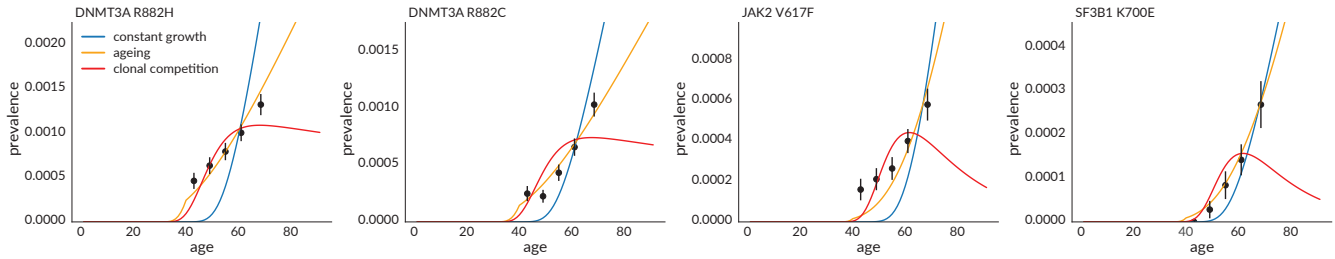

Fig. 5. Age-dependent prevalence of CH variants. Observed overall prevalence of the four CH hotspots in UK Biobank (points) compared with prevalence predicted by the constant growth model (blue), the stem cell ageing model (orange) and the clonal competition model (red).

'clonal competition' and 'stem-cell ageing'), we estimated the expected distribution of growth rates by age and VAF using the parameters inferred for each model from UK Biobank and validated these predictions using simulations (Fig. 6B-C). As with the prevalence predictions, the stem cell ageing model predicts that most variants should continue to grow at a modest rate whereas the competition model predicts that growth should have halted (on average) by late age (Fig. 6D). In all scenarios we predicted that the growth rate should decline linearly with increasing VAF as the clone approaches the point where it has swept the entire HSC population.

The observed longitudinal growth rates of DNMT3A R882H/C clones are slower than would be predicted by a constant growth model (Fig. 6D), supporting the findings from UK Biobank. Among clones with VAF below 10% per year the number of clones growing and shrinking is approximately equal, whereas higher-VAF clones are more likely to grow (7 growing, 2 neutral, 2 shrinking). The wide variability in the growth rates means that it is difficult to distinguish whether the stem cell ageing or clonal competiton models more accurately describe the observations. (Fig. 6E-F). The observed variation in growth rate is greater than would be predicted if variants behaved identically across individuals: for example, the observed variability at the DNMT3A R882H/C hotspot would be consistent with a sequencing depth of around 400 reads (compared to an overall average study depth of 8786 reads) (Supplementary Fig. S8), suggesting that there is variability in expansion rates between individuals carrying the same variant.

Discussion

By analysing the cross-sectional VAF spectra of common CH hotspots across $\sim 400,000$ UK Biobank participants, we demonstrated that clonal expansion does not occur evenly across a person's lifetime. The overabundance of variants seen in young people (relative to the expected frequency under a constant selective advantage) suggests that clonal expansion is slower in the old than the young by at least a factor of three, and that this deceleration occurs mainly before

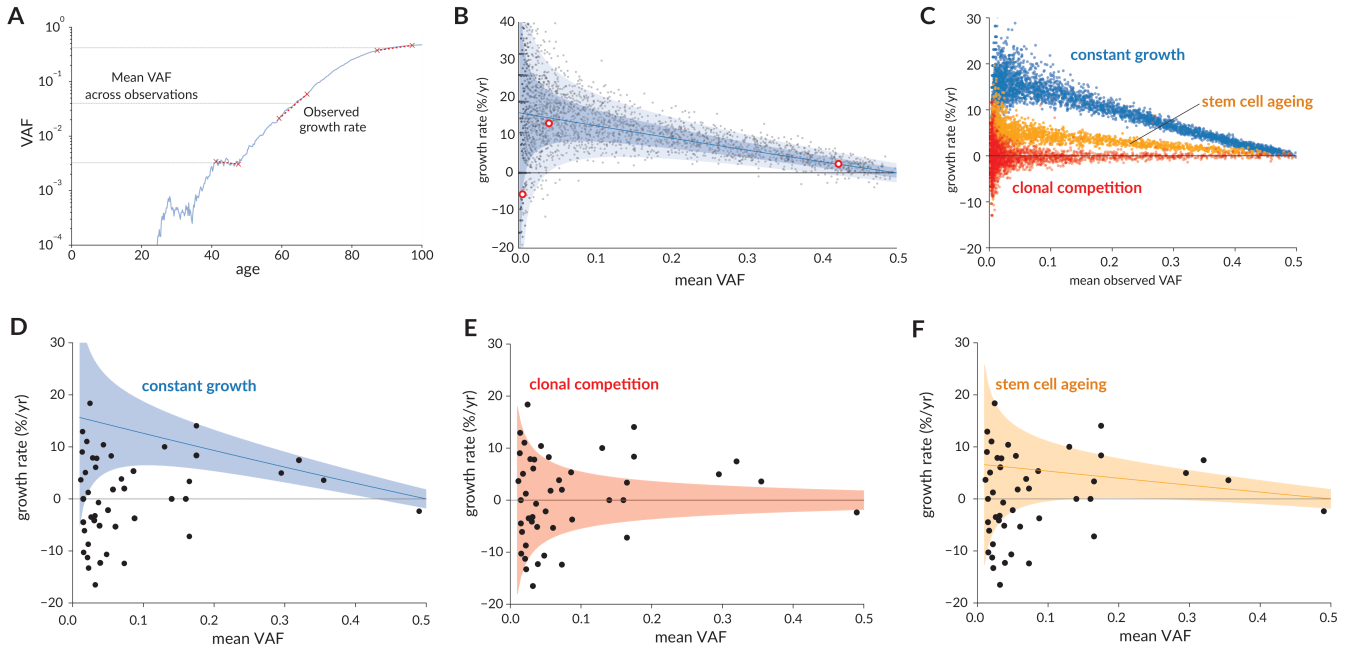

Fig. 6. Longitudinal analysis of clonal expansion. **(A)** A simulated mutant clone with fitness effect $s = 16\%$ per year is sampled three times, each time measuring the growth rate (log-gradient) and mean VAF from two timepoints. **(B)** Simulated distribution of observed growth rates and mean VAF obtained from longitudinal observations of a large number of individuals with variant clones with $s = 16\%$ per year (blue), showing the three example observations from (A) (red) and the analytically derived expectation (blue line) with shaded area showing two standard deviations. **(C)** Comparison of simulated distributions of two-timepoint longitudinal observations under the constant growth model (blue), the stem cell ageing model (orange) and the clonal competition model (red). **(D-F)** Growth rates for individuals carrying DNMT3A R882H/C clones from the Lifelines study compared with expected distribution following the constant growth model (D), the stem cell ageing model (E) and the clonal competition model (F). Shaded area shows two standard deviations based on 'effective' sequencing depth estimated from the data.

the age of 40. The poor performance of the constant growth model was also seen in the age-specific prevalence estimates in UK Biobank, and in the individual growth rates in the longitudinal Lifelines data. This deceleration was observed across all four hotspots studied, spanning three canonical CH genes, suggesting that deceleration may occur more generally in CH.

Our findings are consistent in a number of important ways with existing analysis of CH derived from longitudinal samples in later life (12). Firstly, clonal deceleration would account for the existing disparity between fitness estimates made from cross-sectional VAF distributions and those made from longitudinal samples (e.g. DNMT3A R882H: cross-sectional $s = 14.8\%$ per year (10), longitudinal $s = 5.5\%$ per year (12)). Secondly, longitudinal analysis reveals clones expanding too slowly in later life to explain the observed VAF under an assumption of constant growth, implying early-life deceleration in agreement with our findings. How much this deceleration varies across the genome remains to be seen: other studies have suggested that the slowdown may be more substantial in some genes (e.g. DNMT3A, TP53) than others (e.g. SF3B1, TET2) (12). We found that the magnitude of the slowdown across our four hotspots was not significantly different, although the presence of high-VAF outliers carrying SF3B1 K700E potentially suggest other factors may be involved in the later-life expansion of clones containing that variant. The observation that CH appears at higher VAF in individuals aged 40-50 than would be predicted from current estimates of fitness suggests that clonal growth rates in the young may be considerably faster than previously thought (around 20% per year for DNMT3A R882H/C). This prediction could be tested directly by studying clonal expansion in children or young people, as discussed further below.

We sought to investigate how the fundamental dynamics of HSCs might give rise to clonal deceleration using mathematical models. Some of the key features of clonal dynamics, such as the collapse in clonal diversity at late age, can be shown to be emergent properties of a constant-growth model (9). However, we found that competition between clones growing at a constant rate is unlikely to be sufficient to explain the observed deceleration. In order to produce a significant slowdown in the majority of individuals by age 40, we demonstrated that the mutation rate to 'hidden' drivers would need to be large ($O(10^{-5})$ per year, so that $N\tau\mu \approx 1$, i.e. the average individual has one expanded clone by middle age) and background drivers would need to be sufficiently fit to out-compete the observable clones. Both the mutation rate and fitness are implausibly large compared with existing estimates of missing selection (for example, derived from hitchhiking synonymous variants (13)). Furthermore, the clonal competition model predicts that deceleration should continue until eventually most clones cease to grow entirely, or begin to shrink. This behaviour is not observed in the UK Biobank, where variants continue to grow consistently through life. Positive average growth rates have been observed in individuals up to the age of 100 (12).

The most promising model for the observed slowdown simulates an explicit change in HSC dynamics before the age of 40. The model presented is deliberately simple but illustrates that a period of fast growth followed by a period of slow (but consistent) growth captures the observed features of both VAF distributions and prevalence curves in UK Biobank. Without data from young people we cannot make detailed claims about clonal dynamics in the under-40s beyond an average growth rate, and therefore the nature of the early period is open to interpretation: for example, rapid clonal expansion in childhood followed by a decline during adolescence would produce the same VAF distributions to the model presented here, with suitably adjusted parameters. The slower late-life growth rates predicted by the model align more closely with the rates observed in existing longitudinal studies (12). The model raises the tantalising prospect of a yet-unknown biological change in HSC dynamics occurring in young people. The vast majority of existing CH data is taken from middle-aged and older people; further progress on this problem will be difficult unless we are able to study children and young people directly with deep sequencing (since CH clones are rarer and occur at low VAF in the young). Existing work has identified decreasing of stem cell division rates with age in colon, oesophagus and other non-blood tissues in humans (21), and other mechanisms have been proposed that could give rise to apparent age-related deceleration: for instance, that differential telomere lengths between cell lineages may cause some lineages to go extinct sooner than others (22, 23).

Longitudinal data from the Lifelines study suggests that clonal expansion rates can vary considerably between individuals even in clones ostensibly driven by the same variant, an effect seen in other longitudinal studies (6, 12, 24), Our modelling shows that this variation alone cannot account for the apparent deceleration observed in the UK Biobank. Such variation is likely influenced by germline genetics (15, 16), clonal competition, the presence of multiple-mutant clones and the effects of extrinsic environmental factors. Integrating these effects into models of somatic evolution is an important area for future work.

REFERENCES

1. Iñigo Martincorena, Amit Roshan, Moritz Gerstung, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886, 5 2015. ISSN 0036-8075. doi: 10.1126/SCIENCE.AAA6806.
2. Henry Lee-Six, Sigurgeir Olafsson, Peter Ellis, et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, 574(7779):532–537, 10 2019. ISSN 14764687. doi: 10.1038/S41586-019-1672-7.
3. Iñigo Martincorena, Joanna C. Fowler, Agnieszka Wabik, et al. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 11 2018. ISSN 10959203. doi: 10.1126/SCIENCE.AAU3879.
4. Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine*, 371(26), 12 2014. ISSN 0028-4793. doi: 10.1056/NEJMoa1408617.
5. Giulio Genovese, Anna K. Kähler, Robert E. Handsaker, et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*, 371(26):2477–2487, 12 2014.
6. Andrew L. Young, Grant A. Challen, Brenda M. Birmann, and Todd E. Druley. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications 2016 7:1*, 7(1):1–7, 8 2016. ISSN 2041-1723. doi: 10.1038/ncomms12484.
7. Siddhartha Jaiswal, Pradeep Natarajan, Alexander J. Silver, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine*, 377(2):111–121, 7 2017. ISSN 0028-4793. doi: 10.1056/NEJMOA1701719.
8. Lena Dorsheimer, Birgit Assmus, Tina Rasper, et al. Association of mutations contributing to clonal hematopoiesis with prognosis in chronic ischemic heart failure. *JAMA Cardiol.*, 4 (1):25–33, 1 2019. ISSN 23806591. doi: 10.1001/jamacardio.2018.3965.
9. Emily Mitchell, Michael Spencer Chapman, Nicholas Williams, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature 2022 606:7913*, 606(7913):343–350, 6 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04786-y.
10. Caroline J. Watson, A. L. Papula, Gladys Y.P. Poon, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485), 2020. ISSN 10959203. doi: 10.1126/science.aay9333.
11. Caroline J. Watson and Jamie R. Blundell. Mutation rates and fitness consequences of mosaic chromosomal alterations in blood. *Nature Genetics 2023*, pages 1–9, 9 2023. ISSN 1546-1718. doi: 10.1038/s41588-023-01490-z.
12. Margarete A. Fabre, José Guilherme de Almeida, Edoardo Fiorillo, et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature 2022 606:7913*, 606(7913): 335–342, 6 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04785-z.
13. Gladys Y.P. Poon, Caroline J. Watson, Daniel S. Fisher, and Jamie R. Blundell. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nature Genetics 2021 53:11*, 53(11):1597–1605, 11 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00957-1.
14. David A. Hinds, Kimberly E. Barnholt, Ruben A. Mesa, et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood*, 128 (8):1121–1128, 8 2016. ISSN 0006-4971. doi: 10.1182/BLOOD-2015-06-652941.
15. Siddhartha P. Kar, Pedro M. Quiros, Muxin Gu, et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nature Genetics 2022 54:8*, 54(8):1155–1166, 7 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01121-z.
16. Alexander G. Bick, Joshua S. Weinstock, Satish K. Nandakumar, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature 2020 586:7831*, 586(7831): 763–768, 10 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2819-2.
17. Jk Gopakumar, Joshua Weinstock, Bala B Burugula, et al. Clonal Hematopoiesis Is Driven By Aberrant Activation of TCL1A. *Blood*, 138(Supplement 1):597, 11 2021. ISSN 0006-4971. doi: 10.1182/BLOOD-2021-153242.
18. Florian Zink, Simon N. Stacey, Gudmundur L. Norddahl, et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, 130(6):742–752, 8 2017. ISSN 0006-4971. doi: 10.1182/BLOOD-2017-02-769869.
19. Henry Lee-Six, Nina Friesgaard Øbro, Mairi S. Shepherd, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature 2018 561:7724*, 561(7724): 473–478, 9 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0497-0.
20. Isabelle A. van Zeventer, Aniek O. de Graaf, Jonas B. Salzbrunn, et al. Evolutionary landscape of clonal hematopoiesis in 3,359 individuals from the general population. *Cancer Cell*, 41(6):1017–1031, 6 2023. ISSN 1535-6108. doi: 10.1016/J.CCELL.2023.04.006.
21. Cristian Tomasetti, Justin Poling, Nicholas J. Roberts, et al. Cell division rates decrease with age, providing a potential explanation for the age-dependent deceleration in cancer incidence. *Proceedings of the National Academy of Sciences of the United States of America*, 116(41):20482–20488, 10 2019. ISSN 10916490. doi: 10.1073/PNAS.1905722116/SUPPL{_}FILE/PNAS.1905722116.SD11.XLSX.
22. Kristen E. Schratz, Valeriya Gaysinskaya, Zoe L. Cosner, et al. Somatic reversion impacts myelodysplastic syndromes and acute myeloid leukemia evolution in the short telomere disorders. *Journal of Clinical Investigation*, 131(18), 9 2021. ISSN 15588238. doi: 10.1172/JCI147598.
23. Emily A. DeBoy, Michael G. Tassia, Kristen E. Schratz, et al. Familial Clonal Hematopoiesis in a Long Telomere Syndrome. *New England Journal of Medicine*, 388(26):2422–2433, 6 2023. ISSN 0028-4793. doi: 10.1056/NEJMOA2300503/SUPPL{_}FILE/NEJMOA2300503{_}DISCLOSURES.PDF.
24. Thomas McKerrell, Naomi Park, Jianxiang Chi, et al. JAK2 V617F hematopoietic clones are present several years prior to MPN diagnosis and follow different expansion kinetics. *Blood Advances*, 1(14):968–971, 6 2017. ISSN 2473-9529. doi: 10.1182/BLOODADVANCES.2017007047.
25. Cathie Sudlow, John Gallacher, Naomi Allen, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, 3 2015. ISSN 15491676. doi: 10.1371/journal.pmed.1001779.
26. Rory Collins. What makes UK Biobank special? *Lancet*, 379(9822):1173–1174, 2012. ISSN 1474547X. doi: 10.1016/s0140-6736(12)60404-8.
27. Joseph D. Szustakowski, Suganthi Balasubramanian, Erika Kvikstad, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.*, 53(7):942–948, 7 2021. ISSN 15461718. doi: 10.1038/s41588-021-00885-0.
28. Joshua D. Backman, Alexander H. Li, Anthony Marcketta, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886):628–634, 11 2021. ISSN 14764687. doi: 10.1038/s41586-021-04103-z.
29. Kelly L. Bolton, Ryan N. Ptashkin, Teng Gao, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature Genetics*, 52(11):1219–1226, 11 2020. ISSN 15461718. doi: 10.1038/s41588-020-00710-0.
30. Robert Kralovics, Francesco Passamonti, Andreas S. Buser, et al. A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders . *New England Journal of Medicine*, 352 (17):1779–1790, 4 2005. ISSN 0028-4793. doi: 10.1056/NEJMOA051113/SUPPL{_}FILE/1779SA1.PDF.
31. Damla Olcaydu, Ashot Harutyunyan, Roland Jäger, et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nature Genetics 2009 41:4*, 41(4):450–454, 3 2009. ISSN 1546-1718. doi: 10.1038/ng.341.
32. Po Ru Loh, Giulio Genovese, and Steven A. McCarroll. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature 2020 584:7819*, 584(7819):136–141, 6 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2430-6.
33. Po Ru Loh, Giulio Genovese, Robert E. Handsaker, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*, 559(7714):350–355, 7 2018. ISSN 14764687. doi: 10.1038/S41586-018-0321-X.
34. Salome Scholtens, Nynke Smidt, Morris A. Swertz, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International Journal of Epidemiology*, 44(4):1172–

1180, 8 2015. ISSN 0300-5771. doi: 10.1093/IJE/DYU229.
35. Bart Klijs, Salome Scholtens, Jornt J. Mandemakers, et al. Representativeness of the LifeLines Cohort Study. *PLOS ONE*, 10(9):e0137203, 9 2015. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0137203.
36. Sasha F. Levy, Jamie R. Blundell, Sandeep Venkataram, et al. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature 2015 519:7542*, 519(7542):181–186, 2 2015. ISSN 1476-4687. doi: 10.1038/nature14279.
37. J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1 1965. ISSN 0010-4620. doi: 10.1093/COMJNL/7.4.308.

ACKNOWLEDGMENTS

We thank Arnold Levine, Daniel Fisher and Dan Landau for helpful discussions. We thank all members of the Blundell lab for input. This research has been conducted using the UKB Resource under application number 28126. The Lifelines Biobank initiative has been made possible by subsidy from the Dutch Ministry of Health, Welfare and Sport, the Dutch Ministry of Economic Affairs, the University Medical Center Groningen (UMCG the Netherlands), University Groningen, and the Northern Provinces of the Netherlands. J.R.B is funded by a UKRI Future Leaders Fellowship and by the CRUK Cambridge Cancer Centre. H.A.J.M is supported by the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK (C14478/A29329), Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester.

Supplementary Note 1: Data used in this analysis

A. UK Biobank. Exome sequencing of 450,000 people was performed as part of the UK Biobank (25, 26), a prospective cohort study of 500,000 people in the UK. All UKB participants signed a written informed consent form at enrolment. Ethical approval was given by the North West Multi-Centre Research Ethics Committee (REC 21/NW/0157).

Filtering and quality control performed on these samples is discussed elsewhere (27). These filters included a minimum variant read depth threshold of 7 and a requirement that at least one sample per site exceeded 0.15 VAF (for SNVs). Filtered population VCF (pVCF) files were accessed through the UKB DNANexus platform (28). The variable 'Age.when.attended.assessment.centre_v0' was taken as the age at blood draw. Individuals where this age was unknown were excluded. Individuals with a history of cancer (except basal cell carcinoma) were also excluded on the grounds that systemic cancer treatment (such as radiotherapy and certain kinds of chemotherapy) has been shown to influence clonal expansion in blood (29). The average observed sequencing depth at each position for the four variants was DNMT3A R882H/C: 75, JAK2 V617F: 24, SF3B1 K700E: 56. The lowest observable VAF was 0.12 for all variants.

B. CH hotspots. Existing studies have identified individual SNVs commonly found in expanded clones in blood, known as CH 'hotspots' (10). These sites are useful for studying CH since (1) mutations at these sites are much more likely to be somatic than germline, even at high VAF and (2) these mutations are likely to confer the same fitness effect across different people. Of the hotspots identified by Watson et al. (2020) (10), four were sufficiently prevalent in UKB to explicitly model how VAF distribution stratifies by age: DNMT3A R882H: n=373, DNMT3A R882C: n=236, JAK2 V617F: n=318, SF3B1 K700E: n=50.

C. JAK2 V617F and mosaic loss of heterozygosity. Cells carrying the JAK2 V617F variant (usually somatically acquired) frequently also experience subsequent mosaic copy-neutral loss of heterozygosity (CN-LOH) on chromosome 9p (containing JAK2) associated with later development of myoproliferative neoplasms (MPN) (30, 31). Using calls from Loh et al. (32), we identified 550 mosaic chromosome alterations overlapping the JAK2 V617 locus in UKB, of which 353 were CN-LOH (119 unknown, 75 gain, 3 loss). Of the 318 JAK2 V617 carriers, 133 also carried an overlapping mCA (114 LOH, 13 gain, 4 unknown), indicating a very clear association in the UKB as expected. (Fisher's exact $p < 1 \times 10^{-10}$).

Furthermore, a number of JAK2 V617 carriers without LOH calls exhibited very high VAF (Fig. S1A). Normal CH variants rarely exceed 0.5 VAF (since this is the VAF achieved by a heterozygous mutation at fixation, i.e. present in every cell), but if the variant is copied across to the other allele by a LOH event then higher VAFs are achievable. The upper limit of detection for CN-LOH was a cell fraction of 0.54 (32), and any CN-LOH above this would be undetected so it is likely that these high-VAF JAK2 carriers also carry undetected overlapping CN-LOH. Assuming that any JAK2 V617 carrier with VAF > 0.54 was also a CN-LOH carrier, it is clear from observing the VAF distributions that overlapping CN-LOH confers an additional change in fitness effect (Fig. S1B). We therefore excluded any JAK2 V617 carrier with any confirmed chromosomal alteration or with VAF > 0.54. The number of JAK2 V617F carriers remaining was 145.

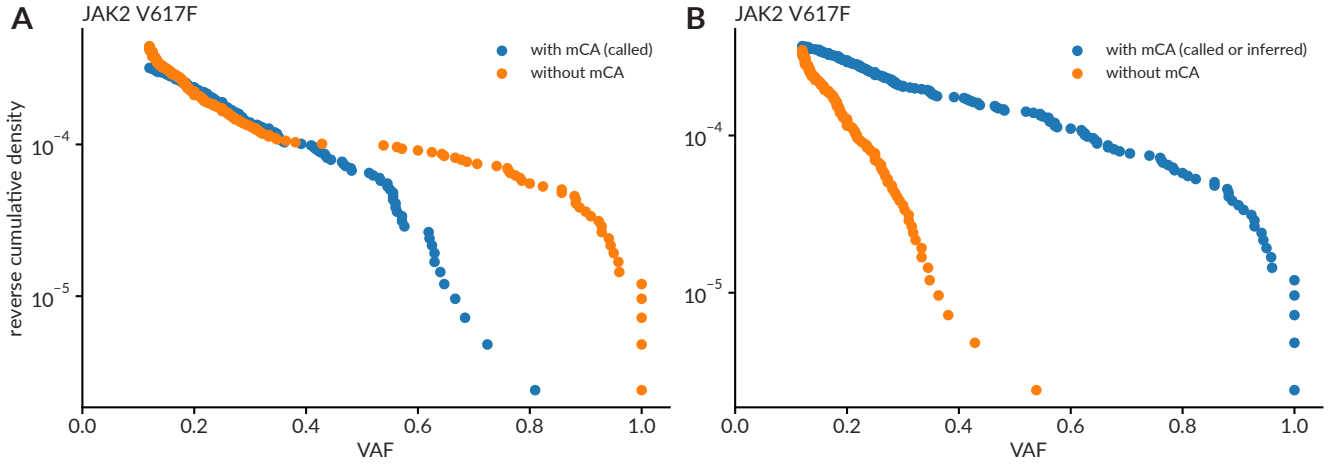

Fig. S1. Effect of co-occurring CN-LOH on JAK2 V617F fitness: Reverse cumulative VAF density of individuals with JAK2 V617F, with accompanying CN-LOH (orange) and without CN-LOH (blue). **(A)** includes only LOH calls called directly by Loh et al. (33), **(B)** with the additional assumption that any variants with JAK2 V617F VAF > 0.54 also carry a CN-LOH.

D. Lifelines longitudinal cohort. Lifelines is a prospective cohort study of 167,729 individuals representative of the population of the northern Netherlands (34, 35). Sequencing data for 3,359 individuals, originally analysed by van Zeventer et al. (20), was accessed via the European Genome-Phenome Archive. This cohort underwent error-corrected (single-strand consensus) blood sequencing at two timepoints with a median time between samples of 43 months, yielding 7,045 samples. The mean

sequencing depth was 8,786 consensus reads, and variants were called with thresholds of ≥ 0.01 VAF in at least one timepoint and ≥ 10 consensus reads. The cohort chosen for sequencing was enriched for blood count abnormalities (individuals with anaemia, thrombocytopenia or neutropenia total n = 1074, cytotsis n = 266). Participants were not screened for haematological malignancies or other cancers. The original analysis of this dataset found no association between blood count abnormalities at baseline and clonal dynamics, and we made no adjustments for these abnormalities. Further information on the Lifelines dataset and sequencing methods can be found in van Zeventer et al. (20).

Supplementary Note 2: Modelling CH

A. Constant-growth model of CH. The work presented here builds on the constant-growth model of CH developed by Watson et al. (2020) (10), and full details of the model and its derivation are found in that publication. In summary, a population of HSCs of size N divides self-symmetrically with average division time τ. Mutations enter this population stochastically at a constant rate μ per cell per year. A cell with a mutation experiences a constant (variant-specific) fitness effect s which biases it towards self-renewal, such that an established clone of mutant cells grows exponentially at rate $s\%$ per year. From this framework it is possible to derive the expected distribution of clone sizes n across a population of individuals at age t years (36):

$$\rho(n,t,\{a_i\})dn = n^{\theta-1}\frac{e^{-n/\tilde{n}(t)}}{\Gamma(\theta)\tilde{n}(t)^{\theta}}dn \tag{1}$$

$$\text{where} \quad \theta = N\tau\mu \quad \text{and} \quad \tilde{n}(t) = \frac{\exp(st)-1}{s\tau}$$

$\tilde{n}(t)$ is interpreted as the characteristic scale of the largest clones in a person of a given age t, in units of number of cells. For the constant-growth model the parameters $\{a_i\} = \{\mu, s\}$; in the alternative growth models these are replaced by new parameters. A change of variables ($2f = \frac{n}{N+n}$) gives the expected distribution of VAF f :

$$\rho(f,t,\{a_i\})df = \frac{(2N)^{\theta}}{f^{1-\theta}(1-2f)^{1+\theta}}\exp\left(\frac{-2Nf}{\tilde{n}(t)(1-2f)}\right)\frac{1}{\Gamma(\theta)\tilde{n}(t)^{\theta}}df \tag{2}$$

The expression for density integrates to 1 over the interval $0 < f < 0.5$ (the maximum attainable VAF for a heterozygous mutant clone assuming that an identical second mutation does not occur on the other allele), meaning that it can be treated as a probability density function for finding a variant at VAF $f \to f + df$ in an individual of age t.

The 'prevalence' of a CH variant was defined as the total fraction of individuals in whom a given CH variant was observed, regardless of VAF. Predictions of prevalence were obtained by integrating the VAF distribution from the lower detection limit in UKB, 0.12, to 0.5:

$$\text{Prev}(t,\{a_i\}) = \int_{0.12}^{0.5}\rho(f,t,\{a_i\})df$$

Predictions for variant density and prevalence for a population with diverse ages were obtained using a weighted sum of single-age predictions.

B. Adjusting for sequencing depth. Sequencing to any finite depth introduces counting noise, which increases as the sequencing depth decreases. We adjusted eq. 2 to account for the sequencing depth in UKB, modelling sequencing to depth D as a binomial random process $\text{Bin}(n,p)$ with $n = D$ and $p = f$. The (discrete) probability of finding r variant reads is then

$$\rho_d(r|D,t,\{a_i\}) = \int_{0}^{0.5}P(r|D,f)\,\rho(f,t)df \tag{3}$$

Where P is the binomial probability:

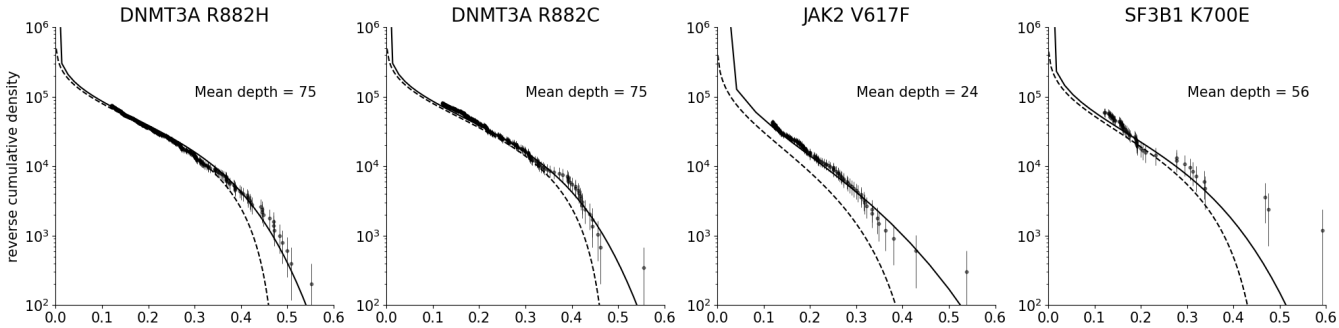

Fig. S2. Effect of finite sequencing depth on predictions of variant distributions. Comparison of predictions of variant density using the constant-growth model before accounting for sequencing effects (dashed lines) and after adjusting for finite sequencing depth (solid lines). The adjustment was done using the mean observed sequencing depth for variant carriers (depth shown in each plot). Lower depth results in a more pronounced sequencing adjustment.

$$P(r|D,f) = \binom{D}{r} f^r (1-f)^{D-r}$$

The effect of sequencing on each of the four hotspot variants (under the constant growth model) is shown in Fig. S2. Notably, accounting for sequencing depth predicts that some variants should appear at $f > 0.5$ due to sampling noise. For each variant, the value of D was taken to be the average sequencing depth for variants at that position.

The distributions shown in all figures in the main text are *reverse cumulative* distributions, normalised by mutation rate – this format was chosen to remove the need to sort variants into bins and to achieve greater visual separation between age groups. The reverse cumulative density $\mathrm{C}(r|D,t)$ is calculated by summing the discrete distribution $\rho_d(r|D,t)$ in reverse order from $r = D$:

$$\mathrm{C}(r|D,t) = \sum_{r'=D}^{r} \rho_d(r'|D,t)$$

C. Age-adjusted maximum likelihood parameter estimation. Together, the age of a variant carrier and the VAF of the variant carry contain joint information about the likely mutation rate and fitness of the variant. Estimating model parameters using a least-squares fit to the overall VAF distribution (Supplementary Fig. S4A), as in previous studies (10), does not utilise this joint information. In order to fit a model across all ages, we estimated the parameters $\{a_i\}$ for each model using a maximum-likelihood approach that explicitly accounted for the age of each individual carrying a variant (Supplementary Fig. S4B). For a single individual j with a CH variant, the likelihood of the parameters $\{a_i\}$ (e.g. for the constant-growth model, $\{a_i\} = \{\mu, s\}$) given the data is simply the discrete probability of finding the variant at its observed read depth r_j, given the age of the participant:

$$\mathcal{L}_j(\{a_i\}|r_j, D_j, t_j) = \rho_d(r_j|D_j, t_j, \{a_i\})$$

There is also a contribution to the likelihood from individuals *without* variants. An individual with no observed variant could (1) truly have no variant or (2) carry a variant below the detection limit of the UKB data, which was 0.12 VAF (empirically). The contribution to the likelihood of an individual k without a variant is:

$$\mathcal{L}_k(\{a_i\}|t_k) = \int_0^{0.12} \rho(f|t_k, \{a_i\}) df$$

using the continuous VAF distribution for the appropriate model (e.g. eq. 2).

The total log-likelihood is therefore:

$$\log\mathcal{L}_{\text{total}}(\{a_i\}|\text{data}) = \sum_{j \text{ in } V} \log\mathcal{L}_j(\{a_i\}|r_j, D_j, t_j) + \sum_{k \text{ not in } V} \log\mathcal{L}_k(\{a_i\}|t_k) \quad \textbf{(4)}$$

Where V is the set of individuals in whom variants are observed. Parameters $\{a_i\}$ were estimated by minimising this log-likelihood using a Nelder-Mead simplex algorithm (37).

The maximum likelihood estimation (MLE) approach has a number of advantages over least-squares fitting. By using the joint information from VAF and age, the model implicitly fits VAF distributions to all ages at once. This is demonstrated in Supplementary Fig. S4C-D: with least-squares fitting (C) the model appears to fit well to the overall density, but when stratified by age the model fits poorly. With maximum-likelihood estimation (D), the model is attempting to fit parameters that explain the distribution at all ages, and it is clear both on aggregate and when stratified by age that the constant-growth model is a poor fit for the data. Furthermore, by considering the likelihood at an individual-variant level, the MLE approach removes the need for arbitrary binning by VAF or by age. The three age tertiles presented here were chosen solely for visualisation purposes – the model fitting approach is agnostic to this choice.

D. Developmental mutations. The above modelling describes a mature, static haematopoietic system with a steady HSC population ($N\tau = 100,000$ cell-years). It is also possible to acquire somatic clones in development (while the HSC population is growing) – if sufficiently prevalent, developmental clones could contribute to a higher 'baseline prevalence' of CH in young people and give the appearance of deceleration with age. Poon et al. (13) calculated the density distribution of these clones (without logistic correction) as:

$$\rho(f) = \frac{\mu_d}{\log 2} \frac{1}{2f^2}$$

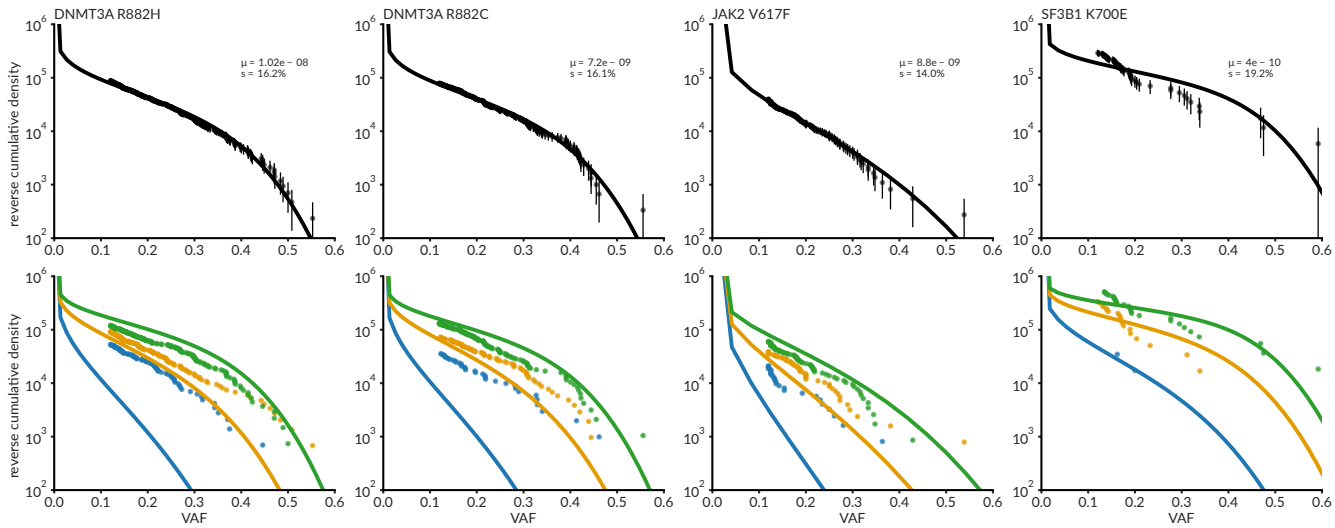

Fig. S3. Least-squares fitted constant growth models. (top row) Predicted VAF distributions for the four CH hotspot variants under a constant exponential growth model (lines) and observed cumulative VAF distributions in UKB (points). Parameters (μ, s) were estimated using least-squares fitting to the aggregate cumulative UKB VAF distributions. **(bottom row)** The same variants and predictions stratified into three age tertiles, (37-52 (blue), 53-61 (yellow), 61-73 (green)), using the same parameters as inferred from the aggregate data.

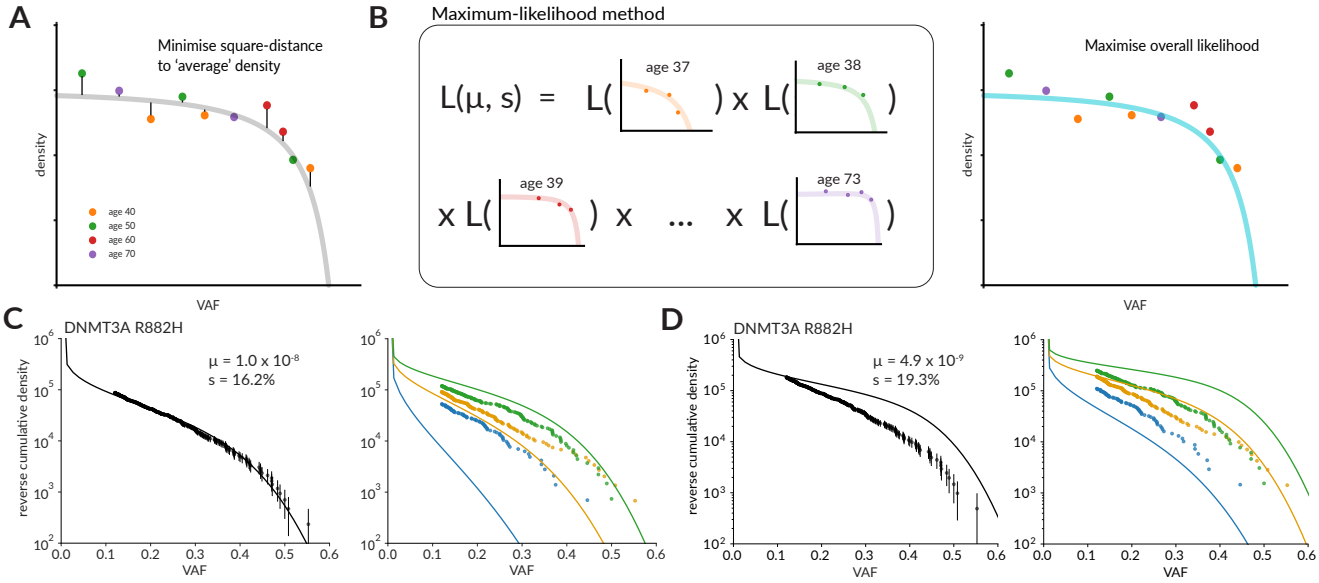

Fig. S4. Age-adjusted maximum likelihood parameter estimation (A) Simple least squares fitting estimation ignores the information from the ages of the variant-carrying individuals **(B)** A maximum-likelihood approach, evaluating the likelihood for each individual variant given the (integer) age of the carrier, utilises the joint information from age and VAF. **(C)** Predicted VAF distribution for UKB DNMT3A R882H, using a constant growth model fitted to the overall VAF distribution with a least-squares approach (overall (left) and stratified into three age tertiles (right). **(D)** Predicted VAF distribution for UKB DNMT3A R882H, using a constant growth model fitted using age-adjusted maximum likelihood (overall (left) and stratified into three age tertiles (right).

where μ_d is the developmental mutation rate. This is estimated in the same publication to be $2-4$ mutations per cell per doubling (across the whole genome), corresponding to $\sim 10^{-9}$ at a single position. Considering that the real density falls to zero as $f \to 0.5$, the reverse cumulative density at the VAF detection limit f^* is approximately

$$\frac{\mu_d}{\log 2}\frac{1}{2f^*}$$

For $f^* = 0.12$, this corresponds to a density of the order $1-10$ (after normalising by total mutation rate as in the plots presented here). Even allowing for some variation in developmental mutation rate, this value is substantially lower than the variant densities observed in UKB, and suggests that developmental mutations are unlikely to play a major role in the age-related deceleration.

Supplementary Note 3: Age-dependent models of clonal dynamics

We explored three models of clonal dynamics in response to the slowdown in clonal expansion with age observed in UKB:

1. **Variation in fitness between individuals:** The fitness effect conferred by a CH hotspot variant is variable across the population, producing different growth rates in different individuals
2. **Clonal competition:** Clonal expansion is constrained by the expansion of other clones experiencing positive selection in the same individual.
3. **Stem cell ageing:** Changes in underlying HSC dynamics cause a general slowing of all clones.

For each model we adapted eq. 2 to derive a density distribution with alternative parameters $\{a_i\}$. Details of the derivation and assumptions used in each model are given here.

1. Variation in fitness between individuals. This model is motivated by the idea that across a population of individuals carrying the same variant clone, the fitness effect of that clone may vary. We modelled this scenario by drawing the fitness effect s for each individual from a distribution of fitness effects (DFE). To see how this could give rise to age-related slowdown, it is instructive to consider a simple example: a DFE consisting of two delta functions at $s = 0.14$ and $s = 0.18$, with the higher-fitness variant ten times rarer (Fig. S5A):

$$\text{DFE}(s) \propto \delta(s-0.14) + \frac{1}{10}\delta(s-0.18)$$

The expected VAF distribution for this DFE is the weighted sum of the two VAF distributions for each delta function, obtained using eq. 2 (since a delta-function DFE is equivalent to constant exponential growth). The resulting VAFs are shown in Fig. S5B – at younger ages, the contribution from the higher-fitness variant dominates (as the lower-fitness clones have not had time to grow yet), but as age increases, the lower-fitness clones increasingly contribute to the observed distribution. If observed at a particular VAF window, the result can appear similar to a slowdown as the distribution becomes increasingly dominated by slower-growing variants. The VAF distribution also changes shape, with the characteristic fall-off at $n \approx \tilde{n}$ becoming more gradual due to the presence of multiple different values of $\tilde{n}$ withing the population.

To evaluate where variation in fitness could explain the UK Biobank data, we fitted a more plausible model with a continuous distribution of fitness effects. Specifically, we used a family of stretched exponential distributions:

$$\text{DFE}(s) \propto \exp\left[-\left(\frac{s}{\bar{s}}\right)^p\right] \tag{5}$$

with one additional parameter compared to the constant-growth model: the constant fitness s is replaced by a characteristic fitness $\bar{s}$ and a 'stretching parameter' p. This distribution is consistent with the expectation that higher-fitness variants are likely to be rarer than lower-fitness variants, and is used in other studies for the same purpose (10, 13). A smaller p implies a more 'spread out' distribution, with the DFE approaching a step function at $\bar{s}$ as $p \to \infty$. The model implies a large number of low-fitness variants, but these are not relevant since variants with $s < \sim 10\%$ per year would almost never reach detectable levels in UKB over a human lifespan and so make no contribution to the observable VAF density. The overall density is obtained by integrating the constant-growth density (eq. 2) over this distribution, and normalising:

$$\rho(f,t,\{\mu,\bar{s},p\}) \propto \int_0^\infty \rho(f,t,\{\mu,s\})\ \exp\left[-\left(\frac{s}{\bar{s}}\right)^p\right] ds \tag{6}$$

As p decreases, resulting in more variation in fitness, the predicted deceleration with age increases, more consistent with the data. However, the shape of the VAF distribution also changes, with a a more gradual decline in density (i.e. a larger frequency of high VAF variants) as illustrated in S5C-F. We fitted eq. 6 to the UK Biobank data, inferring $\bar{s}$ and p independently for each CH hotspot variant using the maxmimum-likelihood approach described above. We found that there were no parameters that produced VAF distributions that fit the data well – the model was unable to match both the deceleration and the shape of the VAF distribution (Main Fig. 2.

2. Clonal competition. Clonal expansion is constrained by the expansion of other clones experiencing positive selection in the same individual. This is a diverse scenario that could encompass competition between numerous clones driven by mutations, copy number events or other unknown drivers of clonal expansion. Our aim was to demonstrate the plausibility that competition can produce variant distributions consistent with those observed in UK Biobank, so the exact details of the competing variants is not of particular concern. We use a simple model: an observable variant (e.g. *DNMT3A* R882H) with parameters $\{\mu_1, s_1\}$ competes with a generic 'background', which we treat as a single event with a mutation rate μ_2 (which is large) and fitness effect s_2. The background need not be observed. For further simplicity, multiple separate mutation events (observed or background) may occur in an individual but they may *not* occur in the same clone (i.e. no double-mutant clones).

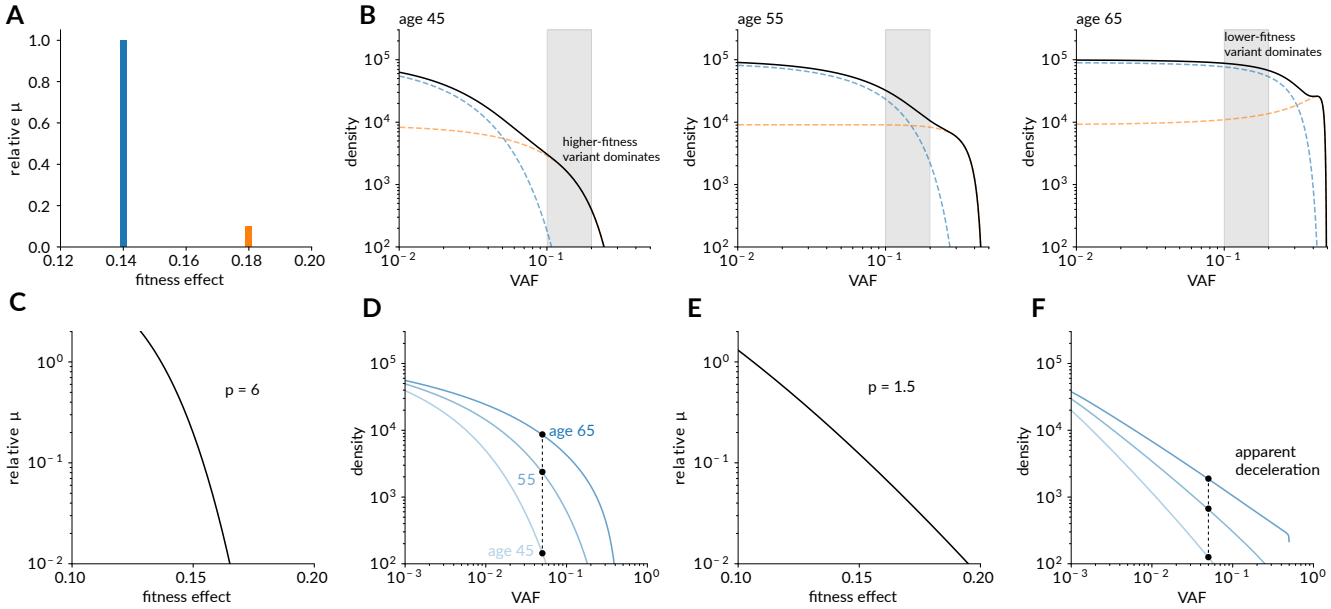

Fig. S5. Distribution of fitness effect influences VAF distribution (A) Example DFE constructed from two delta functions at $s = 0.14$ and $s = 0.18$. **(B)** Predictions of VAF density distribution for the two-delta-function DFE, showing the contribution of each delta function (dashed lines) to the overall distribution (black line) at ages 45, 55, and 65. Examining the shaded VAF range demonstrates how the distribution is dominated by high-fitness variants at younger ages, and by lower-fitness variants at older ages. **(C-D)** Example DFE for a high stretching parameter ($p = 6$), corresponding to a step-function-like DFE (C). The resulting VAF distribution (D) is similar in form to the constant-fitness DFE, with no significant age-related deceleration. **(E-F)** Example DFE for a low stretching parameter ($p = 1.5$), corresponding to a gradually declining DFE containing some high-fitness variants (E). The resulting VAF distribution (F) is different in shape, and there is some apparent deceleration (compared with the high-p DFE).

An expression for the expected variant density is obtained by calculating the frequency of the observed clone in the presence of the expanding background clone. The expected distribution of total observed clone size n_1 in a population *without* competition is given by eq. 1, repeated here:

$$\rho_n(n_1|t,\{\mu,s\})dn = n_1^{\theta-1}\frac{\exp(-n_1/\tilde{n_1}(t))}{\Gamma(\theta)\tilde{n_1}(t)^{\theta}}dn_1$$

with θ and $\tilde{n}(t)$ as before. Considering an HSC population that started with N wild-type cells, with an additional observable clone of size n_1 and background clone of size n_2, the observable VAF f_1 is defined as

$$2f_1 = \frac{n_1}{N+n_1+n_2}$$

Changing variables in eq. 1 from n_1 to f_1 gives

$$\rho_f(f_1)df = \frac{(2(N+n_2))^{\theta}}{f^{1-\theta}(1-2f)^{1+\theta}}\exp\left(\frac{-2(N+n_2)f}{\tilde{n}(t)(1-2f)}\right)\frac{1}{\Gamma(\theta)\tilde{n}(t)^{\theta}}df \tag{7}$$

The expected density of observable clones is obtained by integrating eq. 7 over the expected density of background clone sizes n_2, given again by eq. 1 with parameters $\{\mu_2, s_2\}$:

$$\rho(f_1|t,\{\mu_1,s_1,\mu_2,s_2\})df_1 = \int_0^{\infty}\rho_f(f_1,n_2)\,\rho_n(n_2)\,dn_2\,df_1 \tag{8}$$

Integrating eq. 8 to infinity was numerically unstable, so the upper limit was chosen as $50\tilde{n}(t)$ in practice, a size larger than any background clone would be expected to reach. This model has two additional parameters compared with the constant growth model: the background mutation rate and fitness $\{\mu_2, s_2\}$. Since the mutation rate of the background is large compared to the mutation rate of any individual variant, we expect the background parameters $\{\mu_2, s_2\}$ to be the same for all observable variants. We therefore estimated the parameters for the four CH hotspots jointly, minimising the sum of the log-likelihoods for all four variants. The total log-likelihood was minimised using the Nelder-Mead simplex algorithm, giving estimates for the ten relevant parameters ($\{\mu_1, s_1\}$ for each of the four variants and $\{\mu_2, s_2\}$ for the background).

.

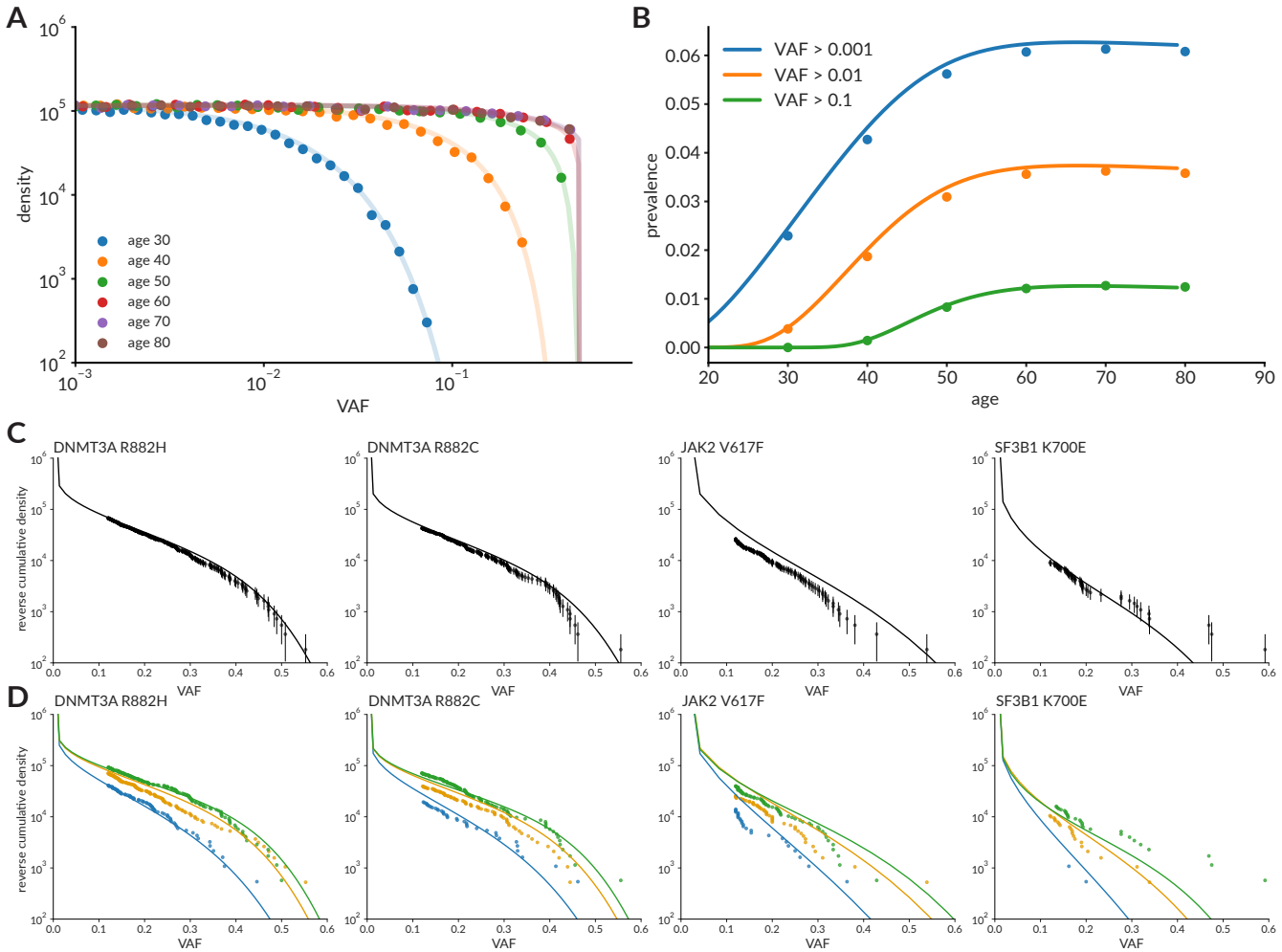

Fig. S6. Clonal competition model. (A) Simulated VAF distributions for clonal competition model (points) compared with analytical expectations (lines), for a population identical in size to the UK Biobank cohort (n=416,171) sampled every 10 years from age 30 to 80). Mutation rate, fitness and are identical to those inferred for DNMT3A R882H. **(B)** Simulated overall prevalence of CH under the clonal competition model (points) compared with analytical prediction (lines), using the same parameters as (A). The lower VAF threshold was varied: VAF > 0.001 (blue); VAF > 0.01 (orange); VAF > 0.1 (green) **(C-D)** Predictions of the clonal competition model (lines) and observed variant distributions in UK Biobank, overall (C) and stratified into three age tertiles (D) for each CH hotspot variant.

3. Stem cell ageing. Under the stem cell ageing hypothesis, the division rate τ of the HSCs changes over an individual's lifetime causing changes in clonal dynamics. We assume that this change is identical in all individuals. From a mathematical standpoint, an age-dependent stem cell division time τ is identical to an inversely varying age-dependent fitness effect s ($s(t) \sim 1/\tau(t)$) and is simpler to analyse this way.

The youngest individual in the UKB cohort is 37, and there are only 3 individuals below the age of 40. As a result, inferences about clonal dynamics before age 40 are limited to estimating an average growth rate up to this point. As such, we model stem cell ageing as a two-stage process with a constant fitness effect s_1 until $t = 40$ yrs and a second, lower, fitness effect s_2 thereafter.

An expression for the variant density in a scenario with variable clonal fitness can be obtained by modifying the characteristic 'maximum clone size' $\tilde{n}$ to account for the history of the growth rate (similarly to Watson et al. (10), Supplementary Material 3):

$$\tilde{n}(t) = \frac{\exp(st) - 1}{s} \quad \rightarrow \quad \frac{\exp(\int_0^t s(t')dt') - 1}{s(t=0)}$$

which simplifies to the constant-growth expression for constant s. The new $\tilde{n}(t)$ can be substituted into the constant-growth density expression (eq. 2) and was validated with simulations (Fig. S7A-B). The resulting expression has one additional free parameter: $\{\mu, s\} \rightarrow \{\mu, s_1, s_2\}$.

Similarly to the clonal competition model, we assumed that the magnitude of the HSC slowdown was the same across all variants. We again estimated the parameters jointly, constraining the model to keep the value of s_1/s_2 the same and minimising

the total log-likelihood. The transition age was fixed at 40 years and we estimated the remaining nine parameters ($\{\mu, s_1\} \times 4$ hotspots and common s_1/s_2).

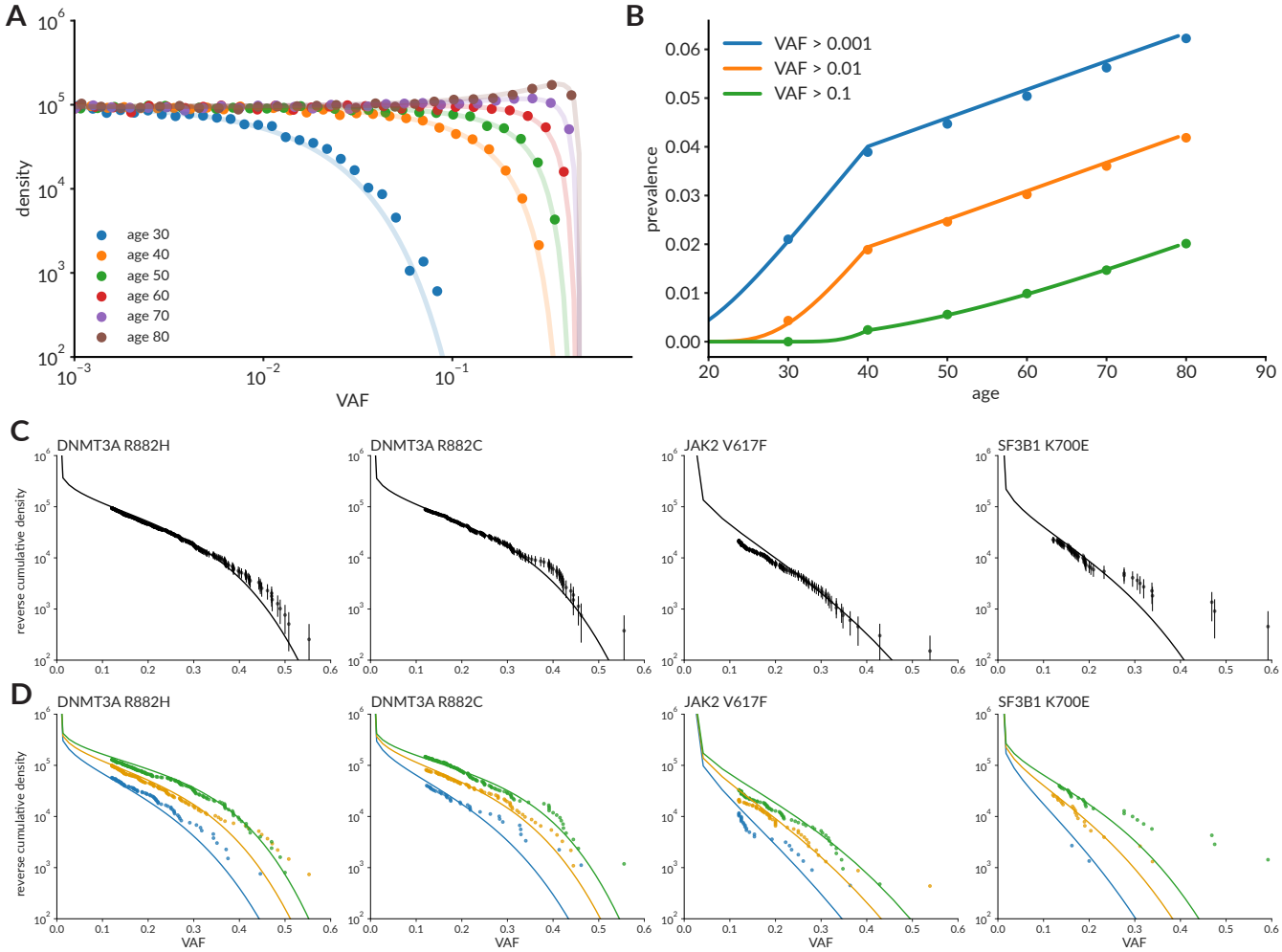

Fig. S7. Stem cell ageing model. (A) Simulated VAF distributions for stem cell ageing model (points) compared with analytical expectations (lines), for a population identical in size to the UK Biobank cohort (n=416,171) sampled every 10 years from age 30 to 80). Mutation rate and fitness parameters are identical to those inferred for DNMT3A R882H. **(B)** Simulated overall prevalence of CH under the stem cell ageing model (points) compared with analytical prediction (lines), using the same parameters as (A). The lower VAF threshold was varied: VAF > 0.001 (blue); VAF > 0.01 (orange); VAF > 0.1 (green) **(C-D)** Predictions of the stem cell ageing model (lines) and observed variant distributions in UK Biobank, overall (C) and stratified into three age tertiles (D) for each CH hotspot variant.

Supplementary Note 4: Longitudinal analysis of CH

We derived an expression for the expected longitudinal growth rate as follows, starting with the constant-growth case. An established clone of n HSCs (VAF f) in a population N with fitness effect s (% per year) grows as

$$n = n_0(1+s)^t = n_0 e^{t\log(1+s)} \qquad 2f = \frac{n}{n+N}$$

We can obtain s directly by differentiating the log-clone size with respect to time:

$$\frac{d}{dt}\log n = \log(1+s) \quad \rightarrow \quad s = \exp(\frac{d}{dt}\log n) - 1$$

However, only f can be measured directly, not n. We can obtain an analogous expression by considering the log-VAF gradient:

$$\begin{aligned}\frac{d}{dt}\log f &= \frac{d}{df}\log f \cdot \frac{df}{dn}\cdot\frac{dn}{dt} \\ &= \frac{1}{f}\cdot\frac{(1-2f)^2}{2N}\cdot n\log(1+s) \\ &= (1-2f)\log(1+s)\end{aligned}$$

The estimated fitness effect s' per year is then (by analogy with above)

$$\begin{aligned}s' &= \exp(\frac{d}{dt}\log f) - 1 \\ &= \exp\left((1-2f)\log(1+s)\right) - 1 \\ &\approx s(1-2f) \quad \text{(small } s\text{)}\end{aligned}$$

We therefore expect that apparent growth rates decline linearly with increasing VAF (f), reaching zero at $f = 0.5$, in line with the intuition that growth stops when the variant clone reaches fixation. For a constant-growth model, this expression is not explicitly dependent on age - a plot of growth rate against VAF should lie along this line at all ages. Usefully, this analysis only considers the instantaneous growth rate of the clone and makes no assumptions about previous growth, unlike similar analysis in previous work. This is essential for our alternative growth models, where growth rates change over time.

To compare this theory with realistic longitudinal VAF measurements, we must also consider two additional factors: (1) sequencing noise, and (2) that we are unable to measure the growth rate accurately at a single time point (i.e. where our two longitudinal measurements are sufficiently close together that they are well approximated by the instantaneous gradient $d(\log f)/dt$). Sequencing noise introduces a spread of growth rates around the expected linear decline. We derived an expression for the width of the expected spread, modelling sequencing as a binomial process. Sequencing a variant of VAF f to a depth D gives an expected variant depth $r = fD$ and a standard deviation $\Delta r = \sqrt{Df(1-f)}$, so for each sequencing measurement

$$\Delta f = \sqrt{\frac{f(1-f)}{D}}$$

The corresponding error in $\log f$ is $\Delta(\log f) = \sqrt{\frac{(1-f)}{Df}}$

If we calculate the log-gradient m from two timepoints:

$$m = \frac{(\log f)_2 - (\log f)_1}{t_2 - t_1}$$

$$\Delta m = \frac{\sqrt{2}\Delta(\log f)}{t_2 - t_1}$$

assuming sequencing errors are i.i.d. and no uncertainty in t. The observed growth rate $s' = e^m - 1$ (as before), so

$$\Delta s' = e^m \Delta m = \frac{1}{t_2 - t_1}\exp[(1-2f)\log(1+s)]\sqrt{\frac{2(1-f)}{fD}}$$

In the Lifelines study, the median $t_2 - t_1$ is 46 months (20). This expression validates well against simulations (Supplementary Fig. S8A).

The only measure of read depth available for the Lifelines data was the average consensus read depth for all samples across the whole panel (D = 8786). The observed variability in growth rates for DNMT3A R882H/C far exceeded what would be

expected from noise alone at this depth (Supplementary Fig. S8C), and was most consistent with a sequencing depth of around 400 reads (fig S8B, D). This suggests that either the consensus read depth at the DNMT3A R882H/C position (DNMT3A c.2710) is significantly below the average, or that there is a considerable degree of real variability in growth rates between individuals, or both.

Since VAF is measured at two discrete time points, when plotting the data f was taken to be the mean of the two VAF measurements. This introduces some additional deviation from the theoretical expectation, especially at low VAF (for example, the simulations in Supplementary Fig. S8A show a small bias towards higher growth rates), but the small deviation is negligible compared with the effect of sequencing noise described above.

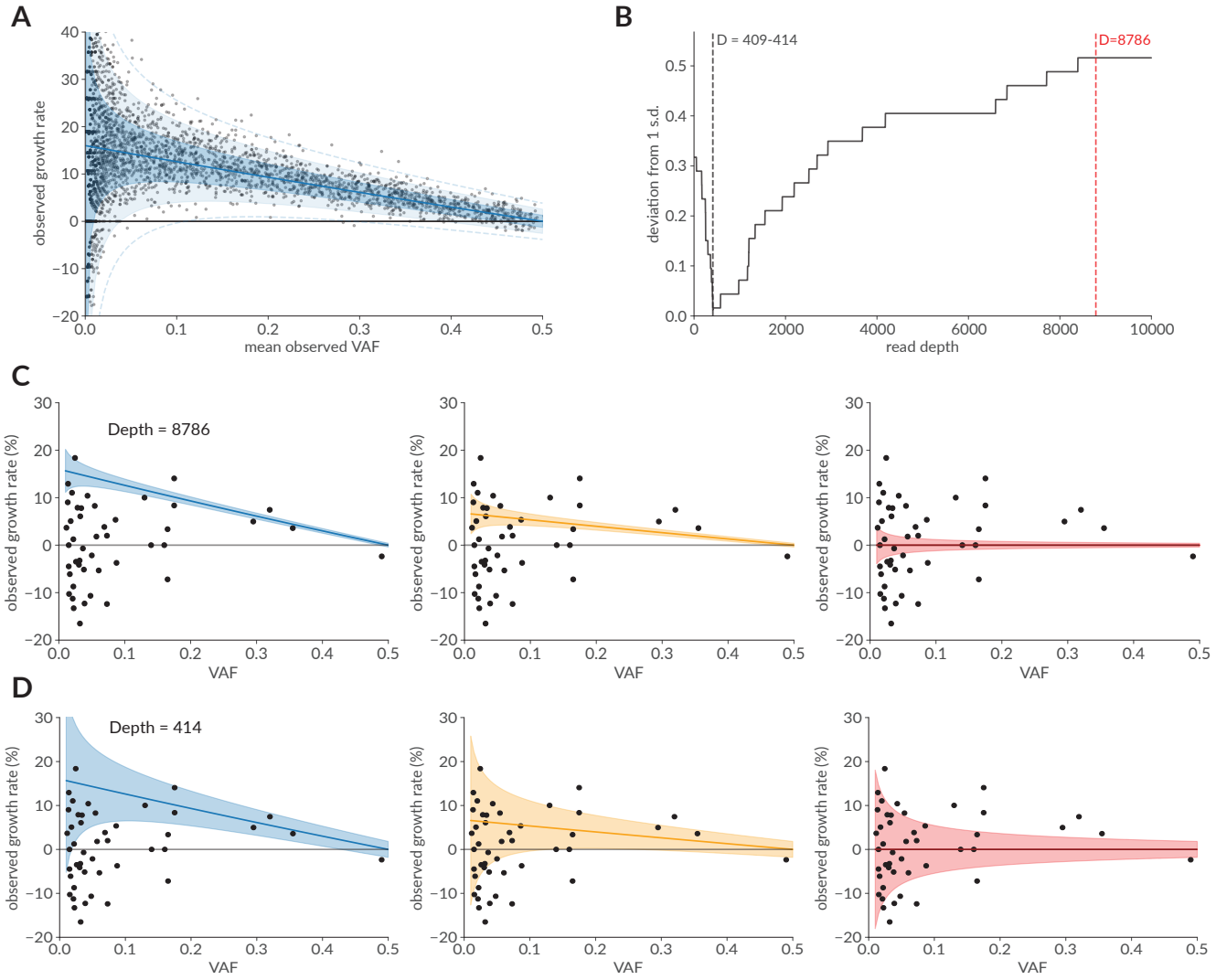

Fig. S8. Longitudinal analysis of Lifelines data. (A) Simulations of a 3,482 individuals aged 60-91 with variant clones, sequenced to depth of 1000 reads at two time points separated by an average of 3.8 years. Variants confer a constant fitness effect of 16% per year. Simulated growth rates and mean VAFs (points) agree with analytical expectations (blue line, shaded areas and dashed line represent 1, 2, 3 standard deviations) **(B)** 'Empirical' sequencing depth inferred from the observed spread of growth rates for DNMT3A R882H/C in Lifelines dataset. Y axis shows the measured deviation from the expected fraction of points within 1 standard deviation from the expected s' as sequencing depth varies. **(C-D)** Growth rates and mean VAFs observed in Lifelines data (DNMT3A R882H/C) (points) compared with expectations from three models of clonal dynamics (constant growth: blue, stem cell ageing: orange, clonal competition:, red), for the quoted average study depth 8786 (C) and the best-fit 'empirical' sequencing depth 414 (D)

We adapted the longitudinal analysis for the clonal competition and stem cell ageing models, using the parameters estimated for each variant from the UKB analysis.

Stem cell ageing. At the ages of the Lifelines participants (60-91), we expect all variants to be in the late-life regime (slower growth). Therefore, the longitudinal predictions remain as above, with s replaced with the late-life fitness effect (s_2) estimated from UKB.

Clonal competition. Under the clonal competition model with the parameters inferred for DNMT3A R882H, clonal expansion has essentially halted by age 60 due to the comparable fitness between the hotspot variant and the background. (Fig. S6B). The

expected growth rate therefore is zero to first approximation; small fluctuations around this would be dwarfed by the effect of sequencing error in the observed data.