CLONAL EXPANSION

The evolutionary dynamics and fitness landscape of clonal hematopoiesis

Caroline J. Watson[1,2]*, A. L. Papula[3], Gladys Y. P. Poon[1,2], Wing H. Wong[4], Andrew L. Young[4], Todd E. Druley[4], Daniel S. Fisher[3], Jamie R. Blundell[1,2]*

Somatic mutations acquired in healthy tissues as we age are major determinants of cancer risk. Whether variants confer a fitness advantage or rise to detectable frequencies by chance remains largely unknown. Blood sequencing data from ~50,000 individuals reveal how mutation, genetic drift, and fitness shape the genetic diversity of healthy blood (clonal hematopoiesis). We show that positive selection, not drift, is the major force shaping clonal hematopoiesis, provide bounds on the number of hematopoietic stem cells, and quantify the fitness advantages of key pathogenic variants, at single-nucleotide resolution, as well as the distribution of fitness effects (fitness landscape) within commonly mutated driver genes. These data are consistent with clonal hematopoiesis being driven by a continuing risk of mutations and clonal expansions that become increasingly detectable with age.

As we age, physiologically healthy tissues such as skin (*1*, *2*), colon (*3*, *4*), esophagus (*5*, *6*), and blood (*7*–*18*) acquire mutations in cancer-associated genes. In blood, this phenomenon, termed clonal hematopoiesis (CH), increases in prevalence with age (*7*–*18*), becoming almost ubiquitous in those over the age of 65 (*10*, *15*). The majority of CH mutations are thought to arise in hematopoietic stem cells (HSCs) (*10*, *19*) and typically fall within the genes *DNMT3A*, *TET2*, *ASXL1*, *JAK2*, and *TP53* and spliceosome genes, although chromosomal alterations are also observed (*17*). Because CH is associated with an increased risk of blood cancers (*7*, *8*, *19*) and the genes affected are commonly mutated in preleukemic stem cells (*20*–*24*), CH has emerged as an important precancerous state, for which a quantitative understanding would accelerate risk stratification and improve our understanding of normal hematopoiesis.

The risk of progressing to a blood cancer depends on the gene in which a variant falls (*14*, *18*). However, our ability to stratify specific variants and their relative risk remains crude. If variants confer a fitness advantage to HSCs, they are more likely to expand over time. Furthermore, higher variant allele frequencies (VAFs) are predictors of acute myeloid leukemia (AML) development (*14*, *18*). It stands to reason, therefore, that by analyzing the spectrum of VAFs, one might be able to infer the fitness advantage conferred by specific variants from a static "snapshot." This would enable us to generate a comprehensive map between specific variants and their fitness consequences, allowing risk to be stratified with greater resolution.

A major challenge to using VAFs to risk stratify variants is that the spectrum of VAFs, even at the level of a specific variant, is considerably broad (*10*). Whether these differences in VAFs are a result of cell-intrinsic fitness advantages (*25*), cell-extrinsic perturbations (*26*), or sheer chance (*13*) remains unclear. To identify the most highly fit variants, we first need to understand how mutation, genetic drift, and differences in fitness (selection) combine to produce the spectrum of VAFs observed in CH.

Results

The VAF distribution from ~50,000 individuals

Insights from evolutionary theory were applied to the VAF spectra of somatic mutations detected in the blood from ~50,000 blood cancer–free individuals from nine publicly available blood sequencing datasets (*7*–*15*) [see (*27*)] to tease apart the effects of mutation, drift, and selection. Using single blood sample snapshots, we quantified the fitness advantages of key pathogenic single-nucleotide variants (SNVs) as well as the spectrum of fitness effects (fitness landscape) of the most commonly mutated driver genes. VAF measurements in bone marrow and peripheral blood show good concordance (*28*), so peripheral blood VAF measurements are used as a proxy to reflect clonal composition at the level of the bone marrow HSCs. The nine studies we analyzed varied in their number of participants and sequencing depth (Fig. 1A). Most large-scale studies were limited by standard sequencing error rates and were only able to detect VAFs >3% (*7*, *8*), whereas smaller studies, which used error-correcting techniques, were able to detect VAFs as low as 0.03% (*10*, *12*, *15*). VAFs varied by more than three orders of magnitude across individuals even within the same gene, as exemplified by *DNMT3A*, the most commonly mutated CH gene (Fig. 1B). The distribution of variants was strongly skewed to low VAFs. Variants were observed far more frequently at certain sites [e.g., *DNMT3A* R882 (Arg^{882}) hotspot codon; red data in Fig. 1B] and were almost exclusively putatively functional (nonsynonymous and frameshifts); synonymous variants were rare and restricted to low VAFs.

A branching model of stem cell dynamics

To reveal the relative contributions of genetic drift, mutation rate differences, and cell-intrinsic fitness effects on the observed variation in VAFs, we considered a simple stochastic branching model of HSC dynamics built on classic population genetic models (*29*–*33*), adapted to include a spectrum of ages and fitness effects [see (*27*)]. The model is of an HSC population of N diploid cells that stochastically self-renew or differentiate symmetrically or asymmetrically (Fig. 1C) and describes a variety of biologically plausible scenarios, including HSCs occupying a fixed number of spatially constrained niches [see (*27*)]. Mutations are acquired stochastically at a constant rate μ per year. The fate of a new mutation depends on its influence on stochastic cell fate decisions through a fitness effect, s, which is the average growth rate per year of that variant relative to the average growth rate of normal HSCs. Neutral mutations ($s = 0$) do not alter the balance between self-renewal and differentiation, which both occur at rate $1/\tau$. Thus, neutral mutations usually rapidly go extinct or, owing to random fluctuations, grow slowly and remain at low VAFs (orange trajectories in Fig. 1D). Beneficial mutations ($s > 0$) increase the rate of self-renewal relative to symmetric differentiation and, provided they escape stochastic extinction, eventually grow exponentially at rate s per year (red and blue trajectories in Fig. 1D). This relative increase in the rate of self-renewal can be achieved by biasing cell fates alone [increasing the probability of self-renewal (*34*) (red plus sign in Fig. 1C) or decreasing differentiation or apoptosis (*35*) (red minus sign in Fig. 1C)] or by a combination of cell fate bias and an increase in division rate.

Variants with a high fitness effect or those acquired early in life are expected to reach high VAFs (trajectories labeled 1 and 2 in Fig. 1D), whereas variants with a low fitness effect or those acquired late in life are restricted to low VAFs (trajectories labeled 3 and 4 in Fig. 1D). This variation in both the age acquired and fitness effect of variants produces a characteristic spectrum of VAFs that can be measured in a single blood sample (insets of Fig. 1D). How these distributions change with age (t) is determined by the fitness effect of variants (s), their mutation rate

[1]Department of Oncology, University of Cambridge, Cambridge, UK. [2]Early Detection Programme, CRUK Cambridge Cancer Centre, University of Cambridge, Cambridge, UK. [3]Department of Applied Physics, Stanford University, Stanford, CA, USA. [4]Department of Pediatrics, Division of Hematology and Oncology, Washington University School of Medicine, St. Louis, MO, USA.
*Corresponding author. Email: cw672@cam.ac.uk (C.J.W.); jrb75@cam.ac.uk (J.R.B.)

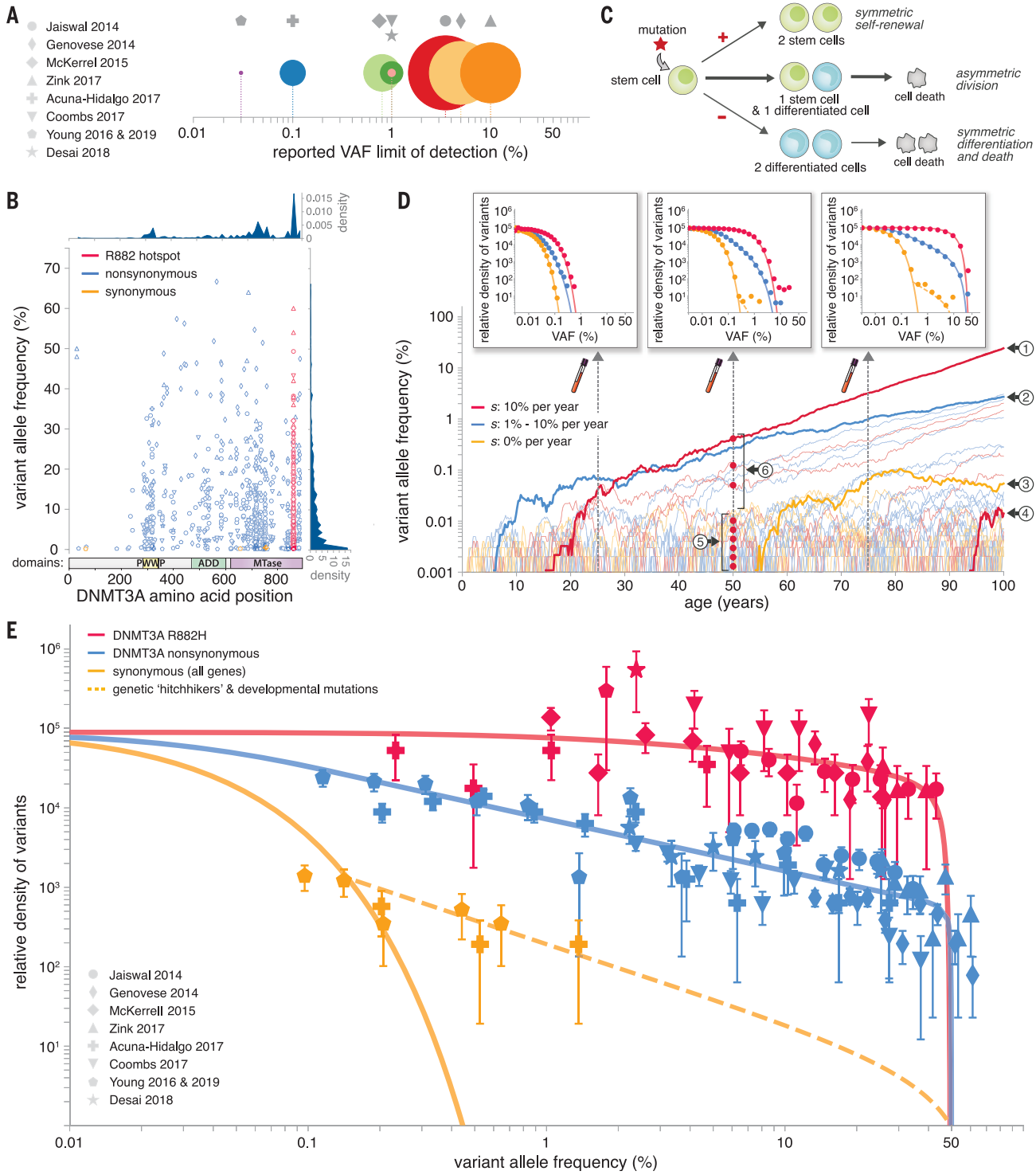

Fig. 1. A branching model of HSC dynamics explains the observed VAF distribution for variants in healthy blood. (**A**) Studies used in this analysis varied in the number of participants (indicated by relative circle size) and reported VAF detection thresholds. (**B**) The density of variants in *DNMT3A* varies widely by VAF (>3 logs) and position in the gene. (**C**) A branching model of HSC dynamics. Mutations with a positive fitness effect (red star) cause an imbalance in stochastic cell fates toward self-renewal. This can be an increase in the rate of self-renewal (red plus sign), a decrease in differentiation or apoptosis (red minus sign), or a combination of the two, resulting in clonal expansions. (**D**) Simulations of HSC populations under a branching model show how differences in fitness effect and age produce VAF spectra (insets) in close agreement with observed data [shown in (E)]. The vertical dashed lines indicate the timings of the blood samples that produce the VAF spectra shown in the insets. The numbered features are explained in the main text. The red dots labeled 5 and 6 highlight where the red trajectories cross the vertical dashed line. (**E**) Plotting all VAF measurements of *DNMT3A* variants as log-binned histograms normalized by mutation rates (data points) demonstrates the consistency with the theoretical predictions of the branching model (lines). The theoretical predictions account for a distribution of ages in the studies. The density of high-frequency synonymous variants is consistent with the predicted density of genetic hitchhikers and early developmental mutations [dashed orange line; see (*27*)]. Error bars represent sampling noise.

(μ), the population size of HSCs (N), and the time (τ) in years between successive symmetric cell differentiation divisions according to the following expression for the probability density as a function of l = log(VAF) [full derivation in (*27*)]:

$$\rho(l) = \theta \exp\left(-\frac{e^l}{\phi}\right) \quad (1)$$

where l = log(VAF), $\theta = 2N\tau\mu$, and $\phi = \frac{e^{st}-1}{2N\tau s}$.

To develop an intuition for the two key features of this distribution, consider variants with a fitness advantage entering the HSC population uniformly at a rate θ/τ per year and growing exponentially. The exponential growth means that variant trajectories, plotted on a log-VAF scale, are uniformly spaced straight lines (red dots labeled 5 in Fig. 1D), producing a flat density with y intercept of θ. Dividing the density of variants by the mutation rate (measured per year), the y intercept therefore provides an estimate for $N\tau$ [insets of Fig. 1D, (*27*)]. Because the age of the oldest surviving variant cannot exceed the age of the individual, there is a characteristic maximum VAF, ϕ, a variant can reach, which increases with fitness effect, s, and age, t. To reach VAFs >ϕ requires a variant to both occur early in life and stochastically drift to high frequencies, which is unlikely. Therefore, the density falls off exponentially for VAFs >ϕ (red dots labeled 6 in Fig. 1D). The sharp density falloff at 50% VAF occurs because even a variant that is present in a very large proportion of total HSCs will tend toward 50% VAF because the cells are diploid.

HSC numbers and division times

To infer HSC numbers and test the predictions of our model, we plotted log-VAF distributions for SNVs from all the studies (*7*–*15*) [see (*27*)]. Studies differed in their number of participants as well as their panel size, both of which affect the number of variants detected. Therefore, to combine the data from all the studies, we normalized the number of observed variants by their study size and total study-specific mutation rate (for variant or gene of interest), controlling for trinucleotide contexts of mutations [see (*27*)]. For a given specific position in the genome, mutation rates are low enough that, over a human life span, clones acquiring multiple driver mutations are rare and thus variants can uniquely mark clones [see (*27*)].

We first focused on mutations in the gene *DNMT3A* (Fig. 1E). The most commonly observed variant in *DNMT3A* is the missense variant R882H (Arg^{882}→His; red data in Fig. 1E). Because fitness effects are expected to be variant-specific (*36*), all R882H variants should confer the same fitness effect and so serve as a useful check on the model. Consistent with our predictions, the density of R882H variants is flat over almost the entire frequency range (VAFs <15%) with a y intercept of $N\tau \approx$ 100,000 ± 30,000 years (figs. S9 and S11). Encouragingly, this number is in agreement with that inferred from single HSC phylogenies (*37*). It is important to note that population genetic analyses can only reliably infer the combination $N\tau$ and not N or τ separately. Early developmental mutations indicate that HSCs accrue ≈1.2 mutations per cell division (*37*), which, combined with an HSC mutation rate in adulthood of ≈16 per cell per year (*37*), suggests that HSCs divide ≈13 times per year. Although symmetric divisions are harder to estimate, this provides an upper bound on the number of HSCs, suggesting that <1.3 million HSCs maintain the peripheral blood. Because $\tau < 1/s_{max}$ [see (*27*)], the maximum inferred $s \approx 25\%$ suggests that $\tau < 4$ years, providing a lower bound of 25,000 on the number of HSCs.

To validate our estimates for $N\tau$, we turned to the distribution of all synonymous variants (orange data in Fig. 1E). Because synonymous variants are generally expected to be functionally neutral, the characteristic VAF of the biggest synonymous variants (ϕ) increases only linearly with age because it is driven by drift alone (see Eq. 1), and $N\tau$ is the time it would take for a neutral mutation to drift to fixation

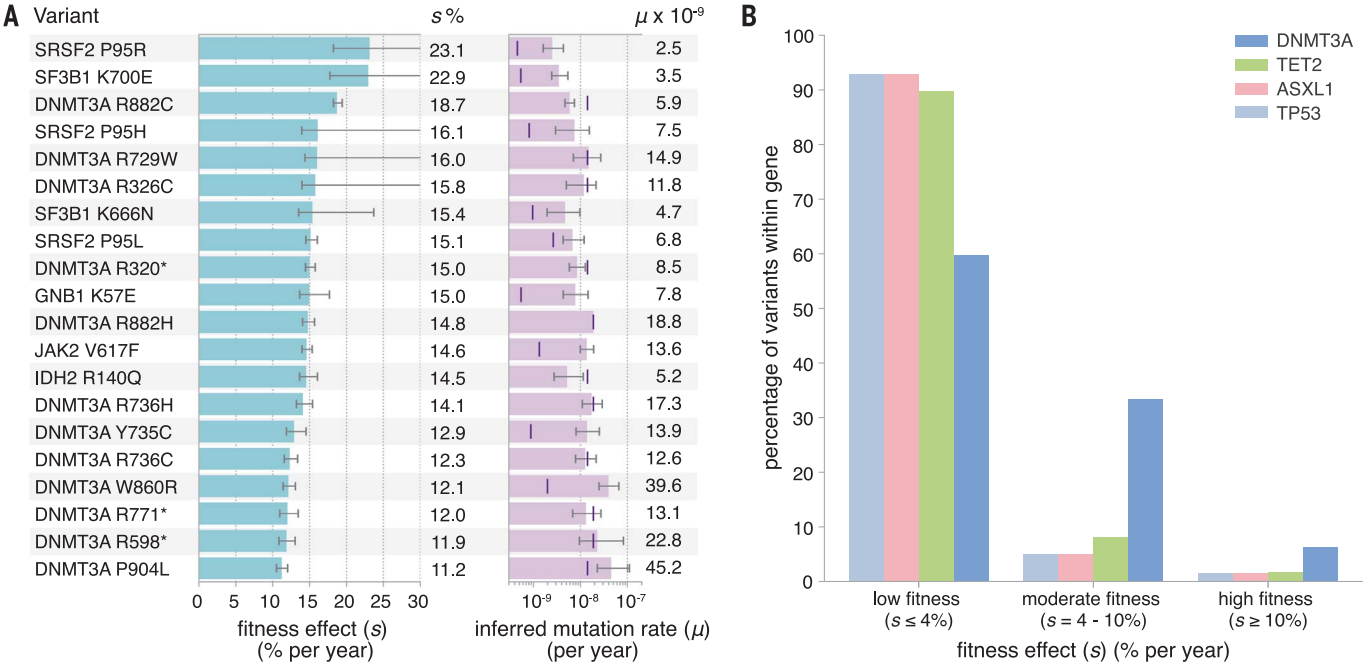

Fig. 2. The fitness landscape of CH variants and genes. (**A**) Inferred fitness effects and mutation rates for the top 20 most commonly observed CH variants. Error bars represent 95% confidence intervals. Purple vertical lines indicate site-specific mutation rates inferred from trinucleotide context [see (*27*)]. (**B**) The distribution of fitness effects of nonsynonymous variants in key CH driver genes, inferred by fitting a stretched exponential distribution and dividing this into three fitness classes (low, moderate, and high) [see (*27*)]. These distributions reveal many low-fitness and few high-fitness variants. Over a human life span, variants with fitness effects <4% expand only a modest factor more than a neutral variant (low fitness), variants with fitness effects of 4 to 10% per year expand by substantial factors (moderate fitness), and variants with fitness effects >10% per year can expand enough to overwhelm the marrow (high fitness).

by chance. The synonymous variants provide a crucial validation of the model because it predicts that the majority of synonymous variants should be found at very low VAFs. Quantitatively, if our inferred value of $N\tau \approx 100{,}000$ years from DNMT3A R882H variants is correct, it would predict that the majority of synonymous mutations should be restricted to VAFs below $\phi = t/2N\tau \approx 0.025\%$ at age 50. This prediction broadly agrees with the data, where the maximum likelihood inferred $\phi \approx 0.03 \pm 0.005\%$ [see (*27*)]. This internal consistency check indicates that both synonymous and *DNMT3A* R882H variants point toward similar values of $N\tau$. Synonymous variants with VAFs $>>\phi$ are rare (orange dashed line in Fig. 1E) and are consistent with having hitchhiked to high frequencies on the back of an expanding clone that had already acquired a fit variant [see (*27*)], although it is also possible that a handful are developmental in origin; have a functional consequence themselves, for example, owing to codon usage bias; or are in fact nonsynonymous in an alternatively spliced transcript.

The fitness landscape of CH

Because the characteristic maximum VAF, ϕ, depends on the fitness effect, s, by estimating ϕ from the VAF spectrum, we can infer a variant's fitness. We illustrate this approach using *DNMT3A* R882H variants. As predicted by the model, the density of R882H variants does indeed begin to fall off exponentially for VAFs >12% [red data in Fig. 1E; see (*27*)]. This suggests that R882H variants provide HSCs with a large selective advantage ($s \approx 15 \pm 1\%$ per year) because, over the course of ≈55 years (mean age across all studies), they have expanded to VAFs ≈12%, although some have reached VAFs as high as 50%.

To reveal the fitness landscape of other highly fit and possibly pathogenic variants, we applied this analysis to each of the 20 most commonly observed variants across all studies (Fig. 2A). Variants in the spliceosome genes *SF3B1* and *SRSF2* are some of the fittest in CH, with fitness effects as high as $s \approx 23\%$ per year, but are relatively rare owing to low mutation rates. *DNMT3A* R882H is the most common CH variant, because it is both highly fit and has a high mutation rate owing to its CpG context. The *DNMT3A* R882C ($Arg^{882} \rightarrow Cys$) variant is notably fitter than R882H ($s \approx 19 \pm 1\%$ versus $s \approx 15 \pm 1\%$ per year) but is observed less frequently because of its lower mutation rate [see (*27*)]. The potential of our analyses is underscored by the *GNB1* K57E ($Lys^{57} \rightarrow Glu$) variant. Although this variant has received little attention in CH, it is highly fit and strongly associated with myeloid cancers and represents a potentially targetable variant (*38*).

To reveal the overall fitness landscapes of key CH driver genes, we considered the VAF distribution of all nonsynonymous variants in each of the genes *DNMT3A*, *TET2*, *ASXL1*, and *TP53* (Fig. 2B). For *DNMT3A*, the density of nonsynonymous variants at low VAFs is broadly consistent with the same $N\tau \approx 100{,}000$ years inferred from R882H variants (blue data in Fig. 1E). However, with increasing VAF, the density of variants declines, consistent with a spectrum of ϕ and thus a spectrum of fitness effects. Performing a maximum likelihood fit to a family of stretched exponential distributions, we found that the spectrum of fitness effects for nonsynonymous variants in *DNMT3A* is very broad, with ≈40% of variants conferring moderate to high fitness effects [$s > 4\%$ per year, Fig. 2B; see (*27*)]. By contrast, the genes *TET2*, *ASXL1*, and *TP53* have a spectrum that is more skewed toward low fitness effects, with only ≈7 to 10% of all possible nonsynonymous variants in these genes conferring moderate or high fitness effects. These distributions highlight that, in these CH genes, most nonsynonymous variants have a low enough fitness that they are effectively neutral, whereas an important minority expand fast enough to become pathogenic and overwhelm the marrow over a human life span.

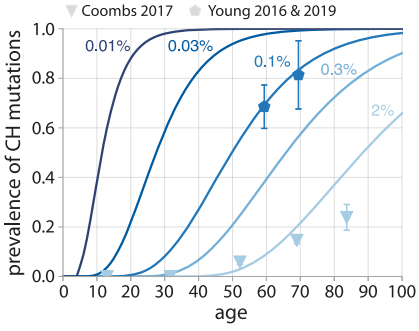

Fig. 3. Predicted prevalence of CH mutations as a function of age for different detection thresholds. Prevalence is predicted for individuals to have acquired at least one variant within 10 of the most commonly mutated CH genes (*DNMT3A*, *TET2*, *ASXL1*, *JAK2*, *TP53*, *CBL*, *SF3B1*, *SRSF2*, *IDH2*, and *KRAS*), taking into account the distribution of fitness effects across these genes [see (*27*)]. The actual prevalence of variants within these genes, as a function of age, is shown for (*10*, *15*) (pentagons, VAF limit of detection ≈ 0.1%) and (*11*) (triangles, VAF limit of detection ≈ 2%). Error bars represent sampling noise.

Highly fit variants confer an increased risk of AML

We next asked whether high-fitness variants confer an increased risk of AML development. By considering the pre-AML and control samples from three studies (*14*, *15*, *18*), we found that individuals harboring one or more of the 20 highly fit variants we identified (Fig. 2A) are ≈4-fold more likely to develop AML compared with those harboring lower-fitness variants [one-sided Fisher's exact test, $p < 10^{-5}$; see (*27*)].

Age dependence of CH

A key prediction of the model is that, because variants enter the HSC population at a constant rate, the apparent prevalence of a specific variant, at a defined sequencing sensitivity, is predicted to increase roughly linearly with age at rate $2N\tau\mu s$ [see (*27*)]. We confirmed this prediction using *DNMT3A* R882H and R882C variants, which, when combined, had enough data to be broken down by age group (fig. S18). In agreement with predictions, the age prevalence of these variants does increase linearly with age, consistent with the age dependence of CH being driven by the expansion of clones that become more detectable in individuals of older ages. The rate of this increase provides an independent way to validate estimates of fitness effects and, in this case, the rate of increase is consistent with a fitness effect of $s \approx 14\%$ per year, which is in agreement with estimates inferred from the VAF distribution (Fig. 2A).

By inferring the spectrum of fitness effects across 10 of the most commonly mutated CH genes, we can predict how common CH will be as a function of both age and sequencing sensitivity [Fig. 3 and (*27*)]. With sensitive-enough sequencing (VAFs ≥0.01%), CH variants will be detectable even in young adults and almost ubiquitous in people aged over 50 years. Our framework also enables us to predict the emergence of clones harboring multiple driver mutations. Although this depends on the cooperativity between mutations, under the assumption of additive fitness effects, we predict that, at a VAF detection limit of 0.01%, <15% of individuals aged 80 years will harbor clones with two or more mutations within the same cell [see (*27*)].

Discussion

A simple framework explains CH

Analyzing the VAF spectra from nine publicly available clonal hematopoiesis datasets in light of evolutionary theory points to a simple and consistent picture of how HSC population dynamics shape the genetic diversity of blood. The very wide variation in VAFs observed among people can be largely explained by the combined effects of chance (when a mutation arises) and fitness differences (how fast they expand). Our framework produces quantitative predictions for the number of HSCs, the prevalence of CH across ages, and how the number of somatic variants scale with VAF. These predictions are in agreement with available data and, in the case of HSC numbers, have been independently validated by an orthogonal method (*37*).

Implicit to our analysis is the assumption that many of the CH mutations drive cell-intrinsic increases in fitness. However, fitness is always context dependent, and therefore, cell-extrinsic effects are likely crucial in some

cases. It is also possible that the fitness effect of variants themselves changes over time, for example, owing to a slow but steady loss or gain of epigenetic marks due to mutations in epigenetic regulators (*39*, *40*). Changes in bone marrow environment driven by aging (*41*, *42*), chemotherapy (*11*, *26*, *35*, *43*), acute infection (*44*, *45*), and inflammation (*46*) could all shape the fitness effects of some variants. Indeed, specific variants (e.g., *PPM1D*, *TP53*, *CHEK2*, and *ASXL1*) are known to be strongly influenced by external factors (*26*, *35*, *47*). Taken together, however, the data from healthy individuals over a broad range of ages are quantitatively consistent with cell-intrinsic fitness differences playing a major role in shaping the variation in HSC clone sizes.

Although it might seem surprising that a simple model captures many quantitative aspects of CH data, more complex scenarios, including spatially partitioned niches, yield the same effective model for the multiyear development of CH; although in these scenarios, N and τ have more complex meanings [see (*27*)]. These include models with HSCs switching between active and quiescent states and models with progenitors occasionally reverting to HSCs. But there are important observations that the model cannot fully explain, including a considerably broader than expected distribution in the number of variants observed in different individuals, although this could be attributed to variations in mutation rates across individuals or environment-specific effects. Distinguishing between these scenarios and teasing apart the relative contributions of cell-intrinsic versus cell-extrinsic influences on cellular fitness will likely require longitudinal data and is an important area for future work.

In HSCs, fitness dominates drift

The relative roles of mutation, drift, and selection in shaping the somatic mutational diversity observed in human tissues has been the subject of much recent debate, especially regarding the conflicting interpretations from the ratio of nonsynonymous to synonymous mutations (dN/dS) (*1*, *5*, *48*) and clone size statistics (*32*, *49*, *50*). In blood, the two measures are in quantitative agreement; nonsynonymous variants are under strong positive selection, and most synonymous variants fluctuate by means of neutral drift.

Our inference of the large HSC population size ($N\tau \approx 100{,}000$ years) has an important interpretation: On average, it would take 100,000 years for a variant to reach VAFs of 50% by drift alone and >2000 years to be detectable by standard sequencing (VAF > 1%). Therefore, the vast majority of CH variants reaching VAFs >0.1% over a human life span likely do so because of positive selection. However, this is not to say that variants with VAFs <0.1% are not potentially pathogenic. Indeed, most highly fit variants exist at low VAFs simply because not enough time has yet passed for them to expand, although they are less likely to acquire subsequent driver mutations while they are at low VAFs.

More than 2500 variants confer moderate to high fitness

By considering the VAF spectrum across 10 of the most commonly mutated CH genes, we have inferred that mutations conferring fitness effects $s > 4\%$ per year occur at a rate of $\approx 4 \times 10^{-6}$ per year [see (*27*)]. Given that the average site-specific mutation rate in HSCs is 1.6×10^{-9} per year [see (*27*)], this implies that there are ≥2500 variants within these genes conferring moderate to high selective advantages. Our framework, in combination with broader coverage sequencing outside of known hotspot regions, could facilitate the discovery of these preleukemic drivers. However, targeting specific preleukemic clones may be clinically challenging, especially because the targeted therapy may alter the clonal dynamics of other variants. Although there is direct evidence from longitudinal data (*18*) and indirect evidence from age-prevalence patterns [see (*27*)] that variants at many of these moderate- and high-fitness sites expand at a roughly constant rate, other variants, notably *JAK2* V617F ($Val^{617} \rightarrow Phe$), might exhibit more complex dynamics given the small exponential growth rates observed in longitudinal data (*51*). It is likely that specific mutations achieve their selective advantages in different ways. Some will simply cause a bias toward self-renewal (*34*, *52*), whereas others may cause a bias as well as an increase in the intrinsic cell division rate. Distinguishing between these scenarios will require important future functional studies.

The variants commonly observed in CH are not necessarily the most fit but are both sufficiently fit and sufficiently frequently mutated. To reveal variants that are infrequently mutated yet potentially highly fit, we considered all variants in *DNMT3A*, *TET2*, *ASXL1*, and *TP53* that were detected at least twice across all nine studies and estimated their fitness effects by determining what fitness effect would be needed to produce the number of observed variants [see (*27*)]. Although the lack of data at infrequently mutated sites and the crudeness of this counting method necessarily lead to large uncertainties, there appear to be at least some highly fit yet infrequently mutated variants which, although individually rare, could be collectively common [see (*27*)]. We note that the high-fitness variants identified in *TP53* are strongly enriched for missense variants in the DNA binding domain (figs. S24 and S25), in agreement with recent functional and clinical data (*53*).

Given the average site-specific mutation rate of 1.6×10^{-9} per year (table S4), a comprehensive map between variant and fitness effect for all sites that confer a selective advantage large enough to expand substantially over a human life span ($s > 4\%$) could be achieved with the current sample size by increasing sequencing sensitivity to detect variants at VAFs >0.04% (fig. S26B). However, because sites can mutate at rates as low as $\mu \sim 10^{-10}$ per year (table S4), to quantify all variants, even rare ones, would require both a 6-fold increase in sample size as well as sequencing sensitivities as low as 0.01% VAF [see (*27*)]. Nonetheless, even with small study sizes, there are major advantages to being sensitive to very low VAFs (*10*, *12*, *15*), particularly in relation to synonymous variants, which, when grouped together, provide important information on $N\tau$ and genetic hitchhikers (Fig. 1E).

The near absence of variants in known AML drivers, such as *FLT3* and *NPM1*, across the nine studies suggests that mutations in these genes do not confer an unconditional selective advantage to HSCs, consistent with studies in mice and humans showing that they are late occurring and possibly cooperating mutations necessary for transformation to AML (*20*, *23*).

Future directions

CH has associated risks with cardiovascular disease (*7*, *54*) and progression to blood cancers (*7*, *8*, *14*, *18*) and consequences in the study of circulating tumor DNA (*55*, *56*), aplastic anemia (*57*), response to chemotherapies (*58*, *59*), and bone marrow transplant (*43*, *60*, *61*). A major challenge is to develop a predictive understanding of how variants and their VAFs affect disease risk. Recent studies show that both gene identity and VAF are predictive of progression to AML (*14*, *18*). The framework presented here provides a rational basis for quantifying the fitness effects of these variants and understanding VAF variations. Using this framework, we demonstrate that fitness estimates can be used to stratify AML risk. Because higher VAFs are strong predictors of AML development (*14*, *18*) and fitter variants are more likely to reach higher VAFs, it is perhaps not surprising that high-fitness variants are able to stratify AML risk. However, fitness predicts which variants are likely to reach high VAF and thus ought to have increased predictive power. Combining this framework with studies that longitudinally track individuals over time will shed light on how these initiating mutations acquire further mutations that drive overt disease. More sensitive sequencing techniques, broader sampling of the genome (e.g., regulatory regions), and the study of environmental factors that alter the fitness of mutations will improve our quantitative understanding of native human hematopoiesis and accelerate the development of risk predictors.

REFERENCES AND NOTES

1. I. Martincorena *et al*., *Science* **348**, 880–886 (2015).
2. A. S. Jonason *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14025–14029 (1996).
3. F. Blokzijl *et al*., *Nature* **538**, 260–264 (2016).
4. H. Lee-Six *et al*., *Nature* **574**, 532–537 (2019).
5. I. Martincorena *et al*., *Science* **362**, 911–917 (2018).
6. A. Yokoyama *et al*., *Nature* **565**, 312–317 (2019).
7. S. Jaiswal *et al*., *N. Engl. J. Med.* **371**, 2488–2498 (2014).
8. G. Genovese *et al*., *N. Engl. J. Med.* **371**, 2477–2487 (2014).
9. T. McKerrell *et al*., *Cell Rep.* **10**, 1239–1245 (2015).
10. A. L. Young, G. A. Challen, B. M. Birmann, T. E. Druley, *Nat. Commun.* **7**, 12484 (2016).
11. C. C. Coombs *et al*., *Cell Stem Cell* **21**, 374–382.e4 (2017).
12. R. Acuna-Hidalgo *et al*., *Am. J. Hum. Genet.* **101**, 50–64 (2017).
13. F. Zink *et al*., *Blood* **130**, 742–752 (2017).
14. P. Desai *et al*., *Nat. Med.* **24**, 1015–1023 (2018).
15. A. L. Young, R. S. Tong, B. M. Birmann, T. E. Druley, *Haematologica* **104**, 2410–2417 (2019).
16. M. Xie *et al*., *Nat. Med.* **20**, 1472–1478 (2014).
17. P.-R. Loh *et al*., *Nature* **559**, 350–355 (2018).
18. S. Abelson *et al*., *Nature* **559**, 400–404 (2018).
19. D. P. Steensma *et al*., *Blood* **126**, 9–16 (2015).
20. M. Jan *et al*., *Sci. Transl. Med.* **4**, 149ra118 (2012).
21. L. I. Shlush *et al*., *Blood* **122**, 487 (2013).
22. M. R. Corces-Zimmerman, R. Majeti, *Leukemia* **28**, 2276–2282 (2014).
23. L. I. Shlush *et al*., *Nature* **506**, 328–333 (2014).
24. L. I. Shlush *et al*., *Nature* **547**, 104–108 (2017).
25. L. I. Zon, *Nature* **453**, 306–313 (2008).
26. K. L. Bolton *et al*., bioRxiv 848739 [Preprint]. 20 November 2019.
27. See supplementary methods.
28. S. M. Hwang *et al*., *Leuk. Res.* **71**, 92–94 (2018).
29. E. Clayton *et al*., *Nature* **446**, 185–189 (2007).
30. A. M. Klein, D. P. Doupé, P. H. Jones, B. D. Simons, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **76**, 021910 (2007).
31. M. M. Desai, D. S. Fisher, *Genetics* **176**, 1759–1798 (2007).
32. B. D. Simons, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 128–133 (2016).
33. J. R. Blundell *et al*., *Nat. Ecol. Evol.* **3**, 293–301 (2019).
34. M. Jeong *et al*., *Cell Rep.* **23**, 1–10 (2018).
35. J. I. Hsu *et al*., *Cell Stem Cell* **23**, 700–713.e6 (2018).
36. S. Venkataram *et al*., *Cell* **166**, 1585–1596.e22 (2016).
37. H. Lee-Six *et al*., *Nature* **561**, 473–478 (2018).
38. A. Yoda *et al*., *Blood* **124**, 3567 (2014).
39. G. A. Challen *et al*., *Nat. Genet.* **44**, 23–31 (2011).
40. N. A. Robertson *et al*., *Curr. Biol.* **29**, R786–R787 (2019).
41. A. I. Rozhok, J. DeGregori, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8914–8921 (2015).
42. T. McKerrell, G. S. Vassiliou, *Sci. Transl. Med.* **7**, 306fs38 (2015).
43. T. N. Wong *et al*., *Nature* **518**, 552–555 (2015).
44. H. Takizawa, S. Boettcher, M. G. Manz, *Blood* **119**, 2991–3002 (2012).
45. M. Meisel *et al*., *Nature* **557**, 580–584 (2018).
46. K. Y. King, M. A. Goodell, *Nat. Rev. Immunol.* **11**, 685–692 (2011).
47. K. Murai *et al*., *Cell Stem Cell* **23**, 687–699.e8 (2018).
48. I. Martincorena *et al*., *Cell* **171**, 1029–1041.e21 (2017).
49. M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, A. Sottoriva, *Nat. Genet.* **48**, 238–244 (2016).
50. M. J. Williams *et al*., *Nat. Genet.* **50**, 895–903 (2018).
51. C. Nielsen, S. E. Bojesen, B. G. Nordestgaard, K. F. Kofoed, H. S. Birgens, *Haematologica* **99**, 1448–1455 (2014).
52. S. M. Chan, R. Majeti, *Int. J. Hematol.* **98**, 648–657 (2013).
53. S. Boettcher *et al*., *Science* **365**, 599–604 (2019).
54. S. Jaiswal *et al*., *N. Engl. J. Med.* **377**, 111–121 (2017).
55. J. Liu *et al*., *Ann. Oncol.* **30**, 464–470 (2019).
56. C. Swanton *et al*., *J. Clin. Oncol.* **36**, 12003–12003 (2018).
57. T. Yoshizato *et al*., *N. Engl. J. Med.* **373**, 35–47 (2015).
58. K. Takahashi *et al*., *Lancet Oncol.* **18**, 100–111 (2017).
59. K. L. Bolton *et al*., *J. Clin. Oncol.* **37**, 7–11 (2019).
60. T. N. Wong *et al*., *Nat. Commun.* **9**, 455 (2018).
61. R. C. Lindsley *et al*., *N. Engl. J. Med.* **376**, 536–547 (2017).
62. C. J. Watson, blundelllab/ClonalHematopoiesis: The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Zenodo (2020); doi:10.5281/zenodo.3706791.
63. C. J. Watson, The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Dryad (2020); https://doi.org/10.5061/dryad.83bk3j9mw.

ACKNOWLEDGMENTS

We thank all members of the Blundell, Fisher, and Druley labs. We thank S. Levy, I. Cvijovic, D. Petrov, B. Simons, M. Gerstung, B. Huntly, I. Martincorena, R. Levine, A. Levine, S. Jaiswal, and R. Majeti for helpful comments. **Funding:** C.J.W. is funded by a CRUK Cambridge Centre Clinical Research Fellowship. A.L.P. is supported by the National Science Foundation GRFP. G.Y.P.P. is funded by the CRUK Cambridge Centre Early Detection Programme and the Bei Shan Tang Foundation. W.H.W., A.L.Y., and T.E.D. are supported by NIH/NCI 1R01CA211711. D.S.F. and J.R.B. are supported by the Stand Up to Cancer Foundation and the National Science Foundation through PHY-1545840. J.R.B. is funded by the CRUK Cambridge Centre Early Detection Programme and by a UKRI Future Leaders Fellowship. **Author contributions:** J.R.B. and D.S.F. conceived the project. C.J.W., J.R.B., D.S.F., and A.L.P. developed the theory. C.J.W. and J.R.B. developed the data analysis methods and performed the data analysis and numerical simulations with input from D.S.F., A.L.P., and G.Y.P.P. D.S.F. and A.L.P. developed and analyzed the alternative models. G.Y.P.P. developed and analyzed the hitchhiking and multiple-mutant model and performed its data analysis. W.H.W., A.L.Y., and T.E.D. provided data. C.J.W. and J.R.B. wrote the manuscript. All authors provided comments and edits on the manuscript. **Competing interests:** C.J.W., A.L.P., G.Y.P.P., W.H.W., D.S.F., and J.R.B. have no competing interests. T.E.D. is the Chief Medical Officer for ArcherDX, Inc. As such, he holds ownership and receives salary. T.E.D. and A.L.Y. are co-inventors on patent application no. 62/106,967 submitted by Washington University School of Medicine in St. Louis, MO, USA, that covers an error-corrected sequencing method. This patent has been licensed by Canopy Biosciences, who were not involved in any data generation within this manuscript. **Data and materials availability:** All code used in this study is available at Zenodo (*62*) and accompanying data are available on Dryad (*63*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6485/1449/suppl/DC1
Supplementary Methods
Figs. S1 to S26
Tables S1 to S11
References (*64*–*69*)

View/request a protocol for this paper from *Bio-protocol*.

31 July 2019; accepted 24 January 2020
10.1126/science.aay9333